RECHERCHE • TECHNOLOGIE • APPLICATIONS

Informatique et SI

La qualité et la gouvernance des données

au service de la performance des entreprises

sous la direction de
Laure Berti-Equille

Hermes Lavoisier

La qualité et la gouvernance des données

LAVOISIER
14, rue de Provigny
94236 Cachan Cedex

www.hermes-science.com
www.lavoisier.fr

ISBN 978-2-7462-2510-7
ISSN 2111-0360

La qualité et la gouvernance des données

au service de la performance des entreprises

sous la direction de

Laure Berti-Equille

hermes science publications

Lavoisier

SÉRIE INFORMATIQUE ET SI
Sous la direction de Jean-Charles Pomerol

Jean-Louis Boulanger, *Outils de mise en œuvre industrielle des techniques formelles*, 2012.

Jean-Louis Boulanger, *Techniques industrielles de modélisation formelle pour le transport*, 2011.

Jean-Louis Boulanger, *Utilisations industrielles des techniques formelles : interprétation abstraite*, 2011.

Christophe Kolski, *Interaction homme-machine dans les transports : information voyageur, personnalisation et assistance*, 2010.

Liste des auteurs

Jacky AKOKA
CNAM
Paris

Nathalie BARTHÉLÉMY
GDF SUEZ
Paris

Soumaya BEN HASSINE-GUETARI
AID
Versailles

Laure BERTI-EQUILLE
IRD
Aix-en-Provence

Nicole BUSSAT
FIRMENICH
Genève
Suisse

Philippe CAPET
THALES
Colombes

Vincent CISELET
REVER
Charleroi
Belgique

Delphine CLÉMENT
AID
Versailles

Sébastien COEUGNIET
AID
Versailles

Isabelle COMYN-WATTIAU
CNAM
Paris

Idriss COOWAR
AID
Versailles

Thomas DELAVALLADE
THALES
Colombes

Thierry DÉLEZ
FIRMENICH
Genève
Suisse

Nunzio DI RUOCCO
Etat de Genève
Genève
Suisse

Walid EL ABED
GDE
Genève
Suisse

Jean HENRARD
REVER
Charleroi
Belgique

Melanie HERSCHEL
Université de Paris 6
Paris

Jean-Marc HICK
REVER
Charleroi
Belgique

Brigitte LABOISSE
AID
Versailles

Frumence MAYALA
REVER
Charleroi
Belgique

Sylvaine NUGIER
EDF
Clamart

Dominique ORBAN
REVER
Charleroi
Belgique

Christine PASCAL-SUISSE
Etat de Genève
Genève
Suisse

Didier ROLAND
REVER
Charleroi
Belgique

Evelyne ROSSIN
EDF
Saint-Denis

Samira SI-SAÏD CHERFI
CNAM
Paris

Jean-Michel SCHEIWILER
Etat de Genève
Genève
Suisse

Anastasiya SOTNYKOVA
Etat de Genève
Genève
Suisse

Virginie THION-GOASDOUE
CNAM
Paris

Yvan ZERMATTEN
Groupe Mutuel
Martigny
Suisse

Table des matières

Préface

A tous ceux qui pensent que l'excellence est une affaire de qualité !

Toujours plus de données échangées, toujours plus d'informations accessibles, toujours plus complexes et hétérogènes, les générations *baby-boom* qui partent massivement en retraite avec leurs connaissances... L'excellence dans la gouvernance des données en entreprise et dans les institutions n'est plus l'ambition que de quelques précurseurs, c'est maintenant un impératif économique, une exigence d'efficacité.

C'est dans ce contexte que depuis trois ans, une soixantaine de personnalités, toutes actrices de la qualité et de la gestion des données, notamment en leurs aspects théoriques, techniques, appliqués et sociaux, partagent expériences et connaissances au sein d'ExQI, association pour la promotion de la culture d'excellence en matière de qualité de l'information.

Véritable carrefour d'apprentissages, d'échanges et de rencontres associant entreprises, universitaires, éditeurs de logiciels, sociétés de services et de conseils et consommateurs de données, ExQI s'adresse aux dirigeants, responsables métier, analystes des données et informaticiens qui partagent tous le même objectif : maximiser la valeur des données de l'entreprise tout en s'alignant avec les enjeux économiques, politiques et environnementaux.

Les articles rassemblés dans ce volume sont tous issus des travaux des groupes de travail animés par ExQI, des recherches universitaires et de l'analyse des meilleures pratiques d'entreprises par AID, BDQS, Bouygues Télécom, EDF, l'Etat de Genève, Firmenich, GDE, GDF SUEZ, Groupe Mutuel, Michelin, REVER ou Thalès.

Ce premier recueil édité en français permet de révéler à tous les publics la richesse du domaine de la qualité et de la gouvernance des données et le foisonnement des idées pour comprendre, analyser et améliorer la valeur de ces données.

Avant de vous souhaiter bonne lecture, je tiens à remercier vivement Laure Berti-Equille, une des fondatrices de l'association ExQI, d'avoir proposé et coordonné avec tant d'énergie la rédaction de cet ouvrage et je félicite chaque auteur pour son enthousiasme et sa remarquable contribution.

Excellente lecture.

Sylvaine NUGIER
Présidente de l'association ExQI

Avant-propos

Faisant suite au succès des deux premières éditions des journées DEP (*data excellence paris*) organisées en décembre 2010 et 2011, l'association ExQI (*excellence qualité information*) propose l'édition d'un ouvrage collectif sur la qualité et la gouvernance des données en entreprise.

Cet ouvrage se veut être le premier ouvrage de référence en langue française sur les thèmes suivants :

– la qualité des données : concepts, définitions, méthodes, techniques et outils pour le nettoyage, la transformation, réconciliation et consolidation des données ;

– la gouvernance des données : outillage managérial et méthodologique, cas pratiques ;

– la migration de données et conversion en cas d'acquisition, fusion, changement organisationnel ou harmonisation de processus ;

– la mise en place d'un référentiel unique ;

– la mise en cohérence de données multi-sources ;

– la cotation et recommandation d'informations fiables ;

– des retours d'expérience sur des programmes de gouvernance des données.

Des contributions industrielles et académiques ont été apportées sur ces sujets d'actualité de telle sorte que l'ouvrage est constitué de douze chapitres qui présentent un panorama précis du domaine, aussi bien dans le domaine de la recherche que sur le plan opérationnel avec des études de cas et des retours d'expérience détaillés.

La bonne qualité des données est aujourd'hui fondamentale pour toute organisation. La gestion et l'amélioration de cette qualité est une tâche coûteuse, difficile et de longue haleine, mais néanmoins incontournable.

Le chapitre 1 intitulé « La qualité des données : concepts de base et techniques d'amélioration », proposé par Nunzio Di Ruocco (Etat de Genève – service gestion des données et de l'information – centre des technologies de l'information), A. Sotnykova (Etat de Genève, service gestion des données et de l'information – centre des technologies de l'information), Jean-Michel Scheiwiler (Etat de Genève, service gestion des données et de l'information – centre des technologies de l'information), dresse une typologie des anomalies sur les données : doublons, données manquantes, valeurs non-standardisées, inconsistances, valeurs inexactes ou encore données inutiles. Il présente les principaux indicateurs de la qualité des données : la pertinence, l'exactitude, la complétude, la consistance, l'opportunité, l'accessibilité et la facilité d'interprétation.

Le premier chapitre énumère ensuite les différentes techniques permettant d'améliorer la qualité des données, que ce soit de manière préventive ou corrective lorsque des problèmes sont détectés. Il montre également comment maintenir cette qualité au fil du temps. Parmi ces techniques, on trouve les modélisations conceptuelle et physique des données, la normalisation, l'utilisation de standards, les métadonnées, l'utilisation de sources de vérité (*master data*), le profilage des données (*data profiling*), le nettoyage des données (*data cleansing*), les alarmes par valeurs (*value alerts*), la saisie assistée, les traitements métiers optimisés et la production d'une documentation complète et mise à jour.

Les difficultés générant les problèmes de qualité des données étant souvent d'ordre humain et organisationnel, le chapitre présente la gouvernance des données (*data governance*) qui est un cadre de contrôle qualité visant à évaluer, à gérer, à exploiter, à optimiser, à contrôler, à entretenir et à protéger les données des entreprises. Il fournit les recommandations que toute application devrait satisfaire pour assurer une bonne qualité de ses données, ceci, afin de sensibiliser tous les acteurs qui travaillent tous les jours à la construction de systèmes d'information. Un processus d'amélioration de la qualité est proposé. Il distingue les bonnes pratiques à appliquer aux nouveaux projets et les actions pouvant être menées sur l'existant.

Le chapitre 2 intitulé « Les critères qualité pour la résolution d'identité », proposé par Nunzio Di Ruocco (Etat de Genève, service gestion des données et de l'information – centre des technologies de l'information) et Christine Pascal-Suisse (Etat de Genève, service gestion des données et de l'information – centre des technologies de l'information), propose une réflexion sur le traitement des données des personnes individuelles et une démarche qualité pour rapprocher/concilier les identités dans le contexte d'une interface entre un référentiel et un système opérationnel contenant des données personnelles. Le besoin est abordé au travers de la problématique connue sous le nom de « résolution d'identité ». Il donne un ensemble de définitions et d'explications sur ce qu'est la « résolution d'identité », au travers des différentes théories connues, en relevant les difficultés qu'elle induit.

Ce chapitre se propose également d'énumérer les anomalies et les conséquences d'un mauvais rapprochement et de présenter un ensemble de techniques pour la résolution d'identité. En outre, le chapitre fournit des recommandations que toute application devrait satisfaire pour assurer des rapprochements fiables, ceci, afin de sensibiliser tous les acteurs qui travaillent tous les jours à la construction de systèmes d'information.

La qualité des données est un sujet important et difficile qui n'a cessé de prendre de l'importance durant les dernières années auprès des chercheurs tout comme des entreprises. Le chapitre 3 intitulé « La qualité des modèles de données », proposé par Samira Si-Saïd Cherfi (maitre de conférences, conservatoire national des arts et métiers, Paris), Jacky Akoka (professeur des universités, conservatoire national des arts et métiers, Paris) et Isabelle Comyn-Wattiau (professeur des universités, conservatoire national des arts et métiers, Paris), aborde la qualité des modèles conceptuels, puisqu'il est maintenant largement reconnu, surtout depuis l'avènement des approches dirigées par les modèles (MDA), que la qualité des données dépend en grande partie de la qualité des modèles conceptuels sous-jacents. Le chapitre trois présente un aperçu des approches existantes avec leurs avantages et leurs limites. Il propose ensuite une approche détaillée pour évaluer et pour améliorer la qualité des modèles conceptuels. Il est important de noter que la majorité des approches de la littérature n'intègrent pas l'amélioration de la qualité et se limitent à son évaluation. Ce chapitre présente également une enquête menée auprès de praticiens. L'exercice de validation a permis de recueillir l'avis et les commentaires des participants sur les caractéristiques qu'il importe de mesurer lorsque l'on s'intéresse à l'amélioration de la qualité des modèles.

Dans le chapitre 4 intitulé « La cotation de l'information : approches conceptuelles et méthodologiques pour un usage stratégique », proposé par Philippe Capet (THALES systèmes C4I de défense et sécurité) et Thomas Delavallade (THALES), étudie l'usage actuel de la cotation de l'information et les besoins associés au sein de divers domaines d'activité, en particulier celui du renseignement. A partir des méthodes pratiquées par la défense nationale et des lacunes doctrinales qu'elles présentent, le chapitre présente une approche plus fondamentale, ainsi que des techniques, pour y remédier. Le projet de recherche CAHORS est présenté en guise d'illustration, ainsi que des applications du processus de cotation à visée tant civile qu'étatique.

Le chapitre 5 intitulé « Application de mesures de distance pour la détection de problèmes de qualité de données », proposé par Melanie Herschel (maitre de conférences, université Paris 6) et Laure Berti-Equille (directeur de recherche, institut de recherche pour le développement, France), a pour objectif de présenter un état de l'art des mesures et des méthodes pour détecter les doublons et les valeurs aberrantes dans le domaine des sciences informatiques et statistiques. Ce panorama

détaille l'éventail des méthodes et de l'outillage technique ; des exemples concrets et des applications pratiques illustrent la présentation.

Le chapitre 6 intitulé « La gestion de données multi-sources : de la théorie à la mise en œuvre dans le cadre d'un référentiel client unique », proposé par Soumaya Ben Hassine-Guetari (doctorante CIFRE AID – ERIC, université Lyon 2), Delphine Clément (consultante qualité et gouvernance de données, AID), Sébastien Coeugniet, (chef de projet qualité de données, AID), Idriss Coowar (chargé de projet qualité de données, AID), Brigitte Laboisse (consultante qualité et gouvernance de données, BDQS), présente les architectures et stratégies de gestion et d'intégration des données multi-sources. En particulier, il décrit dans le détail une démarche opérationnelle de conciliation des données issues de différentes sources dans le cadre de la création d'un référentiel client unique.

La bonne gestion d'une entreprise passe par la connaissance, la compréhension et le meilleur alignement possible de ses processus métier sur les objectifs de l'entreprise. La gestion de ces processus est plus connue sous le terme de BPM (*business process management*) et son intérêt est aujourd'hui bien reconnu de toutes les entreprises. Le chapitre 7 intitulé « L'évaluation de la qualité d'un processus métier – Enjeux, cas d'étude et bonnes pratiques », proposé par Virginie Thion-Goasdoué (maitre de conférences, conservatoire national des arts et métiers, Paris) et Samira Si-Saïd Cherfi (maitre de conférences, conservatoire national des arts et métiers, Paris), s'intéresse à l'étude des processus métier qui occupe donc aujourd'hui une place très importante dans le domaine de l'étude des systèmes d'information. Un processus métier est défini comme étant un ensemble de tâches reliées logiquement et effectuées afin d'atteindre un objectif opérationnel. Des exemples classiques de processus métier sont les processus de développement d'un nouveau produit ou de prise en charge d'une commande client. Un ensemble de processus métier permet ainsi de représenter le fonctionnement (d'une partie) des activités d'une entreprise, faisant généralement intervenir plusieurs acteurs, internes ou externes à l'entreprise. Modéliser les processus de l'entreprise permet de les comprendre, pour ensuite les analyser et les améliorer si besoin est.

Le chapitre 8 intitulé « L'excellence des données : valorisation et gouvernance », proposé par W. El Abed (PDG, *Global Data Excellence*), présente une méthodologie permettant de maximiser la valeur d'utilisation des données de l'entreprise et d'optimiser leur coût de gestion. Ce chapitre décrit les étapes et les stratégies de gouvernance des données qui permettront d'aligner les impératifs métiers avec la gestion des données et les projets informatiques. La mise en place du modèle de gouvernance est détaillée et opérationnalisée par l'évaluation des indicateurs-clés de valeur qui reflètent l'impact des données erronées sur l'entreprise.

Le chapitre 9 intitulé « Retour d'expérience sur un programme de gouvernance de données » proposé par Yvan Zermatten (expert gouvernance des données, DSI, groupe mutuel, Suisse) illustre, dans le domaine de l'assurance, la mise en œuvre vécue de la méthode de l'excellence des données (MED) proposée dans le chapitre précédent. Ce témoignage décrit l'avancement d'un projet de gouvernance avec ses atouts et ses difficultés à la fois techniques et humaines.

En effet, le succès d'un programme de gouvernance des données dépend en grande partie de ses acteurs, de leur rôle, responsabilités et de leur implication tant personnelle que collective. Le chapitre 10 intitulé « Les rôles et responsabilités des acteurs de la gouvernance des données : de la théorie à la pratique », proposé par Evelyne Rossin (chargée de mission politique de la donnée, direction optimisation amont aval trading du groupe EDF), Nathalie Barthélémy (responsable service management de l'information, GDF SUEZ branche global gaz & GNL) et Brigitte Laboisse (consultante qualité et gouvernance de données, BDQS), permet de cartographier et caractériser précisément les acteurs d'après les retours d'expérience de cinq entreprises (Bouygues Télécom, EDF, l'Etat de Genève, GDF SUEZ, Michelin) ayant mis en place une cellule de gouvernance.

Le gestionnaire de données a besoin d'outils et méthodes de valorisation simples et efficaces, permettant de démontrer que la qualité des données génère une valeur tangible pour l'entreprise. Cette valeur est déterminée par la contribution des données de qualité dans la réalisation des activités économiques de l'entreprise. La mesure de l'impact objectif de la qualité des données sur les processus de l'entreprise permet de déterminer des priorités claires afin de mobiliser le management et de concentrer les ressources et les moyens sur les problèmes les plus importants. Une telle approche garantit l'optimisation des services de données (en temps et qualité). La valorisation de la qualité des données est donc un instrument incontournable de gouvernance permettant de justifier la mise en place de processus de gestion efficaces. L'objectif du chapitre 11 intitulé « La valeur de la qualité en gouvernance des données dans la chaîne logistique », proposé par Thierry Délez (master data management director, Firmenich) et Nicole Bussat (Firmenich), est de fournir une méthode simple permettant de répondre à ce challenge.

Depuis quelques années la situation de la modélisation en informatique évolue sous l'effet de deux courants convergents : d'une part, la nécessité pour les organisations de se doter de méthodes et d'outils face à leur difficulté de garder la maîtrise de leurs applications dont la complexité et le nombre ne font qu'augmenter et d'autre part, la présence grandissante de solutions performantes, résultats de projets dans lesquels le code source n'est plus considéré comme l'élément central d'un logiciel, mais comme un élément dérivé de la modélisation.

De par le rôle vital que les données jouent en tant que « carburant » du fonctionnement des organisations, de par les coûts très importants qu'elles engendrent annuellement, de par la valeur patrimoniale et financière qu'elles représentent, la modélisation des données est un moyen pour les organisations de dépasser la simple gestion des données et d'entrer dans la gouvernance des données.

Le chapitre 12 intitulé « La gouvernance des données : apports de l'ingénierie des données dirigée par les modèles », proposé par Vincent Ciselet (ingénieur, REVER), Jean Henrard (Ph. D., REVER), Jean-Marc Hick (Ph. D., REVER), Frumence Mayala (ingénieur, REVER), Dominique Orban (ingénieur, REVER), Didier Roland (Ph. D., REVER), qui s'adresse aux responsables, aux praticiens et à toutes personnes intéressées par la gouvernance des données, montre qu'en considérant les données comme un « écosystème », en proposant des fonctionnalités innovantes telles que la cogénération, la coévolution et la comparaison d'écosystèmes, une démarche d'ingénierie des données dirigée par les modèles (IDDM) contribue aux objectifs de la gouvernance des données.

Chapitre 1

La qualité des données : concepts de base et techniques d'amélioration

1.1. Préambule

Le maintien dans le temps d'un niveau de qualité des données et des informations reste un sujet très complexe, difficilement maîtrisé et maîtrisable. Nous connaissons tous, les risques et les impacts sur les applications métiers lorsque les informations deviennent erronées.

Ce chapitre se propose donc d'énumérer les causes et les conséquences de la non-qualité des données et de proposer un ensemble de techniques pour :

- l'amélioration de la qualité des données ;
- la prévention des défauts ;
- la préservation d'un bon niveau de qualité des données.

La section « les familles d'indicateurs de la qualité » de ce chapitre est dédiée aux indicateurs de qualité. Parmi les multiples indicateurs de qualité existants, nous avons sélectionné les plus importants. Vous trouverez dans cette section leurs définitions exactes et leurs relations avec la qualité des données.

Dans la dernière section, un processus d'amélioration de la qualité est proposé. Il distingue les bonnes pratiques à appliquer aux nouveaux projets et les actions pouvant être menées sur l'existant.

Chapitre rédigé par Nunzio DI RUOCCO, Jean-Michel SCHEIWILER et Anastasiya SOTNYKOVA.

La bonne qualité des données est aujourd'hui fondamentale pour toute organisation. La gestion et l'amélioration de cette qualité est une tâche coûteuse, difficile et de longue haleine, mais néanmoins incontournable.

La direction de l'entreprise doit être convaincue de l'enjeu de la qualité des données et elle doit s'impliquer, montrer son volontarisme et apporter aide et soutien aux processus et aux structures mis en place pour assurer la qualité des données, leur succès en dépend.

1.2. Définition

Une donnée est de qualité si elle répond parfaitement aux besoins de ses utilisateurs [BRA 05] : la qualité des données est dépendante de leur utilisation.

La compréhension des besoins utilisateur est donc une condition nécessaire à la définition et à l'obtention de données de qualité.

Une difficulté importante est que la mauvaise qualité des données ne se détecte pas facilement. Ce sont souvent des incidents ou des anomalies dans le travail opérationnel qui révèlent çà et là, des inconsistances portant sur les données [BRA 05].

1.3. Les familles d'indicateurs de la qualité

Parmi toutes les familles d'indicateurs de la qualité des données, nous avons retenu les plus significatives : la pertinence, l'exactitude, la complétude, la consistance, l'opportunité, l'accessibilité, la facilité d'interprétation, l'unicité, la cohérence et la conformité à un standard.

1.3.1. *Pertinence*

La pertinence mesure la capacité des données à répondre aux besoins actuels et futurs des utilisateurs. Les données choisies pour l'application doivent assurer l'exécution de tous les processus métier traités par cette application. Pour les besoins actuels, les données ne devraient contenir que les objets utiles et utilisés par les processus métier. Parallèlement, pour les besoins futurs, les données devraient être facilement adaptables à l'évolution potentielle des processus métier.

Par exemple, dans des bases de données anciennes, la composante « année » des dates n'était définie que par les dizaines et les unités (« 74 » au lieu de « 1974 », etc.). Avec l'an 2000, ces attributs, qui étaient pertinents jusque-là, sont devenus inutilisables.

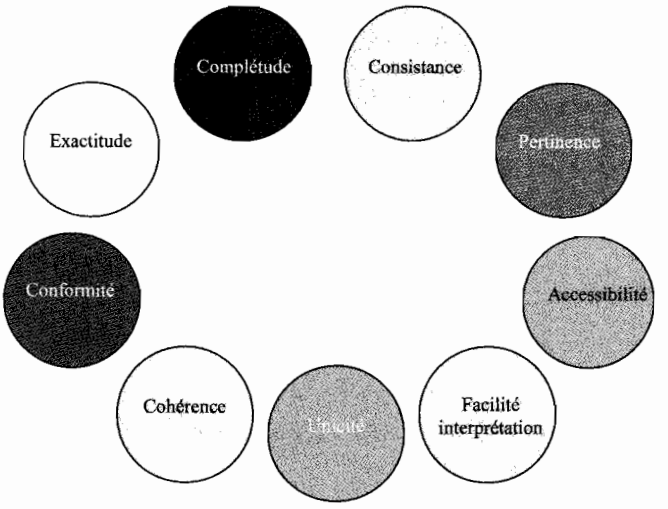

Figure 1.1. *Les familles d'indicateur*

1.3.2. *Exactitude, justesse*

L'exactitude mesure la conformité des données à la réalité : pour être justes, les valeurs en base de données doivent correspondre à leur état dans le monde réel. L'exactitude des données est une caractéristique sensible d'une application. Aucun processus métier ne peut fournir de bons services au client s'il se fonde sur des données erronées.

Par exemple, l'adresse d'un conducteur doit être mise à jour pour refléter le changement de son domicile.

1.3.3. *Complétude*

La complétude des données se juge en fonction des quatre critères suivants.

1.3.3.1. *La complétude des entités*

Cet indicateur est fondamental, il vérifie que tous les objets et les entités nécessaires à l'application sont bien représentés dans les modèles de données.

Par exemple, une base de données des employés est incomplète s'il manque l'entité département.

1.3.3.2. *La complétude des attributs*

Cette complétude est l'indicateur évaluant l'exhaustivité des attributs nécessaires au métier dans les entités du modèle de données.

Par exemple, le métier d'assureur s'appuie sur le nom, le prénom, le lieu de naissance du détenteur de permis de conduire, on doit retrouver tous ces attributs dans la base de données.

1.3.3.3. *La complétude des relations*

La complétude des relations est l'indicateur évaluant l'exhaustivité des associations existant entre les entités du modèle de données.

Par exemple, dans la réalité, une personne peut posséder un ou plusieurs véhicules, notre base de données doit donc comporter une relation « possède » liant les entités « personnes » et les entités « véhicules ».

1.3.3.4. *La complétude des occurrences*

Cette complétude est l'indicateur évaluant l'exhaustivité des occurrences de chaque entité.

Par exemple, dans la table « détenteur de permis », une ligne doit exister pour chaque personne détentrice d'un permis de conduire : personne ne doit avoir été oublié.

1.3.4. *Consistance*

Quand une entité est recopiée et maintenue en plusieurs exemplaires dans une base de données ou dans différentes bases de données, on dit qu'il y a consistance si pour chaque occurrence de cette entité, on retrouve les mêmes valeurs d'attributs dans toutes les bases.

La consistance est une propriété cruciale des données pour leur comparaison, leur rapprochement ou leur agrégation.

Par exemple, l'adresse d'un conducteur doit être la même dans tous les systèmes la référençant.

Si la valeur d'attribut adresse est « chemin Chanoine Latran » dans un enregistrement et « route de Lausanne » dans l'autre, une des adresses est fausse

et tout rapprochement sera impossible. Si l'adresse est mise à jour dans un système, elle devrait l'être dans les autres pour conserver cette consistance.

1.3.5. *Précision temporelle*

Les utilisateurs ont besoin d'avoir les données qui décrivent un système tel qu'il est ou était à un instant précis.

La précision temporelle (*timeliness*) mesure l'exactitude des données par rapport à l'instant qu'elles sont censées représenter.

Par exemple, une base de données est utilisée pour consigner la situation d'une entreprise au 1er janvier 2007. Pour que cette base soit précise temporellement, ses données ne devraient pas être périmées au 1er janvier mais ne devraient pas non plus avoir évolué depuis le 1er janvier.

Le cours de l'action en bourse stocké ne doit pas être celui du 26 décembre, ni du 2 janvier, mais celui de la dernière clôture de l'année le 29 décembre.

1.3.6. *Accessibilité*

L'accessibilité mesure la facilité de localisation et d'accès aux données par l'utilisateur. Elle recouvre également l'accessibilité des métadonnées, de la documentation et des services de support. L'accessibilité est fortement dépendante de l'ergonomie de l'application, ainsi que de l'architecture du système. Une application ergonomique nécessite peu de clics de souris et évite les enchaînements fastidieux d'écrans. Une donnée est souvent utilisée par différents niveaux d'utilisateurs (administrateurs de bases de données, opérateurs alimentant ces bases, utilisateurs finaux se contentant de lire ces données, etc.). L'accessibilité des données devrait être pensée et mesurée pour ces types d'utilisateur différents en prenant en compte leurs différentes méthodes d'accès.

1.3.7. *Facilité d'interprétation*

Cet indicateur décrit la facilité de compréhension des données, de leur analyse et de leur usage. Il est important que l'utilisateur ait une compréhension précise et sans ambigüité des données. Les données qui sont bien documentées et accompagnées de descriptions claires des concepts, de la structure et de leur usage représentent un avantage quant à l'efficacité de leur utilisation. Pour les utilisateurs finaux c'est le

choix des termes et des définitions dans la documentation qui est important, ce choix devrait être basé sur le vocabulaire métier et pas sur des termes techniques. Il est fondamental que les différents types d'utilisateur partagent le même vocabulaire et la même nomenclature pour éviter la confusion, source d'erreur. Les métadonnées sont un outil essentiel pour y arriver (voir section métadonnées).

1.3.8. *Unicité*

L'unicité garantit qu'une entité du monde réel est représentée par un seul et unique objet en l'occurrence un enregistrement, au sein du système. Elle permet d'éviter les doublons.

1.3.9. *Cohérence*

La cohérence assure l'absence d'informations conflictuelles au sein d'un même objet ou entre objets différents. Par exemple, la date de naissance d'un enfant antérieure à la date de naissance des parents. On peut avoir une incohérence temporaire pendant la phase de resynchronisation du système pour des objets propagés.

Par exemple avec des systèmes de cache de données, un statut peut avoir été modifié dans le système opérationnel mais pas encore dans le cache. Cette incohérence perdurera jusqu'au rafraîchissement du cache.

1.3.10. *Conformité vis-à-vis d'un standard, d'un format ou d'une convention de nommage*

La conformité d'un ensemble de données est le respect par celles-ci d'un ensemble de contraintes. Par exemple, l'identifiant d'un équipement doit commencer par deux lettres suivies de trois chiffres.

1.4. Typologie des anomalies sur les données

La quantité de données accumulée par les entreprises modernes est énorme – un *zettaoctet* ou 10^{21} octets. Avec un tel volume de données, le succès d'une entreprise est de plus en plus conditionné par la qualité (et non par la quantité !) et la connaissance de ses données. Vues de l'extérieur, les conséquences de la non-qualité des données se manifestent par :

– des perturbations opérationnelles (exemple : erreurs de traitements sur cas de doublons) ;

– un surcroît du travail dû à la réexécution de processus erronés ;

– l'insatisfaction et la perte de clients ;

– la dévalorisation de l'image de l'entreprise ;

– des difficultés lors de l'évolution ou de la migration de l'application vers une version plus avancée ;

– des problèmes lors de l'intégration des applications et de leurs données dans un *framework*.

Les pertes dues à la mauvaise qualité des données s'élèvent à 20 à 35 % des revenus opérationnels d'une entreprise [ENG 09].

Les causes des anomalies portant sur les données sont variées, mais elles s'expliquent le plus souvent par une sous-estimation générale de l'enjeu des données. Le caractère immatériel des données amplifie cette attitude. Les tâches d'initialisation des données, tant au démarrage d'une nouvelle application (reprise des données en masse), qu'en régime permanent (saisie manuelle au fil de l'eau) sont souvent négligées. Certains défauts de qualité trouvent aussi leur origine dans la conception des applications informatiques et dans des défaillances logicielles (*bugs*).

Enfin, dans un contexte économique en mouvement, l'évolution des systèmes d'information des entreprises est permanente. Fusions, cessions, lancement de nouveaux produits ou services, adaptation au marché, optimisation induisent des transformations sur les données. Avec ces transformations, la phase d'analyse et d'amélioration de la qualité des données sources (et par conséquent des données cibles), devient indispensable.

Avant d'être exploitées par l'utilisateur final, les données d'une application, passent par plusieurs étapes où leur forme et contenu sont définis. Ces étapes sont les suivantes :

– la définition des besoins utilisateurs ;

– la définition des cas d'utilisation ;

– la modélisation conceptuelle ;

– la modélisation physique ;

– la saisie ;

- la mise à jour ;
- le traitement par des processus métier ;
- l'optimisation.

A chaque étape la qualité des données peut être compromise ou, au contraire, améliorée. Avec la croissance des volumes de données produits par une entreprise, les défauts de qualité deviennent difficiles à gérer si l'analyse de qualité ne s'applique pas à chaque étape du cycle de vie d'une application.

Les principaux défauts de données qui peuvent résulter des erreurs humaines ou des défaillances logicielles durant le cycle de vie d'une application sont : les doublons, les données manquantes ou incomplètes, les valeurs non standards, les inconsistances, les valeurs inexactes, les données obsolètes et inutiles.

Pour illustrer ces types de défauts, utilisons l'exemple du tableau 1.1 qui contient des données sur les personnes détentrices du permis de conduire.

Permis ID	Nom	Prénom	Date naiss.	Adresse	Canton	Assurance	Assuré depuis
5600021	Miller, De	David	01/07/65	ch. de Chne Latran 1	Genève	Zurich	01/10/1980
3400093	Rubine	Josiane	26/11/70	avenue de l'Industrie 23	VD		30/01/1989
2901775	De Miller	Christine	22/05/69	chemin Chanoine Latran 1	GE	TSC	01/03/1994
340093	Rubine	Josiane	26/11/70	23, l'Industrie	Vaud	Generali	30/01/1989
0000056	Batini	Osvaldo	24/07/20	32 route de la Liberté	Schwyz/ SZ	KPT	01/09/2005

Tableau 1.1. *Données avec anomalies*

1.4.1. *Doublons*

Les doublons correspondent à deux enregistrements (ou plus) créés dans la base de données mais décrivant une même entité du monde réel. Par exemple, on peut craindre que dans le tableau 1.1, la deuxième ligne et la quatrième ligne décrivent deux fois la même personne – Josiane Rubine – mais avec des valeurs différentes. Les doublons posent des problèmes graves :

– en sélection : si plusieurs enregistrements correspondent au même conducteur, lequel doit être choisi pour être affiché et utilisé ?

– en mise à jour : si l'un des enregistrements est affecté mais pas l'autre, l'un possèdera des données correctes et l'autre non.

Par exemple : dans le cas d'une modification de coordonnées d'un conducteur, un seul enregistrement serait pourvu d'une adresse correcte. Si on utilisait l'enregistrement incorrect, la convocation serait délivrée à une mauvaise adresse.

1.4.2. *Données manquantes*

Ce type d'anomalie correspond à l'absence d'une donnée pourtant attendue. Les données manquantes engendrent des perturbations opérationnelles. La restauration des valeurs manquantes peut demander une surcharge de travail en termes d'échanges humains, d'utilisation de logiciels de nettoyage et de synchronisation des données, et au final, de réexécution du processus métier initial. Par exemple : la donnée « assurance » n'a pas été renseignée dans la deuxième ligne du tableau 1.1.

Si le service des automobiles doit contacter la compagnie d'assurances à propos d'un accident, il en sera incapable même si, dans ce cas précis, la donnée existe pourtant.

1.4.3. *Valeurs non standardisées*

Ce type d'anomalie revient à décrire le même objet de différentes manières. L'utilisation des valeurs non standardisées contribue à la mauvaise compréhension du contenu et du fonctionnement d'une application. Elle peut amener à :

– considérer que deux objets sont différents alors qu'ils sont pourtant identiques ;

– considérer qu'une valeur est fausse alors qu'elle a été exprimée différemment.

EXEMPLE 1.– Les chaînes « Ge » et « Genève » sont utilisées pour décrire le même canton. Si l'utilisateur ou le programme utilisant cette colonne « canton » ne sait pas faire ce rapprochement, ceux-ci seront considérés comme deux cantons différents. De même, si l'algorithme d'un programme vérifie que le canton doit être codé sur deux caractères alors il considèrera la valeur « Genève » comme fausse.

EXEMPLE 2.– « Miller, De » et « De Miller » décrivent le même nom de famille.

EXEMPLE 3.– « ch. de Chne Latran 1 » et « chemin Chanoine Latran 1 » décrivent la même adresse.

1.4.4. *Inconsistances*

Ce type d'anomalie se produit lorsque l'on stocke la même information à différents endroits et que les valeurs diffèrent. Contrairement aux doublons, cette duplication d'information à différents endroits est généralement effectuée en toute connaissance de cause pour des raisons techniques (dénormalisation, etc.). Cependant elle introduit ce risque d'inconsistance.

Par exemple, une base de données source contient la liste de référence des personnes et cette liste est périodiquement recopiée dans la base de données des détenteurs de permis de conduire pour y être utilisée. Entre deux recopies, les informations relatives aux personnes peuvent être modifiées dans la base source sans propagation sur l'autre base. Il y a alors une inconsistance. L'inconsistance peut également apparaître si une erreur se produit durant la recopie.

1.4.5. *Valeurs inexactes*

Il s'agit tout simplement de données erronées par rapport à la réalité. Il est évidemment très dangereux de se fonder sur des données erronées. Par ailleurs, une application alimentée par des valeurs inexactes produira de mauvais résultats qui mécontenteront l'utilisateur. Dans une table, des valeurs peuvent sembler justes mais une analyse basée sur les règles métier peut parfois démontrer l'inverse.

EXEMPLE 1.– Dans la première ligne de notre table (tableau 1.1), la différence entre les dates de naissance et de début d'assurance suggère que la personne a commencé à conduire à quinze ans. Selon les règles métier, cette situation est impossible.

EXEMPLE 2.– Schwytz faisant partie de la Suisse alémanique, les noms des rues du canton devraient être écrits en allemand.

1.4.6. *Données inutiles*

Il s'agit de données stockées dans le système et qui ne sont utilisées par aucun processus ou application. Les données inutiles peuvent provenir :

- d'une modélisation superficielle des exigences des utilisateurs ;
- d'une analyse superficielle des processus métier ;
- de l'évolution de ces derniers ;
- ou de saisies incontrôlées.

Ces données consomment des ressources système diverses (espace disque, espace sur le serveur des sauvegardes). Elles sont peu interprétables, rarement documentées et il est souvent difficile d'avoir la certitude qu'elles sont vraiment inutiles.

Par exemple, suite à une évolution d'application, la base peut être laissée avec des procédures obsolètes et des données égarées qui ne seront plus utilisées par aucun processus métier.

1.5. Techniques servant la qualité des données

Les données de mauvaise qualité ont un impact direct sur la qualité des services fournis aux clients. La qualité de l'information ou des services ne peut pas être meilleure que la qualité de leur matière première – les données. Il existe de nombreuses techniques On peut répartir les techniques d'amélioration de la qualité des données dans les catégories suivantes :

- prévention des défauts dans les données :
  - modélisation conceptuelle, normalisation, modélisation physique ;
  - création et utilisation de standards ;
  - métadonnées ;
  - utilisation de sources de vérité (master data), modèle *publish and subscribe* (abonnement) ;
- découverte et mesure des défauts de données existantes :
  - profilage et nettoyage (*data profiling* et *cleansing*) ;
- préservation de la bonne qualité des données :
  - alertes ;

- assistance à la saisie ;

- planification des traitements ;

- documentation.

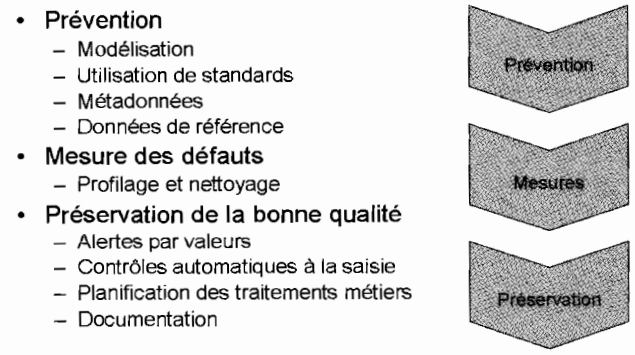

Figure 1.2. *Les techniques d'amélioration*

1.5.1. *Modélisation conceptuelle de données*

Le cycle de vie d'une base de données et d'un produit logiciel en général devrait commencer par l'élaboration des modèles conceptuels des données (MCD) et des processus. Il y a encore dix ans, les volumes de données gérés et la durée de vie des produits logiciels permettaient de se passer de modèles de référence et de compter sur l'expertise des gens ayant développé le logiciel et ses données.

Les modèles conceptuels de données remplissent plusieurs fonctions utiles dans le cycle de vie d'un produit logiciel :

– améliorer l'interopérabilité des structures et fonctions d'un domaine du monde réel, grâce à un modèle basé sur des concepts clairs et intuitifs, proches de ceux utilisés par les acteurs du domaine. Un modèle conceptuel peut être utile pour mieux comprendre la complexité ;

– fournir un vecteur de communication précis entre modélisateurs et développeurs du système d'une part, et spécialistes du domaine d'application et utilisateurs finaux d'autre part ;

– spécifier et guider les prochaines étapes du cycle de vie du système, telles que l'implémentation, la programmation et les optimisations diverses, et constituer une

trace compréhensible par les acteurs du domaine d'application et utilisable pour l'évolution et la maintenance du système en fonctionnement ;

– fournir une documentation de haut niveau, à destination des utilisateurs, basée sur les concepts du domaine d'application.

Le principal objectif d'un modèle conceptuel de données consiste à créer une représentation précise et complète des données métier, de leurs associations et de leurs contraintes, qui soit pertinente dans le contexte de l'application. La modélisation conceptuelle sert à améliorer plusieurs indicateurs de la qualité des données :

– la pertinence car on définit et choisit les entités utiles pour l'application dans le modèle conceptuel de données ;

– la complétude des entités, des attributs et des relations ;

– la facilité d'interprétation car un modèle conceptuel de données est la référence des objets métier pour tout le cycle de vie d'une application.

Mais, comme toutes les activités de modélisation abstraite, la modélisation conceptuelle est une activité créative non automatisable et non déterministe. Un modèle conceptuel de données n'est pas correct ou incorrect, il est plus ou moins adéquat à certains besoins. Un modèle plus détaillé n'est donc pas nécessairement un modèle meilleur [KOL 98].

Un modèle conceptuel de données adéquat pour l'application, et sa traduction vers le modèle physique (par exemple, vers un modèle relationnel) basée sur l'analyse profonde des exigences des utilisateurs, sont les conditions initiales pour concevoir des données de haute qualité. Le coût des défauts de conception est important car la résolution *a posteriori* de ce type de problèmes impacte la chaîne de développement, test, recette, et production.

Les bonnes pratiques de conception de nouvelles applications exigent que l'évolution potentielle du projet soit prise en compte depuis son début.

Outre les exigences métier, un modèle conceptuel de données devrait respecter un certain nombre de règles générales. Ainsi, chaque nom d'objet doit être unique, chaque entité doit comporter au moins un attribut, chaque relation doit être associée à au moins une entité.

Un outil de modélisation de données devrait permettre de vérifier les règles suivantes dans un modèle conceptuel :

– l'unicité du nom et du code d'entité : les noms et les codes d'entité doivent être uniques dans l'espace de nom ;

– la longueur du nom et du code d'entité est limitée ;

– l'existence d'identifiants : chaque entité doit comporter au moins un identifiant ;

– l'existence de liens de relation ou d'association ;

– l'héritage redondant : une entité hérite d'une autre entité plus d'une fois. Cette redondance n'enrichit pas le modèle ;

– l'héritage multiple : une entité est dotée d'un héritage multiple. Cette structure est inhabituelle mais tolérée si vous avez défini ce paramètre de vérification comme un avertissement ;

– le respect des règles de nomenclature en vigueur ;

– une liste de contrôles existe pour les différents objets d'un modèle conceptuel de données tels que les attributs, associations, domaines d'attributs.

Pour créer une base de données, un modèle conceptuel de données est traduit en modèle d'implémentation physique cible, par exemple un modèle relationnel. Ce modèle peut être validé par la technique de normalisation.

1.5.2. *Normalisation*

La technique formelle de validation d'un modèle relationnel est appelée normalisation. Celle-ci est utilisable au stade de conception logique d'une base de données relationnelle [TEO 11]. C'est une technique destinée à produire un ensemble de relations répondant aux règles de normalisation en termes des formes normales. Il existe plusieurs formes normales avec un degré de restriction croissant. Les formes normales plus élevées réduisent la duplication des données et excluent les inconsistances lors de la mise à jour des données.

Le processus d'optimisation peut exiger que ce modèle soit éventuellement dénormalisé. Cette phase d'optimisation peut inclure :

– des processus de dénormalisation de tables pour simplifier les requêtes les plus coûteuses ;

– des ajouts d'index pour accélérer les requêtes les plus fréquentes ;

– et des ajouts de vues qui correspondraient aux requêtes fréquentes.

Au final, on obtiendra une structure de base de données optimisée au fonctionnement de l'application (voir annexe, section 1.7).

1.5.3. *Modélisation physique*

La modélisation physique consiste à transcrire et adapter le modèle logique au système de gestion de base de données cible. Cette étape permet d'exploiter les atouts et spécificités du SGBD (système de gestion de base de données) pour obtenir les meilleures performances.

Les avantages associés à un modèle physique efficace sont :

- une meilleure compréhension de l'utilisation des champs des tables ;
- une meilleure performance ;
- une optimisation de l'espace du stockage ;
- une diminution des efforts de formatage pour la représentation de données côté applicatif.

Lors de la traduction du modèle conceptuel de données en modèle logique et physique, les types de données et leur taille doivent être soigneusement choisis, notamment pour l'exactitude et la justesse de ces données.

1.5.4. *Réutilisation des éléments déjà modélisés*

Un composant réutilisable peut être défini comme une unité du design avec un nom, une structure définie, accompagné de la documentation qui décrit son usage et ses contraintes.

Dans le contexte de la qualité des données, on peut réutiliser des éléments d'un modèle conceptuel ou d'un modèle physique. Par exemple, un même objet « adresse » peut être utilisé dans plusieurs applications aux niveaux conceptuel et physique.

Un objet réutilisable une fois modélisé et implémenté :

- permet de réduire les coûts et les temps de développement ;
- facilite l'interprétation de données (les composants étant connus et reconnus par les intervenants) ;
- contribue à la consistance de données ;
- facilite et fiabilise l'échange des données.

1.5.5. *Utilisation de standards*

Les données doivent être compréhensibles et ne doivent laisser planer aucune ambiguïté quant à leur signification et leur interprétation. Pour cela, elles doivent être lisibles, précises et simples tout en respectant les standards lorsqu'ils existent. Le respect des standards a son importance non seulement pour la facilité d'interprétation (par exemple, la devise euro codifiée par « EUR » est universellement reconnue) mais aussi pour la consistance des données lors de l'intégration de plusieurs bases afin de construire un entrepôt de données (*data warehouse* en anglais) ou pour les échanges de données entre applications dans un système fédéré.

On peut prendre pour exemple l'organisation GS1 (www.gs1.com). Le GS1 est un organisme mondial qui met en œuvre et accompagne l'utilisation des standards dans les chaînes d'approvisionnements (*supply chain*). Les standards de GS1 concernent la codification des produits, des services et des lieux, l'identification automatique (codes à barres et étiquettes radiofréquence – RFID), les langages de communication entre ordinateurs, la classification et la synchronisation des données.

1.5.6. *Métadonnées*

Une métadonnée est une donnée servant à définir ou décrire une autre donnée quel qu'en soit son support (papier ou informatique).

Les métadonnées facilitent :

- la compréhension des données et leur interprétation ;
- leur utilisation ;
- leur gestion ;
- leur partage.

Les métadonnées sont importantes pour la qualité des données, car elles précisent la sémantique, la structure et le format. Elles lèvent toute ambiguïté quant au sens réel de la donnée.

Les métadonnées permettent notamment de documenter et d'expliciter les modèles de données.

De plus, il est très intéressant de les intégrer à un système de données de référence.

Le format XML (*eXtended Markup Language*) est fréquemment employé pour stocker et échanger les métadonnées.

Cependant, celles-ci peuvent être conservées sous d'autres formes – souvent dans des référentiels de type base de données.

1.5.7. *Données de référence (*master data*)*

Une donnée de référence (*master data* en anglais) est un enregistrement unique qui sert de référence à toute l'entreprise. Par exemple, le nom d'un client, le code d'un produit, un numéro de compte sont des données de référence. La méthode de gestion des données de référence (master data management – qu'on retrouve fréquemment sous l'acronyme « MDM ») regroupe l'ensemble de ces données au sein d'un référentiel.

Une telle source de données de référence doit garantir :

– la consistance et la propreté des données (pas de doublons, pas de données manquantes) ;

– la disponibilité des données et leur propagation à travers toute l'entreprise ;

– la précision temporelle des données grâce à la centralisation des mises à jour.

Les données de référence peuvent être utilisées par des systèmes différents pour produire des données métier de qualité. A l'inverse, l'absence de gestion unifiée des données de référence se traduit au quotidien par des pertes d'efficacité opérationnelle qui ont un impact direct sur la performance globale de l'entreprise.

Il existe plusieurs approches de création des données de référence. Les outils du marché de type progiciel de gestion intégré (*enterprise resource planning* – ERP), entrepôt de données (*data warehouse* – DW), gestion de la relation client (*customer relationship management* – CRM), intégration d'applications d'entreprise (*enterprise application integration* – EAI) traitent partiellement cette problématique.

On trouve cependant des produits qui implémentent le concept de gestion de données de référence de manière générique, ce qui permet de traiter tout type de problématique.

Ces outils sont généralement composés de deux parties principales :

– la partie applicative assurant l'interface vers les sources de données hétérogènes, ainsi que la synchronisation et le suivi des règles métier ;

– l'infrastructure maintenant le dépôt des données de référence.

1.5.8. *Publication-souscription (*publish and subscribe*)*

Publication-souscription (*publish and subscribe*) est une technique de notification des changements dans une source de données de référence. Cette technique est appliquée pour maintenir le niveau de qualité des données, et assurer l'exactitude et l'opportunité des données métier. Dans le modèle publication-souscription, les dernières mises à jour sont mises en évidence au moyen d'un flag ou par la valeur d'une colonne, ou enfin, par la génération de fichiers ou de messages contenant les descriptions des changements effectués. Ces changements sont publiés.

Les bases qui utilisent les données de cette source s'abonnent à ces notifications par des procédures de vérification des flags de changements. Chaque changement publié est estampillé par un libellé. Lorsque la base abonnée trouve un changement avec le libellé correspondant au libellé de son abonnement, une procédure de mise à jour est lancée. L'avantage de cette technique est qu'elle est applicable dans les systèmes faiblement liés (*loosely coupled*) où le propriétaire des données ne fait que publier les changements, et ce sont les bases consommatrices qui se chargent des procédures de détection et reprise des changements qui les intéressent.

1.5.9. *Profilage de données (*data profiling*)*

Sur des applications et données existantes, il est trop tard pour mettre en œuvre les techniques de prévention mais, par contre, on appliquera les techniques de découverte et d'évaluation des défauts de données. La qualité des données étant une discipline récente, les approches utilisées ont été héritées d'autres disciplines qui traitent de problèmes similaires – par exemple la statistique qui s'occupe de l'analyse, la description, l'exploration et la déduction sur un ensemble des données. Les approches statistiques sont utilisées dans le profilage des données pour mesurer l'exactitude, la complétude et la consistance des données. Les résultats du profilage peuvent ainsi faciliter l'interprétation des données analysées.

Le profilage de données a pour but le calcul des statistiques et des différentes métadonnées pour l'ensemble des données. Les informations recherchées par le profilage décrivent :

– les domaines : les valeurs d'un attribut doivent se conformer à l'intervalle attendu. Par exemple, l'âge des conducteurs devrait se situer entre 18 et 99 ;

– les types : le type de donnée déclaré pour un attribut doit correspondre au type de donnée réel. Par exemple, pour un attribut de type chaîne de caractères a-t-on bien du texte entre guillemets et non des dates (sans guillemet) ?

– les patrons (*patterns* en anglais) : pour certains attributs, on peut définir des règles syntaxiques auxquels les valeurs doivent se conformer. Par exemple, pour l'attribut « nom abrégé » de la table « cantons », la règle syntaxique des valeurs est de deux majuscules de l'alphabet latin ;

– les fréquences : cette métadonnée est une statistique qui s'appuie sur l'expertise métier. Par exemple, si on applique le profilage sur le service des automobiles du canton de Genève, l'attribut « adresse » doit contenir le mot « Genève » dans la majorité de cas ;

– les dépendances : l'analyse des intra-dépendances dans une table permet de vérifier s'il y a des valeurs qui dépendent d'autres valeurs de la même table, autrement dit, si la table est bien en 3^{e} forme normale (voir annexe, section 1.7) ;

– les redondances : les mêmes valeurs apparaissent dans des tables différentes. Cette mesure révèle des valeurs qui sont potentiellement des clés étrangères. Elle peut être aussi utilisée pour valider les clés étrangères connues.

Les statistiques de profilage de données peuvent être calculées sur les valeurs minimale, maximale et moyenne d'une colonne, la fréquence des valeurs, le nombre des valeurs distinctes, la déviation des valeurs effectives aux valeurs attendues.

Par exemple, le profilage sur l'attribut « date » de la table « accidents » peut montrer que cet attribut contient des dates dans le futur ou des valeurs qui datent de plus de 20 ans. Si la politique du service des automobiles concerné était de garder les données des accidents pour dix ans au maximum, cette découverte suggérerait que certaines valeurs de la colonne « date » sont fausses.

Le profilage est suivi par le nettoyage des données. Il existe sur le marché divers outils de profilage et de nettoyage de données reposant sur différentes techniques [LIN 08, MCG 08, ENG 09]. Le profilage des données pourra ainsi être effectué à l'aide de ces outils ou, manuellement, par comparaison avec la documentation, par des procédures internes à la base de données, par des requêtes SQL et par des consultations auprès des experts du domaine.

1.5.10. *Nettoyage de données (*data cleansing*)*

Le nettoyage des données est le processus de correction des défauts découverts par le profilage des données. A partir des résultats du profilage, on peut mesurer le degré de complexité des défauts sur les données opérationnelles et les ressources nécessaires pour les éliminer. En pratique, il est rare que tous les défauts soient éliminés car les processus de nettoyage complet coûtent cher aux entreprises. Il est

donc important d'analyser les résultats du profilage et de classer les défauts comme critiques, importants ou insignifiants.

Le marché des outils de gestion de la qualité des données ne propose pas d'outils automatiques qui soient capables d'identifier et de corriger toutes les données défectueuses [ADE 05]. Chaque entreprise souhaitant nettoyer ses données doit définir un seuil en dessous duquel la qualité est jugée inacceptable. Le nettoyage peut ensuite être effectué soit directement dans la base opérationnelle, soit lors de la création de l'entrepôt de données qui servira de référence pour l'entreprise.

1.5.11. *Alarmes par valeurs (*value alerts*)*

Une autre technique permettant de préserver l'exactitude des données consiste à ajouter des alarmes pour des données vitales. Une alarme peut être un *trigger* qui se déclenche lorsque la valeur d'un attribut dépasse les limites définies par les règles métier.

1.5.12. *Saisie assistée*

Comme on a dit dans la section causes et conséquences de la non-qualité de données, la source la plus fréquente de données inexactes et inconsistantes est la saisie utilisateur. Même si toutes les personnes participant au processus d'alimentation du système cherchent à réaliser leur travail du mieux possible, des erreurs de saisie seront toujours présentes dans les données. Pour diminuer le taux d'erreurs de ce type, l'application doit disposer – lorsque c'est possible – de fonctions d'assistance à la saisie. On peut citer :

– le contrôle du domaine et du format des valeurs saisies : l'application peut contrôler les valeurs en se fondant sur les règles métier connues. Par exemple, le format de numéro d'un permis de conduire devrait être de douze chiffres. L'application peut s'appuyer sur la définition des objets en base de données (type, taille, contraintes, etc.). Il est donc important de spécifier au mieux les entités dans le SGBD ;

– la suggestion des valeurs possibles : si le spectre des valeurs est connu, l'application peut les proposer à l'utilisateur. Par exemple, pour la saisie de l'âge d'un conducteur, la valeur minimale peut être initialement affichée dans le champ d'âge. Là encore, l'application peut être épaulée par les définitions dans la base de données ;

– le contrôle des doublons : lorsque l'utilisateur saisit les valeurs, l'application peut rechercher les lignes similaires déjà saisies dans la base. Si une ligne avec des

valeurs identiques ou similaires est trouvée, une alarme se déclenche pour éviter la duplication des données.

1.5.13. *Traitements métiers optimisés*

En règle générale, chaque base de données nécessite un nombre de traitements liés aux exigences métier. La synchronisation avec les données de référence (*master data*) assure la consistance et la pertinence des données. Les calculs métiers, comme par exemple les rapports de statistique des accidents routiers par canton et par mois, sont aussi liés à l'opportunité des données.

Pour maintenir une base de données opportune, consistante et accessible, il est important de planifier ces traitements de manière optimisée. Depuis le début du cycle de vie d'un système, ces traitements métier devront être pris en compte et analysés pour optimiser – si nécessaire – la structure physique de la base. En phase d'exploitation, les traitements par lot (*batch*) devront être planifiés pour les périodes creuses d'utilisation de la base, en dehors des heures de travail par exemple. L'exécution des *batches* devrait être surveillée par un système de contrôle incluant une fonction d'alarme en cas d'erreurs durant leur exécution.

1.5.14. *Documentation*

Une documentation complète et compréhensible peut servir plusieurs buts. Son but principal est l'aide à l'interprétation des données métier et du fonctionnement de l'application en général. Une documentation détaillée et accessible à tous les participants du projet facilite la communication entre les participants, et contribue à l'optimisation des processus de transfert des données : intégration, création de l'entrepôt de donnée et des entrepôts spécialisés (*datamarts*). Les bonnes pratiques distinguent plusieurs niveaux de documentation :

- la documentation publique avec une description générale du projet ;
- la documentation restreinte et spécialisée, destinée par exemple aux administrateurs de base de données, contenant toutes les opérations effectuées sur les données.

Un point évident mais pas toujours assuré en pratique est que tous les acteurs du projet doivent connaître l'existence des documents et pouvoir accéder rapidement aux informations souhaitées.

Il est important d'inclure le dictionnaire de données dans cette documentation. En effet, il permet aux interlocuteurs venant d'horizons différents de parler un langage commun, d'éviter les ambigüités et les erreurs de compréhension.

1.6. Stratégies pour améliorer la qualité des données

1.6.1. *Les nouveaux projets*

Chaque nouveau projet offre l'opportunité d'appliquer les techniques de prévention des défauts sur les données. Il s'agira de :

- soigner la phase de modélisation ;
- rechercher l'existence de standards bien établis dans le champ de l'application, les consulter et les exploiter si possible ;
- réutiliser des éléments déjà modélisés ;
- élaborer un modèle conceptuel de données ;
- élaborer un modèle physique de données ;
- produire une documentation complète qui sera tenue à jour dans un référentiel ;
- utiliser des sources de données existantes ou en définir de nouvelles si les données de l'application ont vocation à être réutilisées par d'autres projets avec emploi des méthodologies de gestion des données de référence ;
- définir les métadonnées ;
- implémenter l'aide à la saisie au niveau des applicatifs ;
- planifier des profilages de données à intervalle régulier (par exemple, tous les mois) et procéder à un nettoyage des données. Le projet doit être volontaire et fournir les ressources nécessaires (voir section profilage et nettoyage des données).

1.6.2. *L'existant*

Améliorer la qualité des données d'un projet est beaucoup plus facile et beaucoup moins coûteux dans les phases en amont, c'est-à-dire au moment de la conception.

Malgré tout, le contrôle et l'amélioration de la qualité des données doivent être des processus permanents car de nombreux facteurs dégradent inexorablement cette qualité (obsolescence des données).

Une fois en production, les applications utilisent effectivement les données et il devient délicat de modifier cet existant opérationnel. Dans certaines situations, il est inévitable d'améliorer la qualité des données.

Le projet doit être volontaire pour cette amélioration – on ne peut pas lui imposer contre son gré – et il doit mettre à disposition les ressources nécessaires en temps et en finances pour atteindre cet objectif.

1.6.3. *Profilage et nettoyage des données*

Le profilage et nettoyage de données sont des axes majeurs de l'amélioration de la qualité des données sur des projets existants.

Plusieurs éditeurs reconnus proposent des outils adaptés à ces techniques, parmi eux on peut citer : DataFlux, IBM, Informatica, Sap Business Objects, Trillium Software.

1.6.4. *Actions et recommandations*

Les actions suivantes sont à entreprendre pour améliorer la qualité de données d'un système existant :

– la reconstruction des modèles conceptuel et physique par rétro-ingénierie/rétro-conception (*reverse engineering* en anglais) si nécessaire ;

– la reconstruction des activités du système avec les données effectivement utilisées ;

– l'analyse du domaine métier afin de trouver des contraintes qui ne sont pas implémentées par le modèle physique ; et perfectionnement du modèle conceptuel ;

– la propagation des contraintes au niveau physique et ajout des contraintes manquantes ;

– la comparaison des données avec une référence et correction des valeurs inconsistantes (si possible) ; ajout des clés étrangères vers les données de références pour éviter les inconsistances dans le futur ;

– l'optimisation des requêtes métier pour améliorer la performance et la pertinence des données, ajout des index ;

– l'analyse des données avec un outil de profilage de données ou avec des procédures spécifiques au domaine pour découvrir les doublons, les données manquantes, etc. ;

– l'application des procédures pour nettoyer les anomalies trouvées par le profilage de données ;

– la planification régulière des contrôles et profilages de données pour préserver la qualité des données ;

– la transformation de plusieurs sources de données en une source de vérité ;

– le renforcement des contrôles dans les applications pour éviter les saisies de mauvaise qualité ;

– mener l'effort d'amélioration de la qualité des données en amont, c'est-à-dire dans les systèmes sources (par exemple, corriger la base opérationnelle source plutôt que la base décisionnelle qu'elle alimente), l'amélioration sera ainsi propagée automatiquement en aval.

1.7. Annexes normalisation : théorie et exemple

Une clé est un attribut ou un ensemble d'attributs permettant d'identifier de manière unique une instance de l'entité. Il est possible d'avoir plusieurs clés pour une même entité, on parle alors de clés candidates. Parmi ces clés candidates, une clé sera désignée comme clé primaire, clé principale pour l'identification des instances de l'entité.

Initialement, trois formes normales (*normal form* en anglais) ont été proposées : les premières (1NF), deuxièmes (2NF) et troisièmes formes normales (3NF). Une définition plus contraignante de 3NF proposée par R. Boyce et E.F. Codd est dénommée forme normale Boyce-Codd (BCNF). Les formes normales supérieures : quatrièmes (4NF) et cinquièmes (5NF) sont aussi définies, mais elles ne concernent que des situations assez exceptionnelles et rares. En général il est recommandé de normaliser jusqu'à 3NF.

1NF – première forme normale :

– tout attribut contient une valeur atomique ;

– tous les attributs sont non répétitifs ;

– tous les attributs sont constants dans le temps.

2NF – deuxième forme normale :

– elle respecte la première forme normale ;

– tous les attributs non-clés sont totalement dépendants fonctionnellement de la totalité de la clé primaire.

3NF – troisième forme normale :

- elle respecte la deuxième forme normale ;
- tout attribut n'appartenant pas à une clé ne dépend pas d'un attribut non clé.

BCNF – forme normale de Boyce-Codd : si une entité ou une relation en troisième forme normale a une clé concaténée (multiple), aucune des propriétés élémentaires de cette clé ne doit être en dépendance fonctionnelle d'une autre propriété. Cette normalisation conduit parfois à transformer une entité en relation ou à décomposer une relation en deux relations plus simples.

4NF – quatrième forme normale : pour toute relation de dimension *n* en forme normale de Boyce-Codd, les relations de dimension *n-1* construites sur sa collection doivent avoir un sens. Il ne doit pas être possible de reconstituer les occurrences de la relation de dimension *n* par jointure de deux relations de dimension *n-1*. Cette normalisation conduit parfois à décomposer une relation complexe en deux relations plus simples.

5NF – cinquième forme normale : pour toute relation de dimension *n* (avec *n* supérieur à 2) en quatrième forme normale, il ne doit pas être possible de retrouver l'ensemble de ses occurrences par jointure sur les occurrences des relations partielles prises deux à deux. Cette normalisation conduit parfois à décomposer une relation complexe en plusieurs relations plus simples.

Illustrons le processus de normalisation. Pour le modèle « détenteur de permis », nous disposons des attributs suivants : le nom et prénom du détenteur du permis, le type de permis obtenu (auto, moto, etc.), la date à laquelle le permis a été obtenu, la ville et le canton où il a été obtenu, l'âge du détenteur, son adresse actuelle et un numéro de téléphone auquel on peut le joindre. Enfin, on dispose de la liste des accidents dans lesquels il a été impliqué. On obtient l'entité décrite dans la figure 1.3.

Detenteur de Permis		
Nom	VARCHAR2(50)	<pk>
Prénom	VARCHAR2(50)	<pk>
Type de permis	VARCHAR2(5)	<pk>
Date Obtention	DATE	
Ville Obtention	VARCHAR2(50)	
Canton Obtention	VARCHAR2(50)	
Age Détenteur	NUMBER(3)	
Adresse Détenteur	VARCHAR2(500)	
Téléphone Détenteur	NUMBER(20)	
Accidents	VARCHAR2(2000)	

Figure 1.3. *Entité non normalisée*

1NF : le champ « accidents » est un champ multi-valeurs. Pour rendre cette table en 1NF une nouvelle table contenant les informations sur les accidents est créée, comme le montre la figure 1.4.

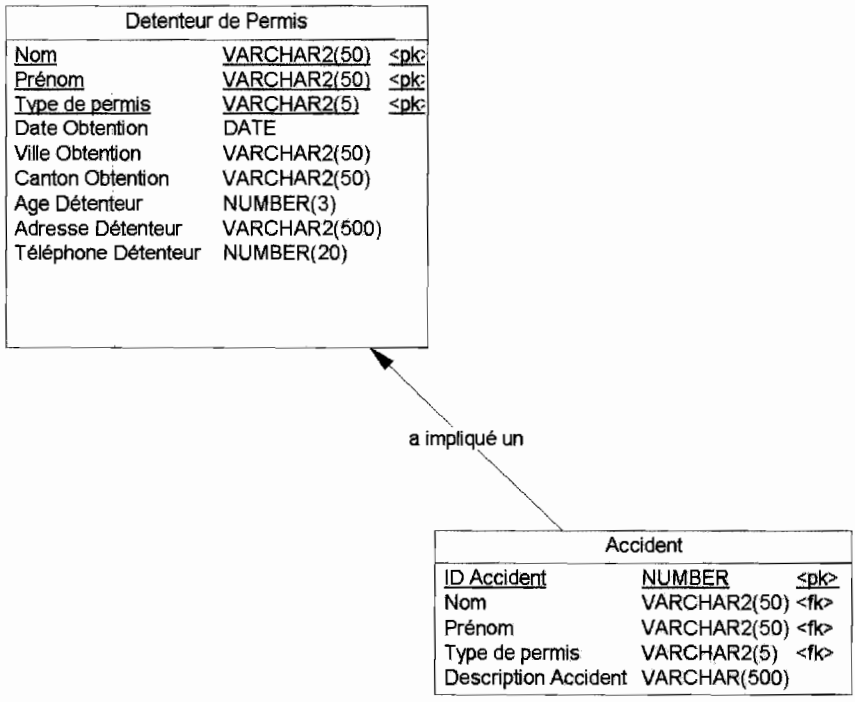

Figure 1.4. *Entité en première forme normale*

2NF : l'âge, l'adresse et le téléphone ne dépendent que du détenteur et pas du type de permis. Ces champs ne dépendent que d'une partie de la clé ; en 2NF, il est nécessaire de les isoler dans une nouvelle table décrivant les personnes. Remarquons que les accidents concernent plutôt la personne que le permis en lui-même, c'est pourquoi la contrainte référentielle des accidents a été déplacée vers la table personne, comme indiqué dans la figure 1.5.

3NF : le canton d'obtention dépend directement de la ville d'obtention. Par exemple, lorsque le permis est obtenu à Lausanne, le canton est VD. Il s'agit d'une dépendance sur un attribut qui ne fait pas partie de la clé primaire, ce que nous devons corriger en 3NF. On extrait ainsi ville/canton à ce stade comme le montre la figure 1.6.

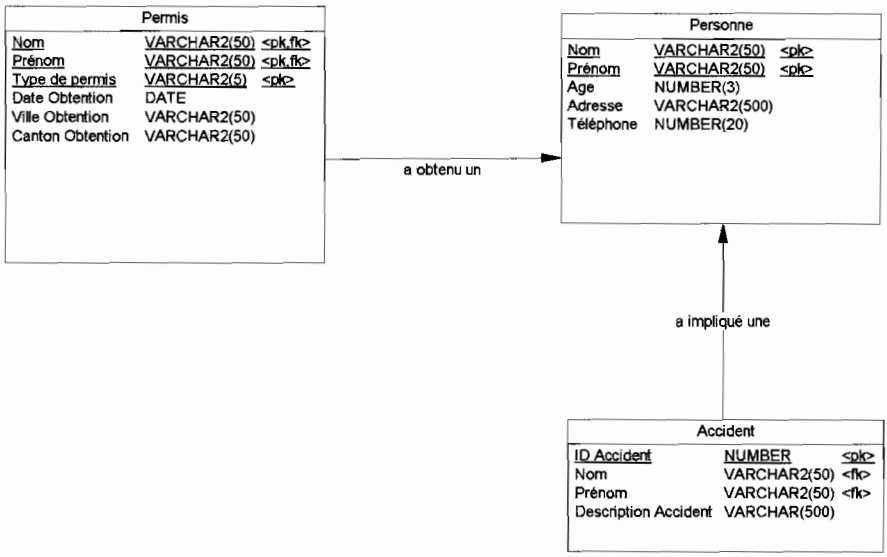

Figure 1.5. *Entité en deuxième forme normale*

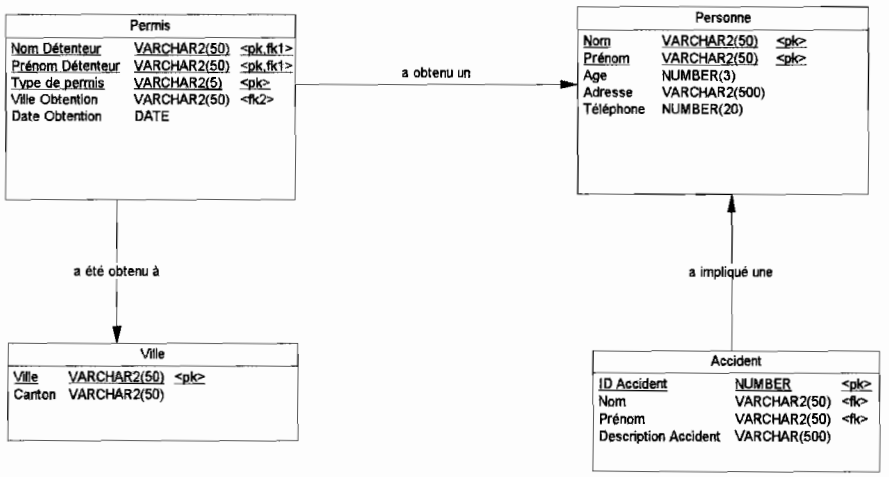

Figure 1.6. *Entité en troisième forme normale (partielle)*

Remarquons que dans notre modèle, un accident ne peut concerner qu'une seule personne à la fois. Pour associer plusieurs personnes à un accident, il faudrait créer une table d'association Personne/Accident.

REMARQUES SUR LA MODÉLISATION PHYSIQUE.– Les erreurs les plus répandues sont :

– l'utilisation du type chaîne de caractères pour des dates ou des heures ;

– l'utilisation du type chaîne de caractères pour stocker des chiffres ;

– l'utilisation de VARCHAR2(4000) ou CHAR(2000) pour stocker toutes les lignes de caractères, avec comme conséquences : un gaspillage de l'espace disque, une perte de performance, une surcharge d'appels à la fonction *trim()* ;

– le stockage du texte dans les champs du type BLOB.

Si on utilise le type chaîne de caractères pour stocker les dates ou les chiffres (par exemple : un champ « âge » en CHAR(10)), on finira invariablement par trouver, dans nos données, des dates qui ne sont pas des dates et des chiffres qui ne sont pas des chiffres. De plus, le choix correct de types améliore la performance des requêtes car le SGBD applique des algorithmes adaptés.

Les règles métier peuvent aussi fournir d'autres contraintes à appliquer dans la base pour assurer la consistance des données. Ces contraintes peuvent être de type :

– NOT NULL ;

– CHECK ;

– la déclaration de valeurs par défaut ;

– la déclaration de clé étrangère permettant de garantir l'intégrité référentielle.

Les contrôles internes à la base doivent être complétés par des contrôles au niveau de l'application, au moment de l'insertion.

Chaque table doit posséder une clé primaire :

– soit une clé métier naturelle ;

– soit une clé synthétique basée sur une séquence.

La clé synthétique peut être intéressante si on n'a aucune clé métier ou si la clé métier est trop compliquée (clé composite de plusieurs champs).

La clé synthétique a l'avantage d'assurer l'existence d'au moins un champ unique sécurisé contre les erreurs de saisie. En effet, dans la plupart des cas, les sources d'inconsistance sont les données saisies par les utilisateurs (ou les applications qui utilisent les données). Idéalement on ne devrait jamais mettre à jour une clé primaire, même métier.

En utilisant des clés synthétiques, notre modèle devient celui représenté dans la figure 1.7.

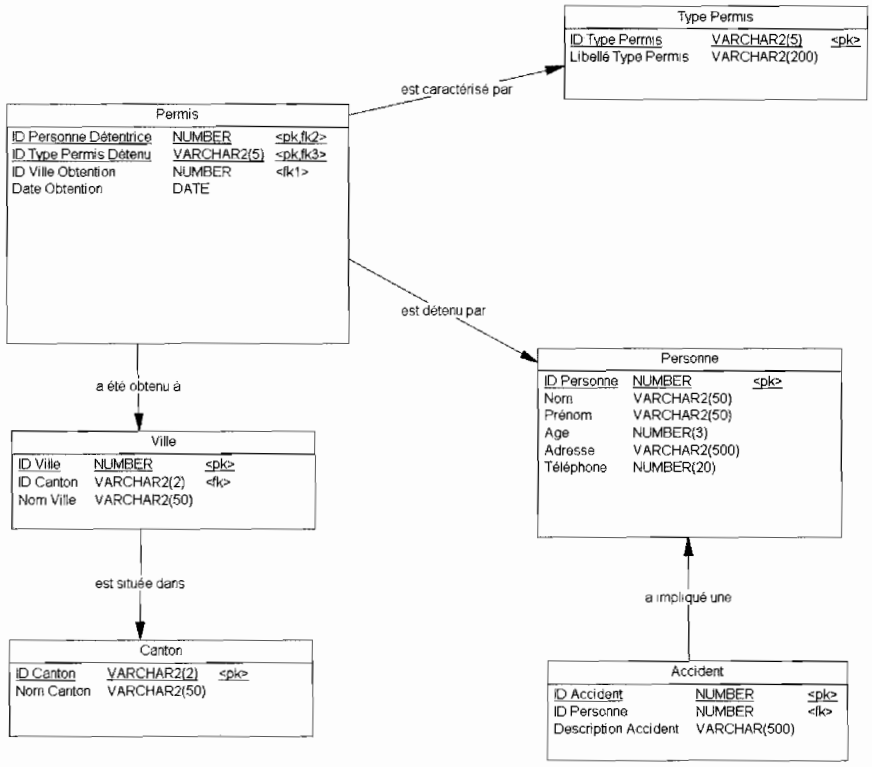

Figure 1.7. *Entité en troisième forme normale*

REMARQUES.–

– Pour le canton et le type de permis, on a une clé alphanumérique car ce sont des tables de référence dont le contenu est quasiment immuable et où il est plus pratique d'utiliser des codes plus parlant que des nombres (exemples : GE, VD, JU pour les cantons, AUTO, MOTO, PLO pour les types de permis).

– Pour la personne, il est utile (voire obligatoire) d'avoir une clé synthétique car le couple nom, prénom ne garantit pas l'unicité à cause des homonymes.

1.8. Bibliographie

[ADE 05] ADELMAN S., *Data Strategy*, Addison Wesley, Upper Saddle River, NJ, 2005.

[BRA 05] BRASSEUR C., *Data Management – qualité des donnés et compétitivité*, Hermès, Paris, 2005.

[ENG 09] ENGLISH P.L., *Information Quality Applied*, Wiley, New York, NY, 2009.

[KOL 98] KOLP M., *Modélisation conceptuelle et systèmes d'information*, Presses Universitaires de Bruxelles, Roelants-Abraham, 1998.

[LIN 08] LINDSEY E., *Three-Dimensional Analysis – Data Profiling Techniques*, Data Profiling LLC, 2008.

[MAY 07] MAYDANCHIK A., *Data Quality Assessment*, Technics Publications LLC, Denville, NJ, 2007.

[MCG 08] MC GILVRAY D., *Executing Data Quality Projects : Ten Steps to Quality Data and Trusted Information*, Morgan Kaufmann, New York, NY, 2008.

[TEO 11] TEOREY J.T., *Database Modeling and Design*, 5e édition, Logical Design, Morgan Kaufmann, New York, NY, 2011.

Chapitre 2

Les critères pour la résolution d'identité appliqués aux personnes physiques

2.1. Contenu et objectifs

Ce chapitre propose une réflexion sur le traitement des données des personnes physiques et une démarche qualité pour la résolution d'identité dans le contexte d'une interface entre un référentiel et un système opérationnel contenant des données de personnes physiques.

Un rapprochement effectué à tort entre deux personnes physiques peut engendrer des dysfonctionnements non négligeables, ainsi que des coûts importants de recherche et de correction d'erreurs. On peut citer, à titre d'exemple, les conséquences, tant pour l'administration que pour l'administré, que peut avoir la facturation d'un service à la mauvaise personne. Il importe également que cette procédure soit standardisée afin de garantir une pratique cohérente entre les différents systèmes d'information (SI). Il existe enfin un potentiel d'économie au niveau de la capitalisation des traitements informatiques nécessaires pour fiabiliser ces rapprochements, ou simplement tendre vers un accroissement de la qualité des données.

D'autre part, tout traitement de données des personnes physiques doit se faire en accord avec la loi (par exemple, LIPAD en Suisse ou informatique et libertés en France).

Plus que technologiques, les difficultés sont d'ordre humain et organisationnel. Sujet transverse, par rapport à l'organisation des structures, il nécessite de convaincre

---

Chapitre rédigé par Nunzio Di Ruocco et Christine Pascal-Suisse.

divers services en expliquant l'intérêt qu'ils peuvent en retirer, alors même que celui-ci n'est pas forcément visible au premier abord. A cela s'ajoute le financement d'une telle entreprise et les difficultés réelles à trouver les bons interlocuteurs, qui doivent avoir à la fois la connaissance et le pouvoir de peser dans les décisions. La direction des SI doit apporter aide et soutien aux processus et structures mis en place pour la réussite de ce type de démarche.

Nous allons aborder notre besoin au travers de la problématique connue sous le nom de « résolution d'identité ».

Ce chapitre se propose donc d'énumérer les anomalies et les conséquences d'un mauvais rapprochement et de lister l'ensemble des techniques pour la résolution d'identité.

En outre, ce chapitre fournit les recommandations que toute application devrait satisfaire pour assurer des rapprochements fiables, ceci, afin de sensibiliser tous les acteurs qui travaillent tous les jours à la construction de systèmes d'information.

La section 2.5 de ce chapitre est dédiée aux critères concernant la classification des anomalies sur les données. Parmi les multiples critères existants, nous avons sélectionné les plus importants pour la résolution d'identité. Vous trouverez dans cette section leurs définitions et leurs relations avec la résolution d'identité.

Dans la dernière section, une proposition concernant la définition des niveaux de fiabilité pour un rapprochement de données est décrite.

2.2. Définitions

2.2.1. *La résolution d'identité*

La résolution d'identité encore appelée « résolution d'entité » est un processus qui permet de mettre en rapport différentes sources de données contenant des données personnelles (*données d'identités*) et d'effectuer un rapprochement entre ces données, même si le lien entre elles n'est pas évident. Il s'appuie sur des schémas de probabilité et analyse les données dans le but d'effectuer un rapprochement ou non, avec un certain degré de fiabilité. Les moteurs de résolution d'identité sont typiquement utilisés pour découvrir les risques, fraudes et conflits d'intérêt mais sont également des outils utiles aux exigences de l'intégration des données client (en anglais, *customer data integration* – CDI) et de la gestion des données de référence (en anglais *master data management* – MDM). Ces moteurs s'appuient sur un ensemble de faits et de règles pour pouvoir réaliser, en appliquant les règles en fonction des faits, la résolution d'identité, comme illustré figure 2.1.

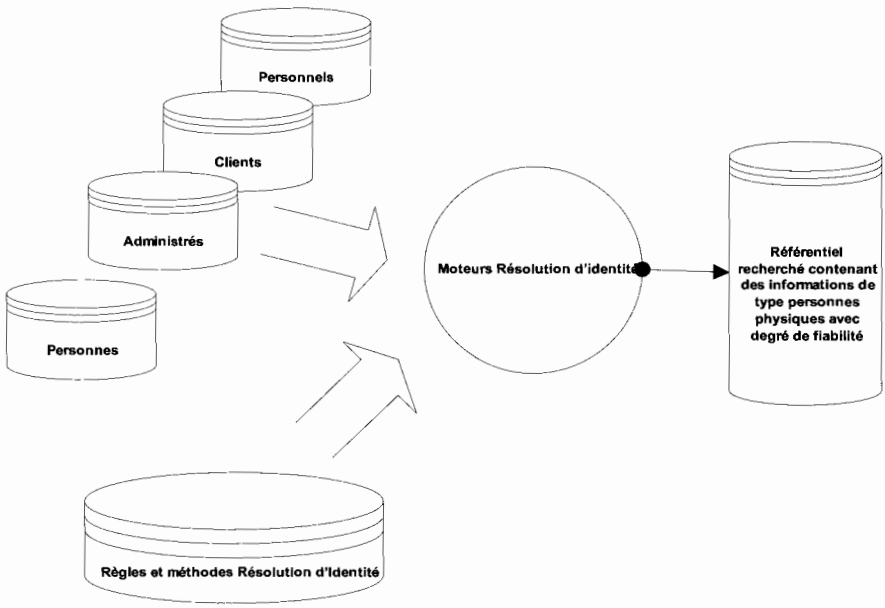

Figure 2.1. *Résolution d'identité*

2.2.2. *Différence entre appariement de données et résolution d'identité*

L'appariement (*matching*) de données est un élément essentiel de la résolution d'identité mais n'est pas la résolution d'identité. En effet, selon les données considérées, celles-ci peuvent effectivement correspondre mais ne sont pas suffisamment discriminantes pour résoudre l'identité des données.

Inversement, les données peuvent ne pas correspondre et pourtant il s'agit bien du même individu. Par exemple, dans l'utilisation du nom de jeune fille et du nom de femme mariée, les noms sont différents mais, en réalité, il s'agit de la même personne.

Quatre opérations permettent de différencier les deux cas selon le professeur John Talburt [TAL 10] :

– travailler avec, à la fois, des données structurées et non structurées ;

– utiliser des règles « métier » et des modèles conceptuels pour arbitrer entre les informations manquantes, conflictuelles et corrompues ;

– utiliser les liens entre les informations pour résoudre le rapprochement ;

– découvrir des relations non évidentes et des réseaux entre les entités.

2.2.3. *Différence entre résolution d'identité et identification d'entité*

L'identification d'entité est considérée comme un cas particulier de la résolution d'identité puisqu'elle nécessite l'application de tous les processus liés à celle-ci, avec la particularité de se référer à des données caractérisant l'identité d'une personne (la personne se trouve alors connue), ce qui n'est pas le cas si l'on considère le périmètre de la résolution d'identité [TAL 10].

En effet, la résolution d'identité permet d'associer deux références sans qu'il soit pour autant nécessaire de connaître l'identité de la personne à laquelle elles se réfèrent. C'est en cela que la résolution d'identité et l'identification d'entité se distinguent.

A titre d'illustration, si l'on considère deux œuvres picturales différentes et que l'on compare ces œuvres, on peut, de par la technique utilisée, l'époque, les pigments et d'autres éléments significatifs, les attribuer à un même artiste (on a alors réalisé la résolution d'identité) sans pour autant pouvoir donner un nom à cet artiste si ce dernier n'est pas, comme Van Gogh ou Raphaël, un artiste reconnu et est resté dans la liste des peintres inconnus. L'identification d'entité ne sera alors pas résolue. En revanche, elle sera effective si une des deux œuvres a déjà pu être attribuée à un artiste connu.

2.3. Problématique de la résolution d'identité

L'utilisation de données d'identité est une nécessité et doit répondre à un degré de fiabilité proche de la perfection, sous peine d'engendrer des conséquences préjudiciables pour la ou les personnes concernées ainsi que pour les organismes se référant à ces données liées à l'identité, en termes de coût, d'image, d'aide à la décision.

Lorsque l'on recherche une personne dans un système, la difficulté est de savoir si les données trouvées correspondent bien à la personne physique, dans le cas où un seul choix est possible. Il s'agit également, de choisir parmi plusieurs propositions, la bonne, si elle s'y trouve.

Il faut donc faire face à :

– des variantes de saisie, à des erreurs de saisie, à l'utilisation de diminutifs, à des permutations de mots, de lettres, ou à des fautes d'orthographe ;

– des différences de structure, de format et de localisation des données ;

– des doublons et des omissions ;

– des différences de langue.

Les principales difficultés sont les suivantes :

– comment connaître la fiabilité d'une source d'information ?

– comment identifier les critères permettant d'évaluer la fiabilité de cette source ?

– comment permettre la correction des sources sans équivoque ?

Même si le concept d'identité est étudié depuis longtemps en sciences humaines et sociales, la résolution d'identité reste difficile à réaliser ; principalement parce que les données manipulées sont ambigües de par leur structure et sujettes à contenir des erreurs ou des variations et parce que l'établissement des liens entre les différentes informations ne va pas de soi.

2.4. Les familles de critères impliqués dans la résolution d'identité

2.4.1. *Identification des familles de données*

La théorie de l'identité personnelle est définie comme étant la perception de soi en tant qu'individu.

Les théories de l'identité sociale divergent selon la vue psychologique ou la vue sociologique.

La théorie de l'identité sociale selon la vue psychologique (*psychologically-based of social identity – PSIT)* repose sur le processus cognitif et psychologique de la perception de soi d'un individu comme un membre de certaines catégories reconnues, telles que la nationalité, la culture, l'origine ethnique, le sexe, l'emploi. [TAJ 86].

La théorie de l'identité sociale selon la vue sociologique (*sociologically-based identity theory – SSIT*) se concentre sur les relations établies entre les acteurs sociaux qui jouent mutuellement des rôles complémentaires tels qu'employé-employeur, patient-médecin [DEA 03]. Le contexte social détermine les rôles spécifiques que chaque individu peut prendre. En effet, une personne peut avoir différents rôles dans son entourage : elle peut être médecin pour ses patients, patiente pour son médecin, employée pour l'hôpital, maire pour son village et citoyenne de son pays. L'identité sociale d'un individu dans ce sens est définie par le rôle basé sur les interactions entre l'individu et les gens environnants [STR 82].

De ces théories émergent donc des directions pour lesquelles nous sommes amenés à proposer différents critères et classifications. En effet, outre le fait que ces critères relèvent de l'une ou l'autre théorie, la représentation même de la donnée

(alphabétique ou numérique) a son importance, ainsi, bien sûr, que son contenu plus ou moins sensible (religion, race, santé, etc.).

2.4.2. *La classification selon les théories de l'identité personnelle et de l'identité sociale*

Issus de ces théories, les critères susceptibles de nous intéresser sont ceux qui vont nous permettre de distinguer un individu d'un autre. Vous trouverez (la liste n'étant pas exhaustive) dans le tableau 2.1, les critères caractérisant l'identité personnelle ainsi que l'identité sociale [XUJ 07, CLA 94].

Théorie	Groupe de critères de données	Critère
Identité personnelle	Les données personnelles propres	nom prénoms date de naissance lieu de naissance nom de jeune fille de la mère adresse numéro de téléphone numéro d'identité national
	Les caractéristiques physiques	poids taille sexe couleur de cheveux couleur des yeux marque visible telle que tatouage, cicatrice, grain de beauté
	Les données biométriques, données uniques par individu	empreintes ADN iris géométrie de la main voix
	Les données biographiques	études parcours professionnel parcours médical
Identité sociale		personne(s) interférant directement avec l'individu
		appartenance à un groupe social
		rôle dans la société

Tableau 2.1. *Classification des critères selon les théories d'identité personnelle et sociale*

2.4.3. *Classification selon la représentation de la donnée*

Certaines données de la personne physique sont des données alphabétiques. Il est difficile de formaliser ce type d'information qui se trouve par le fait, être la source d'un grand nombre d'anomalies (voir la section 2.5) [XUJ 07]. Le tableau 2.2 présente les différentes représentations.

Type d'anomalie	Exemple
Les données numériques	numéro de téléphone numéro d'identité national n° de rue les données biométriques
Les données alphabétiques	nom prénoms adresse lieu de naissance

Tableau 2.2. *Classification des critères selon la représentation de la donnée*

2.4.4. *Classification selon les lois*

Ces lois (par exemple, LIPAD[1] en Suisse ou informatique et libertés[2] en France) ont été établies en vue de régir le traitement des données personnelles par les institutions de manière à protéger les droits fondamentaux des personnes physiques (ou morales).

On distingue les données personnelles sensibles concernant :

– les opinions ou activités religieuses, philosophiques, politiques, syndicales, culturelles ;

– la santé, la sphère intime, l'appartenance à une race ou ethnie ;

– des mesures d'aide sociale ;

– des poursuites ou sanctions pénales ou administratives.

---

1. LIPAD : loi sur l'information du public, l'accès aux documents et la protection des données personnelles A 2 08 (loi du 5 Octobre 2001, entrée en vigueur le 1er Mars 2002, révisée en octobre 2008 pour une entrée en vigueur au 1er Janvier 2010) (Suisse).
2. Informatique et libertés : loi française 78-17 de 1978, révisée en 1991 et 2004, relative à l'informatique, aux fichiers et aux libertés, réglementant la collecte et la conservation d'informations relatives aux personnes.

Ceci dit, toute donnée faisant référence à une personne physique est potentiellement sensible selon l'usage qui peut en être fait.

2.5. Les anomalies

Une des difficultés dans le rapprochement d'identité se situe dans la multitude des possibilités de saisie d'une même information, notamment des données alphabétiques, selon l'utilisation ou non d'abréviations, de codifications (pour un pays, par exemple), de différences d'ordre des mots, d'erreurs de saisie [CLA 94, IST 10]. Le tableau 2.3 présente différents types d'anomalies qui affectent une donnée.

Type d'anomalie	Exemple
variantes de saisie	
erreurs de saisie ou de transcription	
utilisation d'abréviations	avenue ou av. ou ave
utilisation des initiales	Henri Martin ou H. Martin
utilisation de surnoms	Bill pour William, Bob pour Robert
permutation de mots	rue Henri Martin ou rue Martin Henri
permutation de lettres	rue Herni Martin
fautes d'orthographe	
données non-nettoyées	
différences de structure, de format et de localisation des données	
omissions : données non renseignées ou imparfaitement	
doublons	
des différences de langues	Genf et Genève ou jj/mm/aaaa et mm/jj/aaaa
des différences syntaxiques	ordre des mots
des codifications différentes	codification pays sur 3 car ou sur 2 car par exemple Afghanistan AFG ou AF, Afrique du sud ZAF ou ZA

Tableau 2.3. *Les anomalies*

On peut encore souligner, les anomalies résultantes des usages selon les différents pays, que l'on peut illustrer par la problématique du nom et prénom.

Une des données primordiales identifiant une personne physique est bien évidemment son nom. Mais ce dernier reste sujet à des utilisations variées selon les pays et les traditions, avec par exemple utilisation ou non, du nom du mari pour les femmes mariées ou bien ajout du nom du mari à celui du nom de jeune fille ou encore utilisation d'un nom d'usage différent du nom légal ; concernant les prénoms, l'utilisation du 2e ou 3e prénom plutôt que du 1er est possible. Ceci a pour résultat l'enregistrement potentiel de noms différents pour une même personne.

2.6. Techniques appliquées pour la résolution d'identité

Ces techniques sont de deux types :

- les techniques probabilistes ;
- les techniques déterministes.

La réussite de ces techniques dépend très fortement de la qualité des données. En effet, lorsque la qualité des données est faible, due à des erreurs ou des valeurs manquantes, il devient difficile de pouvoir confronter les données de manière efficace avec un résultat satisfaisant, puisqu'elles délivrent des informations incorrectes et insuffisantes [BOO 09]. Mais malheureusement, ces techniques restent d'une efficacité toute relative, également en raison de l'utilisation presque exclusive des données personnelles (celles le plus souvent disponibles), sans pouvoir utiliser de données sociales plus difficilement et plus rarement collectées [XUJ 07].

2.6.1. *Approche probabiliste (via un moteur d'inférence)*

Elle consiste en la construction d'échantillons représentatifs des données pour permettre au système d'apprendre la logique de classification des identités. Sur ces données, le système effectue des statistiques de fréquence et de légitimité pour des schémas d'identités communes et non communes. Ces statistiques sont ensuite utilisées pour déterminer le niveau de confiance que l'on peut accorder à la concordance des identités traitées.

C'est ainsi que des enregistrements comprenant des similarités dans un schéma de données non communes ont une forte probabilité de concerner la même personne.

Une attention toute particulière doit être accordée à la pérennité des schémas au regard des données traitées. En effet, la pertinence de ces évaluations diminue si les données réelles dévient des schémas utilisés pour le traitement. Une adéquation fréquente des schémas est nécessaire [BOO 09].

2.6.2. *Approche déterministe (sur la base d'heuristiques)*

Ces techniques consistent à élaborer des règles qui permettront de déterminer si deux enregistrements concernent une même personne. On peut, en effet, considérer à titre d'exemple, que si le nom de chaque enregistrement est le même, le prénom proche (Anne-Marie et Anna) et la date de naissance identique, il s'agisse bien de la même personne. La limite d'efficacité de telles techniques est atteinte lorsqu'elles reposent sur un nombre élevé de types de données, ce qui complexifie les règles.

De plus, cette méthode génère un taux important de faux négatifs (voir définition dans la section 2.6.3).

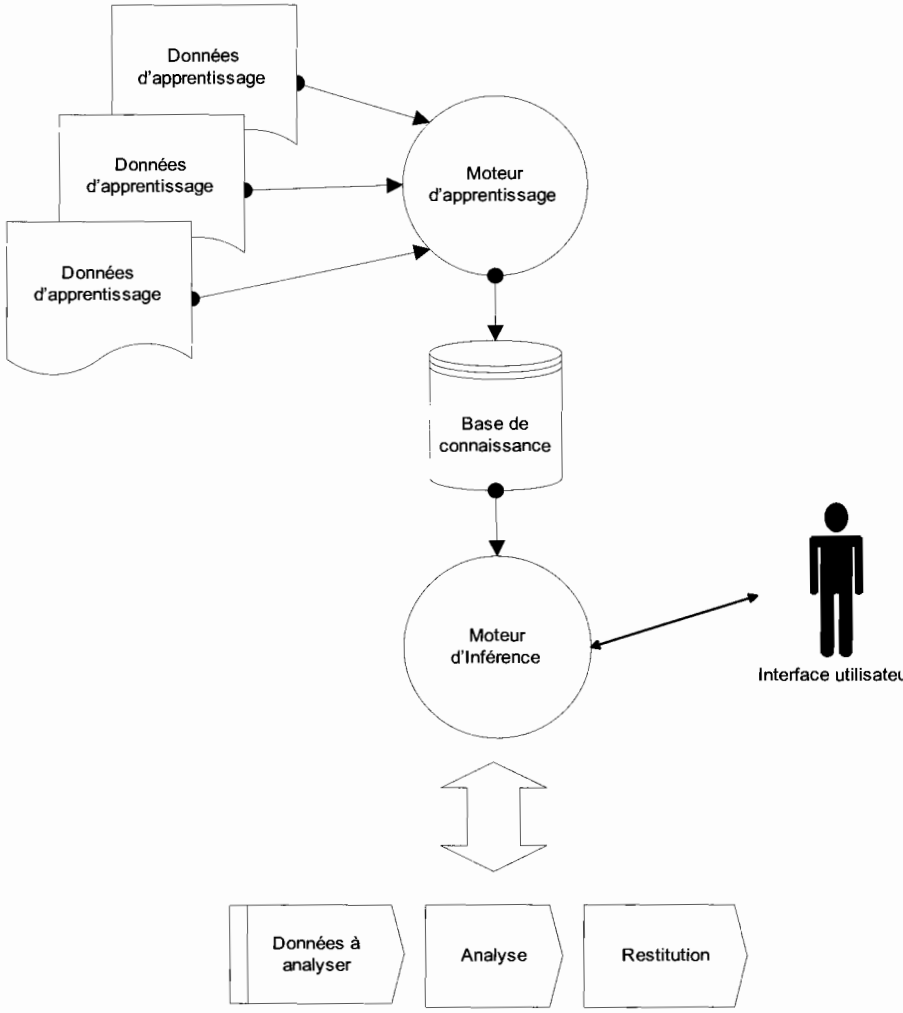

Figure 2.2. *Approche probabiliste*

Bien que les approches probabilistes soient plus efficaces que les approches heuristiques, elles nécessitent un niveau de qualité des données important. Des enregistrements avec des données non renseignées affectent les performances des algorithmes de comparaison [XUJ 07].

2.6.3. *Les résultats d'un appariement en vue de la résolution d'identité*

Le résultat d'un appariement peut correspondre à quatre qualificatifs :

– vrai positif, lorsque l'appariement comporte des enregistrements qui correspondent aux critères de rapprochement et qui sont effectivement conformes à la recherche effectuée ; c'est-à-dire qu'ils correspondent à la personne dans la réalité. C'est le résultat que l'on cherche à obtenir ;

– vrai négatif, lorsque l'appariement ne s'est pas effectué avec des enregistrements qui ne correspondent pas aux critères de rapprochement et qui sont effectivement non conformes à la recherche effectuée ; c'est-à-dire qu'ils ne correspondent pas à la personne dans la réalité. C'est également un résultat attendu ;

– faux positif, lorsque l'appariement comporte des enregistrements qui correspondent aux critères de rapprochement mais qui ne sont pas conformes à la recherche effectuée ; c'est-à-dire qu'en réalité, ils ne concernent pas la personne. Ce résultat va induire des erreurs ;

– faux négatif, lorsque l'appariement ne s'est pas effectué avec des enregistrements qui ne correspondent pas aux critères de rapprochement mais qui sont conformes à la recherche effectuée ; c'est-à-dire qu'en réalité, ils concernent la personne. Ce résultat a occulté le rapprochement et n'est bien évidemment pas attendu.

2.6.4. *Vers la résolution des faux négatifs*

Les faux négatifs sont les enregistrements qui ont été rejetés comme étant des enregistrements ne réalisant pas la correspondance demandée, mais qui en fait font bien référence à une même personne physique [TAL 10].

Par exemple, lorsque les informations étudiées concernent Jade Lebrun, rue de Carouge, 25 et Jade Leblanc, route de Meyrin, 125, on constate que les noms et adresses sont différents mais ces informations concernent néanmoins la même personne si Jade Lebrun a changé de nom en épousant M Leblanc.

Voici deux approches qui permettent de réduire ces faux négatifs :

– possibilité d'agrandir le cadre des critères à comparer, nom des parents par exemple, mais ceci complexifie le processus de mise en correspondance des données. A cela s'ajoutent les questions de savoir si le critère est déterminant, correct, et donne une possibilité supplémentaire d'une valeur manquante ;

– utiliser une source de données autre qui contient explicitement une information déterminante.

2.6.5. *La qualité pour déterminer la donnée à retenir*

Bien que n'étant pas directement une technique pour la détermination d'identité, ce concept contribue à l'évaluation et au choix de la donnée à retenir.

Lorsqu'il y a ambiguïté dans le choix d'une donnée, il va être nécessaire d'avoir des indicateurs de qualité supplémentaires pour prendre une décision. Ces indicateurs peuvent être à trois niveaux [LOS 10].

2.6.5.1. *Qualité de la source*

Différentes sources de données ne présentent pas les mêmes caractéristiques de conformité par rapport aux attentes métiers.

Ces attentes métiers sont définies par des dimensions de qualité des données.

L'évaluation globale de la qualité des sources de données (ou des ensembles de données) permet de les comparer entre elles.

Lorsqu'un enregistrement existe dans deux sources de données, il est raisonnable de sélectionner les valeurs de la source qui a obtenu globalement le meilleur score de qualité des données.

2.6.5.2. *Qualité de l'enregistrement*

De manière similaire, chaque enregistrement peut être vu suivant son propre contexte.

Les règles de qualité des données peuvent être mesurées au niveau de granularité de l'enregistrement (par exemple : complétude, consistance, conformité aux contraintes de validation du domaine, etc.) et ces mesures peuvent fournir une évaluation relative de la qualité d'un enregistrement par rapport à un autre. Aussi, lorsque l'on constate une différence entre deux enregistrements de deux sources de données de qualités égales, il faut évaluer la qualité des enregistrements eux-mêmes.

2.6.5.3. *Qualité de la donnée*

Lorsque les sources sont de qualité équivalente et les enregistrements aussi, le niveau de précision suivant concerne les valeurs elles-mêmes.

Il existe des dimensions intrinsèques rattachées aux valeurs, comme la consistance syntaxique, la consistance sémantique, la validité dans le domaine, etc.

Voici quelques suggestions pour comparer les valeurs recueillies dans des enregistrements correspondants :

– actualisation de la valeur, les valeurs les plus récentes peuvent objectivement être considérées comme meilleures ;

– densité de contenu, les valeurs contenant plus d'information peuvent être considérées comme meilleures. Un prénom complet contient plus d'information qu'une initiale et peut être considéré comme ayant plus de valeur ;

– fréquence de la valeur, lorsque de multiples enregistrements sont suspectés d'être des doublons, les valeurs qu'on retrouve avec la plus grande fréquence peuvent être considérées comme plus fiables que les autres valeurs.

2.6.5.4. *Qualité par la définition de règles*

D'une façon générale sur quoi repose la qualité des données ?

On peut caractériser les données selon les axes suivants :

– complétude, la valeur de la donnée doit être renseignée. Elle doit être liée à toutes les entités pour lesquelles elle est nécessaire ;

– pertinence, la donnée répond aux besoins actuels et futurs des utilisateurs et est utile dans l'exécution des processus métiers. Sa valeur est conforme à l'usage de l'attribut pour lequel il a été défini ;

– unicité ou absence de doublon, une entité du monde réel n'est représentée que par une donnée unique au sein du système ;

– actualité, la donnée correspond à la réalité du moment et n'est pas obsolète ;

– consistance, la donnée, si elle est répliquée, doit contenir les mêmes valeurs ;

– accessibilité, la donnée doit être facile à localiser et à accéder ;

– disponibilité en temps voulu ; la donnée doit être présente et à jour au moment opportun ;

– clarté ou non ambiguïté, la donnée ne doit pas porter à confusion. Il est important que l'utilisateur ait une compréhension précise et sans ambigüité sur la donnée. La donnée bien documentée et accompagnée de descriptions claires des concepts, de sa structure et de son usage représente un avantage quant à l'efficacité de son utilisation.

Une donnée utile doit pouvoir répondre à ces exigences. Elle peut y répondre aux travers de règles définies appartenant à chacun de ces axes.

Il sera possible de définir un niveau de qualité objectif des données au travers de la réalisation des règles définies pour chacun de ces axes.

Par exemple, si on définit la règle suivante : est considéré comme doublon tout enregistrement dont le nom prénom et date de naissance sont identiques, cette règle appartient à l'axe de qualité Unicité. Selon le taux de vérification de cette règle, il faudra définir les seuils pour allouer un niveau de qualité aux sources de données étudiées.

2.7. Traçabilité d'un numéro d'identité national et des attributs liés, exemple en Suisse

En ce qui concerne les personnes physiques, certains pays attribuent un numéro d'identification nationale (par exemple, le numéro de sécurité sociale en France, NAVS13[3] en Suisse, etc.) propre à chaque individu. Certaines règles et recommandations, quant à l'utilisation de ces numéros ont été émises et peuvent permettre de déterminer une résolution d'identité plus évidente. Cependant l'utilisation de ces numéros est très encadrée par la loi. Toutefois le niveau de qualité de la résolution d'identité sera accru si l'utilisation de tels numéros est permise, le processus de traçabilité étant assuré et connu. C'est pourquoi, à titre d'exemple, nous donnons le descriptif des recommandations, afin de préserver le lien entre les données de l'individu et le numéro concerné indépendamment des données « métier ».

Lors de la mise en place de l'attribution d'un numéro d'identité national (NAVS13), la centrale de compensation (CdC[4]) a effectué des recommandations quant aux règles à suivre pour tout utilisateur du numéro et ce, afin de préserver la qualité du lien entre le numéro et la personne physique. Elle enjoint tous les utilisateurs de ce numéro à les suivre. L'une de ces recommandations est de considérer le NAVS13 et les données officielles d'identification comme un bloc indissociable [CDC 10].

Les données qui doivent obligatoirement accompagner le NAVS13 sont :

- le nom de famille officiel ;
- le(s) prénom(s) officiel(s) ;
- la date de naissance ;

3. NAVS13 : numéro assurance vieillesse et survivants à 13 chiffres (Suisse).
4. CdC : centrale de compensation AVS/AI (assurance vieillesse et survivants/assurance invalidité), organe central fédérateur du système AVS/AI, disposant du monopole pour l'attribution et la gestion des NAVS13 en Suisse.

- le sexe ;
- la nationalité.

Bien évidemment, d'autres attributs peuvent être ajoutés.

Le tableau 2.4 représente les données officielles d'identification appartenant au registre national de référence (UPI[5] *Unique Person Identification*), accompagnant le numéro d'identification nationale d'une personne ; le tableau 2.5 dit table « métier ».

Local_ID	NAVS13	Nom (UPI)	Prénom (UPI)	Date Naiss. (UPI)	Sexe (UPI)	Nationalité (UPI)
ID1	756xxxx	Leblanc	Anne-Sophie	12.01.1970	F	8100(CH)
ID2	756yyy	Schmidt	Marc	25.07.1953	M	8207 (D)

Tableau 2.4. *Table de gestion du NAVS13 au sein d'un SI*

Local_ID	Nom	Prénom	Adresse	Assureur	...	
ID1	Leblanc	Anne				
ID2	Schmidt	Marc				

Tableau 2.5. *Exemple d'une table « métier »*

Les tableaux 2.4 et 2.5 illustrent la possibilité de conserver les informations UPI (*unique person identifier database*), en conservant la spécificité des tables « métier », le lien se faisant avec le Local_ID ; l'intérêt étant de conserver une référence fiable entre l'identifiant et les données caractérisant une personne.

Une autre recommandation de la CdC est de s'assurer, au minimum, de la validité actuelle du NAVS13 (qui peut, pour diverses raisons, avoir été invalidée par la CdC) et, dans l'idéal, de l'actualité des données personnelles d'identification.

5. *UPI : unique person identification database*, registre national de référence pour le numéro d'identification national.

Il conviendra donc de tenir compte de la présence ou non de ces informations UPI lors de la définition des niveaux de qualité.

Attention : toute utilisation du NAVS13 est régie par la LIPAD.

2.8. Exemples de cas d'usage

2.8.1. *Le contexte*

Nous disposons de plusieurs sources ayant en commun des données de personnes physiques avec des attributs de données parfois différents. Il faudra pouvoir rapprocher ces données de personnes physiques pour établir un lien entre les différentes sources et n'utiliser ce lien que s'il est très fiable.

2.8.2. *Les données*

Les sources contiennent des données appartenant à la famille des données personnelles, plus particulièrement les données personnelles propres, ainsi qu'à la famille des données sociales. Chacune de ces familles contient des données alphabétiques ou numériques, codifiées ou non, sensibles (selon les lois de protection des données personnelles) ou non. Elles peuvent également être issues de référentiels nationaux des identifications.

Le tableau 2.6 présente des données classées selon différentes familles d'appartenance.

2.8.3. *Cas d'usage 1 : détection des doublons*

Elle consiste en la recherche, au sein d'une même source, des enregistrements concernant une même personne.

Lorsque l'on parle de doublons, on parle d'informations figurant plus d'une fois dans une source. Il existe différents types de doublons :

– les doublons absolus : les informations sont strictement identiques. Ceci ne peut pas arriver lorsqu'une clé est utilisée dans la source car celle-ci ne peut être qu'unique. Ces doublons sont faciles à éliminer sans ambiguïté. Sans doute ces cas devraient être rares ;

	Données personnelles	**Données sociales**	**Sensibles selon la loi** (ex. : LIPAD)	**Données du référentiel national d'identification**
Alphabétiques	Nom	Syndicat	Syndicat	Nom officiel
	Nom de célibataire	Association	Association	Nom de célibataire
	Prénom	Médecin	Médecin	Nom selon passeport
	Lieu de naissance		Ethnie	Autre nom connu
		Nom père	Religion	Prénom(s)
		Nom mère	Condamnation	Nom et prénom du père
			Maladie	Nom et prénom de la mère
	Aide sociale		Opinion Politique	Lieu de naissance
	Maladie			
Numériques	Date de naissance			Date de naissance
	numéro identité national			
	Id registre			
		Id registre Père		
		Id registre Mère		
Codifiées (table référence)	Sexe			Sexe
	Etat civil			
	Pays de naissance			Pays de naissance
	Nationalité			Nationalité
				Autre nationalité

Tableau 2.6. *Famille de données*

n°	**Nom**	**Prénom**	**Date naissance**
ID1	Leblanc	Anne	11101991
ID1	Leblanc	Anne	11101991

Tableau 2.7. *Les doublons absolus*

– les doublons relatifs : les informations sont strictement identiques sauf celles concernant la clé (dans notre exemple du tableau 2.8 : n° identifiant). Si toutes les informations sont identiques, les doublons pourront être éliminés sans ambiguïté, exception faite pourtant si l'identifiant servant de clé se trouve être le numéro d'identification national. Si tel est le cas, se posera alors le problème de savoir quel enregistrement il faut garder, c'est-à-dire quel numéro d'identification national ;

n°	Nom	Prénom	Date naissance
ID1	Leblanc	Anne	11101991
ID6	Leblanc	Anne	11101991

Tableau 2.8. *Les doublons relatifs*

– les doublons approximatifs : les informations sont similaires. Concernant cette catégorie de doublons, il faut déterminer des règles afin de pouvoir déterminer quels sont les enregistrements qui concernent une même personne, pouvoir mesurer un niveau de fiabilité du rapprochement et déterminer selon le niveau de fiabilité si le rapprochement peut être effectif ou non.

N°	Nom	Prénom	Date naissance
ID1	Leblanc	Anne	11101991
ID3	Leblanc	Anne	11101991
ID5	Leblanc	Annie	11101991
ID7	Leblanc	Anne Sophie	11101990
ID9	Leblanc	Sophie	11101990

Tableau 2.9. *Les doublons approximatifs*

2.8.4. *Cas d'usage 2 : rapprochement des enregistrements*

Il s'agit de rapprocher des enregistrements (ou *record linkage*) issus de plusieurs sources de données afin de globaliser les informations pour une même personne.

Il faudra déterminer les règles selon lesquelles il sera possible de rapprocher des informations de deux sources différentes en étant sûre qu'il s'agisse d'informations concernant la même personne. Dans un premier temps il faudra déterminer quelles sont les informations communes. Le numéro d'identification national sera-t-il disponible ? Si oui sera-t-il accompagné des informations du registre national d'identification ? Si non, le rapprochement pourra-t-il se faire sur les seules données issues des différentes sources ?

2.8.5. *Cas d'usage 3 : identification d'une personne*

Il s'agit de rechercher un enregistrement dans une source correspondant à une personne physique.

Le niveau de fiabilité est à déterminer selon le nombre d'attributs correspondants pour des données de types nom, prénom, date de naissance ou bien numéro d'identification national seul.

2.8.6. *Cas d'usage 4 : recherche des données non renseignées*

Il s'agit de rechercher des données non renseignées qui peuvent engendrer un grand nombre de « faux négatifs » qu'il faut chercher à réduire afin d'augmenter le niveau de qualité.

2.9. Les niveaux de fiabilité pour la résolution d'identité

2.9.1. *Proposition*

Il convient tout d'abord de déterminer un ensemble d'attributs suffisant pour pouvoir caractériser un individu, puis d'appliquer un processus de comparaison de ces attributs en conférant à chacun une valeur qui reflète la similarité de la comparaison. Il sera alors possible de déterminer des niveaux de fiabilité de ce rapprochement par la combinaison de différents seuils pour chacun des attributs.

Le tableau 2.10 est une représentation de la « matrice des niveaux », qui est constituée :

– en ligne, des différents niveaux de fiabilité estimés, notés : *N1, N2, N3..., Nm* ;

– en colonne, des attributs caractérisant un individu, notés : *A1, A2, A3, A4... An* ;

– à l'intersection, la valeur minimale (ou seuil) *S(i,j),* de 0 à 100 %, représente une similarité de comparaison pour l'attribut considéré (score). Cette valeur minimale, pour l'attribut *j,* est déterminée selon le degré de réalisation de la règle *R(j)* de comparaison définie et appliquée l'attribut *j*, et en fonction également des seuils des autres attributs.

Niveau	A1	A2	A3	A4	A...	A...	An
N1	S(1,1)	S(1,2)	S(1,3)	S(1,4)	S(,)	S(,)	S(1,n)
N2	S(2,1)	S(2,2)	S(2,3)	S(2,4)	S(,)	S(,)	S(2,n)
N3	S(3,1)	S(3,2)	S(3,3)	S(3,4)	S(,)	S(,)	S(3,n)
N4	S(4,1)	S(4,2)	S(4,3)	S(4,4)	S(,)	S(,)	S(4,n)
N...	S(,1)	S(,2)	S(,3)	S(,4)	S(,)	S(,)	S(,n)
N...	S(,1)	S(,2)	S(,3)	S(,4)	S(,)	S(,)	S(,n)
Nm	S(m,1)	S(m,2)	S(m,3)	S(m,4)	S(m,)	S(m,)	S(m,n)

Tableau 2.10. *Matrice des niveaux de fiabilité*

Un niveau de fiabilité *i*, est caractérisé par l'ensemble des valeurs minimales *S(i,j)* requises pour l'ensemble des *n* attributs. Si le seuil est inexistant, l'attribut n'est alors pas utilisé pour déterminer le niveau de fiabilité. A un même niveau de fiabilité peut correspondre différents ensembles de valeurs minimales requises.

La détermination des règles et des niveaux doit se faire en fonction d'échantillons de données afin d'en vérifier leur cohérence. Elle peut, et sans doute, doit évoluer au fur et à mesure de l'expérience acquise. Celle-ci doit également être accompagnée de mesures quant à la complétude des données, car des données incomplètes réduisent sensiblement les niveaux de qualité.

2.9.2. *Exemple*

Si l'on considère les attributs nom, prénom, date de naissance, lieu de naissance, père, mère comme étant l'ensemble d'attributs caractérisant un individu.

Pour chaque attribut, il faut définir un ensemble de règles de comparaison avec des valeurs reflétant la similarité des comparaisons. Ensuite les différents appariements « attribut – seuil » détermineront les niveaux de fiabilité.

Il existe différentes techniques permettant de comparer des chaînes de caractères, telles que la distance de Hamming, le Bigram ou encore la distance de Jaro. En appliquant, sur chaque attribut de notre ensemble d'attributs caractérisant un individu, la distance de Jaro par exemple, on obtiendra pour chacun d'entre eux des scores de 0 à 100 %, 100 % caractérisant une similarité parfaite.

Après avoir effectué ce travail sur chaque attribut, il faut déterminer les combinaisons d'attribut et de seuil qui échelonneront les niveaux de fiabilité.

Dans notre exemple nous avons considéré six attributs. Ces attributs sont issus des catégories de données personnelles (nom, prénom, date de naissance, lieu de naissance) et sociales (père, mère). Les règles pour chaque attribut :

– nom, prénom, père, mère, lieu de naissance : distance de Jaro ;

– date de naissance : distance de Hamming (processus recommandé pour les chaines numériques).

Nous voulons les quatre niveaux de fiabilité :

– niveau 1 : niveau pour lequel toutes les catégories, donc tous les attributs, sont à 100 %, afin d'être capable d'identifier les enregistrements pour lesquels aucune ambiguïté n'existe ;

– niveau 2 : niveau pour lequel le seuil demandé pour une catégorie d'attributs sera moindre mais compensé par le seuil atteint par les attributs de l'autre catégorie. Ceci permettra d'identifier les enregistrements pour lesquels le rapprochement des données peut être considéré pratiquement comme certain ;

– niveau 3 : niveau d'exigence moindre qui permet d'identifier les enregistrements pour lesquels un rapprochement de données pourrait être effectué ;

– *default* : aucun rapprochement n'est à considérer. Ce sont tous les enregistrements qui ne satisfont pas les seuils des autres niveaux.

Suite à notre définition des niveaux, voici la matrice correspondante dans le tableau 2.11.

Niveau	Nom	Prénom	Date de naissance	Lieu de naissance	Père	Mère
Niveau 1	100	100	100	100	100	100
Niveau 2	100	100	100	100	70	70
	85	100	100		100	100
	100	85	100		100	100
	90	90	80	100	100	100
Niveau 3	70	70	70			
	70	70			80	
Default	0	0	0	0	0	0

Tableau 2.11. *Exemple d'une matrice des niveaux de fiabilité*

Le niveau 1 s'entend comme étant le meilleur niveau de fiabilité, avec un échelonnement vers des niveaux de fiabilité décroissante.

On constate qu'un même échelon de niveau peut être atteint avec des ensembles de combinaisons « attribut-seuil » différents. La diminution d'un seuil pour un attribut étant compensée par l'augmentation du seuil d'un autre attribut.

2.10. Facteurs-clés

La résolution d'identité est un sujet complexe.

Arriver à une résolution d'identité réelle dépend :

– de la détermination des attributs significatifs pour cette résolution ;

– du nombre de ces attributs significatifs ; plus ils seront nombreux, plus la résolution sera fiable mais moins elle pourra être établie ;

– de la qualité de ces attributs ;

– de la qualité des sources ;

– de la détermination des règles à appliquer et de l'évaluation des seuils de réalisations.

2.11. Bibliographie

[BOO 09] BOOKER Q.E., « Identity Resolution in Criminal Justice Data », *Journal of Information assurance and Security*, juin 2009.

[CDC 10] CENTRALE DE COMPENSATION CDC, Gestion du NAVS13 dans les registres Tiers, Confédération suisse, Département fédéral des finances, 2010.

[CLA 94] CLARKE R., « Human Identification in Information Systems : Management Challenges and Public Policy Issues », *Information Technology & People,* vol. 7, n° 4, p. 6-37, décembre 1994.

[DEA 03] DEAUX K., MARTIN D., « Interpersonal networks and social categories : Specifying levels of context in identity processes », *Social Psychology Quarterly*, 2003.

[LOS 10] LOSHIN D., « Identity Resolution, Cleansing, and Survivorship », *Community Of Experts – DataFlux,* avril 2010.

[STR 82] STRYKER S., SERPE R.T., « Commitment, identity salience, and role behavior : Theory and research example », dans W. Ickes et E.S. Knowles (dir.), *Personality, Roles, and Social Behavior*, Springer-Verlap, New York, 1982.

[TAJ 86] TAJFEL H., TURNER J.C., « The social identity theory of inter-group behavior », dans S. Worchel et L.W. Austin (dir.), *Psychology of Intergroup Relations*, Nelson-Hall, Chicago, 1986.

[TAL 10] TALBURT J., *Entity Resolution and Information Quality*, Morgan Kaufmann, Burlington, 2010.

[XUJ 07] XU J., WANG G.A., LI J.. CHAU M., « Complex Problem Solving : Identity Matching Based on Social Contextual Information », *Journal of the Association for Information Systems*, octobre 2007.

Chapitre 3

La qualité des modèles de données

3.1. Introduction

La dépendance des entreprises et des organisations vis-à-vis de leurs systèmes d'information (SI) n'est plus à démontrer. Cette réalité conduit les décideurs à assurer une qualité acceptable des systèmes d'information. Une des caractéristiques principales de la qualité de ces systèmes est sa nature multidimensionnelle. Stylianou et Kumar [STY 00] caractérisent la qualité des systèmes d'information au moyen de six dimensions : l'infrastructure, les logiciels, les données, l'information, l'administration et le service rendu. En particulier, Stylianou distingue la qualité des données à l'entrée du SI de la qualité de l'information en sortie. Cette nuance n'est pas reprise dans les autres approches qui ne différencient pas données et information en termes de qualité.

La qualité de l'infrastructure se définit par la capacité des matériels, des réseaux et des logiciels de base à assurer le niveau de performance et de sécurité requis par les administrateurs réseaux et systèmes et plus généralement par les responsables des systèmes d'information.

La qualité des logiciels se définit, quant à elle, par l'aptitude de ces derniers à satisfaire les besoins des utilisateurs. Elle correspond généralement à une appréciation globale du logiciel fondée sur des facteurs de qualité. De nombreuses dimensions de la qualité des logiciels ont été suggérées. Usrey et Dooley [USR 96] proposent une synthèse des dimensions de tous les modèles de qualité qui font référence, parmi lesquelles, citons la facilité d'accès et d'utilisation du programme,

---

Chapitre rédigé par Samira SI-SAÏD CHERFI, Jacky AKOKA et Isabelle COMYN-WATTIAU.

l'esthétique de l'interface, la flexibilité, la robustesse, etc. Un facteur de qualité peut être mesuré à l'aide de métriques. Gaffney [GAF 81] définit la métrique comme « une mesure mathématique et objective, sensible aux différences inhérentes aux caractéristiques des logiciels ». Ce sont des mesures quantitatives pour un attribut donné. Elles peuvent être des échelles auxquelles sont associées des règles et des méthodes applicables lors de la mise en œuvre du processus de mesure. On différencie le processus de mesure direct (fondé sur le nombre de lignes de code, la taille mémoire, la vitesse d'exécution, etc.), de celui qui est indirect, fondé quant à lui, sur les fonctionnalités, la complexité, la fiabilité, etc. On différencie les métriques externes permettant de mesurer les caractéristiques externes du logiciel (par exemple, le temps moyen entre deux pannes pour mesurer la fiabilité), des métriques internes qui mesurent les caractéristiques internes du logiciel, telles que le pourcentage de changement des spécifications fonctionnelles. Basili, Caldiera et Rombach [BAS 94] proposent un cadre, appelé GQM pour *Goal-Question-Metric*, qui exprime les buts de l'évaluation et génère les questions permettant de s'assurer que ces buts sont atteints. Chaque question est analysée à l'aide de métriques.

La qualité des données ou *qualité de l'information* constitue la troisième dimension. Les organisations s'appuient et dépendent des données pour répondre à leurs besoins opérationnels, stratégiques et réglementaires. Ces données doivent répondre à des exigences de qualité. La qualité des données peut être mesurée selon différents critères objectifs et quantitatifs. Elle peut être représentée et visualisée sous la forme d'un tableau d'indicateurs. De nombreuses dimensions ont été proposées pour caractériser les multiples facettes de la qualité des données, notamment :

– l'exactitude qui se mesure en détectant le taux de valeurs correctes dans la base de données ;

– la complétude ou taux de valeurs non manquantes dans la base ;

– l'actualité traduit le taux de valeurs non obsolètes dans la base de données ;

– la fraîcheur qui se mesure par une comparaison entre la date de saisie et la date courante ;

– la cohérence liée à un ensemble de contraintes (ou règles métier).

La qualité de l'administration du SI, relative au management de la fonction SI, se définit par la capacité de la direction du SI à assurer une gouvernance de ce dernier en alignement avec la politique de l'entreprise. Cela revient à assurer le niveau de qualité requis par les processus de gouvernance des SI, notamment décrits dans le référentiel COBIT [COB 11]. Ce dernier distingue quatre familles de processus SI :

– les processus de planification et d'organisation (évaluation des risques, gestion de la qualité, gestion des investissements, définition d'un plan informatique stratégique, gestion des ressources humaines informatiques, etc.) ;

– les processus d'acquisition et de mise en place (identification des solutions, acquisition et maintenance des logiciels d'application, développement et maintenance des procédures informatiques, installation et validation des systèmes, gestion des changements, etc.) ;

– les processus de distribution et de support (définition des niveaux de service, gestion des services assurés par des tiers, gestion de la performance et de la capacité, gestion des configurations, gestion de la sécurité, gestion de l'exploitation, etc.) ;

– les processus de surveillance (surveillance des processus, évaluer l'adéquation du contrôle interne, etc.).

La qualité de l'administration du SI peut être évaluée en appliquant à ces processus une approche d'évaluation de la qualité de processus.

La qualité de service, au sens de la qualité du service rendu au « client » du SI, peut être mesurée à l'aide de facteurs et de critères de qualité définis par le référentiel ITIL [ITI 11]. Ce dernier propose des règles destinées à améliorer l'efficacité et la bonne utilisation des ressources informatiques, par le regroupement et l'enrichissement de « bonnes pratiques ». De plus, il permet d'aider les directions des systèmes d'information à optimiser leur production de services, afin que ces derniers soient alignés aux besoins des directions utilisatrices. Pour atteindre ce but, ITIL met en œuvre notamment les processus suivants : gestion des niveaux de service, gestion des capacités, gestion financière, gestion de la disponibilité, gestion de la continuité, gestion des problèmes, gestion des incidents, gestion des mises à jour, gestion des configurations, etc. De même, ces processus peuvent être soumis à évaluation.

La plupart des auteurs qui ont travaillé sur la qualité des données et des logiciels ne mentionnent pas la qualité des modèles. Une des raisons est le caractère relativement récent des recherches dans ce domaine. De plus, ces travaux souffrent principalement de deux faiblesses :

– l'absence de normes. En effet, il n'existe pas, à ce jour, de consensus sur le vocabulaire qui entoure la qualité des modèles ;

– le manque de validation des résultats auprès des praticiens. Les seules validations qui existent se font avec des chercheurs ou avec leurs étudiants.

Nous sommes convaincus que la qualité des modèles impacte lourdement la qualité des données. En effet, un système d'information est d'abord un ensemble de modèles de haut niveau que l'on va ensuite traduire en modèles de plus bas niveau

jusqu'à leur implémentation sous forme de données et de programmes. Ainsi, il va de soi que la qualité des modèles est une condition nécessaire pour assurer la qualité des représentations qui en découleront, tels les données et les programmes. A titre d'exemple, les modèles conceptuels de données qu'on décrit souvent à tort comme des représentations abstraites des données contiennent en réalité une modélisation des règles de gestion. Si ces dernières sont représentées de façon erronée ou incomplète, les données qui seront insérées dans le système d'information se conformeront à ces règles erronées ou incomplètes. Peu d'études empiriques existent sur la relation de corrélation entre qualité des modèles et qualité des données. Toutefois, dans [WAN 93], les auteurs définissent la modélisation de la qualité des données comme l'évolution naturelle de la modélisation des données.

Nous présentons, dans la section suivante, un état de l'art détaillé sur la qualité des modèles.

3.2. Etat de l'art sur la qualité des modèles

Les modèles de représentation des systèmes d'information sont des moyens indispensables de spécification du contenu de ces systèmes. Depuis maintenant plus de trente ans, ils se sont imposés comme des éléments incontournables de représentation des différentes facettes de ces derniers. Dès que ces modèles sont devenus des éléments de référence pour l'ensemble du projet de développement du système d'information (SI), l'évaluation de leur qualité s'est avérée pertinente [BAT 92]. Depuis près de vingt ans, se sont succédées de nombreuses approches de la qualité des modèles que nous tentons de synthétiser dans cette partie.

Une première classification peut être extraite de [HOX 98]. Elle inclut la qualité de la représentation (flexibilité, interprétabilité), de la sémantique (concision, contenu, champ, compréhensibilité, cohérence), de la syntaxe (présentation, documentation, cohérence), de l'esthétique (facilité d'utilisation) et du respect des règles de normalisation.

La première approche de la qualité d'un modèle est proposée par Batini *et al.* [BAT 92]. Elle porte sur les modèles entité-relation, mais se transpose facilement aux modèles UML par exemple. Les critères proposés sont la complétude, l'exactitude (*correctness*), l'expressivité, la lisibilité, la minimalité, l'auto-explication, l'extensibilité et la normalité. Un modèle est complet s'il représente tous les aspects pertinents du domaine d'application. Il est correct s'il utilise proprement les concepts du modèle correspondant. On peut distinguer la correction syntaxique et la correction sémantique. Un modèle est minimal si tout élément pertinent y apparaît une fois et une seule. L'expressivité désigne la capacité à représenter naturellement

les exigences et à être compréhensible. La lisibilité est liée au diagramme qui représente le schéma. L'auto-explication caractérise les modèles dont un maximum de propriétés se comprend intuitivement. Un modèle est extensible s'il peut être décomposé. Enfin, il est normal s'il est conforme aux formes normales de la théorie relationnelle.

Lindland *et al.* ont proposé de relier la qualité d'un modèle à la qualité d'un moyen de communication entre les différents acteurs [LIN 94]. A l'image des différents niveaux d'analyse de la langue naturelle, la qualité d'un modèle s'évalue dans sa relation avec :

- le langage utilisé pour décrire le modèle : qualité syntaxique ;
- le domaine modélisé : qualité sémantique ;
- l'audience ou les acteurs intervenant dans le processus : qualité pragmatique.

Krogstie a étendu cette vision en proposant six types de qualités : (1) la qualité physique caractérise l'évaluation du modèle externalisé auprès d'un participant ; (2) la qualité syntaxique proposée par Lindland ; (3) la qualité sémantique de Lindland ; (4) la qualité sémantique perçue par les participants à l'interprétation ; (5) la qualité pragmatique de Lindland ; (6) la qualité sociale est atteinte si les différents acteurs convergent sur une représentation. Le cadre ainsi défini est appelé SEQUAL [KRO 95]. Il intègre les acteurs, le domaine, le langage, conduisant à une vision très complète de la qualité. En 2006, Krogstie a étendu ce cadre pour intégrer aussi les modèles de processus, ce qui suppose une prise en compte de la dynamique [KRO 06].

Levitin *et al.* ont proposé d'organiser les caractéristiques en six catégories classant ainsi quatorze dimensions : contenu (pertinence, non ambiguïté, accessibilité des valeurs), étendue (périmètre nécessaire et suffisant), niveau de détail (granularité des attributs, précision des domaines), composition (caractère naturel des « objets », identifiabilité des occurrences, homogénéité des types), cohérence (sémantique, structurelle), réaction au changement (robustesse, stabilité, flexibilité) [LEV 95].

Esko Marjomaa préconise quatre axes d'analyse de la qualité, respectivement en termes ontologiques, épistémologiques, de valeurs théoriques et pragmatiques [MAR 02]. Toutefois, les quatre aspects se retrouvent dans la réponse aux deux questions : en quoi l'ontologie sous-jacente (l'ontologie définit le choix des concepts du modèle) affecte la forme et le contenu du modèle conceptuel résultant ? En quoi l'habileté (ou son manque d'habileté) du concepteur impacte la formalisation résultante ?

De nombreuses approches ont, parallèlement ou par la suite, proposé des facteurs de qualité propres aux modèles conceptuels et proposé des définitions de la qualité

s'appuyant sur ces facteurs : la compréhensibilité, la complexité, la stabilité. Ainsi, la compréhensibilité est analysée par Serrano *et al.* dans le contexte des entrepôts de données : la compréhensibilité du modèle affecte sa qualité externe et influe sur sa complexité cognitive [SER 07]. Les auteurs font état aussi des corrélations fortes existant entre les différents facteurs : la complexité nuit à sa compréhensibilité. La stabilité d'un modèle est définie comme sa capacité à résister aux changements [MAR 93].

Nous avons proposé un cadre d'analyse de la qualité selon trois différents points de vue [CHE 02] : le point de vue du concepteur englobe tous les éléments évaluant la complétude et la richesse du modèle. Le point de vue de l'utilisateur est déterminé notamment par la facilité de compréhension et la lisibilité du modèle. Enfin, le point de vue du développeur est guidé par la capacité du modèle à être implémenté et à faciliter la maintenance du SI résultant.

En 2005, Moody a produit un état de l'art de la qualité des modèles [MOO 05]. Il a recensé plusieurs dizaines d'approches issues de la recherche ou de la pratique professionnelle et déplore : 1) le peu d'effort de validation des approches théoriques, 2) l'ignorance de ces approches par les professionnels et 3) le manque de standardisation dans le domaine. Il propose de standardiser ce domaine en s'appuyant sur les efforts comparables en qualité du logiciel et rappelle que la qualité des modèles, avec la qualité du logiciel, sont des composants de la qualité du système d'information.

Lange et Chaudron ont proposé un cadre de qualité pour les modèles UML qui organise la qualité en trois couches [LAN 05]. La première couche définit l'utilisation primaire du modèle : maintenance et développement. La deuxième couche décrit les différents objectifs associés à ces utilisations. Ainsi, la maintenance se décompose en trois aspects : la modification, le test et la compréhension. Quant au développement, il recouvre cinq aspects : la communication, l'analyse, la prédiction, l'implémentation et la génération de codes. La troisième couche associe des caractéristiques à ces aspects : la complexité, l'équilibre, la modularité, la capacité à communiquer, la correspondance, la capacité à s'auto-décrire, la concision, la précision, l'esthétique, le niveau de détail, la cohérence et la complétude.

La qualité des modèles ne se limite pas à l'évaluation des modèles conceptuels. En particulier, dans l'approche dirigée par les modèles (MDA pour *Model Driven Approach*), on utilise des modèles à trois niveaux différents d'abstraction. On s'assure que ces modèles sont conformes à un métamodèle et on définit des transformations entre ces modèles. Mohagheghi propose de définir des facteurs de qualité de ces transformations, gages de la qualité des modèles résultants [MOH 07] :

– la préservation de la cohérence peut être vérifiée à l'aide d'outils d'analyse de la cohérence avant et après la transformation ;

– la réutilisabilité peut être facilitée par la modularité, par l'héritage et vérifiée par une inspection des transformations ;

– la simplicité peut être mesurée au travers de la complexité des algorithmes de transformation et de la qualité du modèle résultant ;

– le caractère compact est assuré par le recours à des transformations génériques.

En conclusion, la qualité des modèles est un domaine plus récemment exploré que la qualité des données et, *a fortiori*, la qualité du logiciel. Des efforts importants ont néanmoins été fournis dans la formalisation de cette qualité, dans la mise à disposition de cadres, de facteurs, de métriques pour caractériser et évaluer cette qualité. A notre connaissance, il n'existe pas de travaux de recherche qui ont mesuré le lien entre la qualité des modèles et la qualité des systèmes d'information résultants. Toutefois, personne ne remet en cause l'importance des modèles dans la facilitation du processus de fabrication du logiciel. Au-delà d'une approche supplémentaire de la qualité des modèles, nous proposons de décrire un système de management de la qualité des modèles (SMQM).

3.3. Vers un système de management de la qualité des modèles

L'état de l'art décrit précédemment nous a permis de constater la coexistence de plusieurs cadres d'analyse de la qualité des modèles, sans toutefois qu'aucun ne réalise un consensus. Dans cette partie, nous proposons d'abord une approche de la qualité des modèles, fondée sur les points de vue des parties prenantes. Puis, nous comparons cette approche aux trois cadres de référence les plus pertinents. Enfin, nous montrons comment cette approche s'inscrit dans une démarche cyclique d'amélioration continue de la qualité.

3.3.1. *Notre vision de la qualité des modèles*

La qualité des modèles de systèmes d'information peut être mesurée selon leur capacité à (i) fournir une représentation formelle ou semi-formelle de la réalité observée, (ii) répondre aux besoins des utilisateurs, (iii) servir de base à l'implémentation du système d'information modélisé [CHE 02]. Ce triple but nous a conduits à définir la qualité d'un modèle dans un espace tridimensionnel, à savoir la spécification, l'usage et l'implémentation (figure 3.1), selon les parties prenantes :

– un modèle est une représentation abstraite de la réalité. Cette représentation résulte d'une spécification utilisant un ensemble de notations. La notation fournit la

sémantique formelle qui permet l'analyse. La spécification est liée à la phase de définition de la réalité ;

– la dimension usage mesure la qualité du modèle en fonction de la perception de l'utilisateur. L'usage décrit la facilité d'appréhension de ce modèle par les utilisateurs du système d'information ;

– enfin, la dimension implémentation vise à caractériser l'effort nécessaire pour implémenter un modèle. L'implémentation traduit la réalité modélisée en éléments de description du système d'information à réaliser.

Ces trois dimensions ne sont pas nécessairement orthogonales. Chacune d'elles peut être décrite au moyen de facteurs dont certains (comme la lisibilité, la complétude, la maintenabilité, etc.) sont mentionnés sur la figure 3.1 à titre d'illustration. Chaque facteur fait ensuite l'objet de mesures à l'aide de métriques, prenant leurs valeurs entre 0 et 1. Nous décrivons ci-après quelques facteurs et les métriques associées.

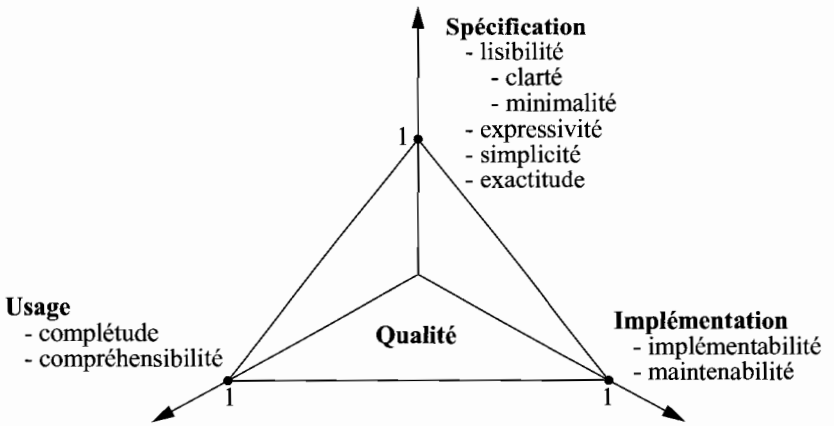

Figure 3.1. *Les trois dimensions de la qualité des modèles [CHE 02]*

Lisibilité : la lisibilité exprime la facilité avec laquelle un modèle peut être lu. La lisibilité peut être décomposée en au moins deux sous-facteurs, la clarté et la minimalité. La clarté s'applique principalement aux représentations graphiques. C'est un critère purement esthétique. Il s'appuie sur l'arrangement graphique des éléments du modèle. Lorsque le nombre de ces éléments croît, on peut aboutir à des croisements d'arc qui nuisent à la lisibilité du modèle [BAT 92]. D'autres caractéristiques graphiques, faciles ou difficiles à mesurer automatiquement, peuvent contribuer à la clarté [SCH 98]. Le second sous-facteur est la minimalité.

Un modèle est minimal si chaque exigence apparaît une fois seulement. En d'autres termes, la non-minimalité traduit un manque de factorisation dans le choix des concepts (objets, entités, classes, associations, etc.) du modèle.

Expressivité : un modèle est expressif s'il représente les exigences des utilisateurs de façon naturelle. On distingue l'expressivité des concepts de l'expressivité du modèle. L'expressivité d'un concept mesure sa capacité à capturer les aspects d'une réalité. Par exemple, le lien d'héritage est plus expressif qu'une simple relation dans le modèle entité-relation étendu. L'expressivité du modèle mesure sa richesse globale.

Simplicité : un modèle est simple s'il contient les concepts minimums. A l'inverse, la complexité d'un modèle croît avec le nombre de liens, tant d'héritage que d'agrégation. Ainsi, un modèle conceptuel est d'autant plus simple qu'il contient plus d'entités que de liens entre les entités [GEN 00].

Exactitude : un modèle est syntaxiquement correct lorsque ses concepts sont correctement définis selon la notation utilisée.

Nous avons défini des métriques pour chaque facteur avec le souci de permettre une automatisation de leur calcul. Ainsi, une métrique de simplicité calculable automatiquement est :

$$Simplicité = \frac{NB(E)}{NB(E) + NB(H) + NB(R)}$$

où *NB(E), NB(H), NB(R)* correspondent respectivement aux nombres d'entités, de liens d'héritage et de liens d'associations dans le modèle.

Complétude : un modèle est complet s'il représente tous les éléments pertinents du domaine d'application. La complétude du modèle est donc le taux de couverture des exigences utilisateurs par le modèle.

Compréhensibilité : c'est la facilité avec laquelle l'utilisateur peut interpréter le modèle. Un modèle compréhensible est plus facile à valider : des noms peu explicites, un haut niveau d'agrégation des concepts et la complexité des contraintes d'intégrité diminuent la compréhensibilité d'un modèle.

Implémentabilité : ce facteur dépend de la proximité entre les concepts du modèle et les concepts de l'environnement de mise en œuvre (système de gestion de base de données, langage de programmation, etc.).

Maintenabilité : la maintenabilité d'un modèle est liée à la cohésion de ses éléments. Plus la cohésion est forte, plus l'impact d'un changement ponctuel sera important sur l'ensemble du système d'information. De nombreuses métriques de cohésion ont été définies dans la modélisation orientée objet par exemple [CHI 94].

3.3.2. *Positionnement de notre approche*

Dans cette partie, nous analysons notre approche de la qualité à la lumière de plusieurs cadres de référence : la norme ISO 9126 [ISO 11], le cadre de Lindland [LIN 94] et le cadre de Krogstie [KRO 05]. En effet, les modèles que nous évaluons sont une représentation abstraite partielle et semi-formelle des logiciels. Aussi, nous pouvons les évaluer comme les premiers artéfacts du logiciel, en utilisant les dimensions préconisées par ISO 9126. La qualité des modèles s'inscrit dans la perspective de la communication entre parties prenantes. En effet, un modèle, quel que soit son niveau d'abstraction et quel que soit son usage, n'est qu'un moyen de communication entre agents, humains ou logiciels, dans le cadre de la gestion d'un système d'information. Ainsi la qualité des modèles s'évalue dans leur capacité à faciliter la communication. C'est pourquoi, nous situons notre approche de qualité dans le cadre de Lindland [LIN 94]. Enfin, un modèle s'appuie sur un langage. La qualité du modèle est donc très dépendante du langage ainsi utilisé. Nous comparons donc notre approche de la qualité des modèles au cadre d'évaluation des langages proposé par Krogstie [KRO 01].

Selon ISO 9126, la qualité d'un logiciel est standardisée au moyen de six caractéristiques qui se veulent exhaustives : i) la fonctionnalité d'un logiciel est sa capacité à fournir les fonctions attendues, ii) la fiabilité réside dans sa capacité à maintenir un niveau de performance, iii) l'utilisabilité décrit la facilité d'utilisation et d'apprentissage du logiciel, iv) l'efficacité compare le rendement du logiciel par rapport aux ressources utilisées, v) la maintenabilité définit la facilité d'évolution du logiciel, vi) enfin, la portabilité décrit sa capacité à être transféré d'un environnement à un autre [ISO 11]. Le modèle est une représentation précoce du logiciel. Toutefois, les neuf facteurs de qualité de ce modèle, illustrant notre approche, préfigurent les six caractéristiques ainsi décrites, comme le montre le tableau (tableau 3.1).Nos facteurs de qualité couvrent aussi les trois axes du cadre de Lindland : la qualité syntaxique est liée à l'absence d'erreurs de syntaxe dans le modèle (exactitude). La qualité sémantique est le résultat de l'évaluation des éléments invalides dans le modèle ainsi que des éléments manquants (minimalité, expressivité, complétude). La qualité pragmatique vise à assurer une compréhension complète du modèle par l'audience, en tout cas une compréhension totale de la partie du modèle qui la concerne (clarté, simplicité, compréhensibilité, implémentabilité, maintenabilité).

Quelques facteurs de qualité	Dimensions de notre approche	Dimensions de Lindland	Dimensions de ISO 9126	Dimensions de Krogstie
Clarté	Spécification	Pragmatique	Maintenabilité	Compréhensibilité
Minimalité	Spécification	Sémantique	Efficacité	Domaine
Expressivité	Spécification	Sémantique	Fonctionnalité	Domaine
Simplicité	Spécification	Pragmatique	Efficacité	Connaissance participant
Exactitude	Spécification	Syntaxique	Fiabilité	Besoins organisationnels
Complétude	Usage	Sémantique	Fonctionnalité	Externalisabilité
Compréhensibilité	Usage	Pragmatique	Utilisabilité	Compréhensibilité
Implémentabilité	Implémentation	Pragmatique	Utilisabilité	Interprétation technique
Maintenabilité	Implémentation	Pragmatique	Maintenabilité	Interprétation technique

Tableau 3.1. *Comparaison des différents cadres de référence*

Enfin, les modèles de données utilisent un langage pour exprimer les concepts pertinents du système d'information. On peut donc aussi situer notre approche dans le cadre d'évaluation de la qualité des langages de Krogstie [KRO 01]. Six dimensions de qualité sont identifiées :

– l'adaptation au domaine qui traduit le fait qu'un langage, une notation, un outil sont plus ou moins adéquats pour représenter un domaine (minimalité, expressivité) ;

– l'adéquation à la connaissance des parties prenantes exprime le fait qu'un modèle correspond plus ou moins à la perception de la réalité (simplicité) ;

– l'externalisabilité de la connaissance permet à celui qui modélise d'exprimer toute sa connaissance (complétude) ;

– la compréhensibilité lie le langage à l'interprétation par les acteurs (clarté, compréhensibilité) ;

– l'interprétation technique permet l'analysabilité, l'exécution automatique, la vérification, etc. (implémentabilité, maintenabilité) ;

– l'adéquation à l'organisation fait référence aux besoins de l'organisation (exactitude).

En conclusion, notre approche de qualité, sans prétendre à l'exhaustivité, couvre de façon équilibrée les différentes facettes établies dans les cadres généraux de qualité proposés dans la littérature. Dans la suite, nous décrivons le processus d'évaluation et d'amélioration de la qualité dans lequel notre approche s'insère.

3.3.3. *Une démarche cyclique d'amélioration de la qualité des modèles*

Un des cadres méthodologiques largement utilisé et reconnu est l'approche PDCA (*Plan-Do-Check-Act*) également connu sous le nom de roue de *Deming* [DEM 86]. Cette approche préconise la gestion de la qualité selon un cycle perpétuel de quatre étapes qui assurent l'amélioration continue de la qualité (figure 3.2).

En utilisant l'approche PDCA, nous avons défini un processus de développement de SI qui permet d'apporter une assistance dirigée par la qualité aux activités d'analyse. La première étape consiste à identifier et à élaborer une stratégie de la qualité du SI (SQSI en cohérence avec la stratégie et les règles du SI considéré. A l'étape suivante, la stratégie de qualité élaborée est appliquée et ses actions sont alors implémentées et exécutées. La troisième étape mesure l'efficacité de la stratégie appliquée. Cette mesure génère un rapport qui servira à l'étape de révision. Cette dernière applique des actions correctives ou préventives. Cette démarche cyclique permet une amélioration continue. Notre processus s'appuie sur un métamodèle décrit dans la figure 3.2.

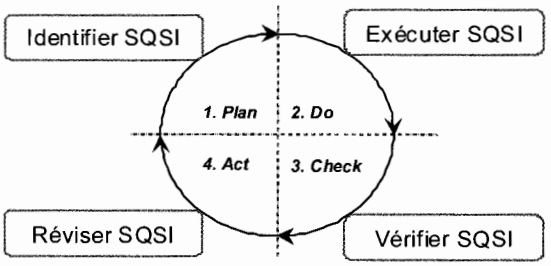

Figure 3.2. *L'approche PDCA adaptée à la stratégie de qualité des SI (SQSI)*

3.3.3.1. *Un métamodèle pour la définition de la qualité*

Notre approche s'appuie sur un métamodèle dont les principaux concepts sont décrits ci-dessous (figure 3.3). Un *attribut de qualité* désigne un groupe de propriétés observables sur le cycle de vie du produit [PRE 01] ou un groupe de propriétés du service rendu par le système à ses utilisateurs. Le service rendu par

un système est son comportement tel qu'il est perçu par ses utilisateurs [BAR 95]. De même, ces derniers lient les attributs de qualité à des repères qui décrivent le comportement du système au sein de l'environnement pour lequel il a été construit. Dans le domaine de l'ingénierie des systèmes, les attributs de qualité peuvent aussi concerner les exigences non fonctionnelles pour évaluer la performance du système. Dans notre approche, un attribut de qualité fournit l'abstraction d'un ensemble de métriques étroitement liées et mesurant la même caractéristique souhaitée du modèle conceptuel (MC). Chaque attribut de qualité doit être générique dans le sens où sa définition ne doit pas dépendre de la notation utilisée pour le MC. Nous avons choisi d'utiliser le terme « attribut de qualité » pour désigner les concepts de la littérature tels que : attribut de qualité, dimension, facteur, caractéristique, sous caractéristique, critère, etc. Quelques exemples d'attributs de qualité sont : la complexité, la complétude, etc.

Les *métriques de qualité* sont les mesures ou les procédures d'évaluation qui attribuent des valeurs numériques ou symboliques afin de caractériser les qualités ou les caractéristiques des *éléments de modèle*. Chaque métrique a une portée définie par l'ensemble des entités et des objets auxquels elle est applicable et une plage de valeurs possibles. La représentation explicite des propriétés des métriques permet de distinguer les métriques les unes par rapport aux autres, surtout lorsqu'elles mesurent le même attribut. Ainsi, l'attribut de qualité « complexité structurelle » a recours à des métriques telles que le nombre d'associations, le nombre d'agrégations, la profondeur maximale de hiérarchies, etc.

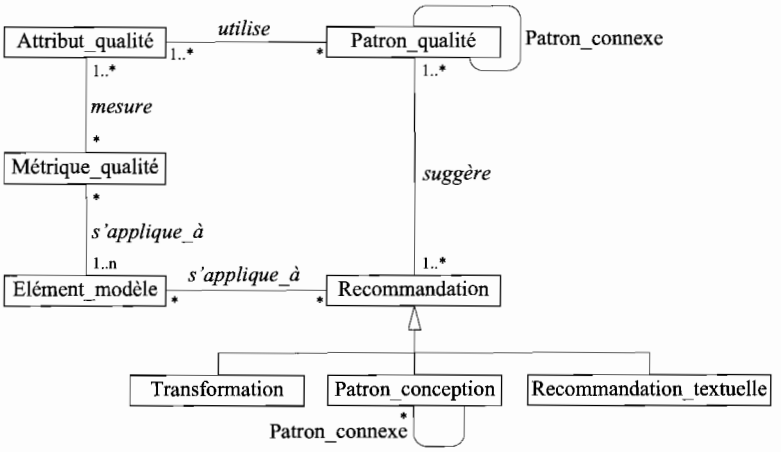

Figure 3.3. *Un métamodèle pour la qualité des SI*

Les *recommandations* sont des suggestions, des propositions ou des conseils sur la façon de modifier les MCD afin d'en améliorer la qualité. Les recommandations sont proposées en fonction des résultats obtenus par l'application des métriques. Les analystes/concepteurs peuvent suivre les recommandations proposées pour améliorer la qualité des MCD. L'objectif est de pouvoir améliorer, si nécessaire et si possible, cette qualité. En effet, la majorité des approches étudiées se concentrent sur l'évaluation de la qualité en proposant des indicateurs et des métriques permettant de cibler les erreurs ou manques dans les MCD. Peu d'approches s'attaquent au problème difficile de l'amélioration de cette qualité [MOO 05].

Les recommandations peuvent être :

– textuelles : elles donnent, dans ce cas, des conseils que l'analyste/concepteur choisit d'appliquer ou non ;

– structurées sous forme de règles de transformation ;

– structurées sous forme de patrons de conception (*design patterns*).

Un *patron de qualité* (*Quality Pattern*) est une proposition spécifique à notre approche. En effet, après l'étude approfondie des travaux de la littérature, nous avons retenu 21 attributs de qualité. Mais ces attributs permettent tout au plus la qualification de la caractéristique à mesurer. L'évaluation de la qualité nécessite la définition d'une ou plusieurs métriques de qualité. Manipuler et faire le choix parmi tous ces concepts nécessite un degré d'expertise qui explique l'utilisation jusqu'ici limitée des approches de qualité proposées dans la littérature. Notre réponse est de proposer un mécanisme qui permet la capitalisation de l'expertise dans ce domaine. Ce mécanisme est le patron de qualité (*Quality Pattern*). Ce concept se rapproche de celui de patron de conception (*Design Pattern*) qui encapsule les bonnes pratiques de conception afin d'en améliorer le résultat [HSU 08]. Le concept de patron de qualité a été proposé pour la première fois par [HOU 97] pour l'ingénierie des logiciels. Dans [CHE 08], nous en avons proposé l'adaptation à la qualité des modèles conceptuels. Chaque patron de qualité cible un problème, possède un contexte d'utilisation et des mots-clés qui permettent de le rapprocher d'un objectif de qualité lors de l'évaluation. Il décrit une solution sous la forme d'un ensemble de critères de qualité à évaluer et englobe les métriques permettant leur évaluation et les recommandations pertinentes d'amélioration de la qualité. A titre d'exemple, nous présentons un patron lié à la complexité des modèles.

Afin de faciliter l'usage du métamodèle de la qualité, nous avons développé une démarche méthodologique s'appuyant sur l'approche PDCA et guidant pas à pas la prise en compte de la qualité lors des premières phases du processus de développement.

Nom du patron : la complexité du modèle.

Contexte : la vérification de la complexité globale d'un modèle conceptuel, en respectant le nombre d'éléments (nombre de classes/entités/attributs, etc.) présents.

Problème : il est largement reconnu que la multiplication des éléments dans un modèle (classes/entités/cas d'utilisation, etc.) peut gêner la lisibilité du modèle. Miller présente une étude défendant l'idée selon laquelle la mémoire humaine (pour un évènement récent) peut contenir 7 ± 2 objets rendant la compréhension d'une plus grande collection d'objets plus difficile [MIL 56]. De même, l'existence de nombreux éléments peut augmenter la complexité. Cette complexité peut nuire à la maintenabilité, parce que maintenir ou corriger un modèle nécessite au préalable sa compréhension. Le patron de complexité propose les métriques suivantes pour un diagramme de classes :

a/ nombre de classes : nombre total de classes dans un modèle ;

b/ nombre d'attributs : nombre d'attributs dans le modèle ;

c/ nombre de méthodes : nombre total de méthodes ou de fonctions dans le modèle ;

d/ nombre de cycles : nombre total de cycles dans un modèle ;

e/ degré de non-redondance : cet indicateur de mesure calcule le rapport entre les concepts non- redondants et les concepts totaux présents dans le modèle ;

f/ nombre d'associations : nombre total d'associations dans un modèle ;

g/ nombre d'agrégations : il calcule le nombre d'agrégations dans un diagramme de classes ;

h/ nombre de compositions : il calcule le nombre de compositions dans un diagramme de classes ;

i/ nombre de généralisations : il calcule le nombre total de généralisations dans un modèle ;

j/ profondeur d'héritage (maximale) : il calcule le plus long chemin de la classe à la racine de la hiérarchie dans une hiérarchie de généralisation.

Solutions :

i) la division des modèles complexes en modèles plus simples peut contribuer à améliorer la compréhension ;

ii) supprimer les éléments redondants ;

iii) fusionner les classes ;

iv) diviser les classes ;

v) appliquer les patrons de conception : GRASP[1] haute cohésion et polymorphisme.

Mots-clés : complexité ; maintenabilité ; modifiabilité ; compréhension ; taille.

Patrons connexes :

i) maintenabilité des modèles (la complexité rend la maintenance des modèles difficile) ;

ii) clarté des modèles (les modèles complexes sont difficiles à lire).

Tableau 3.2. *Exemple de patron de qualité*

3.3.3.2. *Le processus de la gestion de la qualité*

Une lacune constatée dans les approches étudiées est l'absence d'un processus structuré permettant la mise en œuvre des concepts de qualité. Les approches laissent aux analystes/concepteurs la charge de sélectionner les concepts adéquats, de les appliquer et d'en exploiter les résultats.

Pour pallier ce manque, nous proposons un processus détaillé qui guide pas à pas l'analyste dans une démarche de conception de SI dirigée par la qualité. Ce processus est détaillé à la figure 3.4.

A tout instant durant la construction des spécifications (modèles conceptuels) lors de l'analyse, le processus de gestion de la qualité peut être sollicité. L'évaluation se fait sur l'état actuel de la spécification. Le résultat d'une itération de ce processus est un ensemble de mesures de la qualité et d'éventuelles recommandations pour son amélioration. Ces recommandations sont des préconisations quant aux décisions de conception à prendre par l'analyste et s'intègrent donc naturellement dans un processus d'analyse en cours.

1. GRASP : *General Responsibility Assignment Software Patterns* [LAR 04].

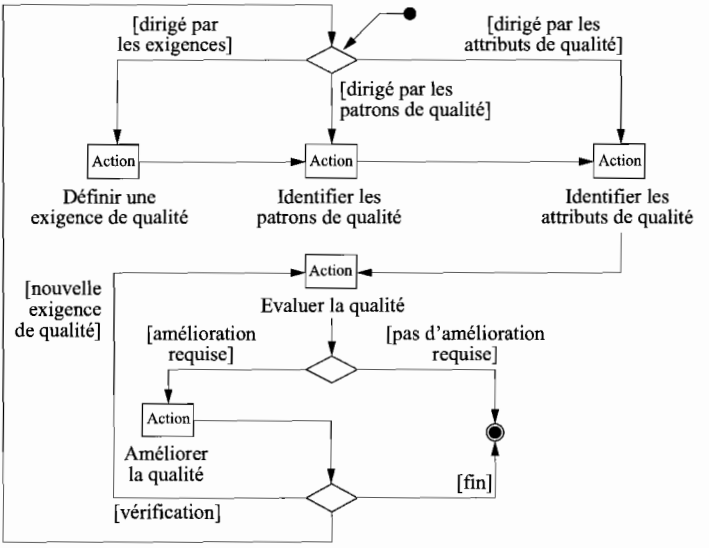

Figure 3.4. *Un processus pour l'amélioration continue de la qualité des SI*

Dans l'approche PDCA, le but de la première étape (étape d'identification – plan) est d'identifier des objectifs de qualité et de planifier les actions à mener en vue d'améliorer la qualité.

Afin d'offrir un maximum de flexibilité à son application et de permettre son adaptation à divers niveaux d'expertise de l'analyste, notre processus prévoit trois modes :

i) le mode « novice » ne présuppose aucune connaissance précise du vocabulaire lié à la qualité ni des patterns de qualité. Dans ce cas, nous proposons d'appliquer l'approche GQM. On commence par l'activité « définir une exigence de qualité » qui aide à la formulation d'un but de qualité en langage naturel. Le processus guide alors l'analyste à l'aide de questions permettant de préciser ce but et de le mener à l'identification d'un patron de qualité ;

ii) le mode « expert confirmé » permet de ne pas utiliser les patrons de qualité. L'expert choisit directement les critères de qualité précisant l'objectif de qualité poursuivi (activité « identifier les attributs de la qualité »). Dans ce cas, le processus guide l'expert dans le choix des métriques adéquates pour mesurer la qualité du SI ;

iii) enfin, le troisième mode est destiné à l'expert souhaitant utiliser le concept de patron de qualité. Une fois le patron désigné, celui-ci guide l'expert dans l'identifi-

cation du critère de qualité à viser et par la suite dans le choix des métriques de qualité à appliquer ;

iv) une fois les métriques de qualité identifiées, le processus calcule les valeurs de qualité (activité : « évaluer la qualité ») correspondant à ces métriques et suggère un ensemble d'actions d'amélioration pouvant aller de la simple recommandation textuelle à une liste d'actions de transformation, comme le montre le métamodèle de la figure 3.3.

La deuxième étape (étape d'exécution – *Do*) consiste à effectuer, si jugées nécessaires, les améliorations préconisées. Cette étape (activité « améliorer la qualité ») est entièrement à la charge de l'analyste, à l'aide des recommandations qui, selon leurs formes et degrés de détail, constituent une aide méthodologique non négligeable.

La troisième étape (étape de vérification – *Check*) suggère de réévaluer la qualité pour mesurer l'efficacité des actions entreprises (retour sur l'activité « évaluer la qualité »). L'approche s'appuie sur une large proposition de métriques de qualité, automatiques, semi-automatiques et manuelles. Notre approche intègre des métriques spécifiques, en complément de celles de la littérature.

La quatrième étape (étape de révision – *Act*) permet de réviser l'objectif de qualité et de le faire évoluer afin d'améliorer la qualité ou d'avancer dans le processus de développement.

La force et l'avantage de ce processus résident dans sa flexibilité, permettant son adaptation au degré d'expertise de l'analyste. Dans la section qui suit, nous avons validé l'approche et l'avons appliquée à un cas d'étude afin d'évaluer son apport et ses limites.

3.4. Validation

Pour garantir la validité de notre approche, nous avons mené deux actions. La première a consisté à confronter les attributs et métriques de qualité à un public d'analystes de niveaux d'expertise variés. La seconde est relative à la réalisation d'un prototype de support de notre démarche qualité.

3.4.1. *Validation par une enquête*

L'enquête décrite dans cette section a servi de base pour évaluer la qualité de différents modèles conceptuels décrivant une même réalité par un ensemble d'acteurs tels que les concepteurs, les utilisateurs et les développeurs. Un des atouts de notre

approche est de permettre une évaluation chiffrée automatique des facteurs de qualité des modèles conceptuels. L'enquête avait pour objectif de vérifier que ces métriques sont en adéquation avec la perception par les parties prenantes de ces mêmes facteurs. Dans notre approche, nous distinguons les utilisateurs finaux des professionnels des SI. Puis, nous procédons à un raffinement de cette dernière catégorie en fonction du domaine de spécialisation. On obtient ainsi des groupes différents composés de spécialistes réseaux, de spécialistes systèmes, de responsables support, de chefs de projets, de concepteurs, etc. Enfin, nous intégrons d'autres caractéristiques telles que le sexe, l'expérience professionnelle, la formation d'origine, etc.

Notre expérimentation prend en compte quatre facteurs de qualité : clarté, minimalité, expressivité, simplicité. Pour ce faire, nous avons recueilli les avis de 113 parties prenantes. Ces dernières représentent à la fois des professionnels des SI (87) et des utilisateurs finaux (26). Le groupe des professionnels des SI est composé de différents spécialistes (réseaux informatiques : 13, systèmes d'exploitation : 15, conception et développement : 48, support : 18, gestion de projet : 23). Ils devaient aussi caractériser leur expérience des différents domaines de l'informatique : base de données (72 ont déclaré une expérience en base de données), programmation (68), système d'information (59) et modélisation conceptuelle (50). De la même façon, on a interrogé les utilisateurs finaux sur leur type d'usage des SI (Internet : 18, bureautique : 19, logiciels métiers : 8). L'échantillon total comporte 88 hommes et 25 femmes. L'âge moyen de l'échantillon est de 31 ans avec un intervalle entre 21 et 57 ans. La richesse de l'échantillon nous permet de discriminer entre les sous-groupes en utilisant différents critères (sexe, expérience professionnelle, profil, âge, formation, etc.).

Au cours de l'expérimentation, chaque participant a reçu huit diagrammes entité-relation représentant le même univers du discours (deux des huit modèles sont fournis en annexe, section 3.6.1, pour plus d'information voir [CHE 10]). Les modèles ont été fournis aux participants avec une information décrivant le contexte. On a demandé aux participants de classer les huit modèles selon quatre facteurs : clarté, simplicité, expressivité, minimalité, après leur avoir expliqué ces facteurs. Au-delà du classement des huit schémas sur la base des facteurs de qualité, ils ont été interrogés sur leur sexe, âge, formation d'origine, expérience professionnelle et usage des SI. Le questionnaire utilisé avait au préalable été testé auprès de quelques participants.

Pour effectuer la validation, nous avons d'abord calculé le classement des huit modèles selon nos facteurs de qualité. Les résultats de l'évaluation par les métriques sont détaillés dans [CHE 02]. Nous avons transformé le classement des participants en une note allant de un à huit. Les résultats obtenus pour chaque modèle, pour chaque facteur et pour l'échantillon complet sont présentés au tableau 3.2 :

- la première colonne indique la valeur moyenne de la qualité perçue ;

- la deuxième colonne indique l'écart type des valeurs perçues ;
- la troisième colonne extrait la valeur perçue la plus fréquente.

	Modèle	Moyenne des valeurs perçues	Ecart-type	Valeur perçue la plus fréquente
Clarté	M1	7,39	1,57	8
	M2	3,50	2,14	2
	M3	4,98	1,77	7
	M4	2,20	1,86	1
	M5	4,08	2,25	4
	M6	4,51	1,55	4
	M7	4,12	1,52	5
	M8	5,00	1,88	7
Simplicité	M1	7,55	1,44	8
	M2	3,29	2,13	2
	M3	5,12	1,73	7
	M4	2,32	1,92	1
	M5	4,17	1,97	6
	M6	4,43	1,66	4
	M7	4,41	1,67	5
	M8	4,47	2,02	4
Expressivité	M1	1,47	1,45	1
	M2	4,43	2,26	2
	M3	3,67	1,74	3
	M4	5,69	2,10	8
	M5	4,43	1,98	2
	M6	5,58	1,57	6
	M7	4,85	1,82	4
	M8	5,92	1,91	8
Minimalité	M1	4,04	2,11	2
	M2	4,20	1,71	3
	M3	5,00	1,44	5
	M4	4,24	1,97	4
	M5	6,25	2,19	8
	M6	4,41	1,84	6
	M7	5,90	2,25	7
	M8	1,86	1,88	1

Tableau 3.3. *Qualité perçue par l'échantillon*

Plusieurs raisons justifient l'utilisation de la valeur la plus fréquente plutôt de la moyenne. La raison la plus importante est le fait que les écarts-types sont très importants. La seconde raison est liée aux difficultés rencontrées par les participants à réaliser un classement unique des huit schémas. Pour ces raisons, nous avons décidé de retenir, pour chaque modèle, les valeurs les plus fréquemment attribuées par les participants. Ce choix présente l'avantage d'écarter systématiquement les questionnaires non remplis de façon rigoureuse.

Nous avons d'abord effectué une analyse statistique basique de la qualité perçue de l'échantillon (tableau 3.4). Nous avons ensuite effectué une analyse factorielle de correspondance. Cela nous a permis d'analyser :

– en ligne, les profils des participants selon leurs caractéristiques (sexe, formation d'origine, expérience professionnelle, etc.) ;

– en colonne, les profils de nos modèles et métriques ;

– globalement les profils de lignes et de colonnes permettant de visualiser l'association entre les niveaux des deux dimensions.

La valeur située à l'intersection d'une ligne et d'une colonne représente le nombre de personnes qui ont classé chaque modèle de la même façon. Ainsi, 80 professionnels de l'informatique ont classé le même modèle comme étant le meilleur en termes de clarté (+) et 42 ont classé un même modèle comme étant le moins bien en termes de clarté (-).

Le but est d'avoir une vue globale des données qui est utile pour analyser la relation entre la qualité perçue et les caractéristiques de l'échantillon. Les résultats sont présentés dans le tableau 3.4.

3.4.1.1. *Analyse statistique de base*

Le modèle ayant la meilleure clarté calculée par notre métrique (M1) est aussi celui perçu comme le plus clair par les participants. 91 % des 113 répondants classent les modèles M1, M3, M7, M8, comme étant les meilleurs en termes de clarté. M1 est le plus mauvais modèle en termes d'expressivité pour 87 % de l'échantillon. 96 % des développeurs et 73 % des utilisateurs finaux le considèrent comme le moins expressif. M1 est aussi le plus mauvais modèle en termes de minimalité pour 85 % de l'échantillon. Ils sont 98 % à penser de même pour les développeurs.

Il est apparemment difficile de comprendre la raison pour laquelle les modèles possédant le plus mauvais score quant à la clarté et la simplicité ne sont pas identiques. Il semble que ces deux facteurs soient difficiles à distinguer. Les personnes composant l'échantillon tendent à confondre la clarté (ou lisibilité) et la

simplicité (en nombre de concepts). Les plus mauvais scores sont ceux relatifs à l'expressivité et à la minimalité. Seulement 26 % considèrent M8 comme étant le meilleur en termes d'expressivité et 29 % considèrent M2 comme étant le plus mauvais en termes de minimalité. L'expressivité et la minimalité requièrent une compréhension de la sémantique des schémas conceptuels. En conséquence, ils sont plus difficiles à comprendre. Les concepts sous-jacents ne facilitent pas l'obtention d'un consensus.

	Clarté +	Clarté –	Simpl +	Simpl –	Express +	Express –	Minim +	Minim –
Professionnel	80	42	77	44	22	76	76	30
Etudiant	21	13	19	11	8	22	20	3
Réseaux	13	7	13	7	2	11	13	1
Systèmes	13	9	14	9	1	13	13	4
Conc.dév.	46	23	44	23	12	46	47	18
Support	14	7	15	11	2	17	15	8
Gestion de projet	22	15	19	14	10	21	18	5
Informaticien	81	41	77	42	23	79	76	29
Utilisateur	20	14	19	13	7	19	20	4
Homme	79	46	75	47	22	75	75	22
Femme	22	9	21	8	8	23	21	9
Internet	14	9	13	8	6	12	13	3
Bureautique	14	9	13	8	5	13	14	3
Logiciels métiers	7	3	6	5	3	4	7	1
Bases de données	66	34	62	37	24	67	62	26
Programmation	62	29	60	31	21	63	59	24
Syst. d'inform.	56	26	52	30	21	55	50	16
Modél. concept.	48	23	42	24	19	47	42	9
Tous	101	55	96	55	29	98	96	33

Tableau 3.4. *Table de contingence*

3.4.1.2. *Résultats de l'analyse de correspondance*

Rappelons que cette technique statistique permet d'analyser les tableaux à double entrée (colonne, ligne) et leurs relations. Ce tableau est composé des classements des schémas conceptuels effectués par les personnes composant l'échantillon. En analysant la table de contingence ainsi obtenue, nous observons les résultats dans le tableau 3.5).

***Chi-square* (Observed value)**	42,306
***Chi-square* (Critical value)**	153,198
DF	126
p-value	1,000
alpha	0,05

Tableau 3.5. *Résultats*

Nous acceptons l'hypothèse H0 (les lignes et les colonnes sont indépendantes) puisque $p > \alpha$. En conséquence, les caractéristiques de l'échantillon (sexe, formation, etc.) sont indépendantes des facteurs utilisés pour évaluer la qualité des schémas.

L'analyse des valeurs propres et du pourcentage d'inertie nous permet d'obtenir les valeurs dans le tableau 3.6).

	F1	**F2**	**F3**
Eigenvalue	0,005	0,003	9,244
Rows depend on column	50,625	30,333	9,244
Cumulative %	50,625	80,958	90,203

Tableau 3.6. *Valeurs propres et pourcentages*

Deux facteurs, appelés F1 et F2, expliquent 80 % de la variance. La contribution du troisième facteur F3 est marginale. En conséquence, il n'est pas pris en compte. Les deux premiers facteurs nous permettent d'interpréter la distance existant entre les caractéristiques de l'échantillon et les facteurs utilisés pour évaluer la qualité perçue des schémas conceptuels. L'axe F1 permet de distinguer les utilisateurs finaux des professionnels de l'informatique dans leur perception de la qualité des schémas. Le premier groupe est essentiellement composé d'utilisateurs ayant une expérience bureautique et d'Internet. Quant au second groupe, il est composé d'informaticiens et de développeurs ayant une grande expérience dans la programmation. L'axe F2

permet de distinguer ces deux mêmes groupes d'un troisième groupe composé de concepteurs, de gestionnaires de projet, de spécialistes des systèmes d'information et, de manière surprenante, d'étudiants. Mentionnons la différence significative existant entre les hommes et les femmes dans leur perception de la qualité des schémas conceptuels.

La figure 3.5 est la représentation graphique en nuage de points des résultats obtenus.

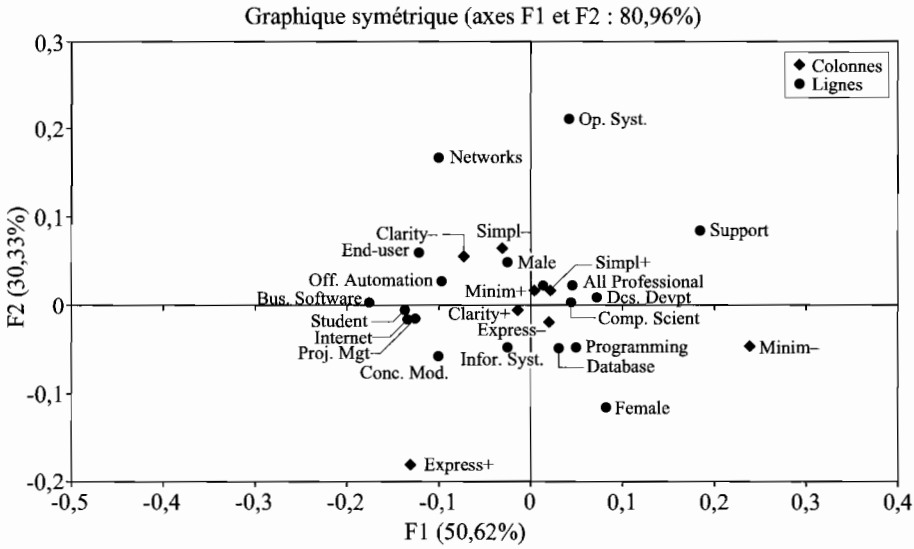

Figure 3.5. *Analyse des axes F1 et F2*

Des résultats plus détaillés de cette analyse peuvent être consultés dans [CHE 10]. Cette étude, réalisée sur un échantillon riche et composée de véritables professionnels, nous a permis une validation significative des métriques d'évaluation de la qualité des modèles. Ainsi, notre ambition de permettre, très tôt dans le processus de conception du SI, de calculer automatiquement des facteurs de qualité du modèle, précurseurs de la qualité du SI futur, nous semble atteignable. Cet objectif d'automatisation passe bien entendu par la conception d'un outil de support de notre approche. Le prototype ainsi conçu est décrit dans la section 3.4.2.

3.4.2. *Prototypage*

Nous avons constaté qu'en dépit de la multiplication des propositions concernant la mesure de la qualité dans la littérature, leur adoption et leur application restent

très limitées. Ainsi, les professionnels avouent accorder beaucoup d'importance à la qualité des spécifications, et pourtant, ils n'adoptent pas les approches fondées sur la qualité. Cela est dû à la difficulté de maîtriser à la fois tous ces concepts et de les appliquer. A cela s'ajoutent la complexité des spécifications elles-mêmes et la diversité des notations utilisées. Les expériences vécues dans d'autres domaines montrent que l'existence d'aides outillées peut aider à l'adoption des méthodes et techniques. Nous avons analysé ce que proposent les ateliers de génie logiciel du marché pour la prise en charge de la qualité [MEH 10]. Cette analyse nous conduit aux observations suivantes :

i) certains ateliers du marché intègrent quelques mesures très simples d'évaluation. Ces mesures ne peuvent cependant pas être modifiées ni enrichies ;

ii) ces outils ne proposent aucune aide à l'interprétation des valeurs obtenues ;

iii) les mesures proposées par ces outils ne sont pas toujours rattachées à des attributs de qualité et sont par conséquent difficiles à utiliser ;

iv) aucun des logiciels ne fournit de recommandations après l'évaluation.

Nous avons proposé, dans notre approche, un outil nommé CM-Quality qui prend en charge la totalité de la démarche proposée [MEH 10]. Il offre un ensemble d'aides outillées pour la définition des concepts de la qualité (attributs, métriques, patrons, recommandations). Il assure aussi la prise en charge du processus d'évaluation et d'amélioration de la qualité dans son intégralité. Il possède enfin l'avantage de pouvoir s'interfacer avec les ateliers de génie logiciel du marché les plus répandus (comme Rational Rose, Objecteering, StarUML, etc.).

CM-Quality est conçu pour deux types d'utilisateurs : les experts qualité et les analystes/concepteurs. L'expert qualité est un utilisateur ayant une connaissance approfondie des concepts de qualité. Il est chargé de définir les concepts de qualité tels que les attributs, les patrons de qualité, etc. L'analyste, quant à lui, a la charge d'établir des spécifications conceptuelles du SI et souhaite en étudier la qualité pour l'améliorer. Les analystes utilisent la connaissance définie par l'expert qualité. CM-Quality contient une base de connaissances qui sert au stockage des différents concepts de qualité. La base de connaissances stocke également des sessions d'évaluation à des fins de trace, en vue notamment de l'amélioration des concepts déjà définis. CM-Quality offre les fonctionnalités suivantes :

i) il met en œuvre et stocke, dans une base de connaissances, une hiérarchie des concepts de qualité, incluant les patrons de qualité, les attributs de qualité, les métriques, les recommandations, etc. Tous ces concepts de qualité peuvent être ajoutés, modifiés, supprimés de la base de connaissances ;

ii) il aide l'analyste/concepteur dans l'identification des critères de qualité pertinents par rapport à l'objectif de qualité formulé ;

iii) il peut être utilisé, par un analyste/concepteur, afin d'évaluer un modèle conformément à un objectif de qualité spécifique ;

iv) il propose des informations post-évaluation sous forme de recommandations pour l'amélioration du modèle ;

v) il permet l'évaluation et la comparaison de plusieurs modèles ;

vi) il peut être utilisé pour évaluer les modèles conçus *via* les ateliers de génie logiciel du marché, tels que Rational Rose, Objecteering, etc.

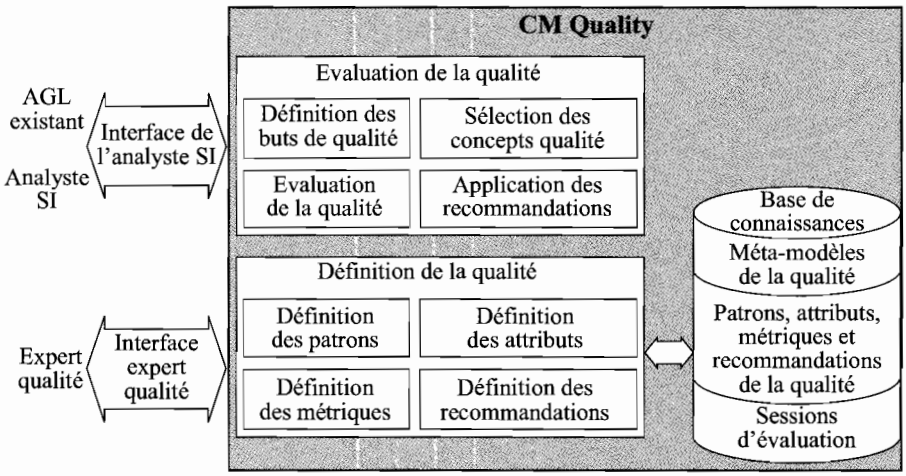

Figure 3.6. *L'architecture de CM-Quality*

L'architecture générale est constituée de deux modules centraux destinés à deux types d'utilisateurs :

– l'expert qualité responsable de la *définition de la stratégie de qualité* souhaitée (attributs de la qualité à privilégier, métriques associées à ces attributs, guides d'améliorations de la qualité) ;

– et l'analyste/concepteur de SI chargé de la conception d'un nouveau système d'information qui peut en *évaluer la qualité*. Les deux modules sont connectés au moyen d'une base de connaissances commune (figure 3.6).

3.4.2.1. *Définition de la stratégie de qualité*

CM-Quality propose quatre utilitaires à l'intention de l'expert : l'outil de définition des patrons de qualité, l'outil de définition des attributs de la qualité, l'outil de définition des métriques de la qualité et enfin, l'outil de définition des recommandations d'amélioration de la qualité.

L'outil de définition des patrons de la qualité nécessite de répondre aux questions suivantes : Dans quel contexte le patron de qualité pourrait être utilisé ? Quel est le problème que permet de résoudre ce patron ? Enfin, quelle solution propose-t-il pour résoudre ce problème ?

L'écran représenté à la figure 3.7 illustre la définition de la signature d'un patron de qualité nommé ici « complexité structurelle ». L'expert qualité doit ensuite en spécifier le détail, en définissant les attributs de qualité qui lui sont associés. C'est à l'aide de ce même outil que l'expert qualité pourra consulter ou modifier un patron déjà présent dans la base de connaissances.

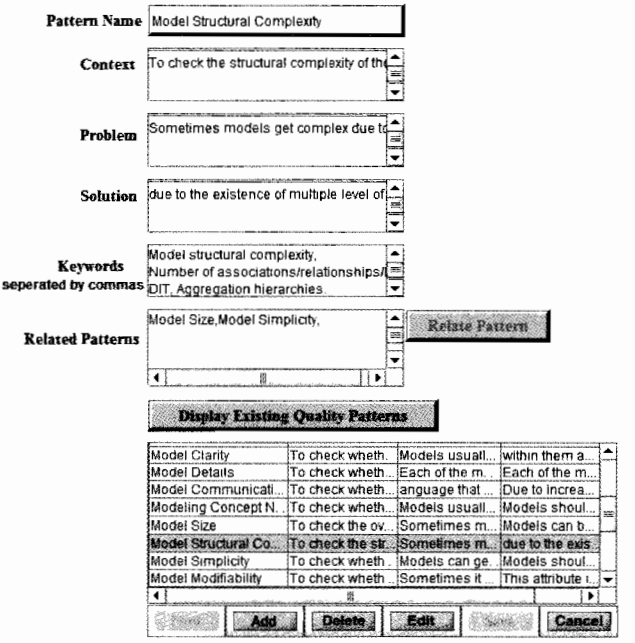

Figure 3.7. *Définition d'un patron de qualité*

L'outil de définition des attributs de la qualité sert à définir un nouvel attribut de qualité. Il permet, comme le montre la figure 3.8, de préciser les noms, description et mots-clés de cet attribut. Il permet également d'accéder aux autres attributs de qualité contenus dans la base de connaissance. Il convient ensuite d'associer l'attribut ainsi défini aux métriques permettant de l'évaluer.

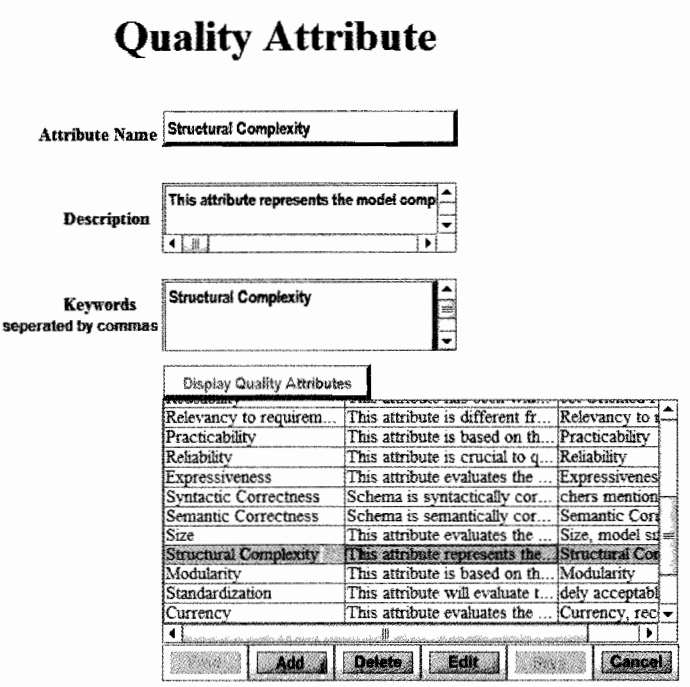

Figure 3.8. *Définition d'un attribut de la qualité*

L'outil de définition des métriques de qualité permet de définir des métriques automatiques ou semi-automatiques, à l'aide d'un langage exécutable de définition de métriques. Un exemple d'expression de métriques est décrit dans le tableau 3.7 [KER 09].

```
<metric name="NB_Classes_Metric" domain="model" >
<description>The number of classes belonging to a model.</description>
<projection globalrelation="true" target="class" condition="id !=“” />
</metric>
```

Tableau 3.7. *Expression de la métrique calculant le nombre de classes*

Le language XML (*eXtensible Markup Language*) a été choisi comme moyen de communication entre l'outil de qualité et les ateliers de génie logiciel (AGL) du marché. En effet, comme nous souhaitons rester compatibles avec les AGL du marché et une grande majorité de ces outils permet de générer des fichiers XML pour stocker les modèles qu'ils génèrent.

Enfin, l'outil de définition des recommandations met à disposition de l'expert qualité un éditeur permettant de définir des guides méthodologiques et des actions d'amélioration qui seront associés aux attributs de qualité.

3.4.2.2. *L'évaluation de la qualité*

Nous illustrons notre approche en déroulant une session d'évaluation de la qualité. Nous considérons pour cela le modèle conceptuel dont la version simplifiée pour des raisons de lisibilité est décrit dans la figure 3.12 en annexe, section 3.6.2 de ce chapitre. Conformément à la démarche méthodologique définie précédemment, l'étape Identifier et Planifier (*Plan*) débute par l'expression d'un objectif (but) de qualité. Dans notre exemple, l'analyste souhaite améliorer la compréhension de son modèle conceptuel. Le processus génère ensuite un ensemble de questions pour préciser ce but. Cette génération de questions exploite les termes de l'objectif qualité et les mots-clés associés aux patrons de qualité (figure 3.9).

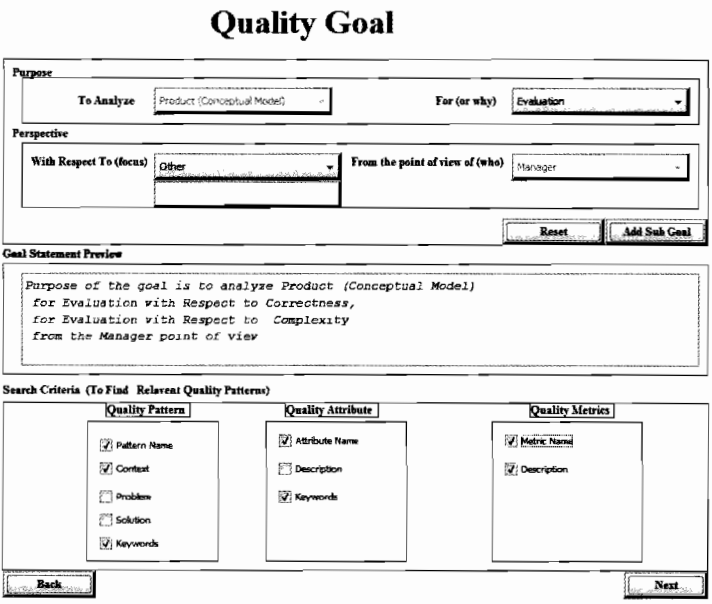

Figure 3.9. *Saisie d'un objectif de qualité*

Suite à la réponse de l'analyste, la liste des patrons de qualité, susceptibles de réaliser au mieux l'objectif ainsi précisé lui est proposée. Dans notre exemple, cette étape aboutit à deux patrons qui sont « complétude du modèle » et « complexité du modèle ».

Les patrons de qualité ainsi ciblés s'appuient sur les attributs suivants :

i) complétude : englobe la complétude syntaxique (par rapport aux exigences de la notation utilisée) et la complétude sémantique du modèle (par rapport à l'expression des exigences des utilisateurs) ;

ii) taille : représente la complexité d'un modèle en tenant compte du nombre d'éléments de modélisation qu'il contient ;

iii) complexité structurelle : définit la complexité en fonction des relations structurelles entre les éléments du modèle (association, héritage, agrégation, etc.) ;

iv) complexité sémantique est liée à la diversité des sujets ou contextes couverts par un modèle.

L'évaluation de cette métrique s'appuie sur une ontologie de domaine. Les éléments du modèle sont appariés aux concepts de l'ontologie. Un regroupement en « clusters » est ensuite fait sur la base de calculs de distances entre les concepts de l'ontologie ainsi identifiés.

L'évaluation des métriques associées à ces attributs produit les valeurs du tableau 3.8.

Les résultats liés à la complétude traduisent le fait que certains besoins des utilisateurs ne sont pas couverts par le modèle. De plus, 18 % seulement des multiplicités ont été définies dans le modèle initial et seulement 50 % des associations sont nommées. Concernant la *taille* et la *complexité structurelle,* les valeurs des métriques sont élevées reflétant une difficulté à comprendre le modèle.

Les recommandations correctives suivantes sont suggérées :

- pour améliorer la complétude :
  - intégrer les besoins omis (les catégories non modélisées) ;
  - compléter la définition des multiplicités ;
  - nommer toutes les associations du modèle ;
- pour améliorer la complexité :
  - mettre en facteur les associations et attributs ;

- appliquer le patron de conception GRASP de cohésion forte ;
- diviser le modèle (pour réduire sa complexité sémantique) ;
- inspecter tous les cycles pour supprimer les éventuelles redondances.

Attributs de qualité	Métrique	Valeur
Complétude	Taux de couverture des besoins	0.7
	Taux de définition des multiplicités	0.1818
	Taux de nommage des associations	0.5
Taille	Nombre de classes	46
	Nombre d'attributs	174
	Nombre d'opérations	120
Complexité structurelle	Nombre d'associations	35
	Nombre de classes-associations	8
	Nombre d'agrégations	1
	Nombre de compositions	6
	Nombre de généralisations	14
	Profondeur maximale des héritages	2
	Nombre de cycles	7
	Degré de non redondance	0.927
Complexité sémantique	Nombre de « clusters »	2

Tableau 3.8. *Valeurs de qualité du modèle initial*

L'étape exécuter (*Do*) consiste à appliquer les recommandations. Cette étape est manuelle. Le résultat est donné à l'annexe, section 3.6.3.

L'étape vérifier (*Check*) réapplique l'évaluation, dont les résultats sont indiqués au tableau 3.9.

Attributs de qualité	Métrique	Valeurs initiales	Valeurs après transformation	
			Module 1	Module 2
Complétude	Taux de couverture des besoins	0.7	1	1
	Taux de définition des multiplicités	0.1818	1	1
	Taux de nommage des associations	0.5	1	1
Taille	Nombre de classes	46	33	23
	Nombre d'attributs	174	77	117
	Nombre d'opérations	120	44	84
Complexité structurelle	Nombre d'associations	35	11	14
	Nombre de classes-associations	8	1	7
	Nombre d'agrégations	1	1	0
	Nombre de compositions	6	3	3
	Nombre de généralisations	14	17	6
	Profondeur maximale des héritages	2	3	1
	Nombre de cycles	7	0	1
	Degré de non redondance	0.927	1	1
Complexité sémantique	Nombre de « clusters »	2		

Tableau 3.9. *Valeurs de qualité du modèle transformé*

La transformation essentielle est celle de la division du modèle selon deux contextes du domaine : la gestion du personnel et la paie. Cette division a conduit aussi à la diminution de la taille et de la complexité structurelle de chaque module. Le modèle a également été complété en ajoutant les sous-classes manquant dans le modèle initial. Les multiplicités et les noms des associations ont été ajoutés.

Lors de l'étape de révision (*Act*), l'analyste entérine ou non les recommandations préconisées et décide ou non d'itérer un nouveau cycle d'évaluation et d'amélioration de la qualité.

Nous avons déroulé l'approche sur différents exemples à l'aide du prototype. Ce dernier s'avère utile, offrant un support utile à l'expert. A ce jour, il est interfacé avec StarUML. Une utilisation dans des conditions réelles sur des exemples de grande taille est prévue.

3.5. Conclusion

Les approches de qualité connaissent un vif succès dans les entreprises. Elles constituent le moyen de fluidifier les processus organisationnels, internes ou externes, et de rassurer les partenaires, clients et fournisseurs. Les systèmes d'information n'échappent pas à cette tendance. La qualité des SI devient un objectif crucial pour les Directions des SI. La nature multidimensionnelle du SI conduit naturellement à une évaluation multiple de sa qualité. Les modèles, notamment conceptuels, sont une préfiguration du SI. La mesure de leur qualité permet d'anticiper celle du SI futur. Les modèles de données sont des garants de la qualité de l'information produite par le SI.

Ainsi, dans ce chapitre, nous avons focalisé notre discours sur la qualité des modèles de données. Un état de l'art a montré les nombreuses recherches sur le sujet mais aussi leur faible appropriation par les experts en entreprise. La description de notre approche a permis d'illustrer concrètement cette recherche foisonnante. Cette dernière a été validée par une enquête en vraie grandeur auprès d'un grand nombre de praticiens d'entreprise et d'utilisateurs. Un prototype support de l'approche a permis d'en attester la faisabilité et l'utilité.

Un modèle de qualité est le garant d'une information fiable. Il permet de maximiser notamment la fiabilité, l'exactitude, l'intégrité et la validité des données.

3.6. Annexes

3.6.1. *Deux des huit modèles conceptuels utilisés pour l'expérimentation*

Dans ce premier modèle (M1), on ne distingue pas les fonctions exercées par les médecins. Ce choix est pertinent lorsque la validation doit être effectue par des utilisateurs non professionnels. Le concept de hiérarchie de généralisations est évité, au moins une première étape.

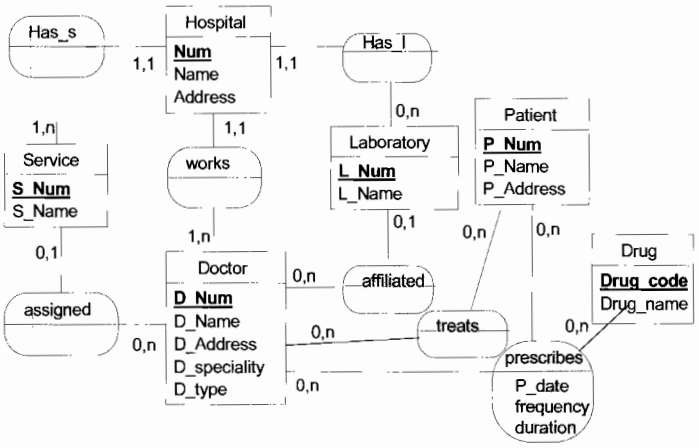

Figure 3.10. *Modèle d'expérimentation M1*

Dans ce troisième schéma (M3), une hiérarchie de généralisation est utilisée pour représenter les trois fonctions de médecins. Ce choix enrichit considérablement l'expressivité du schéma mais conduit aussi à plusieurs réplications d'associations.

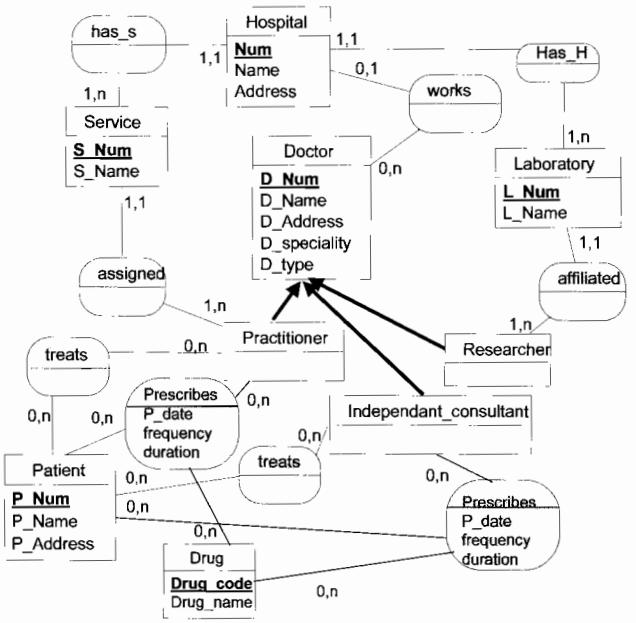

Figure 3.11. *Modèle d'expérimentation M3*

3.6.2. *Le modèle initial*

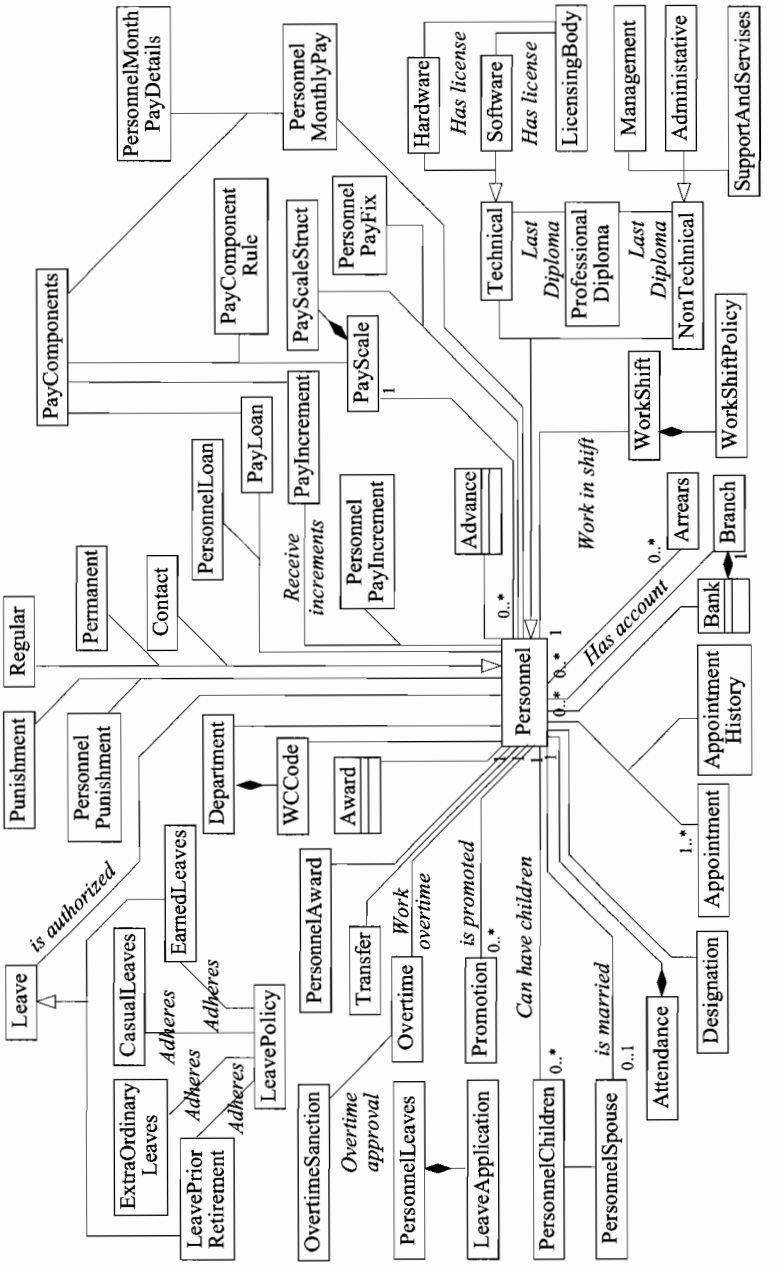

Figure 3.12. *Le modèle initial*

3.6.3. *Le résultat de la transformation*

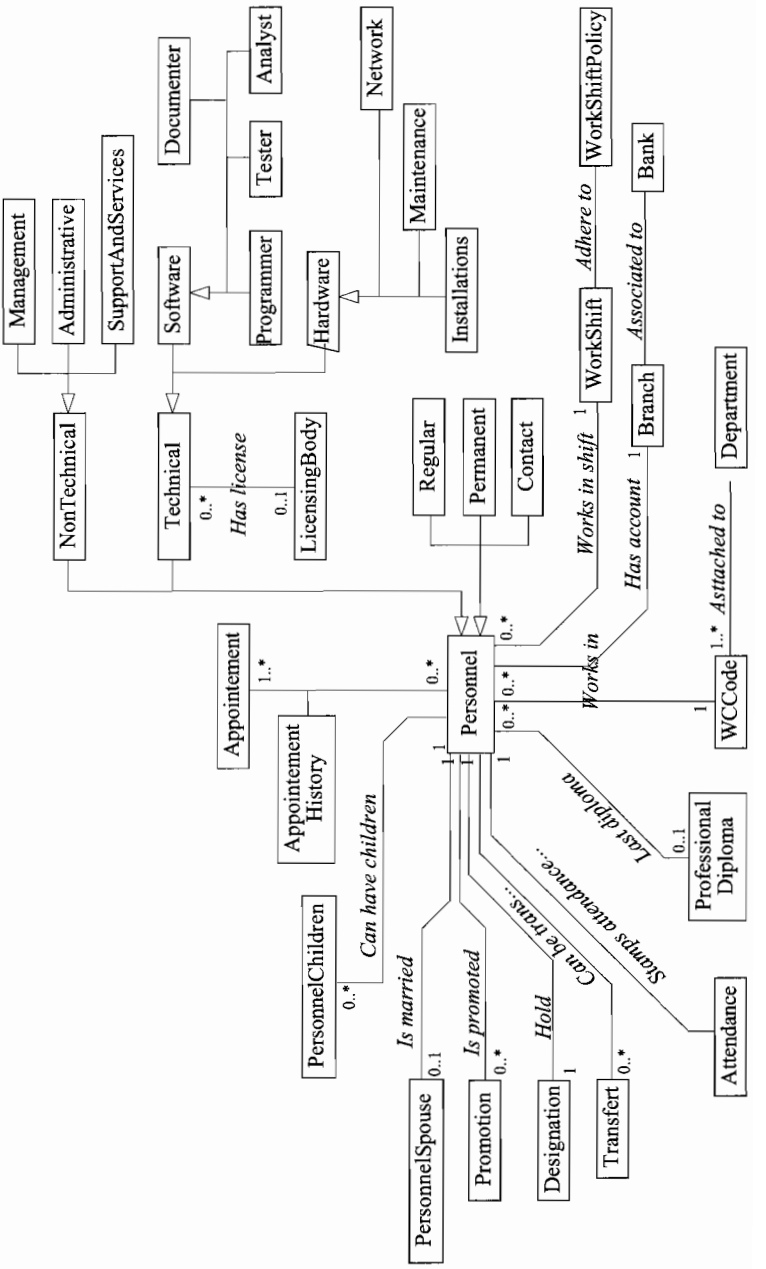

Figure 3.13. *Le module de gestion du personnel*

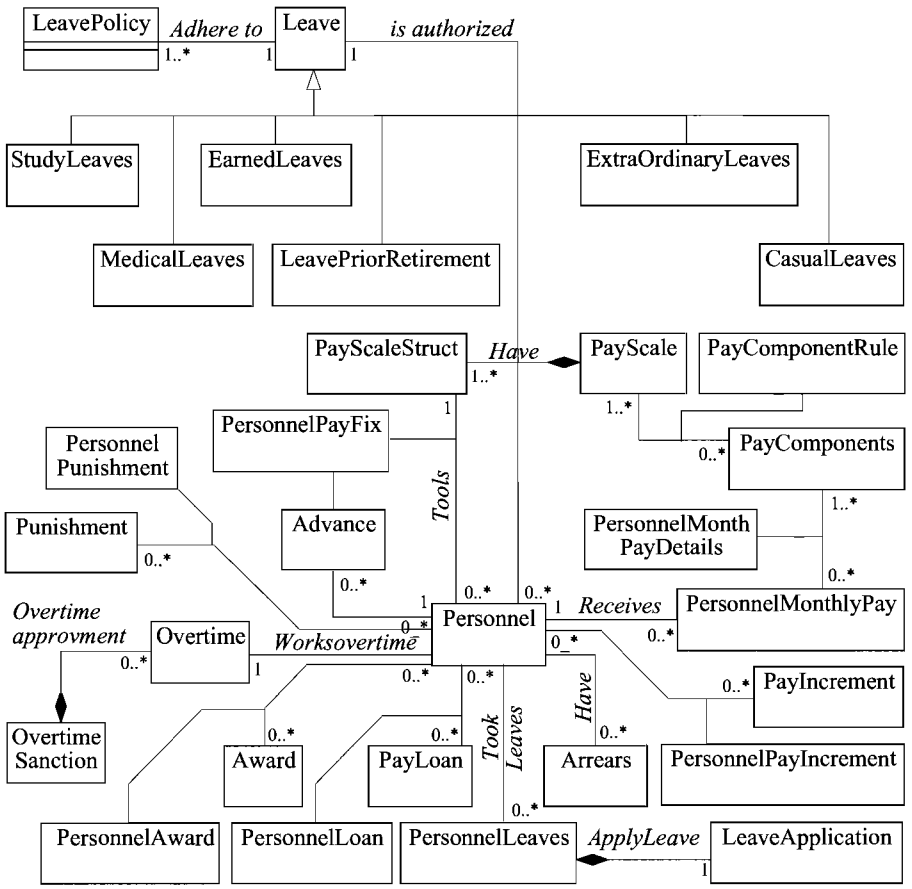

Figure 3.14. *Le module de gestion de la paie*

3.7. Bibliographie

[BAR 95] BARBACCI M., KLEIN M.H., LONGSTAFF T.A., WEINSTOCK C.B., Quality Attributes, Technical Report CMU/SEI-95-TR-021, ESC-TR-95-021, Carnegie Mellon University, 1995.

[BAS 94] BASILI V.R., CALDIERA G., ROMBACH H.D., « The Goal Question Metric Approach », dans J.J. Marciniak (dir.), *Encyclopedia of Software Engineering,* vol. 1. John Wiley and Sons, New York, NY, p. 528-532, 1994.

[BAT 92] BATINI C., CERI S. NAVATHE S.B., *Conceptual database design : An entity relationship approach,* Benjamin Cummings, Redwood City, CA, 1992.

[CHE 02] CHERFI S., AKOKA J., COMYN-WATTIAU I., « Conceptual modeling quality – from EER to UML schemas evaluation », *Proceedings of the 21st International Conference on Conceptual Modeling, ER2002*, S. Spaccapietra, S.T. March, Y. Kambayashi (dir.), Tampere, Finlande, 2002.

[CHE 08] CHERFI S., COMYN-WATTIAU I., AKOKA J., « Quality Patterns for Conceptual Modeling », *Proceedings of the 27st International Conference on Conceptual Modeling ER2008*, Lecture Notes in Computer Science, vol. 5231/2008, p. 142-153, 2008.

[CHE 10] CHERFI S., AKOKA J., COMYN-WATTIAU I., « Qualité perçue des schémas conceptuels – Etude comparée informaticiens *versus* utilisateurs finaux », *Ingénierie des systèmes d'information, RSTI série ISI,* vol. 15, n° 5, 2010.

[CHI 94] CHIDAMBER S.R., KEMERER C.F., « A Metrics Suite for Object Oriented Design », *IEEE Trans. Softw. Eng.*, vol. 20, n° 6, p. 476-493, 1994.

[COB 11] www.isaca.org/Knowledge-Center/COBIT/Pages/Overview.aspx

[DEM 86] DEMING W.E., *Out of the Crisis*, MIT Center for Advanced Engineering Study, 1986.

[GAF 81] GAFFNEY J.E., « Metrics in software quality assurance », *Proceedings of the ACM SIGMETRICS workshop/symposium on measurement and evaluation of software quality*, p. 126-130, 1981.

[GEN 00] GENERO M., JIMÉNEZ L., PIATTINI M., « Measuring the quality of entity relationship diagrams », *Proceedings of 19st International Conference on Conceptual Modeling, ER2000,* vol. 1922, p. 513-526.

[HOU 97] HOUDEK F., KEMPTER H. « Quality Patterns – An Approach to Packaging Software Engineering Experience », *ACM Software Engineering Notes*, n° 22, 1997.

[HOX 98] HOXMEIER J.A., « Typology of database quality factors », *Software Quality Journal*, n° 7, p. 179-193, 1998.

[HSU 08] HSUEH N.L., CHU P.H., CHU W., « A Quantitative Approach for Evaluating the Quality of Design Patterns », *J. Systems and Software*, 2008.

[ISO 11] ISO 9126 : http://www.iso.org/iso/home.html.

[ITI 11] http://www.itil-officialsite.com.

[KER 09] KERSULEC G., CHERFI S.S., AKOKA J., COMYN-WATTIAU I., « Un environnement pour l'évaluation et l'amélioration de la qualité des modèles de systèmes d'information », *Actes du 28ème congrès INFormatique des ORganisations et Systèmes d'Information et de Décision INFORSID'09*, p. 321-344, 2009.

[KRO 95] KROGSTIE J., LINDLAND O.I., SINDRE G., « Towards a Deeper Understanding of Quality in Requirements Engineering », *Proceedings of CAISE 1995*, p. 82-95, 1995.

[KRO 01] KROGSTIE J., « A semiotic approach to quality in requirements specifications », *Proceedings of the IFIP TC8/WG8.1 Working Conference on Organizational Semiotics : Evolving a Science of Information Systems,* Kluwer, B.V. Deventer, The Netherlands, 2001.

[KRO 06] KROGSTIE J., SINDRE G., JORGENSEN H, « Process models representing knowledge for action : a revised quality framework », *European Journal of Information Systems 15*, p. 91-102, 2006.

[LAN 05] LANGE C.F.J., CHAUDRON M.R.V., « Managing Model Quality in UML-based Software Development », *Proceedings of the 13th IEEE International Worksop on Software Technology and Engineering Practice*, 2005.

[LAR 04] LARMAN G., *Applying UML and Patterns – An Introduction to Object-Oriented Analysis and Design and Iterative Development*, 3e édition, Prentice Hall, Upper Saddle River, NJ, 2004.

[LEV 95] LEVITIN A., REDMAN T., « Quality dimensions of a conceptual view », *Information Processing and Management*, vol. 31, n° 1, 1995.

[LIN 94] LINDLAND O.I., SINDRE G., SØLVBERG A., « Understanding Quality in Conceptual Modeling », *IEEE Software*, p. 42-49, 1994.

[MAR 93] MARCHE S., « Measuring the stability of data models », *European Journal of Information Systems*, vol. 2, n° 1, p. 37-47, 1993.

[MAR 02] MARJOMAA E., « Necessary Conditions for High Quality Conceptual Schemata : Two Wicked Problems », *Journal of Conceptual Modeling*, n° 27, 2002.

[MEH 10] MEHMOOD K., A Quality Based Approach for the Analysis and Design of Information Systems, Ph. D Thesis, Cnam Paris, 2010.

[MIL 56] MILLER G., « The Magical Number Seven, Plus or Minus Two : Some Limits on Our Capacity for Processing Information », *The Psychological Review*, vol. 63, p. 81-97, 1956.

[MOH 07] MOHAGHEGHI P., AAGEDAL J., « Evaluating Quality in Model-Driven Engineering », *Proceedings of the International Workshop on Modeling in Software Engineering, IEEE Computer Society*, 2007.

[MOO 05] MOODY D.L., « Theoretical and practical issues in evaluating the quality of conceptual models : current state and future directions », *Data and Knowledge Engineering*, vol. 55, n° 3, p. 243-276, 2005.

[PRE 01] PREISS O., WEGMANN A., « Stakeholder discovery and classification based on systems science principles », *Proceedings 2nd APQSC Asia-Pacific Conference on Quality Software*, IEEE, p. 194-198, 2001.

[SCH 98] SCHUETTE R., ROTTHOWE T., « The Guidelines of Modeling – An Approach to Enhance the Quality in Information Models », *Proceedings of the 17th International Conference on Conceptual Modeling (ER '98)*, p. 240-254, 1998.

[SER 07] SERRANO M.A., TRUJILLO J., CALERO C., PIATTINI M., « Metrics for data warehouse conceptual models understandability », *Information & Software Technology*, vol. 49, n° 8, p. 851-870, 2007.

[STY 00] STYLIANOU A.C., KUMAR R.L., « An integrative framework for IS quality management », *Communications of the ACM*, vol. 43, n° 9, p. 99-104, 2000.

[USR 96] USREY M.W., DOOLEY K.J., « The Dimensions of Software Quality », *Quality Management Journal*, vol. 3, Issue 3, G.S. Easton (dir.), Milwaukee, 1996.

[WAN 93] WANG R.Y., KON H.B., MADNICK S.E., « Data Quality Requirements Analysis and Modeling », *Proceedings of the Ninth International Conference on Data Engineering, IEEE,* 1993.

Chapitre 4

La cotation de l'information : approches conceptuelles et méthodologiques pour un usage stratégique

4.1. Introduction

Dans le contexte stratégique de sécurité globale et principalement de la défense, la comparaison de ce qu'on appelle la *cotation de l'information*, c'est-à-dire de l'évaluation des *fiabilité* et *véracité* de l'information telle qu'elle est reçue par un observateur, avec les travaux portant sur la qualité des données, peut au premier abord paraître assez incongrue. La qualité recouvre des secteurs plus nombreux, et d'une façon plus large et ouverte que ne le fait la cotation, forcément restreinte à certains champs, et limitée à sa seule fonction. Si l'on se réfère par exemple à la définition donnée sur le site de l'*International Association for Information and Data Quality*, on peut y lire à l'entrée *Qualité* : « la totalité des traits d'un produit ou d'un service qui répondent aux besoins exprimés ou implicites. La correspondance aux spécifications, attentes ou exigences d'utilisation. L'absence d'erreurs. »[1] La distance entre cotation et détermination de la qualité des données n'est toutefois qu'apparente. Les *besoins* des services de renseignement, leurs *attentes*, les *erreurs* qu'ils cherchent à éviter passent par l'utilisation d'informations qui doivent être formatées de manière à être exploitables. Or ce formatage suppose à son tour une estimation des *traits* de l'information – son origine et son contenu – qui désignent précisément le rôle de l'opération de cotation.

Chapitre rédigé par Philippe CAPET et Thomas DELAVALLADE.

1. http://iaidq.org/main/glossary.shtml.

A. Revault d'Allonnes précise les relations entre qualité et cotation de l'information : « lorsque la qualité des données qualifie l'information, plutôt que le modèle dont elle est issue, elle représente à quel point celle-ci s'intègre dans ce modèle qui, lui, décrit efficacement la réalité. La cotation déplace l'objet de l'évaluation de l'information vers sa réception et son destinataire » [REV 11, p. 36]. La cotation se plaçant selon l'auteur à l'intersection des représentations de l'incertitude de l'information, de la fusion d'informations et de la qualité des données, la confluence entre qualité et cotation, même si elle ne traduit aucune inclusion de l'une dans l'autre, est ainsi loin d'être négligeable, et son étude peut enrichir chacune des deux activités.

D'autre part, la cotation de l'information n'a pas à rester confinée au strict domaine de la défense. Son extension à d'autres domaines tels que l'intelligence économique, ou la notation de données informationnelles en toute généralité, peut tout à fait lui permettre d'épouser des secteurs aussi étendus que ceux où officie la mesure de qualité des données.

Dans ce chapitre, nous nous appuierons à titre illustratif sur la supposée affaire d'espionnage chez Renault au début de l'année 2011, et sur la masse d'informations successives et contradictoires qui en a rendu compte. Les grandes lignes en seront rappelées, puis nous exposerons les pratiques en usage dans la défense nationale pour coter l'information, étape inscrite dans un processus cyclique que nous présenterons, afin d'en relever des imprécisions et inconvénients. Des applications stratégiques, complexes et contemporaines, s'appuyant sur ces processus, seront alors proposées afin de présenter un ambitieux projet en cours, qui vise à traiter de la cotation et de ses applications : le projet CAHORS.

4.2. Un exemple paradigmatique de cas d'usage : la supposée affaire d'espionnage chez Renault

4.2.1. *L'affaire de départ et son retentissement*

L'affaire du supposé espionnage dont auraient été victime le constructeur Renault, et coupables trois hauts cadres de l'entreprise, a occupé une très importante place dans l'actualité et les médias au premier semestre 2011. Pour rocambolesque et pleine de rebondissements qu'elle ait été, elle mérite une attention soutenue car elle illustre les besoins en matière d'évaluation de l'information, tant selon son degré de véracité que par les valeurs qu'on attribuerait à la fiabilité des divers protagonistes. Véritable cas d'école en cette matière, son caractère condensé dans le temps et abondamment suivi par les journalistes dans la mesure de leurs capacités, en fait également un cas d'usage pour l'étude de la compréhension affinée des flux d'informations.

Nous avons représenté un schéma rappelant les diverses étapes de cette affaire. Construit à partir du moteur de recherche du site du *Monde* (www.lemonde.fr) (moteur développé par la société Sinequa), il propose un histogramme des nombreux articles produits par le site, exprimés selon une abscisse représentant les jours de diffusion en ligne, de début janvier à début mai 2011.

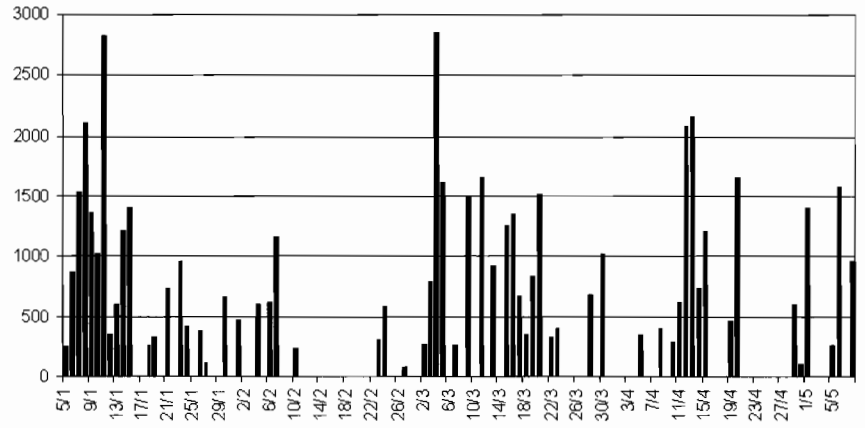

Figure 4.1. *Nombre de mots présents dans les articles du monde.fr sur l'affaire Renault au fil du temps*

Le schéma fait apparaître les principaux rebondissements par l'ampleur qui est alors accordée aux révélations et coups de théâtre qui surviennent en ces quelques mois. Le déclenchement de l'affaire et ses pics correspondent à des événements que nous rappelons ici sous forme de liste des articles du Monde.fr mis en ligne aux différentes périodes, avec à grands traits trois phases séparées par des creux d'actualité : phase 1, du 5 janvier au 10 février, les suites d'accusation de la part de Renault ; phase 2, du 23 février au 30 mars, le retournement de situation et la mise en cause des services de sécurité de Renault ; phase 3, du 5 avril au 6 mai et suivants, les soubresauts des enquêtes, et les commentaires à froid d'éditorialistes, d'universitaires ou de faiseurs d'opinion :

– 5 janvier 2011 : annonce de la mise à pied de trois cadres pour la transmission d'informations sensibles ;

– 6 janvier : le ministre de l'industrie juge sérieuse l'affaire d'espionnage ;

– 7-8 janvier : des soupçons d'espionnage se renforcent, une piste chinoise est évoquée ;

– 9-10 janvier : des comptes dissimulés auraient été ouverts en Suisse pour les présumés coupables ;

– 10 janvier : les cadres sont convoqués pour licenciement ;

– 13-14 janvier : Renault porte plainte avec référence à une puissance étrangère, la Direction centrale du renseignement intérieur (DCRI) est mise à contribution ;

– 23 janvier : le président de Renault affirme être certain qu'il s'agit d'une affaire d'espionnage ;

– fin janvier : perquisition de la DCRI chez Renault, tension de Renault lors les procédures judiciaires ;

– 6 février : nouvelles lumières sur les méthodes d'enquête de Renault ;

– 23-24 février : échec de la commission rogatoire en Suisse pour les supposés comptes des trois cadres ;

– 2 mars : premières questions sur une possible manipulation de Renault ;

– 3 mars : pas de trace d'espionnage selon la DCRI ;

– 4 mars : la direction doute de l'espionnage, rumeurs de démission de dirigeants ;

– 9 mars : l'informateur anonyme de Renault avait reçu de fortes sommes ;

– 11 mars : garde à vue pour des responsables de la sécurité chez Renault ;

– 13 mars : le procureur écarte l'hypothèse d'espionnage, Renault présente ses excuses aux accusés ;

– 15 mars : mise en examen du n° 2 de la sécurité de Renault, le compte en Suisse lui appartenait ;

– 23 mars : il est licencié pour escroquerie ;

– 28 mars : on apprend que, dès la mi-février, Renault doutait d'un espionnage ;

– 11 avril : le numéro deux de Renault quitte ses fonctions, les accusés seront indemnisés.

La conclusion de l'affaire, qui semble n'être pas encore close en avril au vu des articles encore émis sur la figure 4.1 dans les semaines suivantes, consiste en réalité en des bilans rétrospectifs, des tribunes plus ou moins moralisantes, des énumérations de leçons tirées de l'affaire – sans que soit garantie l'absence de réitération d'une future histoire quasi-semblable. L'analyse des divers rebondissements dans le flux d'informations reçues et transmises par la presse et les autres intermédiaires d'information n'aurait-elle pas mené, si elle avait été suffisamment

méthodique, à davantage de circonspection, voire au dégonflement de l'affaire en temps utile ? Plus précisément, quelle serait la méthode d'analyse à appliquer pour circonscrire ce type de crise informationnelle et médiatique avant qu'elle ne dégénère, quels en seraient les concepts constitutifs à comprendre pour la juguler avant qu'elle ne s'envenime ?

L'un des objets de ce chapitre consiste à décrire le type d'analyse et les concepts employés dans une branche bien rodée pour ce genre d'activités, celle du renseignement militaire. Nous verrons également que ces méthodes gagneraient à leur tour à être clarifiées et enrichies.

4.2.2. *Des problèmes de fiabilité et de véracité de l'information reçue et transmise*

Lorsque l'on examine comment s'est déroulée cette affaire telle qu'elle a été rapportée par la presse, et parfois dont les éléments ont été révélés par elle, bien des questions se posent : quelles ont été les sources successives, directes ou indirectes, qui ont accompagné ce feuilleton de quatre mois ? A qui étaient-elles attachées, et vers qui émettaient-elles des informations ? Quelle était leur fiabilité respective aux yeux du récepteur de cette information, et qu'entendrions-nous exactement par cette fiabilité supposée, c'est-à-dire : sur quels critères et par quelle méthode l'établirait-on ? Les successives informations, d'abord considérées comme vraies, potentiellement vraies puis fausses méritent également une classification, et une méthode pour les placer sur une échelle appropriée. En outre, une étude plus approfondie des fiabilités et véracités aurait permis d'aller plus loin dans l'analyse : quels ont été les relais d'information et selon quelles relations ou quels réseaux entre les diverses sources, quelle pourrait être l'intention explicite ou cachée de ceux qui les véhiculent, y a-t-il une tentative possible de désinformation de la part d'une partie prenante, des rumeurs sont-elles utilisées, etc. ? Ces questions complexes ne peuvent être abordées que si l'étape de cotation a été menée à bien avec rigueur.

4.2.3. *Quelle réaction adopter, comment procéder pour s'assurer de la qualité de l'information au préalable et* a posteriori *?*

L'affaire Renault du premier semestre 2011 s'inscrit manifestement dans un registre d'intelligence économique et stratégique. Pour autant, les questions soulevées en cours de route (espionnage, agents doubles, etc.) relèvent également d'un domaine plus coutumier de ces termes : celui du renseignement du ressort des forces armées. Il est un fait que ces deux domaines tendent à se rapprocher voire se confondre, par de multiples aspects : partage de responsabilités, travail en commun pour des investigations sensibles, voire échange de domaine en cours de carrière pour tel ou tel responsable. Il n'en reste pas moins que malgré son essor

et sa professionnalisation, l'intelligence économique ne dispose pas encore des moyens spécifiques, des éléments doctrinaux et de pratiques bien rodées que le renseignement militaire a su développer et perfectionner au cours de son histoire. En ce qui concerne la cotation de l'information en particulier, nous verrons en sections 4.3 et 4.4 comme cette avance est marquée et pourrait inspirer d'autres secteurs.

Nous verrons que dans la doctrine militaire, divers points analytiques sont reliés selon des axes bien précis, chacun des points disposant par ailleurs d'une certaine autonomie et étant défini par des classements *ad hoc*. D'ores et déjà, dans le cadre de l'affaire Renault comme dans bien d'autres qui lui ressemblent, deux points primordiaux peuvent être identifiés : celui de la *source* de l'information, et celui du *contenu* de celle-ci. Comme souligné en section 4.2.2, ces deux notions correspondent à des questions qu'il serait légitime de se poser *a posteriori* d'un tel *imbroglio* tournant au *fiasco*. *A priori*, de même, elles devraient permettre, une fois correctement analysées, de contribuer à y remédier voire à l'anticiper pour prévenir sa survenue.

Quelles sont les diverses sources d'information mentionnées dans la chronologie de l'affaire ? On peut en répertorier d'assez nombreuses et de diverses natures, qui de surcroît évoluent au cours du temps :

– un informateur anonyme de Renault à l'été 2010, qui par la suite est désigné par la presse comme étant responsable de la sécurité chez Renault ;

– des informateurs de ce responsable, qu'il désigne comme étant intermédiaires détenteurs des sommes versées au départ par Renault pour être informé sur les cadres soupçonnés, et qui n'ont semble-t-il jamais existé ;

– des cadres mis en examen qui contestent les faits reprochés et s'adressent aux médias ;

– les avocats des prévenus, citant la plupart du temps leur client, et parfois s'exprimant en leur nom propre ;

– des dirigeants de Renault dûment nommés ;

– des dirigeants anonymes de Renault ;

– des membres gouvernementaux ou de la fonction publique anonymement cités par des journalistes ;

– des membres gouvernementaux ou de la fonction publique nommés (ministres, policiers, magistrats, etc.) ;

– et enfin, les journalistes qui ont appris à leur public l'évolution de l'affaire telle qu'ils la comprenaient au fil du temps.

Notons que la totalité des sources hormis celles qui émanent des journalistes eux-mêmes ne sont évidemment citées, dans ce contexte, que par la presse ; l'intermédiation assurée par elle correspond au sens originel du travail des *médias*. Ceci implique donc que les sources sont successives : un journal ou un site d'information d'Internet rapporte qu'un tel a déclaré, a laissé entendre que, etc. Dans ce cas de figure, nous sommes forcément dans la situation, parfois complexe et ardue à évaluer, de type « A a dit que B a dit que... ». Ce cas de figure n'est cependant pas l'unique à aborder : pour l'analyste qui cherche à mesurer la cotation d'une information, l'informateur à noter peut fort bien s'être intéressé à la source directement – autrement dit, *immédiatement*.

Quant au contenu des informations véhiculées dans l'affaire Renault, il recouvre lui aussi diverses facettes :

– quelle est la valeur qui lui est apportée par la source ? En particulier, des adverbes significatifs (« peut-être », « probablement », « avec certitude », etc.) ou le mode conditionnel ont-ils été utilisés ?

– l'information a-t-elle été recoupée ?

– le niveau de confiance en la source qui l'a produite entre-t-il en ligne de compte aux yeux de qui l'a rapportée ?

Ces différents points, tant sur les variétés de sources que sur les facettes potentielles du contenu de l'information, servent de base aux règles promulguées par la défense.

4.3. Les règles dans le domaine de la défense

Pour le secteur de la défense, la question de la cotation s'inscrit dans le domaine général de la fonction du renseignement militaire. Le renseignement est défini dans de nombreux documents régulièrement remis à jour, dont la plupart font l'objet de doctrines et autres documents officiels, qui dictent la conduite à tenir pour effectuer cette mission [INS 03, TTA 01].

Le renseignement militaire possède plusieurs définitions ; il convient d'abord de distinguer le renseignement comme activité (on parle alors de *fonction* de renseignement) et le renseignement comme résultat de cette activité[2]. L'activité ou la fonction de renseignement rend donc disponibles, abstraction faite de ce qui en est à l'origine, des renseignements obtenus à partir d'*informations*, provenant elles-

2. Le renseignement peut aussi désigner l'ensemble des organismes qui effectuent ces activités.

mêmes de *données*. La pyramide dite du renseignement, fréquemment utilisée, traduit cette chaîne de traitement.

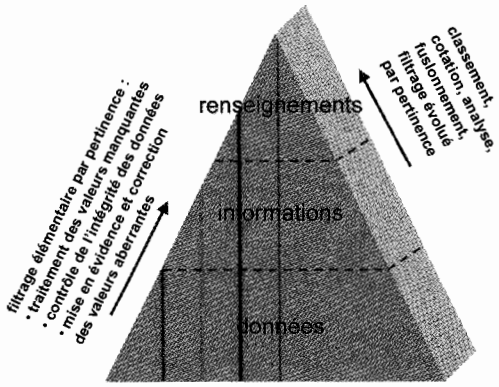

Figure 4.2. *La pyramide du renseignement*

Pour traduire cette partie d'activité ou de fonction, on parle dans la terminologie de l'armée française de *valorisation* de l'information en renseignement.

D'après la doctrine interarmées du renseignement ([INS 03], qui complète le document doctrinal [TTA 01]), la valorisation de l'information en renseignement passe par la réalisation, selon un processus séquentiel, d'un certain nombre de tâches regroupées dans un ensemble de phases. A l'intérieur de la fonction de renseignement et pour ordonner ces dernières, il est défini ce que les doctrines nomment le *cycle* du renseignement. La figure 4.3 dépeint ce cycle dans sa version simplifiée.

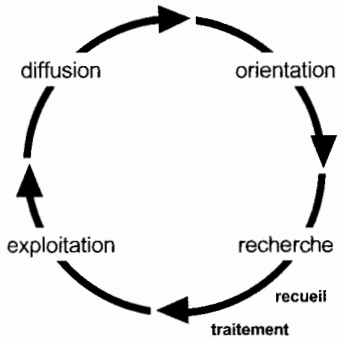

Figure 4.3. *Le cycle du renseignement militaire*

Ce cycle représente une suite d'opérations menées par les acteurs du domaine et animées par un coordinateur dévolu à cette tâche :

– la phase d'orientation regroupe les activités du haut vers le bas, visant à donner des directions de recherche aux organismes concernés, et les activités de détermination des missions de recherche à engager ou de demandes ciblées ;

– la phase de recherche combine l'acquisition et le recueil d'information auprès des sources (c'est-à-dire des *capteurs*) ;

– la phase de diffusion vise à transmettre l'information (une fois valorisée en renseignement) aux personnes qui l'ont demandée, et permet également de reboucler le cycle car cette valorisation invite à de nouvelles recherches.

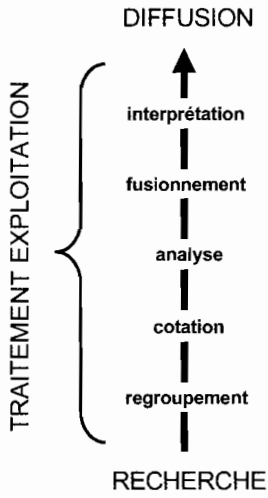

Figure 4.4. *Vue détaillée de la phase d'exploitation dans le cycle du renseignement*

C'est la phase d'exploitation (et de traitement) qui nous intéresse principalement ici. Elle se découpe en cinq tâches qui se suivent :

– le regroupement (classement des informations de même nature) ;

– la cotation (évaluation de la qualité des informations et de leurs sources) ;

– l'analyse (extraction au sein des informations des éléments les plus significatifs) ;

– le fusionnement (intégration dans un schéma commun des éléments informationnels contenus dans différentes informations, de façon à obtenir un contenu enrichi) ;

– l'interprétation (évaluation de l'intérêt et de la portée – ensemble des conséquences induites – de l'information).

Un schéma en découle, sur une portion du cycle du renseignement (figure 4.4).

4.3.1. *Approche doctrinale : classification et pratiques*

Dans le cadre du cycle du renseignement, pour réduire la tâche des analystes en charge de la phase d'exploitation, il est important de réduire au maximum le volume de données traitées. Une façon d'y parvenir consiste à procéder à une sélection des données susceptibles d'être utiles pour répondre à la demande de renseignement avant même de procéder à leur interprétation sémantique fine. Cette sélection, réalisée lors de la phase de recherche du cycle du renseignement, relève bien souvent du simple bon sens. Il s'agit par exemple de ne retenir parmi les données recueillies par des capteurs sur un théâtre d'opérations que celles qui ont été observées dans la zone géographique pertinente au regard du besoin exprimé. On a alors affaire à un filtrage simple des données, qui s'appuie le plus souvent sur les métadonnées associées aux données recueillies.

Une fois que ces données ou informations ont ainsi été produites et pour partie triées, elles sont livrées à exploitation. Dans la tâche de cotation qui s'ouvre alors, après un regroupement qui découle de la phase de recherche, l'analyste en charge est alors amené à utiliser une double échelle pour évaluer source de l'information et contenu de celle-ci. Selon la doctrine, ces deux échelles de six degrés chacune doivent se lire selon le tableau 4.1.

Fiabilité de la source	**Véracité de l'information**
A – Totalement fiable	1 – Corroborée par d'autres sources
B – Habituellement fiable	2 – Probablement vraie
C – Assez fiable	3 – Peut-être vraie
D – Rarement fiable	4 – Véracité douteuse
E – Non fiable	5 – Véracité improbable
F – La fiabilité ne peut être estimée	6 – La véracité ne peut être estimée

Tableau 4.1. *La cotation de l'information dans les armées*

Cette classification officielle en France est identique à celle de l'organisation du traité de l'atlantique nord (OTAN, voir [STA 03]), qui a adopté les termes anglais *reliability* et *credibility*, respectivement francisés en fiabilité et véracité. Auparavant, l'armée française utilisait des termes légèrement distincts des présents, les termes de *qualité* de la source de l'information, et celui de *valeur* du contenu de l'information. Si l'adaptation au contexte allié est aisément compréhensible, l'origine francophone des termes est elle-même fort significative, et en parfaite harmonie avec le présent recueil d'articles : la source est évaluée selon sa qualité ; lors de l'étape de cotation, on aurait fort bien pu concevoir que la qualité recouvre à la fois la source de l'information et son contenu.

La figure 4.5 représente l'exemple très simplifié de l'affaire Renault au fil du temps entre début janvier 2011 et mi-avril 2011, avec pour source étudiée l'agent de sécurité dénonciateur (traits pleins correspondant à l'ordonnée de F à A), et pour information le fait qu'il y ait un espionnage chinois avec le soutien de cadres de l'entreprise (traits en pointillés correspondant à l'ordonnée de 6 à 1).

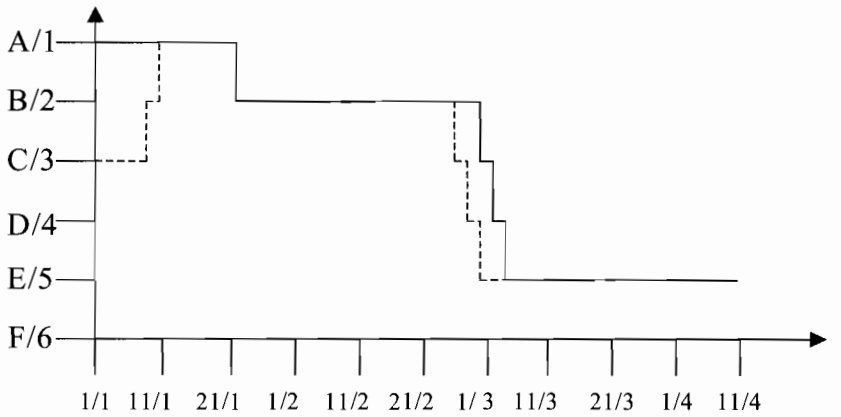

Figure 4.5. *Une représentation simplifiée de la cotation dans l'affaire Renault*

Bien entendu, les deux courbes ne peuvent être que des fonctions créneaux, puisqu'à chaque étape il n'y a que six valeurs possibles pour les deux échelles de cotation. D'autre part, les deux courbes sont très proches l'une de l'autre, ce qui est problématique si l'on veut considérer deux axes censés être indépendants : ici comme dans la plupart des cas, fiabilité et véracité seront, subjectivement, estimées comme étant quasi-semblables quant à la confiance que l'on peut avoir dans l'information véhiculée. Mais là n'est pas le seul problème rencontré par l'usage de cette classification, ainsi que nous le détaillons dans la section 4.3.3.

4.3.2. *Des obstacles logiques et conceptuels aux doctrines de défense*

Dans un louable souci d'analyse et de discrimination des activités, les documents de l'OTAN ou de la France posent clairement que les dimensions de fiabilité et de véracité doivent être « considérées indépendamment l'une de l'autre ». Ceci a pour conséquence que n'importe quelle combinaison des deux éléments lettre/chiffre est *a priori* possible, soit 36 possibilités de ce qui est nommé *bigramme*. Cependant, cette indépendance est-elle souhaitable, ou tout bonnement vraisemblable ?

Nous suivons ici essentiellement le document de référence en usage à l'OTAN [STA 03], qui est le document le plus détaillé sur la question.

4.3.2.1. *Fiabilité de la source*

La fiabilité entre les notes A et E dépend en totalité de l'expérience antérieure qu'a l'analyste des informations fournies par cette source, et de la confiance plus ou moins grande établie à son égard de ce fait. Si une source est nouvelle pour l'analyste, la note F lui est attribuée sans que cela signifie qu'elle est moins fiable qu'une source notée E : il s'agit simplement d'enclencher le processus de cotation de cette source. Il y a au moins deux faits implicites ici :

– une source est réévaluée en fonction des informations qu'elle fournit : par retour d'expérience, la confiance qu'on lui accorde se renforce ou s'étiole et sa cote est donc amenée à varier. L'un des problèmes que pose cette réévaluation continuelle de la source est alors le suivant : sur quels éléments se base-t-on pour juger que la source a été digne ou indigne de confiance ? *A priori*, la réponse la plus simple revient à comparer les affirmations de la source avec la réalité observée indépendamment de celles-ci. Il y aurait alors rétroaction de la véracité en un temps antérieur des informations fournies par la source sur la fiabilité de cette dernière. Les deux processus de cotation ne s'effectuent donc pas en toute indépendance l'un de l'autre, quoiqu'avec un décalage dans le temps. Mais cette dépendance implicite n'est pas expliquée dans les documents doctrinaux de référence ; on ignore comment la mettre en pratique, et les analystes chargés de la cotation doivent *de facto* improviser sa réalisation ;

– la relation de confiance, essentielle pour coter la source, est censée s'établir au vu du passé de la source et des informations plus ou moins véraces fournies par elle. Mais pour fonder sa confiance, l'analyste peut avoir des présupposés sur la source ; il peut également avoir obtenu des recommandations ou des mises en garde concernant cette source, qui, pour modifier son jugement et la confiance accordée, n'entretiennent pour autant aucun rapport logique avec le passé informationnel de la source ; il peut également avoir des relations directes avec cette source et avoir

élaboré une relation de confiance, de méfiance ou de défiance en toute indépendance des informations qu'elle lui a fournies par le passé. En d'autres termes, l'acception de la confiance ici suggérée est par trop réductrice, et de toute manière fort imprécise.

4.3.2.2. *Véracité de l'information*

Même si le STANAG 2511 est le document le plus détaillé concernant les différentes cotes possibles de véracité, la formulation qui en est proposée est très ambiguë et nécessite un effort d'interprétation. L'accent est mis sur la confirmation ou l'infirmation par des sources d'une information transmise par une source originelle, d'où l'importance du recoupement et l'appel aux connaissances antérieures. Mais l'échelle de graduation n'est pas claire dès lors qu'il est fait, par exemple, référence à une source « indépendante » (cote = 2) sans qu'on sache par rapport à quoi. Il semble vraisemblable de penser que cette indépendance réfère à la relation qu'entretient la source avec d'autres sources pourvoyeuses d'information sur le même sujet : un des périls pour l'officier de renseignement, particulièrement accentué avec le renseignement de sources ouvertes, consiste en effet à se laisser intoxiquer par des sources multiples prodiguant la même information, tendant à lui faire accroire que l'information se confirme par recoupement, tandis qu'elle est seulement répétée par des sources secondaires qui copient leurs propos sur des sources primaires (le phénomène de buzz ne naît pas d'autre façon). Suivant cette interprétation, on peut être amené à formuler les clauses de véracité présentes dans la doctrine de la manière suivante :

– note = 1 : information qui confirme ce qui est déjà porté à la connaissance sur le même sujet, apportée par une source indépendante de ce qui est à l'origine de cette connaissance ;

– note = 2 : information dont la probabilité est regardée comme suffisamment établie au regard de la quantité et de la qualité de précédents rapports, mais dont la source n'est pas considérée comme certainement indépendante.

Les notes 3, 4, 5 et 6 sont inchangées, et ne font plus référence à l'indépendance des sources. La note 3 est celle d'une information « neutre » : elle n'apporte aucune nouvelle connaissance, elle n'en supprime pas non plus. De 4 à 5, l'information contredit « de plus en plus » ce qui est déjà connu – reste à savoir comment traduire cette notion. 6 enfin est l'analogue de F pour la fiabilité des sources et est une note à part, qui traduit la circonspection devant un phénomène nouveau, ici un contenu événementiel qui n'a pas d'équivalent connu, là le recours à une source à laquelle on n'a jamais eu affaire.

En acceptant cette reformulation, on constate que l'échelle de cotation est encore plus hétérogène que pour la fiabilité : les deux premières notes (dont la première n'est pas forcément une bonne note : la véracité n'est pas nécessairement renforcée par un recoupement d'autres informations) portent principalement sur l'indépendance des sources prodiguant des informations sur un même sujet, les notes 4 et 5 sur le niveau de contradiction avec les connaissances antérieures, la note 3 est une note « de Normand », comme on parle de réponse de Normand (« peut-être vraie » dit l'intitulé, mais « peut-être fausse » aussi bien…) et la note 6 manifeste la surprise et partant, engendre une certaine indécision. Or tous ces éléments hétérogènes peuvent entretenir des relations plus complexes selon qu'on les combine différemment ; par exemple, que dire si une source, peut-être dépendante des sources antérieures, corrobore pour partie ce que l'on savait et contredit pour une autre partie ce qui était tenu pour acquis ? N'est-il pas pertinent d'estimer si la source entretient une certaine relation de dépendance avec les sources primaires, et les contredits cependant ? Si la relation entre la source secondaire et les sources primaires est notoirement polémique et exécrable, le fait que la source secondaire les contredise pourrait (sous certaines hypothèses supplémentaires, concernant en particulier le niveau de confiance que l'on affecte à toutes ces sources) être interprété comme une corroboration implicite des premières sources, si l'analyste se fie davantage aux primaires qu'à la secondaire. Dans ce cas de figure, la note doit-elle être de deux ou de quatre à cinq ? Ou bien est-ce à dire que cette échelle de véracité opère déjà une agrégation d'évaluations qui masque la complexité des situations possibles ?

4.3.2.3. *Combinaison de fiabilité et véracité*

Outre ces critiques portant sur chacune des deux dimensions, rien n'est dit dans les textes sur l'interprétation à donner à des bigrammes différents et très hétérogènes. Prendra-t-on la même décision si une information est cotée B5 (source habituellement sûre, information d'exactitude peu probable), ou E2 (source pas sûre, information d'exactitude probable) ? Y a-t-il même une décision possible face à un tel cas de figure, hormis celle de réévaluer la fiabilité de la source à la baisse ou à la hausse ? En somme, le bigramme, s'il est réellement établi en toute indépendance pour ses deux dimensions, ne semble pas permettre dans certains cas de se faire une idée sur l'information prodiguée ; dans le cycle du renseignement, elle n'est alors que de maigre utilité.

On le voit, en dépit du caractère strict et rigoureux des documents officiels mûrement réfléchis et pratiqués et d'une méthodologie précise mise en place par eux, de nombreuses questions restent ouvertes et les problèmes soulevés sont complexes ; ils le deviennent d'autant plus lorsque l'analyste est comme noyé sous un flux d'informations croissants.

4.4. Au-delà de la cotation, des applications stratégiques

Nous avons évoqué en section 4.3 certaines fonctions plus évoluées de traitement de l'information en vue de sa valorisation en renseignement et connaissance, qui sont à entamer une fois l'étape de cotation franchie. En sources ouvertes, la veille informationnelle et l'anticipation de crises à l'aide de signaux faibles en font partie, tout comme la surveillance de réseaux sociaux à visée potentiellement criminelle voire terroriste, l'anticipation et la compréhension de rumeurs pernicieuses voire de désinformation clairement belliqueuse.

Ces diverses activités viennent *en aval* de la cotation de l'information : l'estimation de la fiabilité des sources et de la véracité des contenus constituent un préalable nécessaire pour chacune d'entre elles. Examinons les liens de trois d'entre elles avec la cotation qui les précède, à la lumière de l'affaire Renault.

4.4.1. *La détection et la lutte contre la désinformation, le suivi des rumeurs*

Dans l'affaire Renault, le terme en vogue de « désinformation » a été plusieurs fois utilisé par la presse. Il pouvait au départ s'agir d'une désinformation de la part des cadres s'ils avaient été convaincus d'espionnage, dirigée contre leur hiérarchie au profit d'une puissance étrangère. Puis, l'intermédiaire supposé du haut gradé du service de sécurité de l'entreprise a été soupçonné à son tour de désinformation à l'endroit de ce dernier, bien qu'il ait été présenté par ce cadre comme étant à la source d'informations antérieurement corroborées et utiles. Enfin, c'était au tour de ce cadre d'être présenté comme un désinformateur contre ses chefs de Renault.

Comme pour les éléments de la cotation, la première étape à franchir est d'ordre conceptuel et définitionnel : qu'entend-on par « désinformation », en évitant de galvauder le terme ? Les définitions sont variées dans la littérature spécialisée ou non, et le terme est si récent (début du XX^{e} siècle) qu'il manque encore de trouver un consensus. Il n'en reste pas moins qu'il est de plus en plus utilisé, parfois en substitution au terme de mensonge, et que chacun est conscient de l'importance croissante du concept entre autres depuis les développements des technologies de l'information et de la communication (TIC) et la pléthore d'écrits en source ouverte consécutifs à la démocratisation de la pratique d'Internet.

Il est à noter que l'étape préalable de cotation est ici fondamentale. Non que les informations peu véridiques ou émanant de sources douteuses soient à rejeter, contrairement à ce qui peut être fait sous d'autres finalités. Tout à l'inverse, il peut être plus pertinent de surveiller une source d'information régulièrement évaluée comme trompeuse, surtout si ses intentions ont été correctement analysées, afin de possiblement déceler une tentative de désinformation en ses prémices. C'est même

l'un des principes possibles de la désinformation que de propager des informations fausses, et non forcément crédibles, lorsqu'elle peut être assimilée au mensonge. De même, des rumeurs malveillantes en cours de constitution et en passe d'être répandues délibérément par une source ou un ensemble de sources dont la fiabilité est estimée comme faible pourraient être anticipées et circonscrites avant qu'elles prennent de l'ampleur, à condition que l'étape de cotation ait été correctement menée.

4.4.2. *La mesure de gravité informationnelle*

Dans le cadre d'une activité de renseignement, les informations doivent être hiérarchisées en fonction non seulement de leur qualité mais également de leur gravité. Dans l'exemple de l'affaire Renault, chaque information fournie n'a pas forcément le même poids, indépendamment de la cote qu'on lui a attribuée. Certaines sont anecdotiques (déclaration incidente de tel secrétaire d'état) et d'autres potentiellement centrales (rôle de la Chine dans une affaire d'espionnage industriel). Cela est fondamental pour pouvoir évaluer les risques associés à un contenu informationnel donné. Des méthodes de *scoring* et d'agrégation multicritères sont à même d'être utilisées à cet effet. Grâce à l'ontologie du domaine d'étude, c'est-à-dire au système de représentation des connaissances liées à ce domaine, elles tiennent compte d'une part des types d'événements jugés à risque, des acteurs impliqués, et d'autre part des marqueurs textuels de quantité. La notion de gravité étant relative aux intérêts visés, les méthodes à développer doivent être flexibles et paramétrables par un analyste.

4.4.3. *La remontée de réseaux sociaux*

En lien avec les questions de rumeur et de désinformation, l'étude des réseaux sociaux occupe une place croissante dans l'évaluation des risques liés aux TIC. Les utilisateurs de *Facebook*, *Twitter* et autres sites dévolus à la constitution de réseaux peuvent en effet être, pour les utilisateurs, à leur insu ou délibérément, à l'origine de propagations d'informations potentiellement dangereuses. Dans la vie « réelle », les réseaux d'individus existent bien entendu également, et leur compréhension aide à l'anticipation de risques potentiels. Or, la cotation de l'information, et tout particulièrement la fiabilité des sources en son sein, constitue un mode de lecture des réseaux potentiellement riche pour aboutir à une meilleure appréhension des mécanismes sous-jacents qui ont lieu entre les individus reliés. Idéalement, un analyste trouverait avantage à pouvoir visualiser les réseaux sociaux à l'étude sous forme cartographique indiquant les fiabilités des nœuds du graphe correspondant.

Soulignons que ce sont principalement les relations entre sources de fiabilités variées qui jouent ici un rôle central. Or, les relations entre sources d'information ne

sont que marginalement évoquées dans les doctrines militaires du renseignement. Seule la corroboration par plusieurs sources d'une même information s'en rapproche marginalement, encore n'est-ce que pour l'estimation de la véracité de l'information. Et cependant les relations entre sources auxquelles ne serait pas attribué le même degré de fiabilité sont la principale clé de compréhension de propagation de l'information. Dans le cas de l'affaire Renault, les relations entre les divers protagonistes (réels ou fictifs), une fois qu'elles ont fini par être élucidées, ont grandement accru les connaissances qui pouvaient être tirées préalablement du contexte. Dans cet exemple, le cœur du réseau objectif des protagonistes n'est autre que le responsable de la sécurité. Mais dans un premier temps, le réseau (subjectif) est tout autre aux yeux de la direction de l'entreprise puis des médias. Les accusés d'espionnage sont les véritables cibles d'un réseau à remonter jusqu'aux intermédiaires chinois supposés. C'est en tâchant de remonter ce réseau que la direction et les enquêteurs réorganisent la compréhension du réseau, y font apparaître un éventuel intermédiaire mystérieux précédant le chef de la sécurité, puis font sauter ce maillon : tout se retrouve centré sur le nouvel accusé. La remontée de réseau doit permettre de changer les nœuds et leurs relations afin de passer d'une première vision subjective à une vision plus réaliste et objective des faits réels.

Pour approfondir le champ de connaissance ouvert par l'étude de la cartographie des réseaux sociaux, élément essentiel dans l'anticipation de menaces économiques ou stratégiques contemporaines, un observateur doit disposer d'outils d'analyse de graphes au service de la remontée de ces réseaux. La recherche d'éléments du réseau interconnectés avec un élément donné peut s'avérer à cet égard fort utile.

4.5. Des concepts à la technique : le projet CAHORS

Le *Livre blanc* gouvernemental sur la sécurité contre le terrorisme accorde à Internet une attention soutenue : « le Web est le modèle qui figure le mieux l'activité du terrorisme mondial. Il est non seulement en totale adéquation avec la structure de la mouvance terroriste. Mais il est surtout devenu pour elle un vecteur à tout faire. » [LSI 06]. Partant de ce constat, il s'agit pour les services gouvernementaux concernés par la veille anti-terroriste, comme par la criminalité organisée, et plus généralement par la sécurité globale, de se doter de moyens techniques adéquats leur permettant de s'adapter aux mutations technologiques dont Internet est le reflet : fort accroissement des volumes de données disponibles, intensification des flux de données et multiplication des sources d'informations souvent mal répertoriées. Le milieu militaire appelle « renseignement d'origine sources ouvertes », ou ROSO (en anglais OSINT pour *Open Source Intelligence*), l'analyse spécialisée orientée vers ces nouveautés qui occupent de plus en plus les services dévolus au renseignement. Un document doctrinal pour les armées décrit ce qu'est officiellement une telle information de source ouverte [ISO 10], information qu'un spécialiste est chargé de

valoriser en renseignement. L'origine de source ouverte atteint désormais largement l'envergure des autres origines déjà répertoriées et étudiées par les spécialistes – origine électromagnétique, origine humaine, origine image, etc. – et a pris une ampleur stratégique complexe que n'avaient pas toujours ces antécédents. Ce sont ces besoins qui légitiment le projet technique ici dépeint.

4.5.1. *Présentation du projet*

Le projet CAHORS (acronyme de cotation, analyse, hiérarchisation et ontologies pour le renseignement et la sécurité) est financé par l'agence nationale de la recherche, dans le cadre de l'appel à projet concept, systèmes et outils pour la sécurité globale, avec un soutien du pôle de compétitivité cap digital. Commencé en février 2009 et d'une durée de trois ans, il réunit six partenaires dont Thales est le responsable, avec les deux auteurs du présent article pour coordinateurs. Il est suivi par la police nationale pour le ministère de l'Intérieur, et par la délégation générale pour l'armement pour le ministère de la Défense. Comme les termes reflétés par son acronyme l'indiquent, la cotation de l'information y joue un rôle prépondérant, sans négliger les applications qui peuvent en découler. En outre, une attention particulière est portée aux aspects conceptuels fondamentaux, voire philosophiques, associés à la cotation, pour aller vers une modélisation puis des développements informatiques propres à en rendre compte afin de proposer des moyens semi-automatiques pour coter les informations extraites de textes collectés en source ouverte.

Le projet CAHORS s'attache à relever un nombre important de défis techniques, à la fois dans une optique de protection du citoyen face aux menaces terroristes ou criminelles, et pour assurer que la première étape de la gestion de crise potentielle, celle de l'anticipation et de la détection des signes précurseurs, soit idéalement franchie pour une meilleure gestion des risques[3]. Compte tenu des objectifs du projet et des services concernés, l'une des originalités du projet consiste à réunir des partenaires issus de domaines d'études variés : logiciens, philosophes, informaticiens de diverses branches sont réunis pour le mener à bien, ce qui impose de faire un pont entre sciences humaines et sociales et sciences dites dures.

CAHORS se découpe en cinq tâches reliées entre elles :

- extraction d'information ;
- analyse théorique de la cotation ;
- modélisation de la cotation ;

3. Dans la lignée de CAHORS, le rôle de la cotation de l'information dans la lutte anti-terroriste a fait l'objet d'une communication à un symposium de l'OTAN [DEL 09].

- applications pour la sécurité globale ;
- intégration des différents modules.

Ces différentes tâches doivent mener à la construction d'un système détaillé dans le schéma de la figure 4.6.

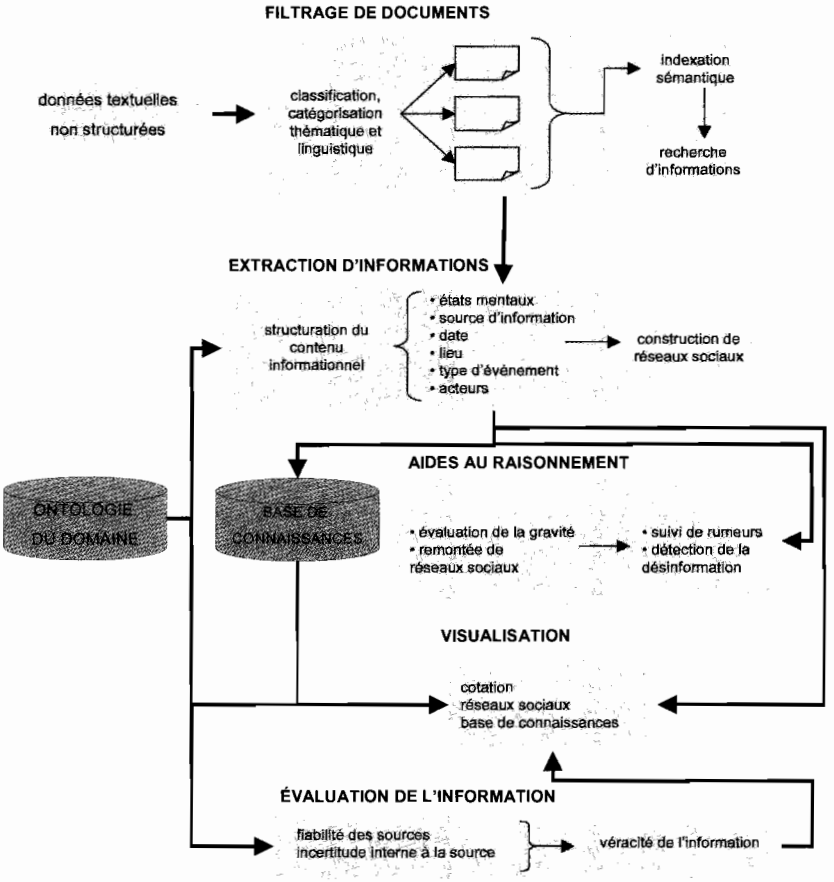

Figure 4.6. *Vue d'ensemble du système CAHORS*

Les principaux points de ce schéma sont expliqués dans les trois sous-sections qui suivent. Pour une approche moins formelle des orientations choisies au cours du temps, une description progressive des travaux effectués dans CAHORS depuis son lancement est retracée dans les comptes-rendus annuels du *workshop* de référence pour la sécurité globale : [CAP 10, CAP 11, CAP 12].

4.5.2. *Fondements épistémologiques et logiques*

Au-delà des critiques que l'on peut faire aux approches doctrinales de la cotation, une première tâche de CAHORS a consisté à mieux comprendre ce qui était à l'œuvre dans le processus souhaitable de la cotation, en complexifiant largement ce qui est jusqu'à présent en usage.

Ainsi, la fiabilité d'une source et la véracité d'une information ne sauraient être estimées dans l'absolu. Il serait par exemple naïf d'accorder la même confiance aux dires d'un ministre des Affaires étrangères selon qu'il s'exprime sur des sujets diplomatiques, économiques, écologiques. Il serait de même risqué d'accorder le même crédit à des informations impliquant certains acteurs selon qu'elles sont délivrées par des sources ayant des relations d'accointance ou d'hostilité vis-à-vis de ces acteurs. Afin d'éviter un relativisme excessif qui conduirait à rendre illusoire tout projet de cotation de l'information, CAHORS s'attache à circonscrire de la manière la plus fine possible la notion de contexte informationnel. Celui-ci inclut en particulier le domaine d'intérêt ainsi que les acteurs auxquels se rapporte l'information traitée. Il faut également tenir compte des intérêts et objectifs de l'utilisateur final chargé de l'évaluation de la cotation, puis de son utilisation. Les exigences de fiabilité doivent ainsi être nettement plus élevées lorsque le domaine auquel se réfère l'information traitée relève des priorités de l'utilisateur.

Concept-clé pour comprendre toute action sociale, la *confiance* reste cependant une notion fort difficile à définir en philosophie et en sciences sociales. Le terme peut avoir au moins trois usages différents : on parle de la confiance comme un état de croyance, de la part d'un individu ou d'une société, à l'égard des actions possibles, mais non certaines d'autrui. Ou comme un acte d'engagement visant à influencer les actions futures des autres à notre égard. Ou encore comme un sentiment, une passion de l'âme, ainsi que la définit Thomas Hobbes : « une passion produite par la croyance ou la foi que nous avons en celui de qui nous attendons ou nous espérons du bien ». Ce que la confiance désigne alors est soit la qualité interpersonnelle à différents niveaux (individuel ou institutionnel), soit l'action même de s'engager dans une situation jugée risquée, soit enfin la motivation, le sentiment qui mène à cette action. Afin de pouvoir user de cette notion dans un système informatisé, le projet s'attache à comprendre de quelle manière elle peut être définie.

L'épistémologie contemporaine a placé la confiance au centre de l'explication des processus de production et de distribution de savoir. La confiance est nécessaire pour assurer la transmission non seulement entre experts et gens ordinaires, mais aussi à l'intérieur des sphères d'expertise. Une notion de confiance épistémique s'est développée dans l'épistémologie et les sciences cognitives, qui vise à comprendre le rôle purement épistémique de la confiance et son rapport avec la notion sociale et morale de confiance. Quels sont les indices de fiabilité et véracité de nos interlo-

cuteurs ? Sont-ils de nature proprement épistémique, ou l'évaluation morale joue-t-elle un rôle dans la crédibilité ? Par exemple, avons-nous un biais à croire les informations qui nous viennent des sources que nous percevons comme bienveillantes ? Les rapports entre confiance épistémique et confiance sociale et affective sont à comprendre en amont afin de pouvoir identifier les critères de fiabilité et véracité que les agents utilisent pour trier l'information qu'ils reçoivent *via* autrui. L'une des partenaires du projet, la philosophe Gloria Origgi, a écrit un ouvrage spécialement dévolu à ce sujet [ORI 08]. On se reportera également à [GAM 88] pour l'étude du rapport entre la confiance et les relations coopératives, et pour une approche plus contemporaine et beaucoup plus vaste à [WIL 06].

4.5.3. *Méthodologies adoptées, algorithmes et techniques logicielles*

A partir de ces considérations épistémologiques de la cotation, la tâche suivante du projet vise à développer des méthodes puis des outils logiciels permettant d'automatiser pour partie la détermination de la cotation. Diverses techniques ont été dans un premier temps examinées, faisant appel aux logiques modales [BOV 10], à la théorie de l'évidence [CHO 10], à d'autres formulations logiques ([DEM 10] ou [DUB 09]), et à bien d'autres branches de disciplines formelles, permettant de choisir parmi elles des briques pour construire une méthodologie formelle générale débouchant sur des algorithmes appropriés.

On l'a vu, l'objectif de la cotation est de qualifier les informations extraites par des mesures qui reflètent, d'une part la fiabilité de leurs sources et d'autre part, leur véracité. Pour y parvenir, divers indicateurs permettant de rendre compte des dimensions identifiées comme caractéristiques de ces deux notions sont à proposer. Cette gamme d'indicateurs s'appuie sur des métriques de réputation et confiance qui existent dans la littérature spécialisée[4]. Celles-ci reposent sur l'historique des contributions des différentes sources, indispensable lors de la confrontation entre les informations d'une même source au cours du temps, ainsi que la confrontation d'informations proches issues de sources distinctes. Des mesures de similarité adaptées doivent être mises au service de cette analyse de l'historique, en s'inspirant des références du domaine telles que [WU 94, MAZ 07]. Les informations concernant les *états mentaux* (croyances, désirs, intentions...) d'une source doivent également être prises en compte pour évaluer le degré de certitude d'une source sur les faits qu'elle rapporte. L'évaluation du degré d'incertitude d'une source permet d'ajuster l'influence de la fiabilité de la source dans l'évaluation de la véracité de l'information, notamment pour les applications stratégiques telles que la détection de la désinformation ou le suivi de rumeurs. Dans la même optique, la neutralité de

4. Sur cette question de métriques de confiance, réputation et autorité, voir [PAG 99] pour l'algorithme *PageRank*, [GYÖ 04] pour *TrustRank*, et [KLE 99] pour *HITS*.

la source vis-à-vis du contexte informationnel doit être modélisée. Ainsi, en fonction du positionnement politique d'un journal et de la thématique abordée, un journaliste sera plus ou moins partial dans son traitement de l'information.

Des contraintes temporelles sont également à modéliser, ce qui n'est pourtant qu'incidemment évoqué dans les doctrines de cotation. Cette tâche de modélisation doit rendre compte d'une partie du processus d'évaluation de l'information, afin de tenir compte de l'évolution de la fiabilité des sources. Comme souligné plus haut, la fiabilité est en effet une variable dynamique qui évolue au cours du temps. D'une part, les stratégies des sources se modifient et s'adaptent à l'évolution de leur contexte. Leur rigueur et leur neutralité dans le traitement de l'information et en conséquence leur fiabilité s'en trouvent alors affectées. La qualité d'un journal quotidien ne saurait être tenue pour éternelle, même si une telle évolution est relativement lente. Les déclarations de certains protagonistes de l'affaire Renault ont également incité à réviser au cours du temps la fiabilité qu'on leur accordait. D'autre part, avec la multiplication des sources, celles qui viennent d'apparaître étant encore peu connues, il sera difficile d'estimer finement leur fiabilité. Contrairement aux sources établies, il convient donc de procéder à des ré-estimations fréquentes de leur fiabilité. Le statut des sources (le degré de connaissance que l'on a sur celles-ci) est en cela un facteur à prendre en compte. De manière plus générale, avec l'arrivée d'éléments nouveaux, certains faits ou informations fournies par d'autres sources peuvent corroborer ou au contraire réfuter des informations émises par une source dans le passé. Il faut prendre en compte ces nouvelles données pour réajuster la fiabilité de la source concernée. La durée de l'intervalle temporel, séparant la production d'une information de l'arrivée des nouveaux éléments, doit influer sur les réajustements de la fiabilité qui sont effectués. Des graphes de contraintes temporelles floues peuvent ainsi être utilisés pour rendre compte de ce type de phénomène. L'introduction du *flou* est ici privilégiée, en recourant à des variables linguistiques aisément interprétables par l'utilisateur et qui traduisent les imprécisions entachant immanquablement l'expression de la durée dans le langage naturel.

Une description plus approfondie de la démarche adoptée et des structures d'algorithmes retenus figure dans [BAE 10] et [DEL 10], avec en particulier l'introduction de la notion d'élément d'information, à comprendre comme un « atome » d'information, et le recours à des variables à prendre en compte dans la chaîne de cotation.

La figure 4.7 représente une vue partielle de la chaîne de traitement de la cotation qui suit la structuration des informations et qui précède des modules applicatifs dévolus à la mesure de la gravité informationnelle, à la remontée de réseaux sociaux et à la détection de rumeurs et de désinformation.

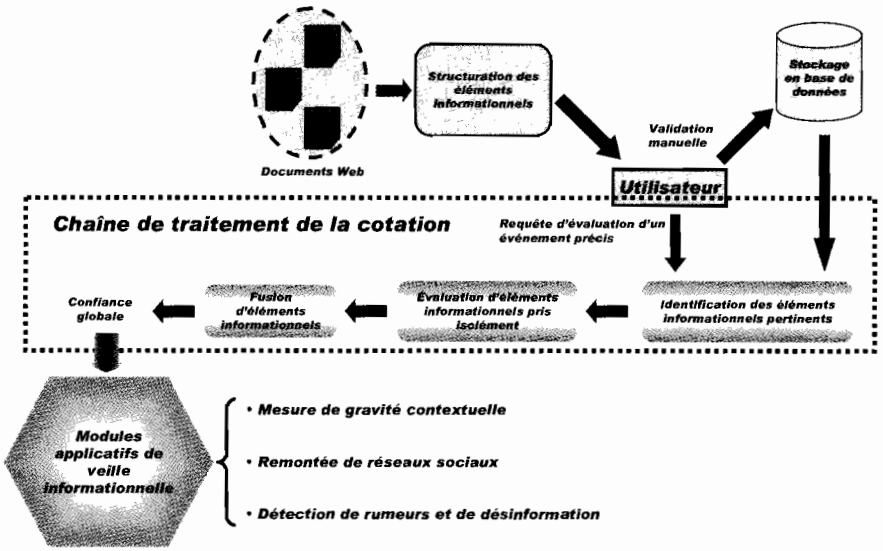

Figure 4.7. *Chaîne de cotation imbriquée dans le processus CAHORS*

Afin de rendre l'outil de cotation aisément utilisable par l'analyste, l'ergonomie doit également être adaptée aux besoins spécifiques du métier. En outre et de manière fondamentale, une évaluation expérimentale de l'outil doit être menée, ce qui a pour l'heure été fait pour le module central de cotation, selon un protocole forgé pour ce faire. La tâche d'évaluation est en règle générale délicate ; elle l'est d'autant plus ici que les méthodes d'évaluation ne préexistent pas pour cette tâche particulière de cotation. Tout un protocole d'évaluation a donc été déterminé pour y répondre. Dans la suite du projet, les modules applicatifs feront l'objet eux aussi d'une telle évaluation, sous la même contrainte de définition d'un protocole adapté.

4.5.4. *Applications*

Le projet CAHORS vise non seulement à développer un outil de cotation, mais encore à aborder les questions de modules applicatifs répondant aux problèmes mentionnés en section 4.4. L'assemblage approprié de ces modules doit permettre de diriger les analystes vers un cycle du renseignement *enrichi.*

La partie du cycle du renseignement comprenant la phase d'exploitation tend en effet à être ainsi élargie ; la tâche de cotation (voir figure 4.4) nourrit en particulier plus directement les tâches d'analyse, de fusionnement et d'interprétation. En outre, ces applications mentionnées en section 4.4 peuvent participer à cet enrichissement de manière jusqu'ici quelque peu négligée lorsque l'ensemble d'informations est

trop volumineux à traiter efficacement, comme dans le cas paradigmatique où les informations sont véhiculées par Internet. Cette portion du cycle pourrait être illustrée comme en figure 4.8.

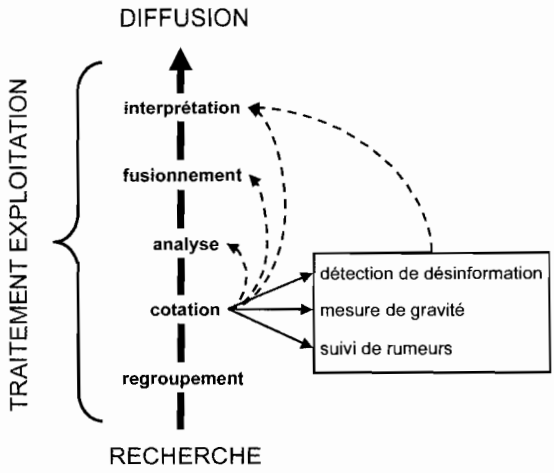

Figure 4.8. *Une phase d'exploitation enrichie, une valorisation amplifiée de l'information en renseignement*

Les flèches en traits pleins indiquent les activités pouvant être dérivées de la tâche de cotation, tandis que les flèches en pointillés désignent les apports multiples de la cotation aux tâches suivantes de la phase de traitement et d'exploitation, directs d'une tâche à l'autre, ou indirects, *via* les activités permises après cotation.

On le voit, le projet CAHORS dans son ensemble s'appuie sur des problèmes posés par les méthodes et pratiques observées par une fonction-pivot inscrite dans le renseignement, et vise par les divers travaux qu'il comprend à reformuler ces questions pour y remédier ; cette démarche passe donc par un certain remaniement des doctrines existantes, et vise également à nourrir, au-delà des questions de Défense *stricto sensu*, d'autres branches stratégiques telles que celle de l'intelligence économique.

4.6. Conclusion

Dans ce chapitre, nous avons insisté sur l'approche pratiquée par la Défense dans le cadre de la cotation de l'information pour aboutir à du renseignement. Si les défauts relevés dans cette approche ont motivé à l'origine le développement du projet CAHORS, il est cependant clair que les résultats de ce projet doivent pouvoir toucher

bien d'autres secteurs que ceux qui s'inscrivent dans la politique de sécurité globale. L'exemple de l'affaire Renault est ainsi emblématique : un processus de cotation de l'information, suivi de l'utilisation d'applications complexes, telles que la détection de la désinformation ou la reconstitution de réseaux sociaux, auraient dans l'idéal permis de désamorcer cette crise artificielle et d'en juguler les retombées néfastes. L'intelligence économique constitue dès lors l'un des nombreux champs où la méthode ici dépeinte trouverait des applications parfaitement idoines. Le renforcement de l'analyse dans ces domaines stratégiques passe indubitablement par cette facette de l'évaluation de la qualité des données qu'est la cotation de l'information.

4.7. Bibliographie

[BAE 10] BAERECKE T., DELAVALLADE T., LESOT M.-J., PICHON F., AKDAG H., BOUCHON-MEUNIER B., CAPET P., CHOLVY L., « Un modèle de cotation pour la veille informationnelle en source ouverte », *6ème colloque Veille Stratégique Scientifique & Technologique, VSST2010*, Toulouse, France, 2010.

[BOV 10] BOVO A., CHOLVY L., « Une approche logique à la plausibilité d'une information rapportée », *Journées d'Intelligence Artificielle Fondamentale (IAF'2010)*, Strasbourg, France, juin 2010.

[CAP 10] CAPET P., DELAVALLADE T., « CAHORS, quelques premiers éléments d'analyse », *Séminaire WISG*, Troyes, France, 2010.

[CAP 11] CAPET P., DELAVALLADE T., « CAHORS, vers une cotation semi-automatisée de l'information », *Séminaire WISG*, Troyes, France, 2011.

[CAP 12] CAPET P., DELAVALLADE T., « CAHORS, applications de la cotation de l'information », *Séminaire WISG*, Troyes, France, 2012.

[CHO 10] CHOLVY L., « Evaluation of Information reported : a model in the Theory of Evidence », *13th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'2010)*, Dartmund, Allemagne, juillet 2010.

[DEL 09] DELAVALLADE T., CAPET P., « Information evaluation as a decision support for counter-terrorism », *Symposium OTAN IST086 C3I for crisis, emergency and response management*, Bucarest, Roumanie, mai 2009.

[DEL 10] DELAVALLADE T., AKDAG H., BAERECKE T., BOUCHON-MEUNIER B., CAPET P., CHOLVY L., LESOT M.-J., PICHON F., « Des données textuelles au renseignement : vers un modèle global de cotation », *Atelier COTA, 21e Journées francophones d'Ingénierie des Connaissances*, Nîmes, France, juin 2010.

[DEM 04] DEMOLOMBE R., « Reasoning about trust : a formal logical framework », *International Conference on iTrus.*, Oxford, Royaume-Uni, 2004.

[DUB 09] DUBOIS D., DENOEUX T., « Pertinence et Sincérité en Fusion d'Informations », *Rencontres Francophones sur la Logique Floue et ses Applications*, Cépaduès-Editions, Annecy, France, 2009.

[GAM 88] GAMBETTA D., *Trust. The Making and Breaking of Cooperative Relations*, Basil Blackwell, Oxford, 1988.

[GYÖ 04] GYÖNGYI Z., GARCIA-MOLINA H., PEDERSEN J., « Combating web spam with TrustRank », *Proceedings of the 30th International Conference on Very Large Data Bases (VLDB)*, Toronto, Canada, septembre 2004.

[INS 03] INSTRUCTION INTERARMÉES SUR LE RENSEIGNEMENT D'INTÉRÊT MILITAIRE, TITRE I, *Doctrine interarmées du renseignement*, PIA 02-200, Paris, 2003.

[ISO 10] RECHERCHE ET EMPLOI DE L'INFORMATION DE SOURCE OUVERTE, *Concept Exploratoire Interarmées*, CEIA – 2.4, Paris, juillet 2010.

[KLE 99] KLEINBERG J., « Authoritative sources in a hyperlinked environment », *Journal of the ACM (JACM)*, vol. 46, n° 5, p. 604-632, 1999.

[LSI 06] *Livre blanc du Gouvernement sur la sécurité intérieure face au terrorisme*, Paris, 2006.

[MAZ 07] MAZUEL L., SABOURET N., « Degré de relation sémantique dans une ontologie pour la commande en langue naturelle », *Plate-Forme AFIA, Ingénierie des Connaissances 2007 (IC 2007)*, 2007.

[NEL 10] NEL F., CAPET P., DELAVALLADE T., « Mesure et anticipation des mouvements informationnels en source ouverte », *Workshop Interdisciplinaire sur la Sécurité Globale*, Troyes, France, 2010.

[NEL 11] NEL F., LESOT M.-J., CAPET P., DELAVALLADE T., « Modélisation de la propagation de l'information sur le Web : de l'extraction des données à la simulation », *Proceedings de la 11ème Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances (EGC)*, 2011.

[ORI 08] ORIGGI G., *Qu'est-ce que la confiance ?*, Vrin, Paris, 2008.

[PAG 99] PAGE L., BRIN S., MOTWANI R., WINOGRAD T., The PageRank Citation Ranking : Bringing Order to the Web, Technical Report, Stanford InfoLab, 1999.

[REV 11] REVAULT D'ALLONNES A., Evaluation sémantique d'informations symboliques : la cotation, Thèse de doctorat, Université de Paris VI, 2011.

[STA 03] STANDARD AGREEMENT, Intelligence Report, STANAG n° 2511, North Atlantic Treaty Organization, 2003.

[TTA 01] TRAITÉ TOUTES ARMES, n° 150, titre VI, *Renseignement*, Paris, 2001.

[WIL 06] WILLIAMS B., *Vérité et véracité*, Gallimard, Paris, 2006.

[WU 94] WU Z., PALMER M., « Verb Semantics and Lexical Selection », *Proceedings of the Annual Meetings of the Associations for Computational Linguistics*, 1994.

Chapitre 5

Application de mesures de distance pour la détection de problèmes de qualité de données

5.1. Introduction

Avec la multiplication des sources d'informations disponibles et l'accroissement des volumes et flux de données potentiellement accessibles, la qualité des données et, au sens large, la qualité des informations n'ont cessé de prendre une place de premier plan tant au niveau académique qu'au sein des entreprises.

Si l'analyse des données, l'extraction de connaissances à partir des données et la prise de décision peuvent être réalisées sur des données inexactes, incomplètes, ambiguës et de qualité médiocre, on peut alors s'interroger sur le sens à donner aux résultats de ces analyses et remettre en cause, à juste titre, la qualité des connaissances ainsi « élaborées », tout comme le bien-fondé des décisions prises.

Aujourd'hui, il n'est donc plus question de négliger les données mais, bien au contraire, d'évaluer et de contrôler leur qualité dans les systèmes d'information, les bases et les entrepôts de données.

Ainsi, ont été proposées de nombreuses mesures objectives, des méthodes et tout un outillage technique pour mener une expertise critique de la qualité des données dans ces systèmes, permettant aux utilisateurs de relativiser la confiance qu'ils pourraient accorder aux données et de leur permettre de mieux en adapter leur usage.

Chapitre rédigé par Melanie HERSCHEL et Laure BERTI-EQUILLE.

L'impact et les coûts de la non-qualité des données (tout comme sa méconnaissance) retentissent à chaque étape d'un processus de traitement des données et de nombreuses techniques peuvent être combinées pour consolider et améliorer la qualité de ces données.

L'objet de ce chapitre est de faire un tour d'horizon des méthodes et des techniques employées pour détecter deux des principaux problèmes de qualité des données que sont les doublons et les données aberrantes, en se concentrant sur les méthodes basées sur des mesures de distance. Nous passerons d'abord en revue les principales sources de problèmes de qualité des données ainsi que les solutions mises en œuvre communément dans la pratique. Ensuite, nous nous consacrerons à la définition des problèmes de détection de doublons et de détection de valeurs aberrantes et nous présenterons les mesures de distances pouvant leur être appliquées. Les approches de détection de doublons et de détection de valeurs aberrantes utilisant ces mesures sont présentées par la suite et elles seront illustrées par des exemples d'application réels.

5.2. Les problèmes de qualité des données et leurs solutions en pratique

5.2.1. *Les principales sources de problèmes*

Les causes des problèmes de qualité des données sont d'origines très diverses selon les différents stades de traitement des données considérés. Les tableaux 5.1 et 5.2 récapitulent à titre indicatif quelques-uns des principaux problèmes de qualité des données pouvant survenir à chacune des étapes de leur traitement. Ils présentent également les solutions mises en œuvre pour y remédier.

Ces solutions peuvent, pour la plupart, être classées selon les grands types d'approches complémentaires suivantes : les approches préventives, les approches diagnostiques, et les approches correctives. Par la suite, nous allons aborder des solutions de diagnostique souvent utilisées dans la pratique et qui font appel à des mesures de distance.

5.2.2. *Des solutions en pratique utilisant des mesures de distance*

Parmi les techniques de détection et de correction des problèmes de qualité des données, celles les plus communément employées dans la pratique sont : 1) la vérification d'après une vérité-terrain ou bien d'après une source de données de référence jugée plus fiable, 2) l'audit des données et, enfin 3) le nettoyage des données.

Par la suite, nous verrons que, dans la majorité des cas de diagnostique, il est nécessaire d'employer des mesures de distance pour détecter les anomalies dans les jeux de données. En particulier, il s'agit d'identifier les données qui, soit sont redondantes et dupliquées, soit s'écartent d'un modèle attendus ou des données prises en référence.

Sources de problèmes de qualité des données	**Solutions potentielles**
Etape de la création des données	
Entrée manuelle : absence de vérifications systématiques des formulaires de saisie Entrée automatique : problèmes de capture OCR, de reconnaissance de la parole Incomplétude, absence de normalisation ou inadéquation de la modélisation conceptuelle des données Entrée de doublons, erreurs de mesure Approximations Contraintes matérielles ou logicielles Corruption des données : faille de sécurité physique et logique des données	**Approche préventive** : Architecture pour la gestion de processus et *workflows* : (audits, intendance des données – « *data stewards* ») **Approche corrective** : Nettoyage de données : élimination des doublons, « *merge/purge* », appariement des noms et adresses, jointure approximative **Approche diagnostique** : détection automatique des erreurs et exceptions
Etape de la collecte ou de l'import des données	
Destruction ou mutilation d'information par des prétraitements inappropriés Perte de données : *buffer overflows*, problèmes de transmission Absence de vérification dans les procédures d'import massif Introduction d'erreurs par les programmes de conversion de données Contraintes matérielles ou logicielles	Utiliser des techniques de fouille de données pour vérifier que toutes les données ont été correctement transmises Vérifier la transmission, l'intégrité des données, leur format Suivi de données Edition de données (*data publishing*) Agrégation de données et construction de résumé de données (*data squashing)*
Etape du stockage des données	
Absence de métadonnées Absence de mise à jour et de rafraîchissement des données obsolètes ou répliquées Modèles et structures de données inappropriés, spécifications incomplètes ou évolution des besoins dans l'analyse et conception du système Modifications *ad hoc*	Métadonnées Planification et personnalisation par domaine Profilage des données, moniteur de navigation dans les données

Etape de l'intégration des données	
Problèmes d'intégration de multiples sources de données ayant des niveaux de qualité et d'agrégation divers Problèmes de synchronisation temporelle Systèmes de données non conventionnels Facteurs sociologiques conduisant à des problèmes d'interprétations et d'intégration des données Jointures *ad hoc* Appariements aléatoires Heuristiques d'appariements des données inappropriées	Exiger des estampilles temporelles précises pour assurer la cohérence logique et temporelle des jeux de données, lignage des données Outils commerciaux pour : la migration des données, le nettoyage de données, le profilage de données Outils académiques pour faire les appariements et les jointures entre jeux de données
Etape de la recherche et de l'analyse des données	
Erreur humaine Contraintes liées à la complexité de calcul Contraintes logicielles, incompatibilité Problèmes de passage à l'échelle, de performances et de confiance dans les résultats Approximations dues aux techniques de réduction des grandes dimensions Utilisation de boîtes noires pour l'analyse Attachement à une famille de modèles statistiques Expertise insuffisante d'un domaine Manque de familiarité avec les données	Planification : évaluer le problème par rapport à l'équipement et *vice versa* Fouille de données exploratoire (*Exploratory Data Mining – EDM*) Plus grande responsabilité des analystes Analyse en continu plutôt que ponctuel Echantillonnage plutôt qu'analyse complète Boucle de rétroaction

Tableau 5.1. *Les principales sources des problèmes de qualité de données et leurs solutions*

5.2.2.1. *La vérification d'après une vérité-terrain ou par comparaison avec des données de référence*

La première technique consiste à comparer les valeurs de données avec leur contrepartie dans le monde réel (vérification d'après la vérité-terrain). Cette méthode est très coûteuse en temps et en moyens et, selon les domaines d'application, elle s'avère difficilement réalisable du fait que la contrepartie réelle est parfois inaccessible ou trop complexe.

Classiquement employé dans le domaine des systèmes d'informations géographiques (SIG), la comparaison par rapport à la vérité-terrain (appelé terrain nominal) permet de réaliser des matrices de confusion entre les données de la base à inspecter et les jeux de données de contrôle [DEV 06]. Des mesures de distance sont donc calculées entre les données et leur contrepartie dans le terrain nominal. De nombreux standards (ISO 19113, ISO 19138) dans ce domaine préconisent des éléments de mesures objectives pour évaluer, en particulier, la cohérence logique, la précision thématique, temporelle, positionnelle et sémantique de la base de données géospatiales [DEV 06], en mesurant les distances entre les objets de la vérité-terrain et les données représentant ces objets.

Une seconde approche appelée consolidation, met en œuvre la comparaison de deux bases de données (ou plus). Les données pertinentes de la base à inspecter sont comparées à leur contrepartie dans l'autre base : les données identiques sont considérées correctes, celles qui ne le sont pas sont signalées pour investigation et correction éventuelle. Dans ce dernier cas, la difficulté réside dans la détermination de la valeur correcte (l'une et l'autre donnée pouvant être fausses). La méthode principalement utilisée pour la correction est le remplacement par imputation : la valeur incorrecte sera remplacée par une valeur jugée « plus correcte ». D'autres méthodes dites de fusion [BLE 08] peuvent être appliquées pour résoudre le problème des conflits de valeurs.

L'inconvénient majeur de l'approche de consolidation multi-source dans la détection des erreurs et leur correction demeure : il n'y a aucune garantie que les données identiques des différentes bases soient correctes. Les données utilisées par comparaison pour détecter les données erronées dans la base à inspecter peuvent être fausses, rendant la recherche d'erreurs difficile ; aussi, cette méthode n'empêche en rien l'introduction de nouvelles erreurs dans les données.

Dans les deux méthodes mentionnées, des mesures de distance sont typiquement utilisées lors de la comparaison des données à la vérité-terrain ou à d'autres données.

5.2.2.2. *L'audit des données*

L'audit des données met en œuvre des programmes chargés de vérifier si les valeurs des données satisfont différents types de contraintes. L'avantage de l'audit des données est sa simplicité de mise en œuvre par rapport aux méthodes précédentes de comparaison. Elle peut se concevoir en même temps que le modèle conceptuel des données et peut utiliser différents outils diagnostiques d'analyse des données. Cependant, elle ne permet pas d'améliorations de la qualité des données. L'audit et l'édition de données visent l'intégrité et la cohérence de celles-ci, c'est-à-dire la conformité à des règles et des contraintes préalablement définies, mais elles ne garantissent en rien l'exactitude des données.

Généralement, l'audit des données s'articule autour des étapes suivantes :

– la définition du périmètre de l'audit dans la base de données, selon les dimensions de qualité à considérer en priorité et pour des utilisateurs-clés identifiés ;

– l'identification des segments de données à analyser (par exemple, les données client, les grands comptes, les PME, etc.) ;

– le choix d'un ensemble représentatif de données (par exemple, par zone géographique) ;

– l'analyse du dictionnaire de données (par exemple, la sélection des attributs, des types de données, des domaines de valeurs, de leur taux de remplissage, etc.) ;

– l'énumération des contraintes : par exemple, l'unicité des clés pour les enregistrements d'une table, le respect des contraintes d'intégrité, le respect de règles syntaxiques pour les valeurs de certains attributs (tels que le numéro de sécurité sociale), le respect du zonage géographique (défini, par exemple, par une règle de cohérence entre la ville et le code postal ou la base d'un dictionnaire de données), etc. ;

– le profilage des données : par exemple, le comptage du taux d'informations non renseignées, du taux d'anomalies de zonage, du taux de données ne respectant pas chacune des contraintes définies, la détection des doublons, la vérification de la cohérence entre la civilité et le prénom, la normalisation des adresses, le taux de NPAI (c'est-à-dire *n'habite pas à l'adresse indiquée*), la vérification syntaxique des numéros de téléphone, etc. ;

– des calculs croisés : par exemple, le calcul du taux d'individus avec le même email, le même nom, la même adresse, le même téléphone, etc. ;

– l'usage de référentiels : par exemple, un dictionnaire des prénoms, la base SIRET ou du référentiel RNVP (restructuration, normalisation et validation postale).

La détection des valeurs aberrantes fait partie notamment de l'étape de profilage, utilisant des mesures de distance pour identifier des valeurs atypiques du domaine.

5.2.2.3. *Le nettoyage des données, leur normalisation et standardisation*

Le nettoyage de données fait partie des stratégies d'amélioration semi-automatique de la qualité des données et il se décompose en trois étapes principales : (1) auditer les données afin de détecter les incohérences et les anomalies, (2) choisir les transformations pour résoudre les problèmes de qualité de données, et enfin, (3) appliquer les transformations choisies au jeu de données selon des règles de priorité.

Le processus de nettoyage des données repose sur un ensemble de transformations qui visent à normaliser les formats de données et à détecter les paires d'enregistrements qui se rapportent le plus probablement au même objet (ou entité du monde réel). Cette étape d'élimination des doublons est appliquée si des données approximativement similaires et donc redondantes sont trouvées. En particulier, un appariement multitable calcule des jointures approximatives entre des données distinctes mais similaires ce qui permet leur consolidation.

Nous observons que des mesures de distance sont appliquées lors du profilage et lors de la détection de doublons. Ces opérations sont mises en œuvre par une multitude d'outils de nettoyage de données, dont le tableau 5.2 ne donne pas une liste exhaustive mais limitée aux outils en open source et aux prototypes de recherche, notamment ceux à l'origine du nettoyage des données dans les années 2000.

Nom de l'outil	Fonctionnalités
Potter's wheel [RAM 01]	S, V, N, D
Ajax [GAL 01]	S, V, N, D
Intelliclean [LOW 01]	D
Bellman [DAS 02]	D, E, P
Febrl [CHR 08] http://sourceforge.net/projects/febrl	S,V, N, D,P, A
D-Dupe [KAN 08] http://www.cs.umd.edu/projects/linqs/ddupe	V, D
XClean [WEI 07]	D, R
Talend Open Profiler and Studio http://www.talend.com	S,V, N, P, A, E
DataCleaner http://datacleaner.eobjects.org/	P, A, E
Pentaho Data Integration http://kettle.pentaho.com/	S, V, N, P, D

Tableau 5.2. *Prototypes de recherche et* open sources

Le tableau 5.2 récapitule les principales forces et fonctionnalités de ces outils qui, pour les *open sources*, sont modulaires et peuvent être étendus : la standardisation (S), la vérification de contraintes (V), le profilage (P), le nettoyage par des opérateurs de transformation (N), l'élimination des doublons (D), la détection et la résolution de conflits (R), l'enrichissement par des métadonnées (E), et l'analyse exploratoire des données (A).

5.3. Approches de détection basées sur des distances

Les méthodes présentées dans cette partie ont en commun l'utilisation d'une mesure de distance ou d'une mesure de similarité[1], quantifiant ainsi jusqu'à quel degré certaines données sont en accord ou en désaccord avec d'autres données. Nous nous intéressons particulièrement aux approches liées aux problèmes de la détection de doublons (*duplicate detection*) et à la détection de valeurs aberrantes (*outlier detection*), relevant tous deux du problème général du diagnostique.

Nous débutons notre discussion sur les méthodes basées sur des distances avec une définition des différents types de problèmes visés. Ensuite, nous décrivons les méthodes ayant été développées pour résoudre ces problèmes. Une présentation de l'outillage technique fait suite à cet aperçu et nous concluons cette partie en donnant des exemples pratiques et en discutant les limitations des approches présentées.

5.3.1. *Types de problèmes visés*

Dans cette section, nous définissons plus formellement les problèmes à résoudre.

5.3.1.1. *Les doublons*

Soient A et B deux ensembles de représentations d'objets (par exemple, tuples relationnels ou éléments XML), $dist(a,b)$ une mesure de distance et θ une valeur de seuil dans le domaine des valeurs possibles pour la distance $dist(\cdot)$. Le but de la détection de doublons est l'identification de chaque paire de représentations d'objets (a,b) dont la distance est inférieure au seuil θ prédéfini c'est-à-dire tel que $a \in A$, $b \in B$, $dist(a,b) \leq \theta$. Chaque paire (a,b) satisfaisant ce critère est un doublon, toute autre paire est qualifiée de non-doublon. Ainsi, le problème de détection de doublons se rapporte à un problème de classification tel que :

$$classification(a,b) = \begin{cases} si\ dist(a,b) \leq \theta \text{ alors } (a,b)\ sont\ des\ doublons \\ sinon (a,b)\ sont\ des\ non\text{-}doublons \end{cases} \quad [5.1]$$

La valeur du seuil θ doit être spécifiée au préalable lors de la configuration par un utilisateur. En pratique, il est très commun d'avoir plus de deux classes, en

1. Dans ce chapitre, on appelle « mesure de distance », une application d qui, à deux points associe un nombre : la distance entre les deux points, qui vérifie les trois propriétés suivantes : (*i*) pour tout x et y, $d(x,y) = d(y,x)$; (*ii*) $d(x,y) = 0$ si et seulement si Identité(x,y) ; (*iii*) pour tout x, y, z, $d(x,z) \leq d(x,y) + d(y,z)$.
On appelle « mesure de similarité », une valeur comprise entre 0 et 1 caractérisant un degré de proximité entre deux objets : plus des objets sont proches, plus leur mesure de similarité tend vers 1 ; plus des objets sont éloignés, plus leur mesure de similarité tend vers 0. La similarité entre deux objets est simplement une relation réflexive et symétrique.

distinguant les doublons certains, les doublons possibles et les non-doublons. Dans ces cas, plusieurs seuils, voire plusieurs mesures de distance, sont utilisés.

5.3.1.2. *Les valeurs atypiques, isolées ou aberrantes*

Les valeurs atypiques ou isolées (*outliers*) ont été étudiées depuis plus d'un siècle en statistiques. Aussi exceptionnelles soient-elles, certaines de ces valeurs sont légitimes et d'autres sont aberrantes et peuvent être considérées comme des anomalies ou des erreurs. Malgré toute l'artillerie des méthodes statistiques disponibles, il demeure difficile, sans connaissance additionnelle, de discerner entre ces deux cas. Une valeur aberrante est définie comme étant *une observation qui semble dévier de façon marquée par rapport à l'ensemble des autres membres de l'échantillon* [GRU 69].

Ainsi le problème de détection d'une valeur aberrante notée a, décrivant un objet o telle que $val(o)= a$, peut se rapporter au calcul d'une distance entre celle-ci et un modèle prédéfini, noté $M(o)$, auquel devrait se conformer la représentation uni ou multivariée de l'objet o, tel que :

$$a \in A,\ M(o) \in A, \text{ si } dist(a,M(o)) \geq \theta \text{ alors } a \text{ est une donnée aberrante}$$

Une multitude de techniques de détection ont été proposées pour définir le modèle de conformité et détecter la déviance des données en anomalie par rapport à ce modèle. Ces techniques sont basées respectivement sur :

– un modèle mathématique (par exemple, la régression linéaire) ;

– une comparaison des caractéristiques des distributions de données avec d'autres échantillons ;

– des méthodes géométriques de mesure de distances de la donnée aberrante au reste des données [KNO 98] ;

– la distribution (ou la densité) des données avec une notion d'exceptions locales (*local outliers*) [BRE 00]. Dans ce dernier cas, il est intéressant de remarquer que certaines méthodes permettront d'identifier des valeurs isolées ou aberrantes de façon globale, sur l'ensemble du jeu de données, alors que d'autres méthodes prendront en compte un autre niveau de granularité plus fin et pourront détecter des données isolées sur des sous-ensembles du jeu de données dans la mesure où ces sous-ensembles n'auront pas la même densité de population.

Des tests de *goodness-of-fit* tels que celui du *Chi2* permettent de vérifier l'indépendance des variables du jeu de données. Dans le cas d'attributs dépendants voire de dépendances fonctionnelles, ces tests permettront d'identifier les combinaisons de valeurs d'attributs qui enfreignent les dépendances attendues. Le test

de Kolmogorov-Smirnov permet de mesurer la distance maximum entre la distribution supposée des données et la distribution empirique calculée à partir des données. Ces tests univariés permettent de valider des techniques d'analyse et des hypothèses sur les modèles employés. D'autres tests multivariés plus complexes tels que le test de Mahalanobis permettant de comparer les distances entre moyennes multivariées, ou encore la mesure Kullback Leibler pour mesurer la distance entre des histogrammes de fréquences, peuvent également être employés. Leur objectif général est de mesurer la différence entre la distribution effective de données et une distribution attendue.

5.3.2. *Mesures de distance pour la détection de doublons*

Dans les deux types de problèmes visés, détection de doublons ou de valeurs aberrantes, la qualité de la détection dépend notamment de la mesure de distance choisie. Cette section se consacre à la présentation des mesures utilisées les plus fréquemment. Ne couvrant pas en détail toutes les mesures existantes, nous recommandons les articles [ELM 07, NAU 10] pour une référence plus complète des mesures proposées pour la détection de doublons.

Avant de passer à la présentation des différentes mesures, il est important de signaler la différence entre une mesure de distance notée *dist(·)*, telle que nous l'utilisons dans la définition de nos deux problèmes, et une mesure de similarité notée *sim(·)*. En effet, tandis que *dist(·)* prend des valeurs plus importantes pour quantifier une plus grande distance entre les éléments comparés, l'inverse est vrai pour *sim(·)*. En admettant que les mesures soient normalisées, ayant un domaine de valeurs entre 0 et 1, une distance $dist(a,b) = 1$ indique la distance maximale (a et b sont complètement différents l'un de l'autre), alors que $sim(a,b) = 1$ indique la similarité maximale (a et b sont considérés identiques). Dans le cas de mesures normalisées, nous pouvons convertir une distance en une similarité telle que :

$$\text{sim}(a,b) = 1 - \text{dist}(a,b)$$

La figure 5.1 catégorise les différentes mesures utilisées lors de la détection de doublons. La majorité de ces mesures considèrent les représentations d'objets comparées comme des chaînes de caractères. Lorsque les représentations sont structurées, nous admettons que ces chaînes de caractères ont été obtenues par concaténation des valeurs individuelles (après les avoir converties en type chaîne de caractères, si nécessaire). Ainsi typiquement, les mesures présentées comparent deux chaînes de caractères *s1* et *s2*.

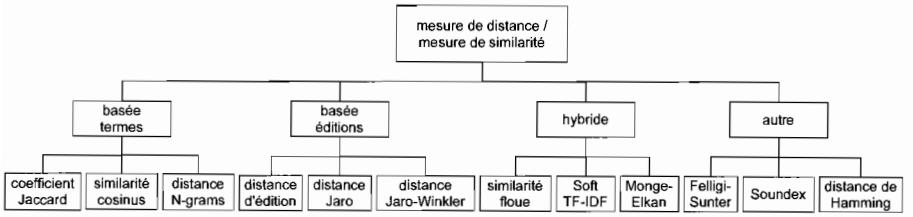

Figure 5.1. *Type de mesures de distance utilisées pour la détection de doublons*

Nous distinguons les principaux types de mesure suivants [NAU 10] :

– *mesures basées sur des termes*, ces mesures comparent deux collections de termes. Ces termes sont obtenus en divisant *s1* et *s2* en termes individuels ;

– *mesures basées édition*, ces mesures comparent *s1* et *s2* et quantifient la distance en nombre d'opérations permettant de transformer *s1* en *s2* ;

– *mesures hybrides*, ces mesures se composent de mesures de types différents, par exemple, des mesures basées sur des termes et des mesures basées édition ;

– *autres mesures*, ces mesures ne peuvent être classées dans l'une des catégories précédentes, par exemple, des distances phonétiques telles que *Soundex*, des distances numériques, dont fait partie la distance de Hamming, ou encore la mesure probabiliste de Fellegi-Sunter.

Les mesures les plus répandues en pratique sont les mesures basées sur des termes, basées édition et les mesures hybrides. Nous limitons la discussion détaillée à des représentants de ces trois types. La définition et les caractéristiques principales des autres mesures sont récapitulées dans le tableau 5.5.

5.3.2.1. *Mesures basées sur des termes*

Parmi les mesures basées sur des termes, le coefficient Jaccard et la similarité cosinus sont présentés car ils sont les plus communément utilisés.

Coefficient de Jaccard : utilisant une fonction *termes(s)* divisant une chaîne de caractères *s* en un ensemble de termes $\{t_1, t_2, ..., t_i, ..., t_n\}$, le coefficient de Jaccard de deux chaînes de caractères *s1* et *s2* est défini par l'équation [5.2] :

$$CoeffJaccard(s1, s2) = \frac{|termes(s1) \cap termes(s2)|}{|termes(s1) \cup termes(s2)|} \qquad [5.2]$$

où $|termes(s)|$ est le cardinal de l'ensemble *termes(s)*.

Similarité cosinus : la similarité cosinus de deux chaînes de caractères *s1* et *s2* est définie par l'équation [5.3] [BIL 03] :

$$SimCosinus(s1,s2) = \frac{V_{s1} \cdot V_{s2}}{\|V_{s1}\| \cdot \|V_{s2}\|} \qquad [5.3]$$

où $\|V_s\| = \sqrt{a^2 + b^2 + c^2 + \cdots}$ pour un vecteur $Vs = [a,b,c,...]$ et le vecteur Vs de dimensionnalité égale au nombre de termes en lesquels la chaîne s peut être décomposée et où le i-ième terme correspond à la i-ième dimension.

Le vecteur Vs contient soit une valeur égale à 0 en i-ième position si s ne contient pas le terme correspondant, soit une valeur correspondant à un poids $w(t_i)$ associé au terme t_i de s.

En pratique, la mesure *TF-IDF* (voir tableau 5.5) est souvent utilisée pour calculer le poids w d'un terme. Afin d'illustrer les mesures basées sur des termes introduites ci-dessus, utilisons une relation décrivant des CDs extraits de *freedb.org* (le 10/08/2011).

CDID	Titre	Artiste	Genre
9e09760c	Britney	Britney	rock
d50bf60e	ngap	britney	misc
9109b10c	Britney	Britney Spears	misc
8d12a51b	Live from Las Vegas	Britney Spears	rock
670dba08	Pink Floyd Live	Pink Floyd	rock
cd0d210f	Try This	Pink	rock
cd0d220f	Try This	Pink	rock

Tableau 5.3. *Exemple extrait de freedb.org*

Dans un premier temps, mesurons le coefficient de Jaccard des deux CDs ayant les identifiants *670dba08* et *cd0d210f*. Nous formons les collections de termes en concaténant les valeurs de l'artiste et du genre avant de diviser la chaîne de

caractères selon les espaces. Ainsi, nous obtenons pour les deux CDs les ensembles de termes suivants :

– *termes(670dba08)* = {Pink, Floyd, rock} ;

– *termes(cd0d210f)* = {Pink, rock}.

Ayant deux termes sur trois en commun, il en résulte un coefficient de Jaccard de 2/3. L'exemple précédent ne prend pas en compte les titres des CDs dans la comparaison, par exemple « Pink Floyd Live » et « Try This » pour les CDs considérés précédemment. En principe, nous obtiendrions les ensembles de termes {Pink, Floyd, Live, rock} et {Try, This, Pink, Rock} où chaque terme n'apparaît qu'une seule fois. En pratique, les termes comparés sont souvent liés à l'attribut d'origine de façon à ce que les ensembles contiennent des paires d'attributs-valeurs. Poursuivant notre exemple, nous obtenons :

– *termes(670dba08)* = {(titre, Pink), (titre, Floyd), (titre, Live), (artiste, Pink), (artiste, Floyd), (genre, rock)} ;

– *termes(cd0d210f)* = {(titre, Try), (titre, This), (artiste, Pink), (genre, rock) }.

Dans cet exemple, deux paires d'attributs-valeurs sont communes aux deux ensembles dont l'union en contient huit, le coefficient Jaccard est donc égal à 2/8.

Dans un deuxième temps, consacrons-nous au calcul de la similarité cosinus des deux CDs 8d12a51b et 670dba08. En admettant qu'uniquement les titres et les genres soient utilisés lors de la comparaison, nous obtenons :

– *termes(8d12a51b)* = {Live, from, Las, Vegas, rock};

– *termes(670dba08)* = {Pink, Floyd, Live, rock}.

Calculons à présent le poids de chaque terme, en utilisant la mesure *TF-IDF* (voir tableau 5.5). Admettons que la relation décrivant les CDs corresponde à la relation intégrale, c'est-à-dire, sept tuples au total d'après le tableau 5.3. Nous observons d'une part que chaque terme n'apparaît qu'une seule fois par tuple, d'où $tf = 1$ dans tous les cas. D'autre part, chaque terme apparaît dans un seul des sept tuples, excepté *Live* apparaissant deux fois et *rock*, qui apparaît dans cinq tuples. Ainsi, nous obtenons les scores *TF-IDF* suivants :

tfidf(Live) = log(1 + 1) * log(7/2) = 0,134

tfidf(from) = log(1 + 1) * log(7/1) = 0,253

tfidf(Las) = log(1 + 1) * log(7/1) = 0,253

tfidf(Vegas) = log(1 + 1) * log(7/1) = 0,253

tfidf(Pink) = log(1 + 1) * log(7/1) = 0,253

tfidf(Floyd) = log(1 + 1) * log(7/1) = 0,253

tfidf(rock) = log(1 + 1) * log(7/5) = 0,044

Dans le tableau 5.4, la première colonne montre les termes du domaine des valeurs dont les termes considérés sont issus. Les deux colonnes suivantes reprennent les vecteurs utilisés pour le calcul de similarité cosinus représentant les deux CDs considérés. Ignorons pour l'instant les colonnes restantes.

Terme du domaine	Vecteur de 8d12a51b	Vecteur de 670dba08	Vecteur de 9e09760c	Vecteur de 9109b10c
Britney	0	0	0,134	0,134
ngap	0	0	0	0
Live	0,134	0,134	0	0
from	0,253	0	0	0
Las	0,253	0	0	0
Vegas	0,253	0	0	0
Pink	0	0,253	0	0
Floyd	0	0,253	0	0
Try	0	0	0	0
This	0	0	0	0
misc	0	0	0	0,134
rock	0,044	0,044	0,044	0

Tableau 5.4. *Exemple des vecteurs représentant les CDs*

Utilisant les vecteurs des CDs *8d12a51b* et *670dba08*, nous obtenons la similarité cosinus suivante :

$$SimCosinus\ (s1,s2) = \frac{0{,}134^2 \cdot 0{,}044^2}{\sqrt{0{,}134^2 + 0{,}253^2 + 0{,}253^2 + 0{,}253^2 + 0{,}044^2} \cdot \sqrt{0{,}253^2 + 0{,}253^2 + 0{,}044^2}}$$
$$= 0{,}00021$$

Nous observons que cette valeur est très proche de 0, indiquant que les deux ensembles de termes sont très différents l'un de l'autre, ce qui est en effet le cas. Notons que le coefficient de Jaccard, dans cet exemple égale à $2/7 \approx 0{,}29$.

Le tableau 5.4 montre également les vecteurs des deux CDs *9e09760c* et *9109b10c*. Leur similarité cosinus est égale à :

$$SimCosinus(s1,s2) = \frac{0{,}134^2}{\sqrt{0{,}134^2 + 0{,}134^2} \cdot \sqrt{0{,}134^2 + 0{,}044^2}}$$
$$= 0{,}672$$

tandis que le coefficient de Jaccard est quant à lui égal à $1/3 \approx 0{,}33$ seulement.

Nous observons dans les deux exemples que le coefficient de Jaccard obtient un résultat similaire aux environs de 0,3, tandis que la similarité cosinus distingue nettement les ensembles très différents dans le premier exemple, et les ensembles similaires dans le second. Ceci est principalement dû au fait que chaque terme est associé à un poids reflétant la force d'identification d'un terme dans un domaine. En effet, tandis que le terme *rock*, commun à beaucoup de tuples dans la relation considérée, n'a pas beaucoup d'impact sur la similarité, des termes plus spécifiques à un CD particulier, tels que les noms d'artistes, ont plus d'influence sur la similarité totale.

Jusqu'à présent, nous avons divisé des chaînes de caractères en ensembles de termes. Il y a cependant d'autres possibilités, notamment la formation de *q-grams* qui est très populaire car elle permet de compenser de petites erreurs typographiques. Etant donnée une chaîne de caractères s, nous obtenons l'ensemble des q-grams en faisant glisser une fenêtre de taille q au-dessus de la chaîne de caractères s, de telle façon que chaque contenu de la fenêtre corresponde à un terme de q caractères. Au début et à la fin de s sont introduits q-1 caractères de complétion (notés #) pour garantir que chaque terme a q caractères.

En générant par exemple des q-grams de longueur 3, c'est-à-dire des trigrammes, nous obtenons pour les valeurs d'artistes « britney » et « Britney » :

- *termes(*britney*)* = {##b, #br, bri, rit, itn, tne, ney, ey#, y#} ;
- *termes(*Britney*)* = {##B, #Br, Bri, rit, itn, tne, ney, ey#, y#}.

Utilisant ces deux ensembles, on obtient : *CoeffJaccard(britney,Britney) = 6/12.* Sans l'utilisation de q-grams, cette similarité aurait été de 0, montrant ainsi que l'utilisation de q-grams peut compenser de petites erreurs telles que la capitalisation différente des deux valeurs comparées.

5.3.2.2. *Mesures de distance d'édition*

Consacrons-nous maintenant aux mesures basées *édition*, notamment la distance de *Levenshtein*, la similarité de *Jaro* et son extension, similarité de *Jaro-Winkler*. D'autres mesures basées édition existent, par exemple la distance de *Smith-Waterman* ou l'utilisation d'espaces affines [NAV 03].

5.3.2.2.1. Distance de Levenshtein et distance d'édition

Soient *s1* et *s2* deux chaînes de caractères à comparer, la distance de Levenshtein notée *DistLevenshtein(s1,s2)* est égale au nombre minimal d'opérations nécessaires à la transformation de *s1* en *s2*, les opérations étant l'ajout, la suppression ou le remplacement d'un caractère.

En général, une distance d'édition est définie par des opérations d'édition ayant chacune un coût associé. La distance de Levenshtein en est un cas particulier, pour lequel chaque opération a un coût égal à 1.

Afin d'obtenir un résultat compris entre 0 et 1, la distance de Levenshtein peut être divisée par la longueur maximale des deux chaînes de caractères, car, au pire, il faut remplacer tous les caractères de la chaîne la plus courte puis la compléter afin d'obtenir la chaîne la plus longue.

La distance de Levenshtein est utile lorsque les chaînes de caractères comparées sont plutôt courtes et ne se distinguent que par quelques erreurs typographiques (oubli d'un caractère, faux caractère, addition d'un caractère supplémentaire, par exemple). En revanche, elle pénalise démesurément les transpositions de caractères ou encore des erreurs en bloc affectant plusieurs caractères, par exemple l'insertion d'un suffixe. Les mesures suivantes ont comme but d'atténuer l'impact des transpositions (similarité de Jaro) et des suffixes non-communs aux deux chaînes de caractères (similarité de Jaro-Winkler).

5.3.2.2.2. Similarités de Jaro et Jaro-Winkler

La similarité de Jaro-Winkler [WIN 91] entre deux chaînes de caractères *s1* et *s2* ayant un préfix commun dénoté ρ se calcule par :

$$SimJaroWinkler(s1, s2) = SimJaro(s1, s2) + |\rho| \cdot f \cdot (1 - SimJaro(s1, s2)) \quad [5.4]$$

Dans cette équation, *f* est un facteur corrigeant la similarité calculée par la similarité de Jaro [JAR 89], notée *SimJaro(s1,s2),* en considérant le préfixe commun ρ entre *s1* et *s2*.

SimJaro(s1, s2) est définie par l'équation [5.5] :

$$SimJaro(s1, s2) = \frac{1}{3} \cdot \left(\frac{|\sigma|}{|s1|} + \frac{|\sigma|}{|s2|} + \frac{|\sigma| - 0.5t}{|\sigma|} \right) \quad [5.5]$$

où σ est l'ensemble des caractères communs entre *s1* et *s2* et *t* est le nombre de transpositions de caractères communs.

Plus formellement, un caractère *c* fait parti de σ si *c* fait partie de *s1* en position *i*, *c* fait partie de *s2* en position *j* et $|i - j| \leq [0.5 \times maximum(|s1|,|s2|)] - 1$. Une transposition existe alors lorsque, en traversant *s1* et *s2*, le *i*-ième caractère commun de *s1* est différent du *i*-ième caractère de *s2*.

L'illustration des mesures basées éditions est fondée sur les mêmes données du domaine des CDs que les exemples des mesures basées sur des termes, notamment sur les données présentées dans le tableau 5.3.

Consacrons-nous tout d'abord à la distance de Levenshtein. Afin d'illustrer toutes les opérations d'édition (ajout, suppression, et remplacement de caractères), nous admettons que nous avons deux valeurs erronées de l'artiste nommée Britney Spears, par exemple *s1* = « britney Spear » et *s2* = « Brit Spears ». Dans ce cas, la distance de Levenshtein mesure 5, car pour transformer *s1* en *s2*, nous devons remplacer *b* par *B*, supprimer trois caractères (*n, e, y*), et ajouter un *s*. La distance normalisée est de $5/maximum(13,11) \approx 0,83$, ce qui peut également être traduit en une similarité 1 – 0,38 = 0,62. La distance de Levenshtein entre « Spears » et « Spaers » mesure 2 et peut être transformée comme décrit précédemment en une similarité égale à 0,67. La transposition de caractères étant une erreur fréquente, un tel impact est souvent démesuré. La distance de Jaro égale à 0,93, dont nous illustrons le calcul ci-après, permet de réduire cet effet.

En comparant s1 = « Spears » et s2 = « Spaers », tous les caractères font partie de l'ensemble de caractères communs, car pour *S, p, r,* et *s,* les positions sont identiques tandis que pour *e* et *a*, les positions varient de 1, ce qui est inférieur à la limite $0.5 \times 6 - 1$. Ainsi, $\sigma = \{S,p,e,a,r,s\}$. En traversant *s1* et *s2*, nous observons que le premier caractère commun (*S*) est également le premier caractère commun de

s2, le second caractère commun de *s1* (*p*) est également le second caractère commun de *s2*, mais le troisième caractère commun de *s1* (*e*) n'est pas le troisième caractère en commun de *s2*. Ceci signifie une première transposition. Une seconde transposition est observée pour le quatrième caractère de *s1* (*a*), les positions restantes sont identiques. Nous avons donc au total un nombre de transpositions $t = 2$. Finalement, nous obtenons :

$$SimJaro(s1, s2) = \frac{1}{3} \cdot \left(\frac{|5|}{|5|} + \frac{|5|}{|5|} + \frac{|5| - 0.5 \cdot 2}{|5|} \right) = 0.93$$

Comparons à présent les CDs ayant les identifiants 9e09760c et 9109b10c en utilisant les noms d'artistes uniquement, c'est-à-dire s1 = « Britney » et s2 = « Britney Spears ».

Dans ce cas, la similarité de Jaro mesure $1/3 \times (7/7 + 7/14 + 1) = 0{,}83$.

La similarité de Jaro-Winkler mesure, pour un choix $f = 0.1$, SimJaroWinkler = $0.83 + 8 \times 0.1 \times (1\text{-}0.83) = 0{,}97$, dû au fait que le préfixe commun est valorisé.

5.3.2.3. *Mesures hybrides*

Les mesures hybrides réutilisent des concepts de plusieurs mesures individuelles, par exemple des mesures basées édition ou des mesures basées sur des termes. Des exemples de mesures hybrides sont la mesure de similarité floue [CHA 03], la similarité Monge-Elkan [MON 96] ou encore la mesure présentée dans [BIL 03]. Dans cette section, nous décrivons brièvement la similarité floue, les autres mesures sont définies dans le tableau 5.5.

5.3.2.3.1. Similarité floue

La similarité floue (*fuzzy match similarity*) divise tout d'abord les deux chaînes de caractères comparées en termes, nous obtenons donc, comme pour les mesures basées sur des termes, deux ensembles de termes termes(s1) et termes(s2). A chaque terme t est associé un coût w(t) égal à sa valeur IDF (voir la définition de la mesure TF-IDF dans le tableau 5.5). La similarité floue correspond alors au coût minimal de transformation de l'ensemble termes(s1) en l'ensemble termes(s2), rappelant une mesure basée édition. Les opérations d'édition sont adaptées aux ensembles de termes et le coût de chaque opération diffère (contrairement à la distance de Levenshtein). Plus précisément :

– le remplacement d'un terme *t1* de *s1* par un terme *t2* de *s2* est associé :à un coût égal à *DistLevenshtein(t1,t2)* × *w(t1)* ;

– l'ajout d'un terme *t* coûte $c_{ins} \times w(t)$ (où c_{ins} est une constante définie au préalable) ;

– l'effacement d'un terme *t* coûte *w(t)* ;

– le coût total de la transformation de *termes(s1)* en *termes(s2)* est noté *tc(s1,s2)*.

En notant *W(s)* la somme des coûts de tous les termes faisant partie de *termes(s)*, la similarité floue est définie par l'équation suivante :

$$FuzzyMatch\ Sim(s1, s2) = 1 - \min\left(\frac{tc(s1, s2)}{W(s1)}, 1\right) \quad [5.6]$$

Appliquée aux deux CDs *d50bf60e* et *9109b10c*, la similarité floue est égale à 0,32 ; ce résultat est obtenu en suivant les étapes de calcul suivantes :

– *termes(d50bf60e)* = {ngap, britney, misc} ;

– *termes(9109b10c)* = {Britney, Britney, Spears, misc}.

En utilisant *idf* pour mesurer le coût d'un terme, nous obtenons :

– *W(d50bf60e)* = *idf*(ngap) + *idf*(britney) + *idf*(misc) ;

– = log(7/1) + log(7/1) + log(7/2) = 2,23.

Le coût total pour transformer *termes(d50bf60e)* en *termes(9109b10c)* correspond à la somme des coûts des opérations suivantes : remplacement de *ngap* par *Britney*, remplacement de *britney* par *Britney* et ajout de *Spears*. Avec $c_{ins} = 1$, nous obtenons :

tc(d50bf60e, 9109b10c) = 1 × idf(ngap) + 1/7 × idf(britney) + idf(Spears)

= log(7/1) + 1/7 log(7/1) + log(7/2) = 1,51

FuzzyMatchSim(d50bf60e, 9109b10c) = 1 – min(1,51/2,23, 1) = 0,32

Calcul de similarité	Définition et principales caractéristiques
Coefficient de Jaccard	Soient deux ensembles de termes S et T : $CoeffJacca\ rd(S,T) = \lvert S \cap T \rvert / \lvert S \cup T \rvert$
Distance de Hamming	Applicable à des champs numériques fixes (n° Sécu, CP) sans prendre en compte les ajout/suppression de caractères.
Distance d'édition	Soient *s1* et *s2* deux chaînes de caractères à apparier, le calcul du coût minimal de conversion de *s1* en *s2* en cumulant le coût unitaire des opérations d'ajout (A), suppression (S) ou remplacement de caractères (R) est tel que : $Edit(s1,s2) = \min\left(\sum A(s1,s2) + S(s1,s2) + R(s1,s2)\right)$
Distance de Jaro	Soient *s1* et *s2*, deux chaînes de caractères de longueur respective *L1* et *L2*, ayant C caractères communs et T transpositions de caractères (utilisé pour les chaînes de caractères courtes) : $Jaro(s1, s2)=(C/L1 + C/L2 + (2C-T)/2C)/3$
Distance de Jaro-Winkler	Soit P la longueur du plus long préfixe commun entre *s1* et *s2* : $Jaro\text{-}Winkler(s1,s2)=Jaro(s1,s2) + max(P,4).(1-Jaro(s1,s2))/10$
Distance N-grams	Somme du nombre de caractères communs sur toutes les sous-chaînes de caractères x de longueur N présents dans les chaînes a et b : $Ngram\ (a,b) = \sqrt{\sum_{\forall x} \lvert f_a(x) - f_b(x) \rvert}$
Soundex	Première lettre du mot puis encodage des consommes sur 3 caractères tel que : B, F, P, V -> 1 ; C, G, J, K, Q, S, X, Z ->2 ; D, T -> 3 ; L -> 4 ; M,N -> 5 ;R -> 6 ; Exemple : « John » et « Jan » sont encodé J500 ; « Dupontel » est encodé D134.
Mesure *TF-IDF*	Soit un terme *s1* et un document d dans un ensemble de documents D, *tf* le nombre d'occurrences du terme *s1* dans le document d et *idf* la fraction du nombre de documents dans D sur le nombre de documents contenant s1 : $Tfidf(s1, d, D)=log(tf(s1,d) + 1)*log(idf(s1, D))$
Mesure probabiliste *IDF* de Fellegi-Sunter	Soient $P_{A\cap B}(s)$ la probabilité que la chaîne de caractère s se retrouve à la fois dans A et dans B (et soit donc identifiée comme doublon) et $P_A(s).P_B(s)$ la probabilité qu'elle ne le soit pas (avec $P_A(s) = P_B(s) = P_{A\cap B}(s)$) $Fellegi\text{-}Sunter\text{-}IDF(s) = \log(P_{A\cap B}(s) / (P_A(s).P_B(s)) = \log(1 / P_A(s))$

Calcul de similarité	Définition et principales caractéristiques
Mesure du cosinus	Soient *a* et *b* deux attributs, *Da* et *Db* les ensembles de termes de chaque attribut, et les scores *Tf-idf* du terme s respectivement dans *Da* et dans *Db* : $SimCosinus(a, b) = \sum_{t \in Da \cap Db} Tfidf(t,D_a).Tfidf(t,D_b)$
Autre distance hybride	Soient $D_a=\{a_1,a_2,...,a_k\}$ et $D_b=\{b_1,b_2,...,b_p\}$ des ensembles de termes, et *s1* et *s2* deux chaînes de caractères à comparer avec distance de similarité $Sim(a_i,b_j)$: $$Hybrid1\ (D_a, D_b) = \frac{1}{k}\sum_{i=1}^{k} \max_{j=1}^{p}\left(Sim\ (a_i, b_j)\right)$$
Similarité floue	Soient *Da* et *Db* deux ensembles de termes, le coût de transformation de *Da* en *Db* est calculé en utilisant la distance d'édition et la mesure *TF-IDF* tel que : $$Cost(D_a, D_b) = \sum_{S_i \in Da} Tfidf\ (s_i, D_a)\Big|_0^1 + Tfidf\ (s_i, D_a) + Edit\ (s_i, s_j) * Tfidf\ (s_i, D_a)$$ $$FuzzyMatch\quad Sim = 1 - \min\left(\left(Cost\ (D_a, D_b)/\sum_{S_i \in Da} Tfidf\ (s_i, D_a)\right),1\right)$$

Tableau 5.5. *Distances de similarité pour comparer les chaînes de caractères et identifier les doublons potentiels*

5.3.3. *Méthodes appliquées pour la détection de valeurs aberrantes*

Les valeurs aberrantes peuvent être classées en trois catégories : les valeurs aberrantes d'amplitude, les valeurs aberrantes spatiales et les valeurs aberrantes relationnelles :

– les valeurs aberrantes d'amplitude sont considérées comme étant trop élevées ou trop basses comparées à l'intervalle des valeurs prises par la majorité des échantillons ;

– les valeurs aberrantes spatiales sont généralement définies comme des observations qui sont extrêmes par rapport aux valeurs voisines ;

– les valeurs aberrantes relationnelles sont définies comme des observations non conformes aux relations (ou corrélations) qui existent entre les variables.

La figure 5.2 catégorise les différentes mesures utilisées lors de la détection de ces différents types de valeurs aberrantes.

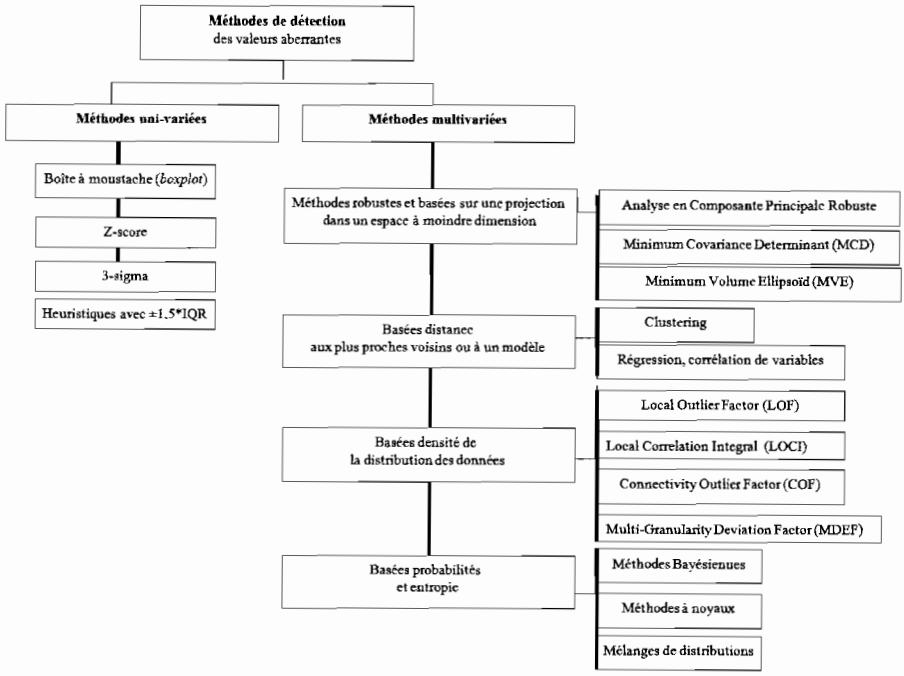

Figure 5.2. *Mesures de distance utilisées lors de la détection de valeurs aberrantes*

5.3.3.1. *Méthodes univariées*

Dans le premier cas, en mode univarié (c'est-à-dire ne considérant les valeurs de qu'un seul attribut), on pourra calculer les indicateurs représentatifs (ou valeurs typiques) telle que la moyenne, la médiane, l'écart-type ; les boîtes à moustaches (*boxplot*) permettront de mettre en évidence les points atypiques par rapport à ces indicateurs. Les extrémités des moustaches sont délimitées par 1,5 fois l'écart interquartile (Q3−Q1) calculé entre les deux quarts inférieurs et supérieurs de la distribution des données par rapport à la médiane. Cette règle permet de déceler l'existence d'un point extrême et elle est plus fiable que la fameuse règle des 3-σqui consiste à isoler les points en-deçà ou au-delà de trois fois l'écart-type autour de la moyenne. En effet, cette règle ne repose pas sur une hypothétique symétrie de la distribution des données et elle utilise des paramètres de localisation (les quartiles) qui, à la différence de la moyenne empirique, sont peu influencés par les points extrêmes. Les principaux inconvénients de ces techniques univariées comme ceux des graphes de contrôle résident d'une part, dans la difficulté à généraliser dans un cadre multidimensionnel pour considérer tous les attributs et d'autre part, à s'affranchir des hypothèses de normalité et de symétrie des jeux de données.

5.3.3.2. *Méthodes multivariées*

Dans le cas multivarié, de nombreuses méthodes non paramétriques ont été proposées. Certaines sont basées sur des projections dans des espaces à moindres dimensions [AGG 01] ou selon des composantes principales comme le montre l'algorithme dans le tableau 5.6. Dans le tableau 5.6, l'algorithme calcule la moyenne et la matrice de variance-covariance du jeu de données et sélectionne les points dont la distance de Mahalanobis à la moyenne est supérieure à la valeur critique de la loi du Chi2 pour le degré de liberté correspondant aux nombre de variables considérées et le seuil de risque choisi.

D'autres méthodes telles que celles proposées par [KNO 98, RAM 00] prennent en compte l'éloignement d'un point à l'ensemble des points de son voisinage selon sa densité ; elles offrent l'avantage de détecter les données isolées localement qui n'auraient pas pu être détectées dans la globalité du jeu de données. A titre d'exemples, [KNO 98] définissent un point O comme *DB(p,d)-outlier* si au moins une fraction p du jeu de données est située à une distance plus grande que la distance d. Selon [RAM 00], les données aberrantes sont les n premiers points dont la distance au k-ième plus proche voisin est la plus grande.

Les méthodes les plus récentes exploitent différents types de corrélations entre les variables et les corrélations spatiales [CHA 05]. Nous invitons le lecteur a consulté le tutorial de Kriegel *et al.* [KRI 10] pour une description détaillée des différentes méthodes disponibles.

Entrée : jeu de données N X D (N lignes, d colonnes)

Sortie : ensemble des valeurs aberrantes candidates noté O

- calcul de la moyenne μ et de la matrice variance-covariance Σ
- soit C, le vecteur-colonne composé de la racine carrée de la distance de Mahalanobis à la moyenne μ tel que :

$$(x-\mu)'\Sigma^{-1}(x-\mu)=(x-\mu)'\begin{bmatrix}\sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d}\\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d}\\ \vdots & \vdots & \ddots & \vdots\\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd}\end{bmatrix}^{-1}(x-\mu)$$

- sélectionner les points O de C dont la valeur est supérieure à la valeur critique de la loi du Chi 2 pour le degré de liberté d à la probabilité de 97.5 % :

$$inv\left(\sqrt{\chi_d^2(.975)}\right)$$

Tableau 5.6. *Méthode multivariée pour la détection des données aberrantes basée sur la distance de Mahalanobis*

Comme l'illustre la figure 5.3, toute la difficulté réside dans le choix des méthodes et de leur paramétrage. Les méthodes dans leur grande diversité divergeront dans leurs résultats de classification des données aberrantes : dans le cas d'une analyse univariée sur chacune des deux variables X et Y de la figure 5.3, la zone de rejet sera définie comme devant être inférieure à 2 % ou supérieure à 98 % du jeu de données en se basant sur une contrainte sur l'écart interquartile, délimitant ainsi un rectangle et excluant tous les points au dehors ; en menant une analyse multivariée combinant les deux variables et employant la distance de Mahalanobis comme l'a présenté l'algorithme du tableau 5.6, la zone des valeurs acceptables sera alors définie par une ellipsoïde. Mais, en superposant les deux analyses, nous constaterons que certains points sont considérés aberrants par l'analyse multivariée alors qu'ils ne le sont pas par l'analyse univariée. De plus, certaines méthodes reposent sur de fortes hypothèses de normalité ou de symétrie du jeu de données qui sont rarement vérifiées dans le cas de données réelles.

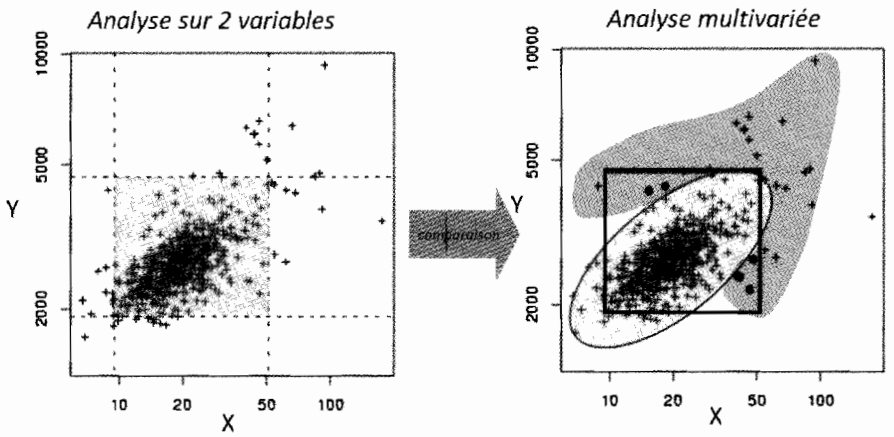

Figure 5.3. *Contradictions entre les méthodes de détection univariées et multivariées*

Les méthodes basées uniquement sur la proximité du voisinage sans considérer la densité de celui-ci n'auront pas la même capacité à détecter les valeurs aberrantes comme l'illustre l'exemple de la figure 5.4. Les points *O1* et *O2* sont à des distances respectives *d1* et *d2* de leurs plus proches voisins respectifs avec $d2 \geq d1$. Une méthode basée sur les plus proches voisins considèrera seulement *O2* comme étant une donnée isolée ou *outlier* (et non *O1)* alors qu'une méthode combinant la distance aux plus proches voisins et la densité de ceux-ci considérera à l'opposé *O1* comme un *outlier* et *O2* n'en sera pas un.

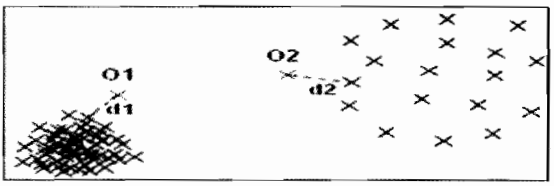

Figure 5.4. *Contradictions soulevées par les méthodes de détection*

Finalement, considérant la diversité des méthodes disponibles et leur pouvoir de détection variable, une approche récente consiste à employer un panel de méthodes à la fois univariées et multivariées, et faire corroborer leurs résultats de détection pour obtenir un consensus sur les valeurs qui apparaissent comme étant aberrantes pour la majorité des méthodes.

5.3.4. *Mise en œuvre des mesures de distance dans des méthodes de détection de doublons*

Ayant discuté les mesures de distances utilisées lors de la détection de doublons et lors la détection de valeurs aberrantes, voyons comment ces mesures sont utilisées et mises en œuvre par différentes méthodes, en particulier, dans le cas de la détection des doublons. Ne pouvant dans ce chapitre détailler toutes les méthodes ayant été proposées, nous nous concentrons sur les méthodes les plus utilisées en pratique pour la détection des doublons (notamment dû à leur applicabilité à de larges volumes de données) et nous invitons nos lecteurs à se reporter aux ouvrages de synthèse [ELM 07, NAU 10] pour un panorama plus complet.

Rappelons que le but de la détection de doublons est de détecter chaque paire de représentations d'objets *(a, b)*, avec $a \in A$ et $b \in B$, dont la distance entre objet est inférieure à un seuil fixé. Cela nécessite la comparaison de toutes les paires issues du produit Cartésien A × B, engendrant une complexité quadratique dans le nombre de représentations d'objets. En pratique, une telle complexité n'est pas admissible pour de larges volumes de données, d'où l'intérêt commun à toutes les méthodes de détection de doublons présentées ci-après à limiter cette complexité en réduisant le nombre de comparaisons nécessaires. Lors de cette démarche, il est cependant indispensable de ne pas réduire la qualité du résultat outre mesure. En effet, tout en omettant certaines comparaisons, il faut minimiser le risque que celles-ci soient en réalité des comparaisons identifiant des doublons, car ceux-ci resteraient alors dans les données après le nettoyage.

Dans le contexte de la réduction de comparaisons de paires de représentations d'objets, deux approches sont souvent citées : les approches basées sur la formation de partitions (*blocking*) [ANA 02, BAX 03, BIL 06] et les méthodes basées sur des fenêtres glissantes (*windowing*) [HER 95, MON 96].

Les méthodes dites *blocking* forment des partitions des ensembles A et B. Le critère de partitionnement appliqué aux ensembles A et B peut être spécifique au domaine considéré (par exemple, des représentations d'objets ayant le même code postal font partie de la même partition) ou basé sur des critères indépendants du domaine. Suite à cette division de A et B en blocs idéalement plus petits que les ensembles mêmes, les représentations d'objets issues des blocs correspondants sur A et B sont comparées (par exemple, les blocs ayant le même code postal).

Plus précisément, étant données l'ensemble des partitions de A noté $\{P_1^A, P_2^A, ..., P_k^A\}$ et l'ensemble des partitions de B noté $\{P_1^B, P_2^B, ...P_k^B\}$, uniquement les paires faisant partie de $(P_1^A \times P_1^B) \cup (P_2^A \times P_2^B) \cup ... \cup (P_k^A \times P_k^B)$ sont comparées.

Le choix du critère de partitionnement est essentiel à l'obtention d'une réduction du nombre de comparaisons satisfaisante, tout en maintenant une haute qualité du résultat. D'une part, il est préférable de choisir un critère qui génère de petites partitions de tailles comparables afin de réduire le nombre de comparaisons. D'autre part, le même critère doit idéalement assurer que des doublons tombent dans des partitions correspondantes.

Concernant les méthodes basées sur des fenêtres glissantes, l'idée générale est de former l'union $D = A \cup B$, de trier D en utilisant une clé de tri adaptée (spécifique au domaine ou générique), de faire glisser une fenêtre d'une certaine taille (fixe ou adaptable) par dessus la séquence triée et de comparer uniquement des représentations d'objets situées dans la même fenêtre. En considérant une taille de fenêtre constante notée w telle que $2 \leq w \leq |D|$, le nombre de comparaisons est réduit de $O(|D|^2)$ à $O(w|N|)$ avec N le nombre d'enregistrements Bien sûr, il faut également prendre en compte la complexité des étapes précédant l'étape de comparaison, dont l'étape de tri qui est la plus complexe en $O(|D| \ log \ |D|)$. Dès lors que w est choisi suffisamment petit, ce qui est le cas en pratique (souvent, des valeurs entre 5 et 30 sont suffisantes), la complexité des méthodes à fenêtres glissantes est en $O(|D| \log |D|)$.

En choisissant une petite valeur de w, le nombre de comparaisons est réduit de manière significative. Mais, afin de détecter les doublons, il faut s'assurer que ceux-ci se trouvent dans la même fenêtre. Le choix de la clé de tri joue un rôle très

important dans ce contexte, car, idéalement, elle permet de trier les doublons proches les uns des autres.

Nous observons que les deux types de méthodes présentés ci-dessus ont chacun un paramètre de configuration dont dépend la qualité du résultat (le critère de partitionnement et la clé de tri, respectivement). En pratique, il est difficile, voir impossible de définir ces paramètres de manière optimale. Pour entraver cet effet, une solution est d'élargir les ensembles de représentations d'objets comparés, par exemple en comparant des représentations de partitions adjacentes ou en augmentant la taille de la fenêtre glissante. Mais cela engendre un plus grand nombre de comparaisons, ce que les méthodes proposées essaient justement de réduire. Une solution souvent choisie en pratique est le choix de plusieurs configurations alternatives (par exemple, une configuration qui délimite les partitions en fonction du code postal, puis une autre utilisant la première lettre du nom de la ville). Ces différentes configurations sont appliquées une à une et les résultats respectifs sont unifiés en formant l'enveloppe transitive de toutes les paires de doublons détectées.

Nous illustrons les deux approches en utilisant des données décrivant des personnes et leurs lieux de résidence, représentées dans le tableau 5.7. Comme précédemment, nous simplifions l'exemple en choisissant $A = B$, nous désirons donc identifier des doublons dans une seule relation. Dans ce cas, le nombre de comparaisons sans l'utilisation des approches présentées est de 21 (en utilisant une mesure de distance symétrique). Admettons que les doublons à détecter soient les paires aux identifiants (7,4), (6,3) et (1,5).

PID	Nom	Code Postal	Ville
1	Pierre	06200	Nice
2	Jean	69001	Lyon
3	Dupont	75002	Paris
4	Didier	75002	Paris
5	Pier	02600	Nice
6	Dupond	75000	Paris 2e
7	Didié	75002	F-Paris

Tableau 5.7. *Exemple d'une liste de personnes et de leurs résidences*

Dans un premier temps, formons des partitions basées sur le code postal. Le résultat est représenté dans le tableau 5.8.

PID	Nom	Code Postal	Ville
1	Pierre	*06200*	Nice
2	Jean	*69001*	Lyon
5	Pier	*02600*	Nice
6	Dupond	*75000*	Paris 2e
7	Didié	*75002*	F-Paris
3	Dupont	*75002*	Paris
4	Didier	*75002*	Paris

Tableau 5.8. *Exemple de liste partionnée selon le code postal*

Nous obtenons cinq partitions, dont une seule contient plus d'un tuple. Lors de la détection de doublons, uniquement les tuples de chaque partition sont comparés entre eux, c'est-à-dire les paires (7,3), (7,4) et (3,4). Le nombre de comparaisons est donc réduit considérablement (3 au lieu de 21). Lors des trois comparaisons exécutées, nous trouvons le doublon (7,4) en utilisant par exemple une mesure de distance d'édition telle que nous l'avons décrite précédemment. Malheureusement, les doublons (6,3) et (1,5) ne sont pas identifiés. Ceci est dû au choix du critère formant les partitions.

Nous pouvons remédier à ce problème en utilisant un second critère de partitionnement, par exemple, en divisant l'ensemble de tuples en fonction de la première lettre de la ville (voir tableau 5.9).

Dans ce cas, nous comparons (1,5), (3,4), (3,6) et (4,6) et trouvons avec ces quatre comparaisons supplémentaires les doublons manquants (mais pas celui trouvé précédemment !).

En unifiant les résultats obtenus de ces deux configurations, nous trouvons tous les doublons avec sept au lieu de 21 comparaisons.

PID	Nom	Code Postal	Ville
1	Pierre	06200	*Nice*
5	Pier	02600	*Nice*
2	Jean	69001	*Lyon*
3	Dupont	75002	*Paris*
4	Didier	75002	*Paris*
6	Dupond	75000	*Paris 2e*
7	Didié	75002	*F-Paris*

Tableau 5.9. *Autre exemple de partionnement de la liste selon la ville de résidence*

Voyons maintenant comment une méthode basée sur une fenêtre glissante se comporte dans cet exemple. Nous devons tout d'abord choisir une clé de tri, par exemple, une clé constituée des deux premiers caractères du nom et des deux premiers chiffres du code postal. Ensuite, les tuples sont triés en ordre croissant de leur clé. Le résultat du tri et les clés y aboutissant sont montrés dans le tableau 5.10.

PID	Nom	Code Postal	Ville	Clé de tri
4	Didier	75002	Paris	*Di75*
7	Didié	75002	F-Paris	*Di75*
3	Dupont	75002	Paris	*Du75*
6	Dupond	75000	Paris 2e	*Du75*
2	Jean	69001	Lyon	*Je69*
5	Pier	02600	Nice	*Pi02*
1	Pierre	06200	Nice	*Pi06*

Tableau 5.10. *Autre exemple de partionnement basé sur une clé de tri*

En choisissant une fenêtre glissante de taille constante égale à 2, nous comparons, dans cet ordre (4,7), (7,3), (3,6), (6,2), (2,5) et (5,1), identifiant les trois paires de doublons avec six comparaisons seulement.

5.3.5. *Exemple d'applications*

Après avoir limité les exemples des sections précédentes à des exemples dédiés à l'illustration des approches, nous donnons ici quelques exemples d'applications réelles. Lors d'un projet de coopération industrielle avec la compagnie SCHUFA Holding AG qui estime la solvabilité des personnes en Allemagne et dont dépend par exemple l'obtention d'un crédit, nous avons détecté des doublons dans leurs données relationnelles décrivant des personnes et des contrats (crédit, téléphone portable, compte bancaire, etc.) Pour cela, des méthodes adaptées au domaine spécifique de SCHUFA ont été développées [WEI 08]. Celles-ci sont cependant basées sur les méthodes discutées dans ce chapitre. Le processus de détection de doublons proposé compare des représentations d'objets (personnes et contrats) en utilisant un profil de comparaison, qui est en principe une séquence des composantes suivantes : filtres utilisant des méthodes *blocking*, classificateurs de doublons basés sur des règles spécifiques au domaine (par exemple, si la date de naissance et le nom de famille sont identiques, et si le prénom a une distance de Levenshtein inférieure à un seuil spécifié, alors les personnes sont des doublons), classificateurs de non-doublons basés sur des règles spécifiques au domaine et une mesure de distance hybride incluant, entre autres, le coefficient de Jaccard et la distance de Levenshtein.

Afin d'éviter l'application de ces composantes à toutes les paires, trop coûteuse étant donné le nombre de tuples à comparer (60 millions de personnes avec les contrats associés), une technique de fenêtre glissante de taille fixe a été utilisée. L'utilisation de la méthode proposée a été évaluée sur une fraction des données SCHUFA mises à notre disposition (dix millions de tuples représentant des personnes et les contrats associés à ces personnes). Le résultat ayant été satisfaisant, les méthodes ont été intégrées dans le système opérationnel de SCHUFA et elles sont utilisées depuis fin 2010.

Nous avons choisi des méthodes similaires pour d'autres applications (par exemple, dans les domaines de données décrivant des CDs ou des films). Dans tous les cas, nous avons observé que des mesures génériques telles que celles discutées ici obtiennent de bons résultats, mais l'inclusion de savoir spécifique au domaine considéré est souvent indispensable à l'obtention d'un résultat de bonne qualité. Cela s'explique d'une part par le meilleur raisonnement classifiant un doublon, mais également par le fait que ce savoir est utilisé pour définir des critères de partitionnement ou des clés de tri.

5.4. Conclusion et perspectives

Ce chapitre a passé en revue les principales méthodes et mesures de distance permettant de détecter les doublons et les données aberrantes. Elles ont été illustrées par de nombreux exemples et leur calcul a été détaillé. De part leur grande diversité et les nombreuses alternatives de paramétrage possible, il n'est pas possible de déclarer qu'une méthode ou une mesure est meilleure qu'une autre ; ce constat dépendra du jeu de données, de ses caractéristiques propres et du domaine d'application. C'est pourquoi il est important d'utiliser et de maîtriser un panel conséquent de méthodes et de mesures afin de corroborer leurs résultats intelligemment. Toutefois, une artillerie de méthodes et de mesures ne sera pas suffisante sans un expert du domaine. Une valeur isolée ou jugée aberrante par un ensemble de méthodes peut toutefois être une exception légitime. Le bénéfice du doute lui sera accordé tant que l'expert ou une vérité-terrain ne la classera pas définitivement comme un problème de qualité de données.

Pour conclure, les multiples problèmes évoqués dans ce chapitre offrent plus que jamais d'intéressantes perspectives de recherche pour les différentes communautés scientifiques travaillant autour de la qualité des données en statistiques, bases de données, ingénierie de la connaissance, gestion de processus. Pour les entreprises et industriels, détenteurs de données en masse, la qualité des données peut demeurer un épineux problème qui se pose de façon récurrente, les guidant ponctuellement vers des choix pragmatiques et souvent à court terme par manque de moyens et d'appuis hiérarchiques. Il est alors clair que pour apporter des solutions concrètes, opérationnelles sur le long terme et théoriquement fondées, des collaborations étroites entre le monde académique et les industriels sont une nécessité.

5.5. Bibliographie

[AGG 01] AGGARWAL C.C., YU P.S., « Outlier detection for high dimensional data », *ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2001)*, 2001.

[ANA 02] ANANTHAKRISHNA R., CHAUDHURI S., GANTI V., « Eliminating fuzzy duplicates in data warehouses », *International Conference on Very Large Databases*, Hong-Kong, Chine, août 2002.

[BAX 03] BAXTER R., CHRISTEN P., CHURCHES T., « A comparison of fast blocking methods for record linkage », *International Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, Washington DC, 2003.

[BIL 03] BILENKO M., MOONEY R.J., COHEN W.W., RAVIKUMAR P.D., FIENBERG S.E., « Adaptive name matching in information integration », *IEEE Intelligent Systems*, vol. 18, n° 5, p. 16-23, 2003.

[BIL 06] BILENKO M., KAMATH B., MOONEY R.J., « Adaptive blocking : Learning to scale up record linkage », *IEEE International Conference on Data Mining*, Las Vegas, Nevada, Etats-Unis, juin 2006.

[BLE 08] BLEIHOLDER J., NAUMANN F., « Data fusion », *ACM Computing Surveys*, vol. 41, n° 1, p. 1:1 – 1:41, 2008.

[BRE 00] BREUNIG M., KRIEGEL H., NG R., SANDER J., « LOF : Identifying density-based local outliers », *International Conference ACM SIGMOD*, p. 93-104, 2000.

[CHA 03] CHAUDHURI S., GANJAM K., GANTI V., MOTWANI R., « Robust and efficient fuzzy match for online data cleaning », *ACM International Conference on the Management of Data*, 2003.

[CHA 05] CHAWLA S., SUN P., « SLOM : a new measure for local spatial outliers », *Knowledge and Information Systems,* 2005.

[CHR 08] CHRISTEN P., « Febrl – An Open Source Data Cleaning, Deduplication and Record Linkage System with a Graphical User Interface », *ACM SIGKDD 2008 Conference*, Las Vegas, août 2008.

[DAS 02] DASU T., JOHNSON T., MUTHUKRISHNAN S., SHKAPENYUK V., « Mining database structure or, How to build a data quality browser », *ACM SIGMOD Conference,* 2002.

[DEV 06] DEVILLERS R., JEANSOULIN R., *Fundamentals of Spatial Data Quality*, Wiley, New York, 2006.

[ELM 07] ELMAGARMID A.K., IPEIROTIS P.G. VERYKIOS V.S., « Duplicate record detection : A survey », *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, n° 1, p. 1-16, 2007.

[GAL 01] GALHARDAS H., FLORESCU D., SHASHA D., SIMON E., SAITA C., « Declarative data cleaning : Language, model, and algorithms », *International Conference on Very Large Databases (VLDB)*, p. 371-380, 2001.

[GRU 69] GRUBBS F.E., « Procedures for detecting outlying observations in samples », *Technometrics* 11, p. 1-21, 1969.

[HER 95] HERNÀNDEZ M.A., STOLFO S.J., « The merge/purge problem for large databases », *International Conference on the Management of Data*, 1995.

[JAR 89] JARO M.A., « Advances in record linking methodology as applied to matching the 1985 census of Tampa Florida », *Journal of the Americal Statistics Association*, vol. 84, n° 406, p. 414-420, 1989.

[KAN 08] KANG H., GETOOR L., SHNEIDERMAN B., BILGIC M., LICAMELE L., « Interactive Entity Resolution in Relational Data : A Visual Analytic Tool and Its Evaluation », *IEEE Transactions on Visualization and Computer Graphics,* vol. 14, n° 5, p. 999-1014, 2008.

[KNO 98] KNORR E., NG R., « Algorithms for mining distance-based outliers in large datasets », *International Conference on Very Large Databases (VLDB)*, p. 392-403, 1998.

[KRI 10] KRIEGEL H.-P., KRÖGER P., ZIMEK A., « Outlier detection techniques », *16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.

[LOW 01] LOW W.L., LEE M.L., LING T.W., « A knowledge-based approach for duplicate elimination in data cleaning », *Information System*, vol. 26, n° 8, 2001.

[LUC 03] LU C.-T., CHEN D., KOU Y., « Algorithms for Spatial Outlier Detection », *IEEE International Conference on Data Mining*, 2003.

[MON 97] MONGE A.E., ELKAN C.P., « An efficient domain-independent algorithm for detecting approximately duplicate database records », *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, Tuscon, Arizona, Etats-Unis, mai 1997.

[NAU 10] NAUMANN F., HERSCHEL M., *An introduction to duplicate detection*, Synthesis Lectures on Data Management, Morgan & Claypool Publishers, 2010.

[NAV 03] NAVARRO G., « A guided tour to approximate string matching », *ACM Computing Surveys,* vol. 33, n° 1, p. 31-88, 2001.

[PAP 03] PAPADIMITRIOU S., KITAGAWA H., GIBBONS P.B., FALOUTSOS C., « LOCI : Fast outlier detection using the local correlation integral », *Proc. 19th IEEE Int. Conf. on Data Engineering (ICDE '03)*, 2003.

[RAM 00] RAMASWAMY S., RASTOGI R., SHIM K., « Efficient Algorithms for Mining Outliers from Large Data Sets », *International Conference ACM SIGMOD*, Dallas, Texas, 2000.

[RAM 01] RAMAN V., HELLERSTEIN J.M., « Potter's Wheel : an Interactive data cleaning system », *International Conference on Very Large Databases (VLDB)*, 2001.

[WEI 07] WEIS M., MANOLESCU I., « Declarative XML Data Cleaning with XClean », *Proceedings of CAiSE 2007*, p. 96-110, 2007.

[WEI 08] WEIS M., NAUMANN F., JEHLE U., LUFTER J., SCHUSTER H., « Industry-scale duplicate detection », *Proceedings of the VLDB*, vol. 1, n° 2, p. 1253-1264, 2008.

[WIN 91] WINKLER W.E., THIBOUDEAU Y., An application of the Felligi Sunter Model of record linkage to the 1990 US Decennial Census, US Bureau of the Census, 1991.

[ZHA 09] ZHANG K., HUTTER M., JIN H., « A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data », *Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD 2009)*, 2009.

Chapitre 6

La gestion de données multi-sources : de la théorie à la mise en œuvre dans le cadre d'un référentiel client unique

6.1. Introduction

Quel est mon degré de connaissance des clients ? Suis-je capable de suivre de bout en bout le parcours de mes clients à travers les différentes activités de mon entreprise, de l'interaction Web en passant par le service après vente (SAV) et la vente en magasin par exemple ? Telles sont les questions que toute direction générale se pose et auxquelles il n'est pas si simple de répondre, tout au moins de façon complète, correcte et pertinente.

Plusieurs causes participent à cet état de fait : chaque secteur de l'entreprise gère son propre système d'information car les processus et objectifs métier diffèrent ; pour la plupart, ces systèmes d'information ne communiquent pas entre eux ; les données, dont les données client, sont ainsi cloisonnées dans leur silo, souvent redondantes entre elles et parfois divergentes, même pour des données aussi sensibles que la gestion des consentements marketing (*opt-in*/*opt-out*[1]). Or, dans un

---

Chapitre rédigé par Soumaya BEN HASSINE-GUETARI, Delphine CLÉMENT, Sébastien COEUGNIET, Idriss COOWAR et Brigitte LABOISSE.

1. *Opt-in* : ce terme signifie que le client ou le prospect accepte de recevoir des communications promotionnelles de la part d'un annonceur.
Opt-out : ce terme signifie que le client ou le prospect n'accepte plus de recevoir de communications promotionnelles de la part d'un annonceur. Aujourd'hui, la législation CNIL sur la protection des données à caractère personnel oblige les annonceurs à collecter ce

marché très concurrentiel, les entreprises n'ont d'autre choix que de placer leur client au centre de leur stratégie marketing et commerciale.

Comment parvenir alors à réconcilier ce patrimoine de données client hétérogènes pour acquérir une vue client unique, une vue à 360°, une seule version de la vérité ?

Telles sont les visées d'un référentiel client unique (RCU), véritable projet stratégique d'entreprise, et dont la garantie de succès passe indubitablement par la prise en compte des aspects de qualité et de gouvernance des données.

Cet article présente un contexte très opérationnel de la gestion des données multi-sources où nous traitons principalement l'angle particulier de son application à la création d'un référentiel client unique. Mais auparavant, une étude bibliographique rappelle l'ensemble des travaux publiés dans le contexte de la gestion des données multi-sources.

N'est pas abordée dans ce chapitre, la notion d'identification ou de déduplication utilisées pour rapprocher les données d'un client X sur une source A, aux données de ce même client X sur une deuxième source B. En effet, une littérature importante couvre déjà le domaine que l'on appelle en jargon métier « customer identification » en d'autres termes, le dédoublonnage des individus ; c'est un domaine bien maitrisé avec de nombreux logiciels disponibles à ce jour. Nous avons volontairement exclu ce pan de ce chapitre, nous focalisant sur l'étape ultérieure qui, lorsque le lien est fait entre les données, consiste à avoir cette fameuse vue à 360 ° du client, gommant ou consolidant les divergences entre sources.

6.2. Mise en œuvre d'un référentiel client unique : le contexte

Parce que la connaissance client se matérialise par l'historique de données stocké dans les différents systèmes d'information de l'entreprise, leur qualité et leur gouvernance constituent les fondamentaux de la réussite d'un projet de référentiel client unique.

Dans notre expérience, les projets RCU qui sont un succès[2], sont des projets qui mettent la donnée au cœur de la problématique : c'est la donnée qui détermine les

consentement préalable pour les communications par email et par SMS. Pour les communications papier et téléphone, l'*opt-in* n'est pas encore obligatoire.

2. Ce projet est qualifié de « succès » lorsque le métier adhère, qu'il a « confiance » dans les données RCU, non pas seulement lorsque la Direction des systèmes d'information (DSI) livre le RCU en temps et en heure et dans le budget imparti ; nous considérons ce dernier critère comme nécessaire mais non suffisant.

spécifications fonctionnelles (détermination de la notion de personne convergente). Ainsi, un sous-projet « données » (*stream data* dans le jargon) sera créé dès le départ dans le projet RCU. Le pilote de ce sous-projet sera pleinement intégré au comité projet (comité de pilotage). Ce *stream data* concerne plus particulièrement les phases-clés :

– du choix des sources de données client ;

– de la valorisation des données client de qualité ;

– des stratégies de rapprochement entre sources et de leur optimisation[3] ;

– de la mise en correspondance (*mapping*) des données des sources vers le modèle de données RCU cible ;

– de la définition de la notion de compte, de client, de foyer, de site ;

– de l'établissement d'une matrice de priorisation entre sources intervenant lors des phases de fusion/consolidation des données hétérogènes en un enregistrement client unique.

Certaines de ces étapes sont représentées de façon macroscopique dans la figure 6.1.

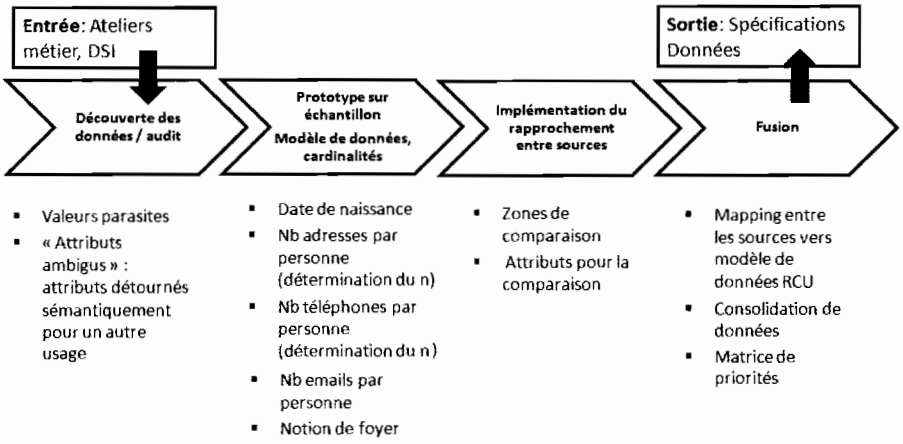

Figure 6.1. *Etapes qualité de données dans un projet RCU*

3. L'optimisation du rapprochement des données multi-sources consiste à trouver le meilleur compromis entre performance et qualité - taux de rapprochés à tort (*overmatch*) et taux de ratés (*undermatch*).

6.3. Partie théorique

Avec l'avènement du traitement distribué et l'utilisation abondante des services Web inter et intra organisationnels alimentée par la disponibilité des connexions réseaux à faibles coûts, les données multi-sources partagées ont de plus en plus envahi les systèmes d'informations. Ceci a induit, dans un premier temps, le changement de leurs architectures qui sont passées de centralisées à distribuées en passant par des architectures de systèmes coopératives et fédérées. Puis, dans un deuxième temps, cela a entrainé une panoplie de problèmes d'exploitation allant du traitement des incohérences des données doubles à la synchronisation des données distribuées avec toutes les opérations de nettoyage de bases et de gestion de l'incertitude (causée par le manque d'information sur la provenance de certaines données) que ces tâches impliquent.

Dans ce contexte, plusieurs travaux de recherche ont été entrepris depuis la fin des années quatre vingt permettant le développement de mécanismes de négociation pour la gestion des incohérences au sein des systèmes coopératifs ainsi qu'une panoplie de stratégies d'intégration et de fusion des données.

Dans cette section, nous nous proposons de présenter les architectures et les stratégies de gestion des données multi-sources qui ont été définies par la littérature durant ces trente dernières années, leur objectif commun étant de trouver un mécanisme de communication fiable entre les sources en question permettant de mieux interpréter les informations qui y circulent. Les solutions proposées varient alors de l'architecture de médiation à l'architecture de fédération (avec toutes leurs variantes) et les domaines impliqués relèvent autant des bases de données que des statistiques et des techniques d'étude de la provenance.

Dans notre étude, nous classifions les architectures de gestion des données multi-sources en deux groupes distincts : ceux qui visent une intégration logique préservant l'autonomie des sources locales moyennant des vues locales ou globales, et ceux qui aspirent à une intégration physique permettant la construction d'une base centralisée complète et cohérente *via* différentes stratégies de fusion des données. Notons que l'intégration logique, qui aboutit à la conception d'un modèle d'intégration, est parfois utilisée comme prémisse à l'étape d'intégration physique, le flux général étant inscrit dans le cadre d'une mise en place d'un processus d'intégration des données multi-sources (figure 6.2).

Nous soulignerons, par la suite, le rôle de la qualité des données dans la mise en place de ces architectures d'intégration et nous finirons par une discussion sur l'applicabilité de ces solutions dans le monde pratique avant d'introduire l'étude de cas.

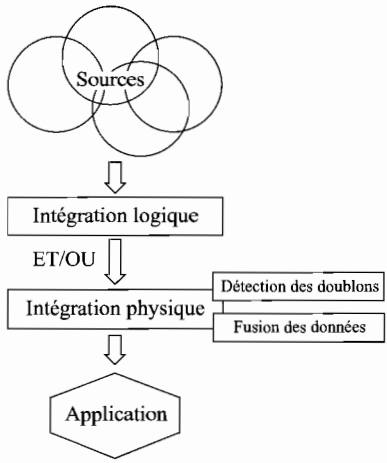

Figure 6.2. *Processus général d'intégration des données [BLE 08]*

6.3.1. *L'approche logique*

6.3.1.1. *Description générale*

L'approche logique conserve l'aspect distribué et autonome des sources de la base et communique avec l'utilisateur (ou plus généralement la couche applicative du modèle) *via* des vues globales ou locales. Le modèle d'intégration type utilisé dans cette approche est le modèle de médiation décrit dans la figure 6.3. Nous verrons dans cette section qu'il existe, cependant, des différences entre les architectures proposées dans le cadre de l'intégration logique des données.

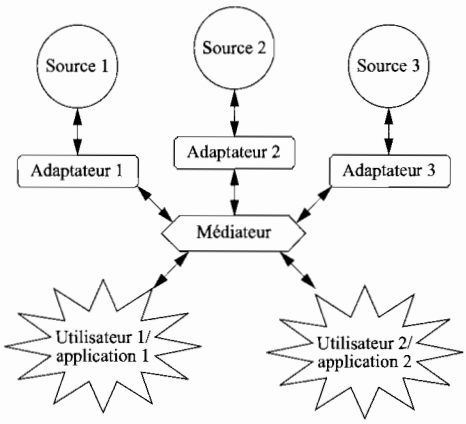

Figure 6.3. *Schéma simplifié d'une architecture de médiation [HAC 04]*

Au sein d'une telle architecture, le lien entre le schéma global de la base et le schéma local au niveau de chacune des sources s'exprime à travers des vues locales ou globales. En pratique, cette architecture est mise en place sous la forme de :

– bases de données fédérées, où l'on distingue entre les systèmes d'information faiblement couplés, les bases de données fortement couplées et les systèmes d'information à base de médiateur. L'analyse de ces différentes variantes, élaborée dans le paragraphe suivant, est basée sur les travaux de [BUS 99] qui comportent une étude détaillée des systèmes fédérés et de leurs différentes variantes ;

– systèmes coopératifs qui se définissent comme des systèmes organisationnels différents, interconnectés, autonomes, géographiquement répartis, interopérables et partageant des objectifs communs [MEC 02]. Du point de vue technique, alors que les systèmes fédérés sont gérés par un langage de requête adapté, les systèmes coopératifs nécessitent des programmes sophistiqués de communication appelés agents informatiques ;

– bases de données interopérables où les utilisateurs requêtent les données sans se soucier de leur emplacement grâce à un ensemble de métadonnées décrivant les différentes sources respectives. Ces métadonnées décrivent la représentation, l'échelle (principalement dans le cadre des données géographiques) ainsi que le niveau de qualité des données [ZHA 03] ;

– bases de données distribuées où les bases de données sont interconnectées *via* un réseau de communication, notamment une architecture en client/serveur.

Toutes ces architectures sont plus ou moins semblables (puisqu'elles réfèrent à des systèmes distribués), elles diffèrent cependant dans la façon avec laquelle les « intégrateurs » construisent leur schéma global d'interfaçage entre les sources et la couche applicative [TAR 98]. Dans ce qui suit, nous détaillons le principe des bases de données fédérées et coopératives en surlignant les différences qui existent entre elles.

L'avantage de ces architectures réside dans la conservation de l'indépendance et l'autonomie des différentes bases locales. En revanche, ces systèmes perdent en transparence au niveau de l'emplacement et du schéma puisque les requêtes doivent spécifier à la fois la source interrogée et l'entité concernée. Aussi, dans ce genre d'architecture, l'utilisateur est responsable de l'intégration des données et doit, de ce fait, gérer tous les problèmes de conflits que cette tâche implique.

Les bases de données fédérées fortement couplées agissent comme des bases de données classiques gérant les accès en lecture/écriture. Les composants de cette base sont ainsi définis comme des sources structurées auxquelles on accède *via* des requêtes et communiquent *via* des schémas d'export et des schémas de

fédération (voir figure 6.4). Ce genre d'architecture offre une transparence au niveau de l'accès et de la localité, contrairement aux bases de données faiblement couplées, mais, perdent en autonomie et en flexibilité d'évolution étant données les contraintes relatives à la signalisation de tout changement de structures ainsi que les contraintes transactionnelles relatives à la gestion des accès multiples concurrents.

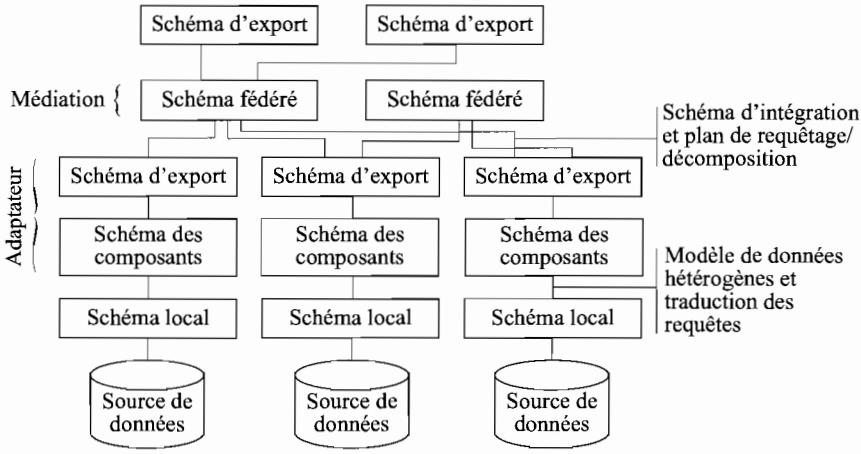

Figure 6.4. *Architecture des bases de données fédérées [BUS 99]*

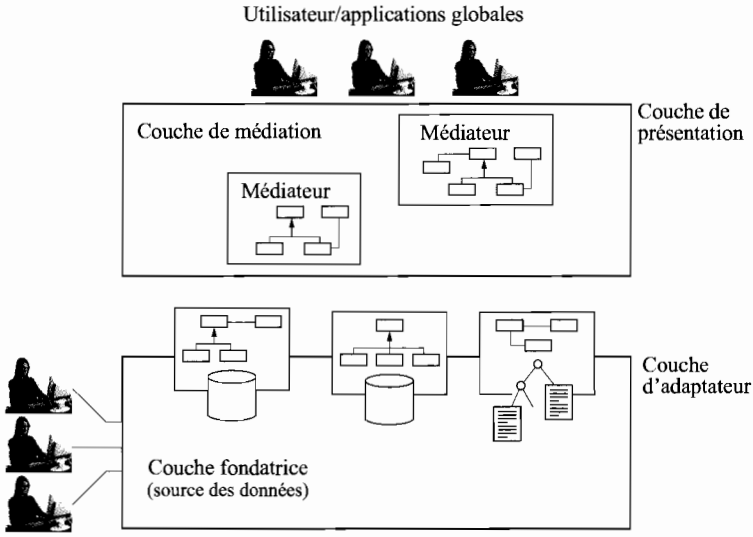

Figure 6.5. *Architecture d'un système d'information à base de médiateur [BUS 99]*

Les systèmes d'information à base de médiateur (voir figure 6.5) sont introduits par Wiederhold en 1993 [WIE 93]. Un médiateur gère l'accès entre l'utilisateur et les sources locales et ils se distinguent des autres systèmes fédérés par la façon avec laquelle le schéma global est construit. Ainsi, il est construit de manière descendante (conformément au schéma *top-down*) dans les systèmes d'information à base de médiateur alors qu'il est construit de manière ascendante (*bottom-up*) dans les bases de données fédérées, par exemple. L'architecture *top-down* permet ainsi aux utilisateurs d'avoir un accès aux données en fonction des informations dont ils ont besoin. C'est pour cela que le médiateur est assimilé, dans ce genre d'architecture, à un service avec tous les avantages que ceci implique, à savoir : la flexibilité d'évolution, la réutilisabilité et la facilité relative de gestion en comparaison aux bases de données fédérées (car ne nécessitant pas un schéma global figé complet et minimaliste).

Par ailleurs, les systèmes coopératifs se présentent comme des systèmes dont la communication nécessite des programmes sophistiqués basés sur la notion de services appelés agents informatiques formant la couche coopérative. Ces systèmes partagent, alors, des services plutôt que des données [MEC 02, TAR 98]. Les agents en question assurent une autonomie flexible au niveau des sources ainsi que la sécurité requise et les mécanismes appropriés pour la compréhension sémantique des processus et des services offerts par les sources locales. Dans ce contexte, [TAR 98] proposent une architecture à base d'agents pour la gestion de ce genre de système. Une telle architecture est représentée dans la figure 6.6. Cette architecture est composée :

– d'agents de coordination pour identifier et répartir les requêtes provenant de la couche applicative aux agents d'exécution adéquats ;

– d'agents spécialisés qui utilisent les agents des bases de données pour leur fournir les informations requises pour exécuter les requêtes des utilisateurs ;

– d'agent d'encapsulation qui se trouvent au niveau des sources locales et offrent un environnement coopératif de partage d'information et de gestion de la sécurité des services avec différents niveaux d'autonomie.

Outre les bases fédérées et les bases coopératives, de nouvelles formes de bases de données multi-sources ont été proposées. Ainsi, en 2008, les bases de données fédérées distribuées et dynamiques ont vu le jour [BEN 08]. Cette catégorie de bases définit un ensemble d'opérations ad hoc permettant un accès efficace aux sources de données distribuées lorsque les applications nécessitant les données en question ignorent leur localisation dans le réseau. Concrètement, cet accès ad hoc est réalisé grâce au mécanisme *« stocke localement, interroge partout »* (« store locally, query anywhere ») qui permet un accès global aux données quelle que soit la source émettrice de la requête dans le réseau. En effet, les données sont stockées dans des

tables locales situées au niveau de chaque source. Ces tables sont accessibles depuis n'importe quelle autre source du réseau *via* des requêtes de type SQL (*structured query language*) ainsi que des processus distribués stockés sous forme de procédures. La requête se propage à travers le réseau et retourne le résultat à la source émettrice.

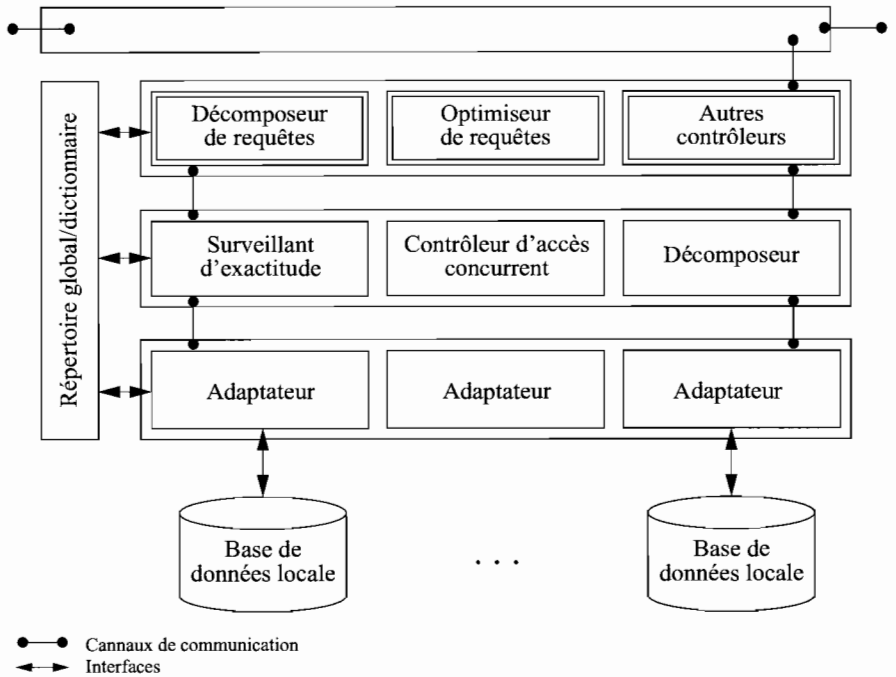

Figure 6.6. *Architecture de coopération proposée par [TAR 98]*

Une autre forme de bases de données multi-sources se décrit par les bases de données pair à pair (*peer-to-peer*) caractérisées par la préservation de l'autonomie et l'hétérogénéité des différentes sources. En effet, ces sources utilisent des schémas de mise en correspondance (*mapping*) et de médiation différents au lieu d'un unique schéma global pour communiquer entre elles [CAL 05].

Ainsi, nous avons défini, tout au long de cette section, le principe des architectures distribuées les plus rencontrées dans la littérature. Nous avons montré qu'elles diffèrent uniquement par la manière avec laquelle leurs schémas d'intégration sont définis. Nous pouvons distinguer généralement deux techniques d'intégration qui sont détaillées, ainsi que leurs variantes, dans ce qui suit.

6.3.1.2. *Les techniques d'intégration logique*

Les méthodes d'intégration par modèle de médiation sont généralement mises en place par deux techniques :

– l'utilisation de vues locales sur le schéma global où l'on fixe le schéma global et l'on décrit les sources par rapport à ce schéma. Cette méthode est connue sous le nom anglophone de *local-as-view* (LAV) et est plutôt descendante (part du schéma global et descend vers les sources) ;

– l'utilisation de vues globales exprimée en fonction des schémas locaux des différentes sources. Cette approche connue sous le nom de *global-as-view* (GAV) est plutôt ascendante puisqu'on part des sources pour produire le schéma global.

Ainsi, dans la méthode LAV, chaque source propose la vue avec laquelle elle communique avec la couche de médiation, alors que dans la méthode GAV, le schéma de médiation propose une vue globale à l'ensemble des sources que ces dernières utilisent pour communiquer avec la couche applicative.

Chacune de ces méthodes a ses avantages et ses inconvénients. Ainsi, l'ajout des sources se fait facilement dans une architecture LAV n'impliquant que le rajout de la vue locale de la source en question ; en revanche, la transformation des requêtes de la couche applicative est beaucoup plus fastidieuse en l'absence d'un référencement global. Du point de vue des techniques GAV, la réécriture des requêtes des utilisateurs est assez intuitive et simple alors que le rajout d'une nouvelle source remet en question le schéma global du médiateur et nécessite sa mise à jour.

De nouveaux schémas d'intégration profitant des avantages des techniques LAV et GAV ont été proposés par la suite. Ainsi, la méthode *both-as-view* (BAV) se base sur un schéma d'intégration hybride en utilisant un schéma de transformation de séquences réversible. Dans cette architecture, les relations sémantiques entre les sources des données ainsi que les relations spécifiées au sein du schéma de médiation sont exprimées à travers un langage logique [BRI 03].

Toujours dans le cadre des modèles hybrides, Rizopoulos propose en 2010 un schéma de *mapping* (mise en correspondance, association) qui prend en considération l'incertitude au niveau de la compatibilité de l'association ainsi que l'incertitude au niveau de l'association sémantique moyennant un modèle de données sous forme d'hypergraphe (*HDM : Hypergraph Data Model*). Dans ce modèle, l'incertitude est gérée moyennant des scores de classification qui classent l'ensemble des schémas de *mapping* possibles entre chaque paire d'objets selon leur probabilité de vraisemblance. De cette manière, les paires les plus probables sont les premières explorées dans l'objectif de fournir des réponses correctes aux requêtes des utilisateurs (couche applicative) [RIZ 10].

6.3.2. *L'approche physique*

6.3.2.1. *Description de l'approche*

La différence entre les systèmes d'intégration physique et les systèmes d'intégration logique réside dans le fait que les données rencontrées dans les systèmes d'exploitation sont matérialisées, dans le cadre de l'approche physique, dans un entrepôt de données ou une base de données centralisée, contrairement à l'approche logique où les données sont distribuées conservant leur aspect multi-sources.

L'architecture générale d'une approche d'intégration physique est résumée dans la figure 6.7.

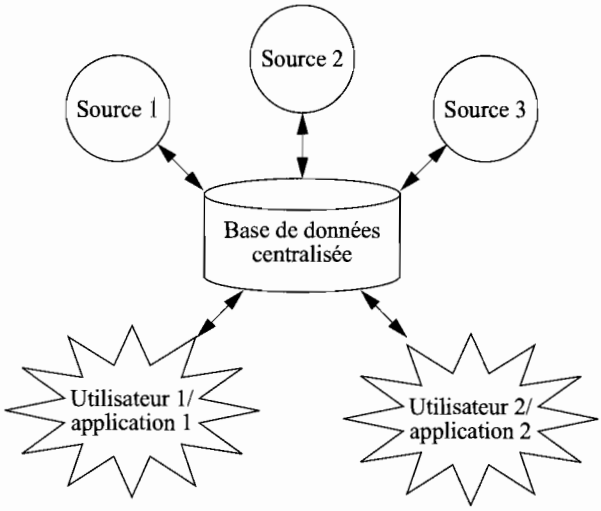

Figure 6.7. *Intégration physique*

Avant de spécifier les différents aspects fonctionnels et techniques de l'approche physique de l'intégration des données, commençons par spécifier la différence entre fusion des données et intégration des données. En effet, Naumann et Bleiholder [BLE 98] définissent la fusion des données comme une étape intervenant dans le processus général de l'intégration physique qui intervient après le *mapping* des schémas et la détection des doublons. Ainsi, l'intégration des données est un processus basé sur les étapes suivantes :

– *mise en forme des données*, principalement le formatage des résultats des requêtes des utilisateurs en un schéma unique. Deux approches sont alors utilisées : intégration des schémas et *mapping* (mise en correspondance) des schémas.

L'intégration des schémas nécessite une bonne connaissance des sources de données et se réalise par l'analyse de l'ensemble des schémas locaux puis par la génération, dans la mesure du possible, d'un nouveau schéma qui soit à la fois complet et minimal et, correct et compréhensible, tout en respectant les schémas initiaux des sources. Le *mapping* des schémas consiste en la mise en correspondance d'un ensemble donné de schémas locaux étant donné un schéma cible fixé. Ces techniques sont utilisées, par ailleurs, lors de l'intégration par l'approche logique GAV. La différence entre elles consiste dans le fait que le *mapping* se fait par rapport à un schéma cible qui n'est pas l'union de tous les schémas locaux de la base, ce qui est le cas dans l'intégration ;

– *détection des doublons* ou encore couplage des données (*record linkage*), réconciliation des références (*reference reconciliation*), etc., où l'objectif est d'identifier les représentations multiples du même objet réel, étape primordiale prémisse de la fusion des données. Cette tâche est requise aussi dans l'intégration logique, notamment les approches LAV and GAV ;

– *intégration des données*, où l'objectif est d'avoir un schéma à la fois complet et concis. Ainsi, dans [BLE 08], l'intégration des données vise généralement à atteindre deux objectifs : *améliorer la complétude* des données dans l'ensemble du système d'information en assurant, en même temps, la *minimalité* du schéma global sous jacent. L'amélioration de la complétude se fait en rajoutant, au niveau du système central, de nouvelles sources (de nouveaux objets, de nouveaux attributs décrivant les objets) ; la *minimalité*, quant à elle, se fait en éliminant les données redondantes, en fusionnant les doublons et en groupant les attributs semblables dans un attribut unique et global.

Nous nous intéressons principalement à l'étape de mise en correspondance des données et nous détaillons, dans ce qui suit, les différentes stratégies et techniques proposées par la littérature pour gérer les incohérences sous-jacentes.

6.3.2.2. *Les techniques d'intégration physique : gestion des conflits dans le cadre de la fusion*

En 2001, Wang, Ziad et Lee définissent un modèle de gestion des conflits, induit par la fusion des données hétérogènes, basé sur l'étude de la provenance des données. Ce modèle est considéré parmi les premiers modèles offrant une solution à base de métadonnées pour la gestion de l'hétérogénéité des systèmes multi-sources. Il analyse la crédibilité des données en se basant sur deux métadonnées principales : la connaissance des sources de données. La résolution de ces deux marquages permet d'avoir une idée sur la crédibilité des données utilisée comme pilier de la détermination de l'exactitude des données conflictuelles [WAN 01].

Toujours dans le contexte d'étude de la provenance, [DON 09a] propose un modèle probabiliste de « *découverte de vérité* » (*truth discovery*) permettant de déterminer la source la plus fiable et donc, de résoudre les conflits des données multi-sources étant donné un ensemble de sources Web. Ce modèle analyse la confiance que l'on peut accorder aux données intégrées en évaluant, d'une part, l'exactitude de leurs sources ainsi que les dépendances qui peuvent exister entre ces différentes sources, et utilisant, d'autre part, des modèles Bayésiens afin d'estimer la probabilité d'exactitude des données. Par ailleurs, l'analyse des dépendances entre les sources prend tout son sens lorsque deux sources partagent de fausses informations. Des modèles Bayésiens sont utilisés pour estimer la probabilité que deux sources soient dépendantes. Ensuite, étant donnée l'information sur la dépendance, la résolution des conflits se définit en privilégiant la source la plus correcte (délivrant un maximum d'informations correctes). L'expérimentation de cette approche sur des données réelles montre sa robustesse notamment pour la détection des « *copiages indirects* » (*indirect copying*).

Récemment, avec l'avènement des réseaux sociaux et leur utilisation pour l'enrichissement des données clients et prospects dans le domaine du marketing par exemple, des problèmes d'efficacité au niveau des campagnes marketing ont été détectés poussant de plus en plus les chercheurs à investiguer dans le domaine de la résolution de conflits. Ainsi, [TAL 11] propose un modèle d'analyse d'associations basé sur la méthode de provenance des données. Ces associations sont modélisées sous forme d'un réseau de graphes dans lequel les références sont représentées par des nœuds et où les arcs définissent les associations entre ces références.

Outre l'étude de la provenance, des modèles de *mapping* (mise en correspondance) ont été définis. Ainsi, en 2006, Kolovos, Paige et Polack définissent un modèle basé sur le langage EML (*Epsilon Merging Language*) [KOL 06]. Il s'agit d'un modèle à base de règles établissant la mise en correspondance selon un processus à quatre étapes :

1. la comparaison des composants : basée sur le langage de comparaison des modèles (ECL pour *epsilon comparison language* utilisant des règles de correspondance (*matching rules*) où chaque règle compare les paires d'instances des composants du modèle et renvoie l'information sur leur conformité ;

2. l'étude de conformité qui examine les éléments identifiés comme « *correspondants* » lors de la première étape en identifiant le potentiel des conflits ;

3. la mise en correspondance *(mapping)* où chaque règle de mise en correspondance définit les éléments qui peuvent s'associer ainsi que l'incidence de cette association dans le modèle résultant. De plus, des règles de transformation définissent les types d'instances transformables ainsi que son incidence dans le modèle résultat. Ceci s'effectue grâce à deux langages : le langage de transformation

des modèles (ETL pour *epsilon model transformation language*) et le langage de transformation des textes (EGL) ;

4. la réconciliation : qui consiste à résoudre les conflits éventuels induits par le *mapping*.

Ce langage EML se distingue par son extensibilité, sa possibilité d'enrichissement et surtout par son approche sémantique de gestion des conflits laquelle utilise des règles de comparaison, des règles de transformation, des règles de *mapping* et des règles d'étude de conformité. De plus, ce langage permet de gérer l'hétérogénéité des modèles au sein d'un même métamodèle, ce qui n'est pas le cas des autres approches proposées dans ce même contexte. Cependant, ce modèle a l'inconvénient de solliciter l'utilisateur pour la prise de décision (afin de minimiser sa complexité), solution plutôt coûteuse et surtout approximative dans la pratique dans le sens où on est confronté au *mapping* de centaines entités dont certaines lui seraient peu connues faisant partie du système *patrimonial* de l'entreprise (*legacy system*).

Ainsi, afin d'automatiser la résolution des conflits, plusieurs approches ont été proposées. Bleiholder et Naumann [BLE 08] les regroupent dans les catégories suivantes :

– les stratégies d'ignorance des conflits (*conflict ignorance*) où aucune décision n'est prise. Dans cette catégorie, nous citons l'approche « *ignorer* » (*pass it on*) où la décision est laissée à l'utilisateur. Aussi, nous citons l'approche « *considérer toutes les possibilités* » (*consider all possibilities*) où une liste énumérant les différentes éventualités est établie et fournie à l'utilisateur pour qu'il prenne sa décision ;

– les stratégies d'évitement des conflits (*conflict avoidance*) où l'on distingue principalement deux approches :

- l'approche à base d'instances où aucune décision n'est prise. Cette approche est basée sur le principe « *considérer l'information* » (*take information*) qui prend en compte l'ensemble des informations disponibles en filtrant les valeurs nulles. Aussi, nous citons l'utilisation du principe « *ne pas déformer l'information* » (*no gossiping*) qui considère uniquement les valeurs cohérentes et plausibles ;

- l'approche à base de métadonnées utilisant le principe « *fait confiance à tes amis* » (*trust your friends*) où les données d'une source sont privilégiées étant donnés des critères tels que la fiabilité, le volume des données fournies et d'autres critères qualité ;

– les stratégies de résolution des conflits et d'intégration des informations où nous citons par exemple :

- la stratégie décisionnelle résolvant les conflits en étudiant la provenance des données ;

- la stratégie de médiation utilisant les algorithmes de compromis et d'autres algorithmes privilégiant la récence des données. Dans ce genre de stratégie, la donnée imputée au système intégré résultant peut être différente des données sujet du conflit si la stratégie utilisée est « *choisir le médian* » (*meet in the middle*). En effet, dans le cas où les données sont relatives à l'attribut « nombre des employés » et que les propositions conflictuelles sont 30 et 40, le résultat de l'étape de résolution des conflits utilisant cette stratégie est de 35. Ceci dit, cette méthode ne doit pas être utilisée quelles que soient les valeurs, par exemple une valeur de 54 résolvant les conflits de l'attribut « Age » 9 et 99 ne peut pas être satisfaisante ;

- une autre stratégie utilisée est la stratégie « *à jour* » (*keep up-to-date*) qui préconise la valeur la plus récente.

Toujours dans le cadre de la gestion des conflits par l'étude de la qualité des données, des approches se basent sur l'étude de l'exactitude des données où l'exactitude est évaluée par extrapolation à partir d'un échantillon d'analyse [TAL 11] ou par des méthodes plus sophistiquées utilisant les modèles probabilistes [DON 09b].

D'autres approches utilisent une méthode de *scoring* privilégiant une source (un site) sur une autre étant donné le nombre de visites impliquant l'ordre d'apparition dans le navigateur [WU 11] et ce dans le contexte de la gestion des conflits des données non-structurées (notamment les données Web).

6.3.3. *La qualité des données : outil de gestion des données multi-sources*

Nous avons vu dans les sections précédentes (voir 6.3.1 et 6.3.2) les stratégies proposées par la littérature pour l'intégration des données multi-sources ainsi que certaines techniques de gestion des conflits que cette intégration implique. Nous remarquons que la plupart de ces techniques (notamment celles basées sur la provenance) utilisent les méthodes basées sur la qualité des données principalement pour analyser la qualité des sources. Ainsi les dimensions de crédibilité, d'exactitude et de fiabilité sont les principales utilisées. Par exemple, [CHO 04] utilise un processus d'évaluation basé sur le modèle STANAG 2022 (standardization agreements 2022) qui analyse la valeur du couple (*fiabilité de la source* ; *crédibilité de la source*) déduites des utilisations et des sollicitations antérieures de la source en question. La fiabilité de la source se définit ainsi par le degré de confiance qu'on peut lui attribuer étant donné son historique d'utilisation, et la crédibilité se définit par l'appréciation générale de cette source recueillie auprès d'autres utilisateurs. Cette approche, bien que pertinente (car prenant en compte l'historique), pénalise, à cause de l'importance de cet historique, les nouvelles sources qui sont considérées comme non fiables et non crédibles.

Par ailleurs, les méthodes de fusion basées sur l'étude des instances s'intéressent à l'évaluation de la qualité des données intrinsèques. Ainsi on privilégie les données récentes, cohérentes et syntaxiquement exactes.

La qualité des données intervient aussi dans l'évaluation des résultats des requêtes de médiation telles que l'approche proposée par [KOS 05] qui définit un outil capable d'évaluer la qualité des résultats produits par des requêtes alternatives générées pour un objet de médiation et de les confronter aux préférences des utilisateurs afin de délivrer des résultats adaptés à celles-ci. Les dimensions qualité choisies sont la fraîcheur, le délai et le coût.

Aussi, nous citons l'approche proposée par [BAT 07] qui définit, toujours dans le cadre de l'amélioration de la qualité des résultats dans les systèmes d'intégration des données, un ensemble de métriques permettant d'analyser :

– la complétude du schéma : c'est le pourcentage des concepts du domaine représentés dans le schéma d'intégration par rapport à l'ensemble des concepts représentés dans l'ensemble des différentes sources ;

– la minimalité (ou concision) du schéma : basée sur la redondance des entités du schéma et la redondance des relations ;

– la cohérence des types incluant la cohérence des types de données gérées dans le schéma, la cohérence des attributs du schéma et la cohérence de la représentation du type des données dans le schéma.

Par ailleurs, Batista et Salgado définissent un algorithme d'amélioration de la minimalité du schéma basé sur l'élimination de la redondance au niveau des entités, des relations et des attributs et proposent la solution *IQ manager* comme une concrétisation de cette solution en démontrant son efficacité par une expérimentation.

6.3.4. *Discussion*

Les sections précédentes décrivent les technologies et méthodologies proposées dans la littérature pour l'intégration des données disparates formant la base de données multi-sources. Ces approches sont diverses et variées et leur utilisation en pratique dépend de la stratégie d'intégration privilégiée par le métier. Généralement, le métier préfère se fier aux retours d'expérience et suit les bonnes pratiques publiées dans le contexte de la gestion des données multi-sources telles que les règles de gestion des incohérences et les règles de priorisation des sources.

Nous spécifions dans ce qui suit un cas pratique de gestion des données multi-sources, notamment, un cas pratique de fusion de données.

6.4. Gestion de données multi-sources : nos outils de paramétrage

Dans les années 2000, lorsque l'on parlait de fusion de données, quelques conclusions ont été mises en évidence :

– les données évoluent impliquant l'amélioration ou la dégradation de leur qualité. Ainsi, les règles métier utilisées pour la fusion/intégration des données doivent offrir la flexibilité nécessaire pour faire face à ce genre de changements ;

– la gestion de la relation client (customer relationship management ou CRM en anglais) est utilisé par les marketeurs dans le cadre des campagnes de marketing direct. Ces marketeurs n'ont pas la capacité de choisir les meilleures valeurs durant l'étape de ciblage : même si nous avons à notre disposition les différentes valeurs concurrentes, nous avons toujours besoin de calculer un « enregistrement consolidé » (« consolidated record », appelé encore enregistrement en or « golden record »[4], ou encore enregistrement survivant « survivor record »). Ainsi, étant donné un enregistrement A et un enregistrement B représentant deux états différents d'une même personne, un enregistrement C (que nous appelons « enregistrement consolidé ») est construit à partir des valeurs de A et B. Les instructions sont en général la combinaison des valeurs non nulles de A et B ; cela présuppose que la dimension qualité complétude est plus importante que la dimension exactitude, hypothèse que nous rejetons.

6.4.1. *Gestion des attributs*

Soit la matrice de données M, avec n individus et p colonnes (nom, prénom, date de naissance, Adr1, Adr2, Adr3, CP, Ville, etc.). Nous appelons, attribut, une colonne et l'ensemble des valeurs prises par les individus sur cette colonne ; ainsi sur la figure 6.8, un exemple d'attribut est *date de naissance.*

Nous distinguons deux types d'attributs :

– *les attributs indépendants* (date de naissance, par exemple) ;

4. Ce terme est plutôt utilisé dans le processus de gestion des données de référence, plus connu sous son nom anglophone master data management (MDM), terme créé par *Loshin* pour définir l'enregistrement créé suite à l'intégration des données multi-sources. Personnellement, nous n'approuvons pas le terme abusif de *« golden record »,* nous préférons le terme « enregistrement consolidé ».

– *les attributs dépendants ou liés :* par exemple, une adresse physique. Une adresse est composée de plusieurs attributs représentés en général, en France, dans une base de données par : Adr1, Adr2, Adr3, CP, ville. Ainsi, les attributs *Adr1* et *ville* sont dépendants dans le sens où la modification ou la validation de l'un remet en cause la modification ou la validation de l'autre.

	Nom	Prenom	Date de naissance	Adr1	Adr2	Adr3	CP	Ville
Individu 1	X	X	X	X	X	X	X	X
...	...	...	...	...	...	☐	☐	☐
...	...	...	...	...	☐	☐	☐	☐
Individu n	X	X	X	X	X	X	X	X

Figure 6.8. *Exemple de matrice des données*

Dans un contexte de fusion d'enregistrements, l'ensemble des attributs du bloc fonctionnel « adresse » est traité de manière non dissociable. La figure 6.9 en représente un exemple.

Enregistrement 1	Enregistrement 2	Fusion
Bâtiment 2	Rue Danton	Bâtiment 2
4 rue Henri le Sidaner	92500 Rueil Malmaison	4 rue Danton
78000 Versailles		92500 RUEIL MALMAISON

Figure 6.9. *Fusion des attributs dépendants*

6.4.2. *Règles de priorité*

Dans un monde parfait, la décision pour choisir une donnée étant donné des valeurs divergentes serait de pouvoir contrôler, mesurer l'exactitude et décider au cas par cas. Les méthodes proposées dans la littérature pour résoudre les conflits sont ainsi de :

– fournir les divergences à l'utilisateur pour qu'il tranche ;

– prendre la moyenne ou calculer une fonction d'agrégation (solution applicable aux données numériques) ;

– mesurer l'exactitude d'une source sur un échantillon et l'extrapoler à la source toute entière ;

– utiliser la date de la donnée pour privilégier la donnée la plus récente.

Quand nous avons créé les règles en 2000, nous avons privilégié les choix suivants :

– un processus complètement automatique, sans décisions dépendantes de l'utilisateur (car trop coûteux) ;

– une prise en compte de la notion « temps » et « obsolescence » des données ;

– une mesure d'exactitude ou plutôt de cohérence en automatique dès que possible.

Ces choix se sont traduits par la mise en place des paramètres de décision suivants :

– *source de la donnée :* fournisseur, canal de collecte (Web, formulaire de carte de fidélité, etc.) ;

– *date* : la date relative à la donnée qui constitue un élément souvent difficile à obtenir. La définition du paramètre date dépend de l'attribut que ce paramètre décrit :

- pour un effectif entreprise : la date à laquelle l'effectif est constaté, déclaré. Par exemple, dans le fichier SIRENE de l'INSEE, la valeur publiée en janvier 2011 correspond à une déclaration automatisée des données sociales (DADS) de décembre 2009, indiquée dans la variable DEFEN de la notice. La date dont nous parlons est décembre 2009 ;

- pour une adresse, un email, il s'agit de la date à laquelle la personne a fourni la donnée. Simple sur le principe, c'est en réalité une donnée souvent peu ou mal capturée dans les systèmes d'information : on trouve dans le meilleur des cas une date correspondant à l'accès à une fiche client en modification sans savoir quelle(s) rubrique(s) ont été modifiée(s) ou validée(s). Très souvent, la date peut correspondre à un simple accès en lecture à la fiche client. Cette donnée est souvent critique pour les attributs volatiles comme l'adresse ou l'effectif en opposition à des attributs comme la date de naissance ;

– *code de contrôle* : *control code,* ou mesure d'exactitude. Le terme de mesure d'exactitude est en réalité mal adapté ; il ne s'agit pas de la distance de la valeur à la valeur réelle (voir la dimension *exactitude* présentée plus haut) mais plutôt des résultats de contrôle de cohérence qui peuvent être faits. Par exemple, le code normalisation d'adresse est obtenu par la comparaison avec les référentiels postaux de l'adresse et l'indication du bon référencement de l'adresse ou non (ville, rue, numéro dans la rue selon la précision du référentiel). En revanche, ce code ne préjuge pas du fait que la personne habite bien à cette adresse. Un autre exemple concerne le contrôle de cohérence entre l'effectif de l'établissement et l'effectif de l'entreprise (l'établissement ayant pour siège cette même entreprise) ;

– *code d'usage :* il s'agit d'un code retour obtenu après l'utilisation d'une donnée, contrairement au code de contrôle qui est calculé *a priori*. Par exemple, pour une donnée comme l'email, on pourra distinguer :

- email utilisé et ouvert ;
- email utilisé et *hard bounce*[5] ;
- email utilisé sans retour (ni ouverture, ni *bounce*[6]) ;
- email non utilisé.

6.4.3. *Paramètres*

Les paramètres utilisés sont illustrés dans le tableau 6.1.

Bloc	Nom du bloc fonctionnel
Origine	Origine de la donnée
Code de traitement	Code de traitement
Obsolescence	Obsolescence de la donnée
Période lune de miel	Une période d'état de grâce de n jours suite à une mise à jour est considérée comme « lune de miel » : aucune autre donnée n'est plus prioritair
Date de péremption	Passés n jours, la donnée est considérée comme invalide
Score	Niveau de priorité : meilleur niveau à 0

Tableau 6.1. *Paramètres de calcul de priorité*

5. *Hard bounce* : un email est *hard bounce* lorsqu'il n'est pas délivré à son destinataire. Exemples de cause de *hard bounce* : adresse email incorrecte, nom de domaine inconnu, utilisateur inconnu. Les emails *hard bounce* ont un code SMTP de type 5.
6. Le terme *bounce* couvre à la fois le « *hard bounce* » et le « *soft bounce* ». Le *soft bounce*, contrairement au *hard bounce*, est un problème de *délivrabilité* temporaire d'un email à son destinataire.

6.4.4. *Illustration du paramétrage sur un attribut multi-source*[7]

Nous proposons d'illustrer le paramétrage sur un exemple. Soit trois sources de données :

– source A : un CRM, où l'information est principalement mise à jour par des commerciaux ;

– source B : un fournisseur de données que nous appellerons *ALPHA* pour des raisons de confidentialité ;

– source C : un autre fournisseur de données *BETA*.

L'objectif ici est de mettre en place des règles de priorité relatives à l'attribut « *nombre d'employés* » fourni par les différentes sources. La définition prise pour le nombre d'employés est le nombre de personnes en contrat indéterminé dans le site en fin d'année, donc au niveau *SIRET*. Pour décider, lorsque le fournisseur *ALPHA* donne une valeur 15 et le fournisseur *BETA* 200 quelle valeur retenir, donc comment paramétrer les règles de priorité, nous savons que :

– la donnée fournie par les commerciaux n'est pas nécessairement la plus fiable. Les valeurs ne sont pas nécessairement aberrantes ou obsolètes mais souvent ne respectent pas la définition : il s'agit du nombre d'employés au niveau entreprise (*SIREN*) et non pas site (*SIRET*), ou la valeur inclut les contrats à durée déterminée ;

– politiquement, il est très difficile de ne pas prendre en compte une valeur fournie par les commerciaux et ce, pour deux raisons : si un commercial fait l'effort de mettre à jour une donnée, nous devons avoir de bons arguments pour ne pas la prendre en compte. Par ailleurs, il s'agit d'une donnée stratégique : très souvent, les sociétés sont organisées en départements qui sont dédiés à des segments de clientèle, par exemple grands comptes ou petites entreprises. La segmentation du marché est en partie basée sur l'attribut nombre d'employés : affecter une valeur 10 plutôt que 9 peut déclencher un changement de département si la limite des petites entreprises est fixée à neuf salariés. Nous vous laissons imaginer les conséquences si un commercial indique 10 et les règles de priorité privilégient la source *BETA* à 9, sachant que la commission du commercial est liée en partie au nombre d'entreprises qu'il peut prospecter ;

– le nombre d'employés est une information fournie par le fournisseur *ALPHA* selon en fait deux origines distinctes (sans que le fournisseur n'indique ces origines) : d'une part, les *DADS*[8] faites par les entreprises en fin d'année, d'autre

7. Si l'obsolescence est calculée sur plusieurs attributs, il y aura autant de matrices de paramétrage que d'attributs ou de blocs fonctionnels d'attributs.
8. DADS : déclaration annuelle de données sociales.

part, des enquêtes directes auprès des entreprises. Si la première source est réputée fiable à la fois en termes de respect de la définition ainsi qu'en termes de contrôle de la donnée, on ne sait rien sur les données d'enquête et la manière dont les entreprises remplissent le questionnaire sachant que ces entreprises reçoivent le questionnaire parce qu'elles n'ont pas envoyé de *DADS*. Enfin, d'un point de vue date, le fournisseur *ALPHA* indique une date d'observation du nombre d'employés : cette date peut être ancienne, il est fréquent qu'on reçoive en 2011 les *DADS* de 2009. En revanche, cette date s'est avérée plausible dans les contrôles faits par distribution ;

– le fournisseur *BETA* a également plusieurs origines pour le nombre d'employés :

- le fournisseur *ALPHA* lui-même, qui est un fournisseur officiel ;

- la saisie des bilans publiés par les entreprises ;

- des enquêtes sur des sous-populations.

Comme *ALPHA*, *BETA* n'indique pas l'origine de la donnée qu'il fournit. Lui n'indique pas non plus la date de la donnée, contrairement à *ALPHA*.

Les règles de priorité mises en place sont résumées dans le tableau 6.2 et utilisent les paramètres suivants :

– code de traitement : le seul contrôle mis en place est la cohérence entre le nombre d'employés site et entreprise :

- si *nb_employes_site* > *nb_employes_entreprise*, alors *code_traitement* = KO, sinon OK. Si les valeurs sont cohérentes ou si la valeur entreprise est vide, alors le code de traitement est OK ;

– obsolescence : pour chaque source, nous considérons une obsolescence de trois mois, ce qui signifie que la source perd un point à chaque trimestre. Par exemple, si le code de traitement est correct, la source BETA « démarre » avec une priorité à trois et prend un point (plus le score de priorité est faible, plus la priorité est importante) chaque trimestre comme illustré sur le graphique figure 6.10.

Le fournisseur ALPHA est en exception avec une obsolescence de six mois. La raison est la suivante : le fournisseur ALPHA a souvent un délai de fourniture de la donnée de trois mois, c'est-à-dire qu'il va fournir en mars 2011 des données avec une date à décembre 2010. Hors, comme le fournisseur BETA n'indique pas de date pour ses données, par défaut il lui est attribué la date courante. Cette obsolescence de six mois pour ALPHA est donc un moyen technique de « compenser » et de permettre à ALPHA d'avoir par exemple en mars 2011 la même obsolescence que BETA ;

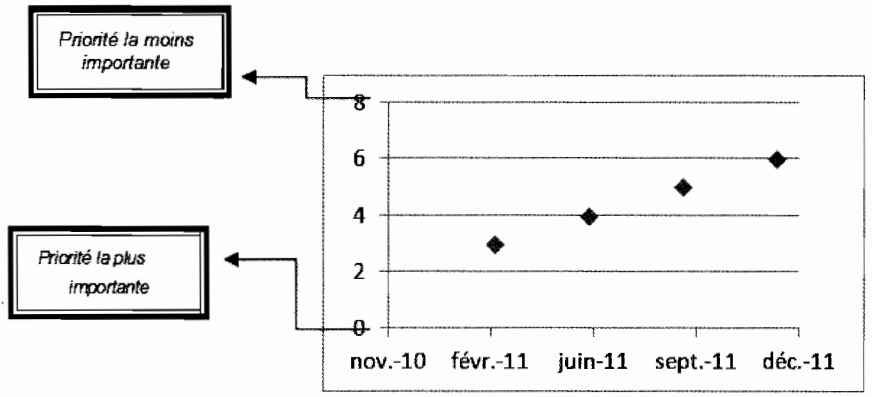

Figure 6.10. *Obsolescence source ALPHA*

– la période « lune de miel » : pour la source CRM, afin de s'assurer que la donnée fournie par le commercial est prioritaire quelles que soient les autres sources, une période de lune de miel est instaurée pendant six mois. La conséquence est que pendant les six mois consécutifs à la mise à jour de l'attribut « nombre d'employés » par le commercial, aucune autre source n'est prise en compte pour cet attribut. Au bout de cette période, la priorité standard est prise en compte. Ce fonctionnement est illustré par le graphique ci-dessous, correspondant à une source CRM en obsolescence d'un point en juin 2011, mise à jour en juillet 2011, en « lune de miel » de juillet à décembre 2011(figure 6.11) ;

– période de péremption : après 24 mois, la donnée devient périmée et la priorité est affectée à 99 (la plus basse) (tableau 6.2).

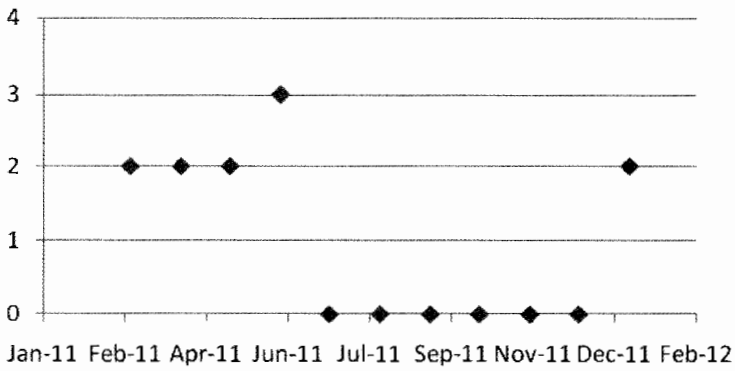

Figure 6.11. *« Lune de miel » source CRM*

Source	Code traitement	Obsolescence	Lune de miel	Péremption	Priorité
CRM	*	3	6	24	2
ALPHA	OK	6	-	24	1
ALPHA	KO	6	-	24	99
BETA	OK	3	-	24	3
BETA	KO	3	-	24	99

Tableau 6.2. *Règles de priorité*

Illustrons le comportement de la fusion sur un exemple. Soit une valeur fournie initialement par le fournisseur ALPHA : la *date indiquée* est de décembre 2010, le *code traitement* est *OK*, le *niveau de priorité* est noté à 1 selon les règles de priorité. La projection de la priorité sur 2011-2012, avec une obsolescence de six mois, est représentée sur la figure 6.12.

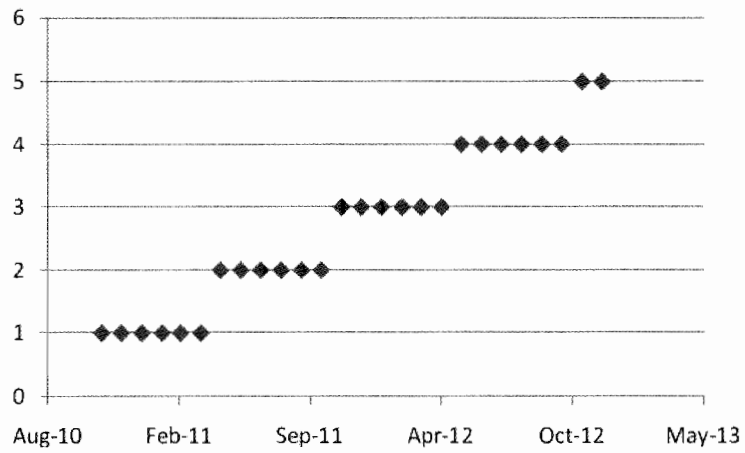

Figure 6.12. *Priorité V1 du CRM initiale (avant l'intégration du Flux1)*

Notons maintenant l'intégration de quatre flux :

– flux 1 : en mai 2011, le CRM fournit une autre donnée. La priorité du CRM est à 2, la priorité consolidée V1 est également à 2. Pour la même priorité, la date de

mise à jour la plus récente l'emporte, ce qui permet de sélectionner la donnée du CRM. La figure 6.13 fournit la version V2 de la consolidation. Notons la période lune de miel de mai à octobre 2011 où aucune donnée externe ne sera prise en compte ;

– flux 2 : en août 2011, le fournisseur *BETA* donne une autre donnée. Cette donnée est considérée comme une donnée datant du mois d'août, même si *BETA* ne fournit pas de date. Le niveau de priorité de *BETA* est à 3 comparée à 0 pour l'enregistrement consolidé : la donnée est ignorée, la version V3 de la projection de la priorité est identique à V2 (figure 6.14) ;

– flux 3 : en juin 2012, le fournisseur *BETA* vient avec une nouvelle donnée. Le niveau de priorité consolidé est à 4, la donnée *BETA* a un niveau de priorité à 3 (code traitement OK), cette dernière est prise en compte. D'où la nouvelle projection V4 sur la figure 6.15 ;

– flux 4 : en juillet 2012, le fournisseur *ALPHA* fournit une nouvelle donnée indiquée en date de décembre. Le niveau de priorité est calculé :

$$\begin{aligned} Priorité &= 1 + Ent\left(\frac{numbre\,de\,mois\,entre\,Décembre\ 2011\ et\ Juillet\ 2012}{\text{obsolescene}}\right) \\ &= 1 + Ent\left(\frac{7}{6}\right) = 1 + 1 = 2 \end{aligned}$$

A la même période, l'enregistrement consolidé a une priorité de 3, ce qui déclenche la prise de compte de *ALPHA*. La nouvelle projection de la priorité est donnée en figure 6.16.

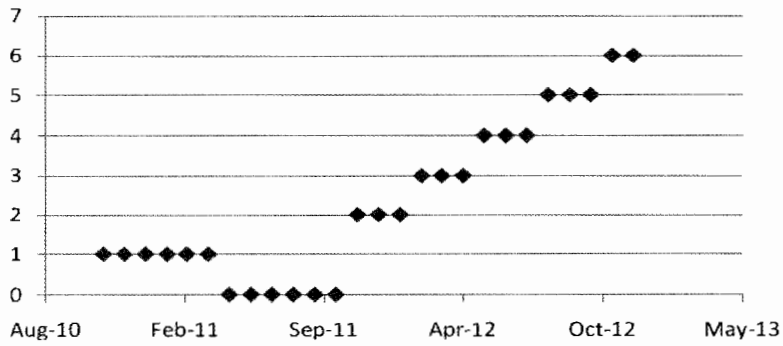

Figure 6.13. *Priorité V2 du CRM après Flux1 et avant intégration du Flux2*

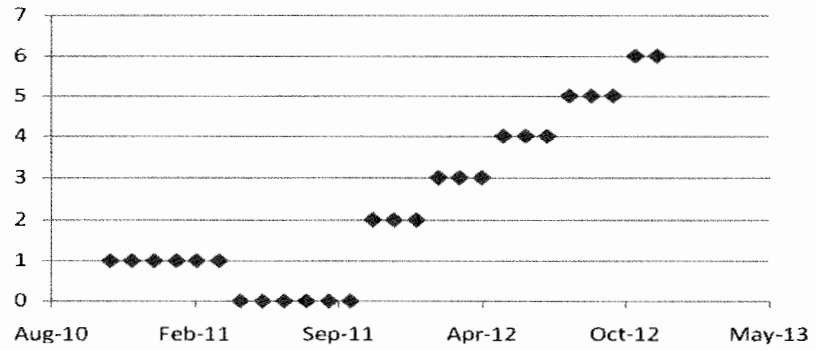

Figure 6.14. *Priorité V3 du CRM après Flux2 et avant Flux3*

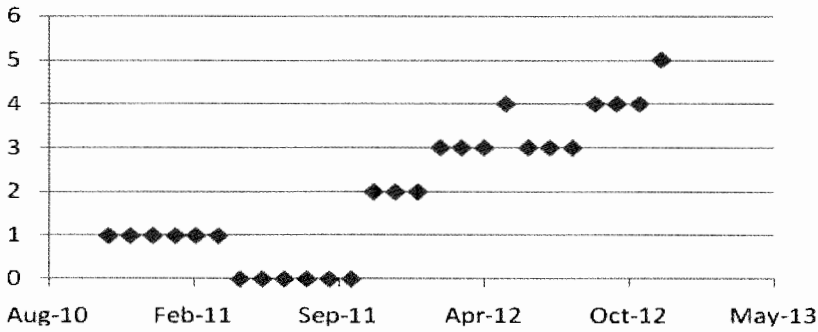

Figure 6.15. *Priorité V4 du CRM après Flux3 et avant Flux 4*

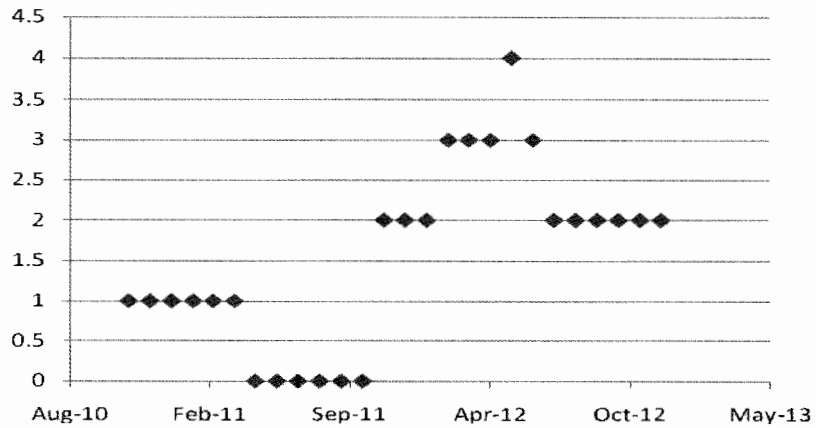

Figure 6.16. *Priorité V5 du CRM après Flux 4*

6.5. Etude de cas : la mise en place d'un référentiel client unique

6.5.1. *La démarche projet : comment décider des priorités ?*

Lors de la phase de mise en correspondance entre les données client des sources et le modèle du RCU cible, le chef de projet qualité de données définit et constitue les blocs logiques ou blocs fonctionnels de données qui, lors de l'étape de fusion aboutissant à la création de l'enregistrement client consolidé, seront indissociables et soumis sur l'ensemble des attributs qui le composent aux mêmes règles de priorité entre sources.

Chaque bloc contiendra le ou les indicateurs qualité pour les données de ces entités logiques appelés « code traitement », un indicateur de traçabilité de la source (d'où proviennent les données) appelé « origine » ainsi qu'un indicateur de fraîcheur appelé « date » indiquant la dernière date de mise à jour des données du bloc logique.

Le chef de projet qualité de données procède en général seul à une première proposition de choix des données qui vont constituer chacun des blocs. Il se base pour ce faire sur son étude des différentes sources de données initialisant puis alimentant le RCU. Une fois cette proposition réalisée, le chef de projet qualité de données la fait valider par les équipes métier.

La figure 6.17 est une illustration type des blocs de données constitués et permettant la gestion de la fusion des données client B2C[9] en un enregistrement client unique. Cette illustration s'articule sur trois niveaux : le niveau table, le niveau bloc et le niveau attributs composant le bloc.

Les blocs de données liées et les données autonomes ou indépendantes étant maintenant définis pour l'ensemble des attributs du RCU, le chef de projet qualité de données va pouvoir orchestrer la fusion entre les différentes sources en procédant au paramétrage des priorités.

Un premier paramétrage sera réalisé par le chef de projet qualité de données sur la base des résultats de l'audit des sources d'une part et l'analyse des processus de collecte et de maintenance des données dans chacune des sources d'autre part. Ainsi, si la date de naissance est collectée lors de l'ouverture d'une carte de fidélité sur présentation obligatoire de la carte d'identité, le chef de projet qualité de données donnera naturellement la priorité à cette source pour ce qui est du bloc « date de naissance ». Par ailleurs, si parmi les sources de données constituant le RCU se trouve une source de données Web, celle-ci primera sur les autres sources pour ce

9. B2C (pour *business to consumer*) : définit une relation commerciale vers des clients particuliers par opposition à B2B (pour *business to business*) qui définit une relation commerciale vers des entreprises.

qui est du bloc « email ». Enfin, si les équipes métier opèrent des traitements de RNVP[10], d'Estocade[11], de Charade[12] et d'Alliage[13] sur leurs adresses CRM afin d'assurer le meilleur acheminement des courriers promotionnels, le chef de projet qualité de données priorisera la source CRM pour le bloc « adresse ». Et ainsi de suite.

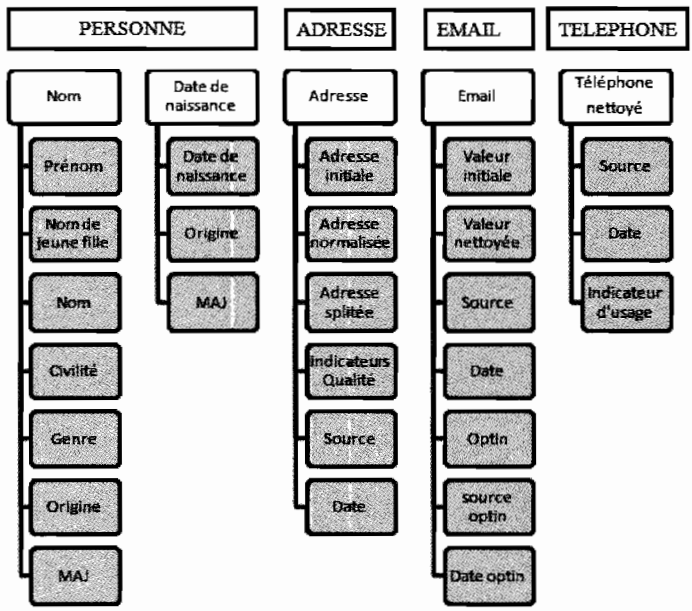

Figure 6.17. *Exemple de blocs logiques de données RCU B2C*

Une seule priorité par bloc et champ autonome peut s'avérer insuffisante pour départager les données. Il convient donc d'établir trois à quatre niveaux de priorité qui s'appliqueront en cascade pour parvenir à une décision non équivoque sur le bloc ou le champ à conserver.

10. RNVP : restructuration normalisation validation postale.
11. Estocade : fichier de La Poste permettant le repérage des adresses déménagées sur historique de 54 mois.
12. Charade : fichier de La Poste permettant l'achat des nouvelles adresses (en cas de déménagement).
13. Alliage : solution de La Poste pour une gestion dématérialisée des PND (pli non délivrable).

Par exemple, si la source CRM est prioritaire sur le bloc « nom » mais que les deux enregistrements client à fusionner proviennent tous de la source CRM, ce niveau de priorité ne suffira pas à la sélection. Le chef de projet qualité de données affectera dans sa matrice un niveau deux de priorité qui peut être l'indicateur qualité du bloc. Cependant, dans notre exemple, si les deux enregistrements sont équivalents sur le plan de la qualité, le chef de projet qualité de données aura besoin d'un niveau trois de priorité qui pourra être la date de mise à jour, et à défaut de décision possible un niveau ultime de priorité qui sera de conserver le bloc de données ou le champ du premier enregistrement restant dans le groupe.

Dans le cadre de la mise en place d'un référentiel client unique sur des données B2C, notre expérience nous permet de recommander de choisir les niveaux dans l'ordre qui suit :

1. indicateur qualité de données : la qualité de la donnée mesurée objectivement (contrôle de cohérence) prime sur toute autre priorité. Ainsi, une date de naissance non crédible (2090 par exemple), même issue d'une source bien réputée sur ce bloc de données, ne doit pas être prioritaire par rapport à une date de naissance crédible issue d'une autre source moins réputée sur ce bloc de données. A noter que certains blocs de données n'ont aucun indicateur qualité rattaché. Si tel est le cas, la matrice comprendra pour ce bloc trois niveaux seulement ;

2. source : la réputation de la source, évaluée de manière objective par l'audit initial de la qualité de données des sources, puis recoupée par l'étude des processus de collecte et de maintenance des données, garantit un certain niveau d'exactitude dans la prise de décision ;

3. date de dernière mise à jour : cette date de dernière mise à jour d'un attribut ou d'un bloc logique de données représente un indicateur de fraîcheur de la donnée, induisant ainsi un niveau de fiabilité du choix sur le mode « plus l'information est fraîche, plus elle est correcte ». Cependant, ces dates sont parfois des dates système qui se mettent à jour de façon technique (date modifiée lors de l'accès en mode lecture à une fiche client par exemple) sans que la donnée ait réellement été modifiée. De plus, les systèmes d'information aujourd'hui n'assurent pas un horodatage des attributs de façon suffisamment fine ; il n'existe bien souvent qu'un seul champ de date de mise à jour pour l'ensemble de la table. Pour ces deux raisons, le chef de projet qualité de données n'utilise pas la date de mise à jour comme priorité plus haute dans la hiérarchie des choix ;

4. premier enregistrement du groupe : cette priorité n'existe que pour permettre de trancher entre les champs ou blocs de données de plusieurs enregistrements à fusionner lorsque les priorités une à trois n'ont pas permis de conduire à une décision.

6.5.2. *Exemple pratique*

Soient la matrice de priorité et l'application de cette matrice sur les données (respectivement les tableaux 6.3 et 6.4).

Bloc	Origine	Priorité
Nom	B	1
	F	2
Date de naissance	B	1
	F	2

Tableau 6.3. *Matrice de priorité niveau source*

Nom	Nom de jeune fille	Prénom	Code civilité	Origine (source) du bloc nom	Date de MAJ du nom
dufort		frédérique	2	F	15/0703
DUFORT	BLANC	FREDERIQUE	2	B	25/02/06
DUFORT BLANC	BLANC	FREDERIQUE	2	B	09/11/06
DUFORT	BLANC	FREDERIQUE	2	B	09/11/06
BLANC DUFORT	BLANC	FREDERIQUE	2	B	09/11/06

Tableau 6.4. *Application de la matrice de priorité sur les données*

La grille d'exécution du processus de consolidation se déroule comme suit. Pour le bloc « NOM », c'est la source B qui est « gagnante » ; ici quatre enregistrements proviennent de cette source B donc il est impossible, sur la base de ce seul critère de la source, de déterminer l'enregistrement à conserver ; la matrice de choix regarde ensuite la date de mise à jour et applique une règle DATM (date de modification la plus récente), ici le 09/11/06 ; trois enregistrements sont encore en lice, la matrice applique alors une règle SYS, c'est-à-dire que le vainqueur est le premier des trois enregistrements restant dans le groupe.

Une fois la matrice de priorité pour chaque bloc de données et champs autonomes créée, le chef de projet qualité de données va exécuter la fusion en mode « bac à sable[14] » – ce bac à sable est constitué de l'ensemble des sources initialisant

14. Le bac à sable définit un pilote, un prototype, dans ce cas du référentiel client unique. Le bac à sable sera alimenté par les données multi-sources consolidées ; le métier marketing

le RCU, en volume total ou en échantillon représentatif. Le chef de projet qualité de données soumettra ensuite le résultat de cette fusion à la recette métier et illustrera les règles par des exemples concrets. A l'issue de cet exercice, la matrice de priorité de fusion sera validée et figée pour le mode initialisation. Elle pourra donc être spécifiée et paramétrée pour la production.

A noter que lors de l'implémentation d'un RCU, l'entreprise a tout intérêt à conserver un patrimoine exhaustif de coordonnées d'adressage – email, adresse, téléphone et mobile. Par conséquent, pour les coordonnées, la fusion et surtout les priorités associées au bloc « coordonnées » seront appliquées si une coordonnée est commune à plusieurs sources et qu'un choix doit s'opérer.

Les données de consentement marketing (*opt-in/opt-out*) suivront en règle générale la coordonnée choisie. Ainsi, si le chef de projet Qualité de Données donne priorité à la source F pour l'email, l'*opt-in* associé à l'email sera celui de la coordonnée choisie.

6.5.3. *Maintenance et mise à jour du RCU : l'impact sur la matrice de priorité*

Lorsque les sources de données vivent en coexistence avec le RCU, ce qui représente souvent une phase projet intermédiaire sur la voie de l'implémentation d'une solution complète de MDM, où les données de référence sont directement créées dans le référentiel puis redescendues vers les sources pour consommation, il est nécessaire de faire évoluer la matrice de priorité entre sources pour prendre en compte les spécificités et les contraintes du mode récurrent.

Dans ce mode de coexistence entre sources de données et RCU, deux options de mise à jour des données du RCU sont envisageables :

– option 1 : le RCU est uniquement un réceptacle qui reçoit les mises à jour et les créations et ne rediffusent pas ;

– option 2 : les requêtes se font sur le RCU ou une image déportée, le *datamart* local[15], pour des raisons de temps de réponse.

---

utilisera ces données pilote pour faire sa recette et aider à la spécification des indicateurs qualité de données pertinents, du bon positionnement des curseurs de rapprochement des enregistrements en paquets de doublons et enfin des règles de priorité pour la consolidation.

15. *Datamart* : un *datamart* est un sous-ensemble d'un entrepôt de données destiné à fournir des données aux utilisateurs, et souvent spécialisé vers un groupe ou un type d'affaire. Techniquement, c'est une base de données relationnelle utilisée en informatique décisionnelle et exploité en entreprise pour restituer des informations ciblées sur un métier spécifique, constituant pour ce dernier un ensemble d'indicateurs utilisés pour le pilotage de l'activité et l'aide à la décision.

Dans l'option 1, les modifications faites par la source F ne sont pas visibles par la source B. Par exemple, un téléphone modifié par F ne sera pas visible par B. Chaque source travaille « à l'aveugle », seul le RCU inclut les différentes mises à jour. Une des conséquences est que, dans le cas de doublons (plusieurs comptes Web de la même personne), l'information restera disparate sur les différents comptes (par exemple, dates de naissance non cohérentes). *Cette option n'est pas recommandée mais peut être transitoire dans la mise en place d'un RCU.* Dans ce cas, les mises à jour du RCU sont similaires en général à l'initialisation, à la profondeur d'historique près (et la variabilité de la qualité de la source).

Dans l'option 2, les mises à jour sont synchronisées, on travaille sur une copie synchronisée. Si la source F met à jour, la source B en bénéficie. Cela signifie également que si F envoie une valeur manquante pour l'attribut TEL (valeur NULL) par exemple, alors qu'il était rempli initialement, ce dernier a volontairement été annulé. Les règles de consolidation sont-elles dépendantes du système d'information mis en place ? Revoyons les deux options proposées dans [LOS 11] et illustrées dans la figure 6.18.

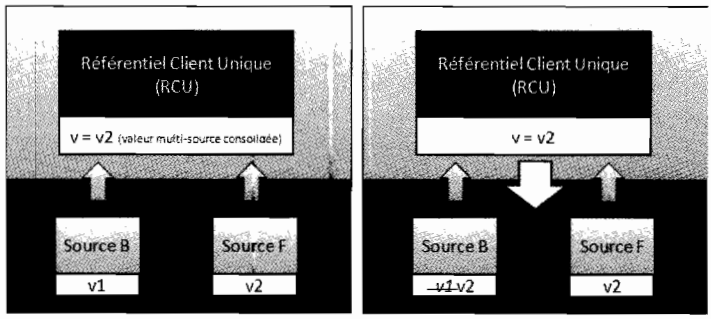

Figure 6.18. *Options de mise à jour des données du RCU*

Pour ce qui est des modalités de gestion, le RCU sera donc mis à jour par la gestion de blocs, de champs autonomes et de priorités. La mise à jour du RCU est traitée par processus de fusion.

Cette solution est dans la prolongation des spécifications faites pour l'initialisation. Elle permet de moduler les règles par groupe de champs, de garder un paramétrage évolutif. En revanche, elle nécessite d'avoir la main sur l'IHM pour :

- faire apparaitre les champs d'un même groupe de manière homogène ;
- gérer les dates de mise à jour, sources, attributs par blocs.

La responsabilité du chef de projet qualité de données sera de mettre en place puis d'affiner le paramétrage de la matrice de priorité pour le mode récurrent, cependant, il conservera les mêmes notions de blocs et de champs.

En mode récurrent, pour les priorités basées sur la source de la donnée, ce sera l'origine de l'enregistrement entrant, nouveau ou modifié, qui sera comparée à l'origine de chaque bloc fonctionnel de la personne existante dans le RCU et retrouvée par dé doublonnage ou jointure sur clé unique.

Le chef de projet qualité de données rajoutera à la matrice de l'initialisation des notions d'obsolescence ou de péremption – d'un délai de trois mois pour commencer. Ces notions serviront à éviter une situation de blocage des mises à jour du RCU, c'est-à-dire à éviter qu'une source, la source B par exemple pour le bloc « nom », soit indéfiniment prioritaire, quel que soit l'âge de ses données bloc « nom » et que le RCU ne bénéficie pas d'une mise à jour par la source F, même si l'information de la source F pour ce même bloc « nom » est plus fraîche que l'information de la source B.

De la même façon que pour l'initialisation, le chef de projet qualité de données simulera ces évolutions et fera valider les règles de priorités pour le mode récurrent par le métier sur la base de jeu de données test. Une fois ces évolutions validées, le chef de Projet pourra spécifier et paramétrer la maintenance.

Une évolution de la matrice de priorité est illustrée dans le tableau 6.5.

Bloc	**Origine (source de la donnée)**	**Code traitement**	**Priorité**	**Péremption**
Nom	B	Non parasite[16]	1	3
	F	Non B2B[17]	2	3
Date de naissance	B	Date nais- OK[18]	1	3
	F		2	3

Tableau 6.5. *Illustration de l'évolution de la matrice de priorités en mode récurrent*

16. L'indicateur « non parasite » pour le bloc nom se réfère dans notre exemple au résultat positif d'une vérification de non appartenance du nom à une liste de valeurs parasites ou injurieuses. Le nom n'est pas fictif.
17. L'indicateur « non B2B » pour le bloc nom se réfère dans notre exemple au résultat positif d'une vérification de non appartenance à une liste de noms de société. Le nom est bien celui d'une personne.
18. La vérification opérée sur la date de naissance consiste dans notre exemple à tester sa plausibilité et à écarter toutes les dates de naissance < 1910 et > 2000.

6.6. Résultats et conclusion

La plateforme de gestion de la fusion de données multi-sources, comprenant la notion de blocs fonctionnels et d'attributs indépendants, les attributs qualité, la matrice de priorité entre sources, est un outil très utile pour guider la démarche du chef de projet Qualité de Données lors de l'implémentation d'un RCU.

L'audit qualité de données des différentes sources composant le RCU ainsi que l'étude des processus de collecte et de maintenance des données client dans chacune des sources vont permettre au chef de projet qualité de données d'optimiser le paramétrage de ce processus pour garantir au mieux l'exactitude[19] et la pertinence des données composant l'enregistrement client unique.

De plus, limite déjà évoquée, les systèmes d'information ne disposent bien souvent pas de date de mise à jour sur l'ensemble des blocs de données ou des attributs indépendants, ces dates étant la plupart du temps au niveau de la table dans sa globalité. Par ailleurs, ces dates peuvent aussi être modifiées par des mises à jour techniques et non des mises à jour des données elles-mêmes. Ces deux facteurs fragilisent la règle de priorité basée sur la date et la traçabilité du module de gestion de la fusion des données multi-sources.

Enfin, le processus de fusion est assez sophistiqué ; dès lors, la compréhension et l'appropriation de l'ensemble de ses règles, paramètres et modalités par le client final (l'utilisateur) n'est pas toujours aisée. Aussi, il sera capital pour la réussite du projet de fusion des données que le chef de projet Qualité de Données propose aux équipes métier un premier niveau de réglage basé sur sa connaissance des sources (audit des données et étude des processus) et l'exécute sur un jeu de données test, représentatif des données composant le RCU. Il sera plus aisé pour les équipes métier, lors de cette phase de recette « données », de se projeter et de commenter pour affiner le paramétrage de ce processus.

Une fois le projet RCU sur les rails, les phases d'initialisation et de récurrence en production, le chef de projet qualité de données passera la main à l'équipe de gouvernance des données de l'entreprise. Cette équipe prendra en charge la surveillance (monitoring) régulière de la qualité des données à la fois des sources et du RCU et les modifications à apporter au paramétrage du processus de fusion des données multi-sources en cas de dérive ou d'ajout d'une nouvelle source dans le modèle du RCU.

---

19. L'exactitude des données de l'enregistrement multi-sources consolidé ne repose que sur des hypothèses qui découlent de la connaissance des processus métier de collecte de l'information, de la réputation des sources et des mesures objectives réalisées sur les données.

6.7. Bibliographie

[BAT 07] BATISTA M.C.M., SALGADO A.C., « Information quality measurement in data integration schemas », *Proccedings of the 33th conference of VLDB*, 2007.

[BEN 08] BENT G., DANTRESSANGLE P., VYVYAN D., MOWSHOWITZ A., MITSOU V., « A dynamic distributed federated database », *Second annual conference of ITA*, Imperial College, Londres, 2008.

[BLE 98] BLEIHOLDER J., NAUMANN F., « Data fusion and data quality », *The new techniques and technologies for statistics seminar (NTTS)*, Sorrento, Italie, 1998.

[BLE 08] BLEIHOLDER J., NAUMANN F., « Data fusion », *ACM computing surveys (CSUR)*, n° 41(1), 2008.

[BRI 03] MC BRIEN P., POULOVASSILIS A., « Data integration by bidirectional schema transformation rules », *19th International conference on data engineering*, p. 227-238, 2003.

[BUS 99] BUSSE S., KUTSCHE R.D., LESER U., WEBER H., « Federated information systems : concepts, terminology and architectures », *Technical report*, n° 99-9, TU Berlin, 1999.

[CAL 05] CALVANESE D., GIACOMO G.D., LEMBO D., LENZERINI M., ROSATI R., « Inconsistency tolerance in P2P data integration : an epistemic logic approach », *International conference on database programming languages (DBPL)*, 2005.

[CHO 04] CHOLVY L., « Information evaluation in fusion : a case study », *Information processing and management of uncertainty*, 2004.

[DON 09a] DONG X.L., BERTI-EQUILLE L., SRIVASTAVA D., « Integrating conflicting data : the role of source dependence », *VLDB endowment*, n° 2(1), 2009.

[DON 09b] DONG X.L., NAUMANN F., « Data Fusion : resolving data conflicts for integration », *VLDB endowment*, n° 2(2), 2009.

[HAC 04] HACID M.S., REYANUD C., « L'intégration de sources de données », *Revue information, interaction, intelligence*, 2004.

[KOL 06] KOLOVOS D.S., PAIGE R.F., POLACK F.A.C., « Merging models with the epsilon merging language (EML) », *Model driven engineering languages and systems*, n° 4199, p. 215-229, 2006.

[KOS 05] KOSTADINOV D., PERALTA V., SOUKANE A., XUE X., « Intégration de données hétérogènes basée sur la qualité », *Actes du 23ème congrès INFORSID*, Grenoble, France, 2005.

[LOS 11] LOSHIN D., « The practitioner's guide to data quality improvement », *Knowledge integrity Inc.*, Morgan Kaufmann Publisher, New York, 2011.

[MEC 02] MECELLA M., SCANNAPIECO M., VIRGILLITO A., BALDONI R., CATARCI T., BATINI C., « Managing data quality in cooperative information systems », *10th International conference on cooperative information systems*, 2002.

[RIZ 10] RIZOPOULOS N., Schema matching and schema merging based on uncertain semantic mappings, Thèse de doctorat, Imperial College London, 2010.

[TAL 11] TALBURT J.R., *Entity resolution and information quality*, Morgan Kaufmann Publishers, New York, 2011.

[TAR 98] TARI Z., ZALAVSKY A., SAVNIK I., « Supporting cooperative databases with distributed objects », *Parallel and distributed systems : theory and applications*, J.L. Aguilar Castro Publisher, 1998.

[WAN 01] WANG R.Y., ZIAD M., LEE Y.W., *Data quality*, Kluwer Academic Publishers, Norwell, MA, 2001.

[WIE 93] WIEDERHOLD G., « Intelligent integration of information », *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, p. 434-437, 1993.

[WU 11] WU M., MARIAN A., « A framework for corroborating answers from multiple web sources », *Information systems journal*, n° 36(2), p. 431-449, 2011.

[ZHA 03] ZHANG C., PENG Z.R., LI W., DAY M.J., « GML-based interoperable geographical databases », *Cartography*, vol. 32, n° 2, décembre 2003.

Chapitre 7

L'évaluation de la qualité d'un processus métier : enjeux, cas d'étude et bonnes pratiques

7.1. Introduction

La bonne gestion d'une entreprise passe par la connaissance, la compréhension et le meilleur alignement possible de ses processus métier sur les objectifs de l'entreprise. La gestion de ces processus est plus connue sous le terme de « *BPM* », *business process management*[1], et son intérêt est aujourd'hui bien reconnu de toutes les entreprises. L'étude des processus métier occupe donc une place très importante dans le domaine de l'étude des systèmes d'information.

Un processus métier est défini par T.H. Davenport et J.E. Short [DAV 90] comme étant un ensemble de tâches reliées logiquement et effectuées afin d'atteindre un objectif opérationnel. Des exemples classiques de processus métier sont les processus de développement d'un nouveau produit ou de prise en charge d'une commande client. Un ensemble de processus métier permet de représenter le fonctionnement (d'une partie) des activités d'une entreprise, faisant généralement intervenir plusieurs acteurs, internes ou externes à l'entreprise.

Processus et données sont étroitement liés puisque les processus métier exploitent des données de l'entreprise et produisent de nouvelles données. Des

---

Chapitre rédigé par Virginie THION-GOASDOUE et Samira SI-SAÏD CHERFI.

1. Traduit littéralement par la gestion des processus métier.

processus de mauvaise qualité engendrent la production de données de mauvaise qualité, et des données de mauvaise qualité peuvent engendrer un mauvais déroulement des processus métier. Gouverner ses données, implique donc également gérer ses processus. Les normes préconisant une approche orientée processus (par exemple [ISO 00]) recommandent d'ailleurs de régulièrement évaluer et améliorer la qualité des processus [MOR 07]. Dans cette vision, un processus métier peut être vu comme une unité d'analyse du (ou des) système(s) d'information de l'entreprise.

Mais qu'est-ce que la qualité d'un processus métier, et comment la mesurer ? Voici les questions auxquelles ce chapitre tente d'apporter quelques réponses.

Le chapitre est organisé comme suit. Dans la section 7.2, nous commençons par dresser un état de l'art de différentes méthodes et métriques existantes permettant de mener l'évaluation de la qualité d'un processus. Nous nous intéressons ensuite, dans la section 7.3, à un cas d'application réel : l'évaluation de la qualité du processus de changement de prestataire dans un projet informatique sous-traité, mis en œuvre dans l'un des services d'un grand EPST (établissement public à caractère scientifique et technologique) français. Ce cas réel permet de comprendre, concrètement, les problèmes qu'adresse une démarche qualité des processus métier et les solutions qu'elle peut apporter. La présentation de ce cas pratique est aussi l'occasion pour nous de proposer quelques bonnes pratiques et conseils issus de notre expérience.

7.2. Evaluation de la qualité des processus métiers : des métriques et des méthodes

Nous présentons ci-après une revue de la littérature du domaine de l'évaluation de la qualité des processus métier. Dans ces travaux, on peut différencier les méthodes (démarches à suivre) pour évaluer la qualité qui s'appuient sur des principes généraux et guides méthodologiques, des travaux proposant des ensembles de métriques candidates à la qualification de la qualité d'un processus métier.

7.2.1. *Des méthodes pour l'évaluation de la qualité d'un processus*

Des méthodes de haut niveau[2] telles que PDCA ou DMAIC permettent de guider une évaluation de la qualité d'un processus métier, en en définissant les grandes étapes.

L'une des méthodes les plus connues est la méthode dite cycle de Deming [DEM 86] aussi appelée cycle de Shewhart [SHE 80] ou encore *Plan-Do-Check-Act*

2. Ces méthodes peuvent d'ailleurs également être utilisées pour l'évaluation de la qualité des données [BAT 09].

(PDCA) qui propose de mettre en œuvre une amélioration continue de la qualité en réitérant continuellement ces quatre étapes : planifier (*Plan*), mettre en œuvre (*Do*), vérifier (*Check*) et agir (*Act*). La première étape (*Plan*) consiste à définir les problèmes de qualité, définir des actions permettant d'améliorer la qualité, et planifier ces actions. L'étape de mise en œuvre (*Do*) consiste en la réalisation de l'amélioration. L'étape suivante de vérification (*Check*) consiste à évaluer l'efficacité de la solution d'amélioration déployée. Enfin, l'évolution de la solution est assurée dans la dernière étape (*Act*).

Le programme Six Sigma a proposé une adaptation de la méthode PDCA appelée méthode DMAIC [DEF 90]. L'acronyme DMAIC est issu des différentes étapes de l'approche, à savoir : définir la qualité (*Define*), la mesurer (*Measure*), analyser les résultats des mesures (*Analyze*), mettre en œuvre des actions d'amélioration du processus (*Improve*) et contrôler les effets des actions d'amélioration mises en œuvre (*Control*).

Partant arbitrairement de la méthode DMAIC tout-à-fait classique, nous décrivons ci-dessous, plus précisément, ses grandes étapes.

L'étape de *définition de la qualité du (des) processus* est fondamentale[3]. Elle consiste à dégager un ensemble de métriques concrètes (et éventuellement des seuils associés), mesurables, permettant de définir la qualité du processus.

Il est bien admis par les communautés étudiant la gestion de la qualité des données et de la qualité des schémas conceptuels [RED 97, WAN 00, MOO 05, BAT 06, BER 07] que la définition de la qualité dépend d'un objectif visé dans un contexte opérationnel particulier : « étant donné mon environnement et mon objectif opérationnel, quels sont mes critères de qualité ? ». Pour présenter les choses simplement, les besoins de qualité (*objectifs qualité*) découlent d'objectifs opérationnels dépendant d'un contexte dans lequel ces objectifs doivent être atteints. En effet, dans notre cas, un certain niveau de qualité des processus métier est nécessaire pour pouvoir atteindre les objectifs opérationnels visés. Pour savoir si un processus métier respecte le niveau de qualité attendu, on cherche à répondre à un ensemble de questions concernant cette qualité (ces questions sont appelées questions qualité). Chaque question qualité est elle-même décrite par un ensemble de métriques qui, une fois mesurées, permettront de répondre – au moins en partie – à la question. Cette approche permettant la décomposition de la qualité est connue sous le nom de paradigme *Goal-Question-Metric (GQM)* [BAS 94]. Ce paradigme fut tout d'abord utilisé pour l'évaluer de la qualité des développements logiciels. Il peut naturellement être adapté au problème de la définition de la qualité d'un processus métier [BAT 09, AVE 04].

3. Dans la suite, de façon à simplifier le discours, nous considérons l'évaluation d'un seul processus métier, et non d'un ensemble de processus métier.

Dans la pratique, les objectifs opérationnels, les objectifs qualité, les questions qualité et les métriques sont définis, de manière consensuelle, par un groupe de travail constitué d'acteurs métier. Dans l'idéal, l'étude qualité devrait être co-pilotée par un expert qualité et un référent métier (ou une personne possédant les deux compétences). L'expert qualité sait mener une étude d'évaluation de la qualité des processus métier (connaît les méthodologies, métriques, bonnes pratiques, et « pièges » éventuels). Le référent métier accompagne l'expert qualité pour toutes les questions relative au métier, et le guide dans l'entreprise (par exemple lui indique quels acteurs métier doivent participer à l'évaluation).

Le résultat de l'étape de définition de la qualité est un ensemble de métriques à évaluer, éventuellement organisées selon la hiérarchie GQM ou, comme cela se fait classiquement pour l'évaluation de la qualité des données, selon des dimensions qualité [BAT 06].

Vient ensuite l'étape de *mesure de la qualité* qui consiste à :

– recueillir les informations nécessaires à l'évaluation des métriques définie à l'étape précédente et ;

– évaluer les métriques.

Cette étape est évidemment très dépendante des métriques choisies. Le recueil des informations nécessaires à l'évaluation des métriques ne doit pas être négligé car la qualité des résultats de l'évaluation du processus dépend en partie de la qualité des informations recueillies.

Une connaissance approfondie du métier supporté par le processus n'est pas forcément nécessaire à cette étape puisque les métriques (et la façon dont elles sont calculées) ont précisément été définies dans l'étape précédente. L'expert qualité peut donc procéder seul aux mesures, éventuellement aidé d'outils permettant d'effectuer les mesures ou de personnes portant des compétences complémentaires nécessaires. Nous pouvons citer à titre d'exemples : (i) des statisticiens si les mesures nécessitent des calculs statistiques complexes, manipulant par exemple des outils statistiques tels que SAS [SAS] ou R [R], (ii) des outils de mesure de la consistance d'un processus métier par rapport à un ensemble de contraintes tels que par exemple le *model checker* NuSMV (*New Symbolic Model Verifier*) [CIM 02] si une telle consistance est à vérifier, ou encore (iii) des sociologues comme dans le cas d'étude présenté en section 7.3.

Le résultat de cette étape est un ensemble de mesures résultant de l'évaluation des métriques.

L'étape d'*analyse des résultats des mesures* consiste à analyser les résultats des mesures obtenus à l'issue de l'étape précédente (incluant une éventuelle étude statistique des résultats visant par exemple à exhiber des corrélations entre mesures), et à présenter les résultats aux acteurs métier, puis à un (ou des) décideur(s). La présentation des résultats inclut le rappel des limites de l'étude (par exemple, les métriques non mesurées).

Le cœur de cette analyse est l'interprétation des résultats par un groupe d'acteurs métier : des pistes métier permettant d'expliquer les résultats obtenus sont généralement dégagées. L'entreprise peut ensuite :

– soit décider de mettre en œuvre des actions d'amélioration du processus si la qualité du processus n'est pas satisfaisante. Une phase de contrôle des effets de ces actions suit généralement dans ce cas ;

– soit conclure que le processus métier est satisfaisant tel quel. Dans ce cas, l'entreprise peut soit « en rester là pour le moment », soit planifier une ou plusieurs actions de suivi de la qualité du processus afin de s'assurer que celle-ci ne se dégrade pas dans le temps (par exemple, sous l'effet d'évolutions environnementales).

Il serait idéal à ce stade de pouvoir comparer le coût de l'amélioration du processus par rapport au coût du manque de qualité de celui-ci afin de décider si l'amélioration a intérêt à être mise en œuvre [ENG 99]. Mais ceci est la plupart du temps infaisable car le coût de la non-qualité, même s'il est souvent important, est malheureusement très difficile à apprécier (comme d'ailleurs dans le cas de la mesure de l'impact du manque de qualité des données [RED 98]).

L'étape d'*amélioration du processus* consiste en la mise en œuvre d'actions d'amélioration du processus. Ces actions impactent souvent le métier. Elles peuvent concerner ses procédures, son organisation, le processus en lui-même, ses données ou de façon plus large son système d'information. Les effets de ces modifications pourront être évalués s'il est décidé d'enclencher une action de contrôle (voir ci-après).

Enfin, le *contrôle de la qualité* consiste à évaluer la qualité du processus après que les actions d'amélioration aient été mises en œuvre afin de juger de leur efficacité.

La qualité du processus peut être surveillée de façon régulière afin d'en suivre l'évolution (même sans mise en place d'action d'amélioration). On parle alors de *monitoring de la qualité*.

7.2.2. *Des métriques pour l'évaluation de la qualité des processus métier*

Il existe une large contribution dans la littérature à l'étude et à la mesure de la qualité des données et des modèles de données [GEN 08, GEM 03, PAR 92]. L'étude

de la qualité des processus et des modèles de processus est un sujet qui n'attire que récemment l'intérêt des chercheurs. Nos investigations nous ont conduits à classer les travaux existants dans trois catégories : la qualité des notations, la qualité des modèles et la qualité des processus.

Une notation est définie comme un langage offrant un ensemble de diagrammes et de moyens d'expression permettant la définition de modèles. La multiplication des notations existantes pour la modélisation des processus métier et la variation des modes d'expressions qu'elles offrent d'une part, la diversité des besoins des processus métiers, d'autre part, posent le problème du choix de la notation la plus adéquate pour une organisation. L'hypothèse sous-jacente est que la capacité d'expression d'une notation et sa facilité d'utilisation et d'interprétation ont un impact direct sur la qualité des modèles de processus utilisant cette notation et sur l'interprétation et l'exécution faite de ces modèles donc les processus eux-mêmes.

Dans [WAH 05] les auteurs présentent une évaluation analytique des notations en utilisant deux cadres connus dans la littérature que sont l'ontologie de BWW[4] [WAN 90] et le modèle sémiotique [LIN 94]. Dans [WOH 06, MEN 08] les auteurs discutent l'adéquation de BPMN [WHI 04], largement adopté par les chercheurs et par les entreprises, à la modélisation des processus métier. Les auteurs dans les deux articles soulignent *via* une étude comparative de diverses notations, les limites de certaines dans l'expression de certains aspects des processus métier.

Concernant les modèles de processus, d'après de rapport Gartner [GAR 05] les entreprises qui rencontrent plus de succès dans l'implémentation de leurs processus métier sont celles qui consacrent plus de 40 % de la durée totale des projets à la découverte et à la modélisation de leurs processus métier. Aujourd'hui, la modélisation des processus métier apparaît comme une des technologies stratégiques pour une entreprise dans les années à venir. Une approche permettant l'amélioration de la qualité des modèles des processus consiste à proposer des guides qui assistent la construction de ces modèles. Dans [BEC 00] les auteurs discutent de l'impact de la complexité sur la qualité des modèles des processus métier. Ils proposent des définitions pour certains critères de qualité, considérés comme majeurs, tels que la correction syntaxique ou sémantique et la pertinence, ainsi que des guides méthodologiques associés. Dans [MEN 10] les auteurs proposent un ensemble de guides méthodologiques pour améliorer la qualité des modèles de processus produits. Les auteurs de [ERI 00] proposent une extension du langage UML pour la modélisation des processus métier et une démarche qui guide cette modélisation depuis les objectifs métier de l'entreprise.

4. BWW : Bunge-Wand-Weber.

Cependant, l'amélioration de la qualité passe aussi par la détection de la non-qualité. Cette mesure nécessite le développement de concepts autour de la définition et de la mesure de la qualité des modèles conceptuels. Les auteurs dans [HMR 08] ont définit un cadre général pour la qualification des dimensions de la qualité d'un processus métier. Ce cadre prévoit quatre axes pour l'analyse de la qualité des processus métier : les fonctions, les entrées-sorties, les ressources à caractère non humain et les ressources à caractère humain. Ils associent à chacun de ces axes un ensemble de critères ou de dimensions de la qualité. Dans [SAN 10] les auteurs présentent un ensemble de métriques structurelles pour la mesure de la qualité de modèles de processus. Les auteurs dans [VAN 07] présentent un ensemble de métriques issues du domaine du génie logiciel telles que la cohésion, la taille ou le couplage. Ces métriques sont appliquées à la mesure de la qualité des processus métier et sont implémentées dans un outil *open source* appelé ProM. D'autres travaux qui s'appuient sur l'adaptation des métriques issues du génie logiciel sont proposés [MAK 10, LAU 06].

Enfin, il est possible de définir des métriques permettant d'évaluer en partie les résultats de l'application de processus métier, sans forcément s'intéresser au détail du processus, en mesurant par exemple les transformations induites sur l'environnement par l'application du processus métier ([HAS 09] en est un exemple). Evidemment, le choix des métriques dans ce cas est extrêmement dépendant les objectifs de l'entreprise (nous verrons quelques exemples de telles métriques dans le cas d'étude de la section suivante).

Cependant tout comme d'autres travaux dans d'autres domaines tels que la qualité des données ou la qualité des modèles conceptuels, les critères et métriques souffrent d'un manque de validation des propositions. On trouve dans [CAN 05 ; MEN 08] diverses expérimentations menés notamment sur la compréhensibilité et la complexité des modèles de processus métier. Une récente contribution [LAU 11] démontre cependant la limite de ce genre d'expérimentation fortement dépendante du public interrogé et surtout des questions posées et de la façon dont elles sont posées. Les auteurs ont mené l'exercice intéressant de rejouer des expérimentations en modifiant la façon dont les questions sont posées de manière à les rendre plus accessible aux participants à l'expérimentation.

7.3. Un cas concret : évaluation de la qualité du processus de transition d'un projet informatique sous-traité

Dans cette section, nous proposons un cas concret d'évaluation de processus métier. Le processus est l'un des processus métier de pilotage d'un grand EPST (établissement public à caractère scientifique et technologique) français. Il concerne le pilotage de la phase dite de transition d'un projet informatique sous-traité, phase

durant laquelle une SSII travaillant sur le projet « passe la main », en cours de projet, à une autre SSII.

7.3.1. *Présentation du cas d'étude et de son contexte*

Nous présentons tout d'abord le cas d'étude et son contexte opérationnel pour l'évaluation (repris de [GRI 10]). Depuis une dizaine d'années, les établissements publics à caractère scientifique et technologique (EPST) se concentrent sur leur cœur de métier, à savoir la recherche, et externalisent certains métiers support tels que les métiers des directions des ressources humaines, des affaires financières, ou encore du système d'information. Le métier de l'entité interne de direction des systèmes d'information (DSI), centré sur la conception et le développement de nouvelles applications, outils et logiciels, s'en trouve modifié. La DSI gère maintenant des projets dits externalisés (au sens de [LAC 93]). Elle est toujours responsable de chaque projet de développement mais sous-traite certaines tâches de développement ou de maintenance à une société de services externe[5] aussi appelée prestataire, choisie à l'issue d'un appel d'offre. Cet acte de sous-traitance est appelé tierce maintenance applicative (TMA). Dans les organisations publiques, telles que les EPST, les règles du marché public concernant la sous-traitance obligent à mettre (ou remettre) en concurrence les marchés tous les trois ans. Des appels d'offre sont donc (re-)lancés au moins tous les trois ans et peuvent amener à changer de prestataire au cours d'un projet. Ainsi, périodiquement, une équipe sortante transfère le projet en cours vers une équipe entrante. Ce transfert constitue l'une des phases du processus de gestion d'un projet externalisé. Elle est appelée transition due à un changement de prestataire (tout simplement transition dans la suite). La transition, placée sous la responsabilité du chef de projet (membre de la DSI) a pour principal objectif le transfert des documentations, logiciels et connaissances liées au projet du prestataire sortant au prestataire entrant. Dans l'EPST avec laquelle nous fûmes en contact, un processus métier de type processus de pilotage [MOR 07], décrivant une procédure à suivre, est défini pour piloter cette phase de transition. Il comporte six activités :

– (activité 1) l'initialisation marquant le début officiel (contractuel) de la phase de transition ;

– (activité 2) l'arrêt et la restitution de la tierce maintenance applicative (TMA) durant laquelle sont inventoriés les documents et codes constituant le projet de développement ;

– (activité 3) la rédaction et la validation du plan de transfert ;

– (activité 4) le « transfert de connaissances ». Sont concernés par cette activité les logiciels, les modules (les codes de façon large) et les connaissances explicitées

5. Ou plusieurs sociétés de services mais dans notre cas une seule.

sous forme de documents. Ces documents et codes sont pour la plupart rédigés par le prestataire sortant. La DSI est peu impliquée dans cette activité ;

– (activité 5) la maintenance en coopération durant laquelle les prestataires sortant et entrant assurent ensemble la maintenance de l'application ;

– (activité 6) le transfert de responsabilités, marquant le départ officiel du prestataire sortant.

Chacune des activités est constituée d'un ensemble organisé de tâches. Une contrainte forte est imposée à l'équipe projet : la transition doit être assurée en vingt jours ouvrés. Comme la plupart des processus métier, ce processus peut être modélisé. La figure 7.2 présente un (petit) extrait du diagramme d'activités UML associé au modèle du processus du cas d'étude[6].

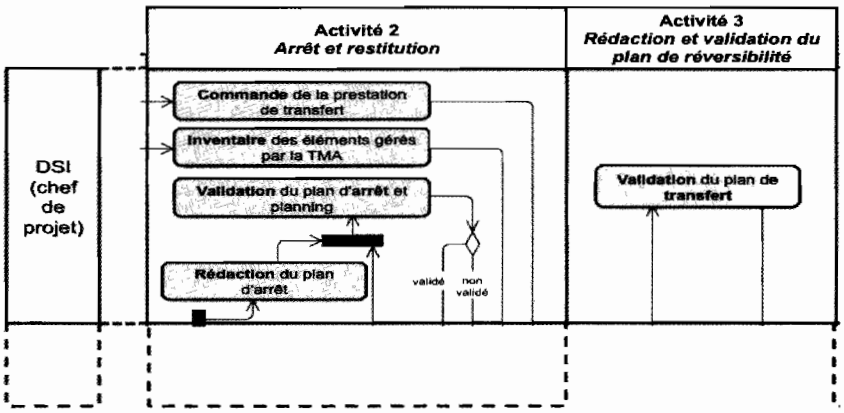

Figure 7.1. *Activités 2 et 3 du processus de transition*

Le chef de projet de l'EPST pilotant la transition a pour objectif opérationnel de réussir au mieux le transfert de documentations, logiciels et connaissances liées au projet de l'équipe sortante vers l'équipe entrante. Avant l'évaluation du processus, il a pu dresser deux constats.

CONSTAT 1.– Le processus de transition fait intervenir bien plus de personnes en interne (au sein de la DSI) que le chef de projet seul qui est pourtant le seul acteur à

6. Le diagramme d'activité UML associé au modèle du processus du cas d'étude comporte un grand nombre de tâches. Nous ne présentons donc ici qu'un extrait du diagramme d'activités focalisé sur les activités deux et trois qui se révèleront être des activités d'intérêt au moment de l'analyse des résultats de l'évaluation.

apparaître officiellement sur la procédure de transition. En effet, un réseau social d'aide informelle est mis en place entre membres de la DSI afin de réussir au mieux la phase de transition. Par exemple, le chef de projet est responsable de la tâche d'*inventaire des éléments gérés par la TMA* faisant partie de l'activité *arrêt et restitution* (activité 2).

Afin de valider ou compléter son inventaire, de sa propre initiative, le chef de projet demande l'aide informelle de différentes personnes :

– l'expert technique de l'architecture applicative ;

– l'expert architecture logicielle, qui fait lui-même appel à l'administrateur de bases de données, à un membre du front office, et à un développeur JAVA ;

– l'expert architecture matérielle, qui fait lui-même appel à deux ingénieurs système et réseau ;

– le référent fonctionnel, qui fait lui-même appel à trois experts métier.

Ainsi, cette tâche qui semble « toute simple » selon le processus métier décrit au départ et qui semblait n'impliquer que le chef de projet implique en fait treize personnes dont douze de façon informelle. Aux dires du chef de projet, les connaissances de ces personnes sont importantes pour que le livrable de la tâche (un document lisant les éléments à transférer) soit de bonne qualité.

Du fait de la forte contrainte de temps imposée (vingt jours ouvrés), le processus pourrait donc être sensible à l'absence de personnes qui n'apparaissent pourtant pas sur le processus métier modélisé au départ.

CONSTAT 2.– Il semble au chef de projet que le transfert de connaissances ne s'effectue pas au mieux car le prestataire entrant est souvent peu autonome lorsqu'il prend la main sur le projet à l'issue de la phase de transition. Ce constat indique qu'une partie de la connaissance nécessaire au projet manque au prestataire entrant. Pourtant, tous les documents écrits, logiciels et codes nécessaires au prestataire entrant sont bien transmis par le prestataire sortant. Le chef de projet en a déduit que la connaissance nécessaire au projet est mal transférée de l'équipe sortante au prestataire entrant.

Cette notion de connaissance demande à ce stade un éclaircissement. Classiquement, on distingue deux types de connaissances [NON 91] : les connaissances explicites qui peuvent être écrites ou encodées pour être transmises et les connaissances tacites non explicitables (par exemple les intuitions, les tours de main, les expériences physiques, les connaissances environnementales). Il a été démontré que le partage des connaissances, non seulement explicites mais aussi tacites au cours d'un projet externalisé, joue un rôle fondamental dans la performance de ce

projet [LEE 01] et évidemment, plus particulièrement lors d'une phase de transition [ALA 10]. Ainsi, pour le chef de projet, transférer au mieux le projet signifie également transférer au mieux les connaissances explicites et les connaissances tacites nécessaires au projet. Enfin, il est important de comprendre que le transfert des connaissances ne se limite pas à leur transmission de l'équipe sortante au prestataire entrant mais nécessite également l'absorption de la connaissance par le prestataire entrant. Concrètement, les connaissances transmises verbalement ou dans un document ne sont transférées que lorsque l'équipe entrante a reçu puis assimilé les connaissances portées par le document au point de savoir mettre en application ces connaissances ; on parle alors d'absorption de la connaissance [DAV 98].

Forts de ces constats, il a paru évident qu'évaluer (et éventuellement améliorer) la qualité du processus métier de transition serait une bonne chose. Cette évaluation constitue le sujet du cas d'étude que nous présentons ci-après. L'évaluation que nous présentons ci-dessous est centrée sur la qualité du processus métier en termes de transfert de connaissance (l'objectif de l'évaluation n'était pas une évaluation la plus exhaustive possible, ce qui n'est d'ailleurs généralement pas possible car trop coûteux).

Pour mener cette évaluation, les étapes du cycle classique DMAIC présentées précédemment ont été suivies, à savoir : définir la qualité (*Define*), l'évaluer (*Measure*) et analyser les résultats (*Analyze*) pour ensuite choisir et mettre en œuvre des actions d'amélioration du processus métier (*Improve*), et enfin contrôler l'efficacité de ces actions d'amélioration (*Control*). Ainsi, deux batteries de mesures ont été effectuées, l'une dans l'étape *Measure* et l'autre dans l'étape *Control*.

7.3.2. *Définition de la qualité du processus* (Define)

Après avoir décidé du besoin d'évaluation de la qualité du processus, il convient de définir ce qu'est la qualité de ce processus. Il s'agit là d'une étape fondamentale. Nous rappelons que :

La définition de la qualité dépend d'un objectif opérationnel et d'un contexte.

Une méthode permettant d'aider à définir la qualité est celle définie dans le paradigme GQM (voir section 7.2.1). Concrètement, un groupe de travail composé d'acteurs métier est constitué[7]. Ce groupe se réunit à plusieurs reprises afin d'appliquer la méthode GQM, jusqu'à aboutir à une définition satisfaisante et consensuelle de la qualité du processus.

7. Ce groupe peut être complété par d'autres compétences. L'évaluation du cas d'étude a, par exemple, nécessité de faire intervenir un *knowledge manager* en raison des questions fortement liées au transfert de connaissances dans le processus.

Ce travail est piloté par un « expert » en évaluation de la qualité (appelé « expert qualité » dans la suite) au fait des problèmes, outils et méthodes liées à l'évaluation de la qualité des processus. Bien sûr, ses connaissances du métier sont parfois limitées, c'est pourquoi une expertise métier doit toujours lui être associée. Dans l'idéal, un référent métier devrait être désigné. Ce référent métier est un point d'entrée dans le monde métier, il guide l'expert qualité dans le choix des personnes à impliquer dans l'étape de définition de la qualité du processus.

Application (cas d'étude) : le paradigme GQM a permis de définir des métriques permettant d'évaluer la qualité du processus métier de transition. Le lecteur notera bien ici que toutes autres métriques pourraient être choisies dans un autre contexte opérationnel.

A l'issue de l'étape de définition, quatre objectifs qualité sont exhibés pour le cas d'application. Les deux premiers sont : (QG_1), identifier les activités les plus complexes (ici plutôt en termes de critères permettant d'apprécier l'effort nécessaire au « pilotage » des activités par la clé de projet) et (QG_2), identifier les activités et tâches les plus sensibles au risque de perte de connaissance afin d'en surveiller le bon déroulement par la suite. Le troisième objectif qualité est (QG_3), assurer un bon transfert des connaissances explicites et tacites lors du processus de transition. Enfin, la contrainte de temps est exprimée par l'objectif (QG_4), effectuer la transition en vingt jours ouvrés. Chacun de ces objectifs qualité est décliné en un ensemble de questions qualité (voir tableau 7.1). Chaque question qualité est elle-même déclinée en un ensemble de métriques (voir tableau 7.2) qui, une fois mesurées, permettent de répondre au moins en partie à la question qualité. Les valeurs visées des métriques (aussi parfois appelées seuils), si elles existent, sont fixées par expertise (qualité ou métier).

Objectifs qualité	**Questions qualité**
QG_1	($QQ_{1.1}$) Quelle est la complexité de chaque activité ?
QG_2	($QQ_{2.1}$) Le réseau informel mis en œuvre pour effectuer la tâche est-il complexe ?
QG_3	($QQ_{3.1}$) Les connaissances explicites sont-elles bien transférées (bien transmises par l'équipe sortante et bien absorbées par le prestataire entrant) ?
	($QQ_{3.2}$) Les connaissances tacites sont-elles bien absorbées par l'équipe entrante ?
	($QQ_{3.3}$) L'équipe entrante a-t-elle bien compris les connaissances liées au projet, au point d'être autonome à la fin de la transition ?
QG_4	($QQ_{4.1}$) En combien de temps la phase de transition est-elle effectuée ?

Tableau 7.1. *Des objectifs qualité aux questions qualité*

		Métrique		Valeur visée
		Nom avec les paramètres t : tâche, a : activité, p : processus	**Description**	
QG_1	$QQ_{1.1}$	complex(a)	Pour chaque **activité** a, $(\|T_a\| + \|D_a\|)/E(a)$ où $\|T_a\|$ est le nombre de tâches de a placées sous la responsabilité de la DSI, $\|D_a\|$ est le nombre de nœuds de décision apparaissant dans l'activité et E(a) est le nombre d'arcs entres les tâches ou nœuds de décision de a.	A (**)
		taille(a)	Pour chaque **activité** a, $\|T_a\|$.	A (**)
		durée(a)	Pour chaque **activité** a, durée moyenne de a (en nombre de jours).	A (**)
		nbTjour(a)	Pour chaque **activité** a, ratio $\|T_a\|$/durée(a).	A (**)
QG_2	$QQ_{2.1}$	tailleRS(t)	Taille (nombre d'individus) du réseau social(*) mis en œuvre au sein de la DSI pour effectuer chaque **tâche** t, notée tailleRS(t).	≤ 5
		tailleRS(a)	Pour chaque **activité** a composée d'un ensemble de tâches T_a : tailleRS(a) = max({tailleRS(t)|t ∈ T_a }).	≤ 5
		tailleChaine(t)	Taille du plus long chemin d'un exécuteur à un contributeur informel dans le réseau social(*) mis en œuvre au sein de la DSI pour effectuer chaque **tâche** t, noté tailleChaine(t).	≤ 2
		tailleChaine(a)	Pour chaque **activité** a composée d'un ensemble de tâches T_a : tailleChaine(a) = max({tailleChaine(t) / t ∈ T_a}).	≤ 6
QG_3	$QQ_{3.1}$	TransKexplicit(p)	Retard moyen (en nombre de jours) de la livraison des documents écrits de l'équipe sortante au prestataire entrant.	0
		AbsorbKexplicit(p)	Mesure subjective, par l'équipe entrante, du niveau de compréhension des documents écrits reçus (connaissance explicite). En moyenne, sur tous les documents et tous les lecteurs, avec l'échelle suivante pour les interviews : 1-bonne compréhension, 2-compréhension moyenne et 3-faible compréhension. Mesure obtenue par interviews de l'équipe entrante.	1

QG_3	$QQ_{3.2}$	AbsorbKtacit(p)	Mesure subjective, par l'équipe entrante, du niveau de compréhension de connaissances implicites (telles que par exemple les routines organisationnelles). En moyenne, sur tous les documents et tous les lecteurs, avec l'échelle suivante pour les interviews : 1-bonne compréhension, 2-compréhension moyenne et 3-faible compréhension. Mesure obtenue par interviews de l'équipe entrante.	≤ 2
	$QQ_{3.3}$	autonomie(p)	Mesure subjective du chef de projet du niveau d'autonomie de l'équipe entrante pour le traitement d'une anomalie sur le projet. Echelle avec quatre valeurs : excellente autonomie, bonne autonomie, autonomie moyenne ou mauvaise autonomie). Mesure obtenue par interview du chef de projet.	bonne ou excellente
QG_4	$QQ_{4.1}$	duréeTrans(p)	Date du transfert officiel de responsabilités (dans l'Activité 6) – date établie lors de l'initialisation (dans l'Activité 1).	≤ 20

(*) Le réseau social évoqué est le réseau de demandes d'aide informelle sous-jacent à chaque tâche ou activité. Le détail de ces métriques peut être trouvé dans [GRI 10].
(**) A = Aucun seuil fixé a priori.

Tableau 7.2. *Des questions qualité aux métriques qualité*

On peut constater une grande « diversité » des métriques choisies. Elles peuvent concerner :

– la qualité du processus modélisé aussi appelé qualité du modèle, sans considérer son exécution (il s'agit ici des métriques *complex* et *taille*) ;

– ou encore, pour une application particulière de ce processus :

- des paramètres de son déroulement (métriques *tailleRS*, *tailleChaine*, *durée* et *nbTjour*, *TransKexplicit*, *dureeTrans*) ;

- ses effets sur l'environnement (métrique autonomie, qualifiant d'ailleurs aussi l'efficacité du processus) ;

- la qualité de ses données et informations produites (métrique *Absorb-Kexplicit*). La mesure *AbsorbKexplicit*, grandement dépendante de la qualité des documents produits par l'équipe sortante, nous rappelle d'ailleurs que :

La qualité des informations ou des données produites par un processus participe généralement à la définition de sa qualité.

Ce type de métrique montre bien que la qualité des données impacte la qualité d'un processus. Il est aussi bien évident qu'un transfert bien effectué donnera lieu à des documents produits par les équipes sortantes et entrantes de meilleure qualité, engendrant ainsi une meilleure qualité des données produites dans l'entreprise. Ceci vaut pour les processus orientés contrôle comme celui du cas d'application présenté ici mais vaut, aussi et surtout, pour les processus orientés données.

Les métriques peuvent aussi concerner différents niveaux de granularité du processus : les tâches, les activités, ou même le processus lui-même. De plus, l'évaluation attendue d'une métrique peut être objective (par exemple ici la mesure du temps d'exécution des activités), ou subjective (s'appuyant éventuellement sur une simple appréciation du chef de projet, comme par exemple la métrique autonomie). Ces notions de mesures subjectives et objectives sont encore une fois à rapprocher de travaux menés dans le cadre de l'évaluation de la qualité des données [PIP 02].

Evidemment, pour une évaluation encore plus complète (et plus coûteuse !), bien d'autres métriques auraient pu être envisagées ([AVE 04] présente par exemple un cas d'utilisation pour lequel sont définies d'autres métriques).

Il est à noter que certains aspects de la qualité ne sont pas mesurables. Dans notre cas, il est bien clair que la motivation du prestataire sortant à transférer ses connaissances à l'équipe entrante joue un rôle important dans la réussite de la phase de transition. Cette motivation pourrait ne pas être au rendez-vous car, rappelons-le, le prestataire sortant n'a pas été reconduit suite à l'appel d'offre et il est sur le point de quitter définitivement le projet. Mais cette motivation est très difficilement mesurable. Il s'agit de l'une des limites d'une telle évaluation.

De plus, le chef de projet ayant commandité l'étude qualité a désiré limiter l'évaluation du processus métier aux tâches pilotées par l'acteur DSI (EPST) afin de ne pas interférer dans les tâches menées par les prestataires indépendamment de la DSI. Ainsi, certaines tâches du processus métier n'ont volontairement pas été évaluées.

L'évaluation de la qualité d'un processus métier est souvent partielle.

Il convient d'en tenir compte au moment de la présentation des résultats de l'évaluation afin que l'interprétation des résultats par les acteurs métier soit la plus juste possible.

7.3.3. *Recueil des informations nécessaires à l'évaluation et calcul des mesures* (Measure)

Le recueil des informations nécessaires à l'évaluation et leur calcul sont des tâches généralement gérées par l'expert qualité. Elles peuvent nécessiter des compétences complémentaires et peuvent impliquer une large participation des acteurs métier.

Application (cas d'étude) : dans le cas d'étude, les métriques *complex* et *taille* sont mesurées sur le processus modélisé indépendamment de toute mise œuvre de celui-ci. Elles sont donc facilement calculées par l'expert qualité seul.

Les métriques *AbsorbKexplicit, AbsorbKtacit* et *autonomie*, sont calculées après entretiens auprès de l'équipe entrante et du chef de projet (ici cinq personnes). La valeur de *durée* est donnée par le chef de projet.

Les métriques *tailleRS*, et *tailleChaine* sont plus compliquées à évaluer car elles sont calculées en fonctions de données issues du réseau social des relations informelles mises en œuvre pour l'exécution du processus. Il convient de capturer ce réseau social. Des sociologues ont donc du être sollicités pour aider l'expert qualité dans la définition et la mise en œuvre de la collecte des informations permettant d'exhiber le réseau. La démarche adoptée intégrait le choix des relations à observer, la définition du périmètre du système, l'acquisition des informations (par entretiens) et la modélisation de ce réseau sous forme de graphe(s) (comme cela se pratique dans le cadre de l'analyse structurelle de réseau social [DEG 94, WAS 94]). Exhiber ce réseau nécessite de mener des entretiens auprès d'acteurs du processus métier (ici une quinzaine d'acteurs métier).

Dans ce cas d'application, l'étape de recueil des informations nécessaires à l'évaluation a donc imposé de mener un grand nombre d'entretiens. Elle est coûteuse en ressources humaines et en temps.

Un écueil classique du point de vue de la gestion du projet d'évaluation de la qualité est de planifier l'étude sans avoir anticipé l'implication nécessaire en termes de ressources humaines, en particulier les acteurs métier. En effet, les acteurs métier sont généralement très sollicités par ailleurs. Si leur participation n'a pas été prévue en amont, leur mobilisation venant en marge de leur travail risque d'être difficile.

L'implication des acteurs métier peut être importante, il est préférable qu'elle soit prévue en amont.

Le cas d'étude a également montré que le processus modélisé au départ utilisant la notation des diagrammes d'activité d'UML ne contenait pas toutes les informations nécessaires à l'évaluation de la qualité du processus. Pourtant, cette

notation était suffisante au suivi du processus avant qu'une évaluation de la qualité ne soit envisagée. Le formalisme à utiliser pour modéliser le processus dépend donc bien de ce qu'il sera fait de la modélisation.

Si le processus doit être modélisé pour être évalué (ce qui n'est d'ailleurs pas systématiquement nécessaire), le choix de la notation à utiliser n'est pas anodin.

Application (cas d'étude) : le tableau 7.3 présente une partie des résultats de l'évaluation de la qualité du processus de transition.

Activité / *Métrique*	*Initialisation*	*Arrêt et restitution*	*Rédaction et validation du plan de transfert*	*Transfert de connaissances*	*Maintenance et coopération*	*Transfert de responsabilités*
complex	0,5	0,7	0,6	0	0,7	0,7
taille	1	4	1	0	2	1
durée	1	1	1	6	10	2
nbTjour	1	4	1	0	0,2	1
tailleRS	5	13	4	NA[8]	4	5
tailleChaine	2	3	2	NA[8]	2	2

TransKexplicit *: 0 (aucun retard dans les livraisons des documents écrits).*
*AbsorbKexplicit : **2,2** (compréhension moyenne par l'équipe entrante des documents qui lui sont transmis).*
*AbsorbKtacit : **3** (faible absorption de la connaissance liée au projet par le prestataire entrant).*
*automomie : autonomie **moyenne** de l'équipe entrante.*
duréeTrans : 20 jours (contrainte respectée).
(Les valeurs grisées sont des valeurs d'intérêt discutées dans la suite.)

Tableau 7.3. *Une partie des résultats de l'évaluation*

8. Métrique non évaluable car l'acteur EPST n'est responsable d'aucune tâche dans cette activité qui concerne le transfert des connaissances de l'équipe prestataire sortante vers l'équipe prestataire entrante.

7.3.4. *Analyse des résultats* (Analyze) *et amélioration du processus métier* (Improve)

L'étape d'analyse des résultats de l'évaluation est réalisée par des acteurs métier. Elle est évidemment très dépendante du domaine applicatif. A l'issue de cette étape, quelques conclusions sont dressées et des actions d'amélioration du processus peuvent être envisagées. Nous illustrons cette étape sur le cas d'étude.

Application (cas d'étude) : l'issue de l'étude des résultats avec l'aide du chef de projet et un *Knowledge manager*, plusieurs conclusions furent dégagées et actions d'amélioration du processus.

Les mesures *tailleRS* et *tailleChaine* ont clairement identifié l'activité *Arrêt et restitution* comme étant l'activité la plus sensible du processus à la perte de connaissance. Au sein de cette activité, deux tâches ont été repérées comme particulièrement sensibles : les tâches *Inventaire des éléments gérés par la tierce maintenance applicative*, et *rédaction du plan d'arrêt*. Plusieurs conclusions ont pu en être tirées :

1- ces tâches sont plus complexes à accomplir que ne le laissaient supposer les informations disponibles avant l'étude, les exécuteurs des tâches de ces activités sollicitant beaucoup d'aide informelle (en dehors de la procédure officielle) ;

2- la qualité d'exécution de ces tâches risque d'être particulièrement impactée en cas d'absence de personnes qui ne sont pas forcément repérées comme exécutrices officielles. Le chef de projet en charge de ces activités portera donc une attention toute particulière à la qualité des livrables produit par ces tâches, en particulier si elles sont effectuées dans une période propice aux absences (par exemple lors d'un pique de grippe saisonnière, lors des congés estivaux ou au moment d'une réorganisation des services de l'entreprise).

Il aurait été possible de faire évoluer le processus métier par exemple en décomposant la tâche d'*inventaire des éléments gérés par la tierce maintenance applicative* en plusieurs sous-tâches faisant apparaître explicitement certains contributeurs. Cette option n'a pas été retenue car elle aurait trop complexifié le processus, le rendant difficile à piloter (et à mettre en œuvre). En termes de métrique, cette option aurait rendu les tâches de l'activité moins sensibles au risque de perte de connaissance avec des mesures de *tailleRS* et *tailleChaine* plus basses mais aurait trop fait croître les mesures de complexité du processus (mesures *complex* et *taille*). Cette situation met en évidence le fait que la qualité est souvent un compromis entre plusieurs choix et que l'amélioration de certaines mesures peut entraîner la dégradation de certaines autres. Cette dégradation peut ne pas être acceptable, comme dans notre cas d'étude pour lequel l'amélioration de la robustesse du processus nécessitait une trop grande augmentation de sa complexité.

Certaines métriques sont interreliées (s'influencent les unes les autres) « négativement ». Améliorer une mesure pour une métrique peut amener à dégrader la mesure d'une autre métrique.

L'analyse de *AbsorbKexplicit et AbsorbKtacit* a montré que certains documents rédigés par l'équipe sortante étaient difficiles à appréhender par l'équipe entrante. Une raison identifiée est que les équipes projet sortantes et entrantes ne se côtoient pas suffisamment tout au long du processus de transition, celui-ci étant très axé sur la production et la transmission de documents écrits. Ceci a amené à faire évoluer le processus en différents points, les deux plus important étant :

1 – prise en main de l'activité de transfert des connaissances par le chef de projet. Il supervise maintenant cette phase. Il organise deux séances de travail à laquelle sont conviés les prestataires entrants et sortants. La première séance de travail dure trois jours, elle est dédiée à la présentation des équipes, de leurs rôles et expériences sur le projet. La seconde séance de travail sur cinq jours est dédiée à la présentation des aspects fonctionnels et techniques du projet. La phase de transfert sous cette forme est résolument plus axée sur le partage de connaissances explicites et tacites alors qu'elle consistait surtout en un transfert de connaissances explicites (transmission de document) avant ;

2 – amélioration de la phase de maintenance en coopération durant laquelle les prestataires sortants et entrants assurent ensemble la maintenance de l'application. Jusqu'ici aucune directive claire n'était donnée concernant le contenu de cette tâche. Elle consiste parfois en une simple phase d'observation du projet par le prestataire entrant. Cette étape, pourtant fondamentale, est souvent délaissée. Elle peut même parfois servir de « variable d'ajustement » : si le temps manque, sa durée est écourtée. Le chef de projet, responsable de cette tâche, encourage maintenant le prestataire entrant à mettre en pratique les connaissances qu'il a reçues en prenant quelques incidents en cours pour tenter de les résoudre. Si des éléments lui manquent pour résoudre des incidents alors cela signifie que certaines connaissances lui manquent ou sont mal absorbées. Effectuer cet exercice à ce moment-là permet de déceler les problèmes de transfert de connaissance avant que le prestataire sortant n'ait quitté le projet (avec sa connaissance !). Ces actions d'amélioration sont présentées dans [GRI 11].

Il est bien connu de la communauté du *Knowledge Management* qu'une bonne assimilation de la connaissance tacite aide à la compréhension de la connaissance explicite et vice versa. On voit donc ici qu'un bon degré de compréhension des documents reçus par l'équipe entrante (*AbsorbKexplicit*) participe à une meilleure assimilation de la connaissance tacite liées au projet et donc, au final à un meilleur degré d'autonomie pour la reprise du projet (en partie mesuré dans *autonomie*). Les métriques *AbsorbKexplicit*et *autonomie* sont donc interreliées. Ceci ne pose pas de

problème particulier, il convient juste d'en tenir compte au moment de l'interprétation des résultats (et également du choix de modification du processus).

Certaines métriques sont interreliées (s'influence les unes les autres) « positivement ». Améliorer la mesure d'une métrique peut amener à améliorer la mesure d'une autre métrique.

7.3.5. *Suivi de la qualité du processus* (Control)

Les améliorations apportées au processus métier ont-elles porté leurs fruits ? Quel est l'impact d'une réorganisation interne sur la qualité du processus métier ? Comment la qualité du processus métier évolue-t-elle dans le temps ? C'est à ce type de questions que peut répondre un suivi de la qualité du processus.

L'action de suivi peut être mise en œuvre à plus ou moins long terme. On peut soit chercher à mesurer l'impact d'actions d'amélioration en effectuant juste une mesure de la qualité après amélioration. On peut aussi décider de mesurer périodiquement la qualité du processus métier. En d'autres termes, selon les besoins, la mesure de la qualité du processus peut être effectuée soit en mode projet (une mise en œuvre unique DMAIC, *Define-Measure-Analyze-Improve-Control*), soit relever d'une mission permanente de surveillance au sein de l'entreprise (surveillance et amélioration continue par suivi du cycle DMAIC par exemple) [MOR 07]. Cette décision dépend des besoins et des moyens de l'entreprise.

Application (cas d'étude) : le cas d'étude va jusqu'au contrôle de la qualité afin de mesurer en partie les effets des actions d'amélioration décidées. Les mesures dépendant des réseaux sociaux n'ont pas été remesurées car trop coûteuses.

Le tableau 7.4 montre des résultats de l'évaluation de contrôle. Evidemment, ces résultats sont à mettre en perspective car cette seconde mesure a été effectuée lors d'une autre phase de transition impliquant des prestataires différents. Néanmoins, de l'avis de la chef de projet habituée à gérer des transitions de projet d'un prestataire à un autre, ces actions ont permis une meilleure communication des prestataires sortants et entrants, ainsi qu'un meilleur transfert des connaissances. L'autonomie du prestataire entrant a été jugée meilleure suite à cette phase transition qu'à la précédente (mesure *autonomie*), signe que le transfert s'est beaucoup mieux déroulé. Ceci a pu être fait tout en respectant la contrainte de temps de vingt jours ouvrés pour la transition (mesure *duréeTrans*), et sans complexifier le processus (et donc sans complexifier son pilotage, comme les montrent les mesures *complex, taille, durée* et *nbTjour*).

Activité / *Métrique*	*Initialisation*	*Arrêt et restitution*	*Rédaction et validation du plan de transfert*	*Transfert de connaissances*	*Maintenance et coopération*	*Transfert de responsabilités*
complex	0,5	0,7	0,6	0,5	0,7	0,7
taille	1	4	1	2	2	1
durée	0,5	0,5	1	8	9	1
nbTjour	1	4	1	0,25	0,2	1
tailleRS	-	-	-	-	-	-
tailleChaine	-	-	-	-	-	-

TransKexplicit *: 0 (aucun retard dans les livraisons des documents écrits).*
*automomie : **bonne** autonomie de l'équipe entrante.*
duréeTrans : 20 jours (contrainte respectée).
Légende :
- : non mesuré car trop coûteux.
Les mesures grisées sont les mesures ayant évolué.

Tableau 7.4. *Une partie des résultats de l'évaluation après amélioration*

Une autre mesure ayant évoluée est la complexité du processus : le chef de projet est maintenant plus impliqué dans la transition, il doit gérer deux tâches supplémentaires dans le processus, sur huit jours (nouvelle activité de *transfert des connaissances*) là où il intervenait assez peu avant. Cela se traduit par plus de temps passé par le chef de projet sur le pilotage de la phase de transition.

7.4. Conclusion

Processus et données sont étroitement liés puisque les processus métier exploitent des données de l'entreprise et produisent de nouvelles données. Des processus de mauvaise qualité engendrent la production de données de mauvaise qualité et des données de mauvaise qualité peuvent engendrer un mauvais déroulement des processus métier. Gouverner les données de l'entreprise implique donc également gérer ses processus, et gérer ses processus implique de régulièrement les évaluer et les améliorer. Nous avons présenté dans ce chapitre un ensemble d'outils méthodologiques pour la gestion de la qualité des processus métier. Nous

avons montré à travers un cas pratique comment divers moyens méthodologiques peuvent être mis en œuvre pour mesurer et améliorer la qualité de ce processus.

Le cas d'étude a permis de mettre en évidence quelques limites d'une telle évaluation ainsi que certains constats d'intérêt sur lesquels nous revenons ici.

Concernant les méthodes de gestion de la qualité dans la littérature, un grand effort reste à fournir par les acteurs qui les mettent en application dans un projet. Notre expérience nous a permis de mettre ainsi en évidence le manque d'aide efficace de ces méthodes qui s'attachent à définir les grandes lignes de l'approche et les concepts sous-jacents mais manquent de guides pouvant fournir une assistance pas à pas.

Nous avons également pu voir que les domaines de l'évaluation de la qualité des données, des modèles et des processus se « nourrissent les uns les autres ». De nombreuses méthodes et métriques sont reprises d'un domaine à un autre. Comme a pu le montrer le cas d'étude, évaluer un processus peut passer par l'évaluation de son modèle et des données qu'il manipule ou produit. De même, certains travaux montrent comment la qualité des données peut être évaluée par la mesure de la qualité du processus qui la produit (par exemple, pour l'évaluation de la fraicheur [PER 06]).

La gestion de la qualité d'un système d'information ne se résume pas uniquement à la qualité des sources de données qu'il utilise et à la qualité des livrables qu'il fournit mais dépend aussi largement des processus métier qui transforment les sources pour produire les livrables.

Le cas d'étude a également fait ressortir le problème de l'interdépendance entre les facteurs de la qualité qui reste un problème difficile peu abordé dans la littérature.

Enfin, il nous parait légitime à travers cette expérimentation de souligner l'aspect lié au coût lié à la mesure et à l'amélioration de la qualité et de se poser la question de comment évaluer le gain apporté par cette gestion de la qualité.

7.5. Bibliographie

[ALA 10] ALARANTA M., JARVENPAA S.L., « Changing IT Providers in Public Sector Outsourcing : Managing the Loss of Experiential Knowledge », *Proceedings of the International Conference on System Sciences (HICSS)*, p. 1-10, 2010.

[AVE 04] AVERSANO L., BODHUIN T., CANFORA G., TORTORELLA M., « A Framework for Measuring Business Processes Based on GQM », *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2004.

[BAS 94] BASILI V.R., CALDIERA G., ROMBACH H.D., « The Goal Question Metric Approach », *Encyclopedia of Software Engineering*, Wiley, New York, 1994.

[BAT 06] BATINI C., SCANNAPIECO M., *Data Quality : Concepts, Methodologies and Techniques*, Springer-Verlag, New York, 2006.

[BAT 09] BATINI C., CAPPIELLO C., FRANCALANCI C., MAURINO A., « Methodologies for data quality assessment and improvement », *ACM Comput. Surv.*, vol. 41, n° 3, p. 1-52, 2009.

[BEC 00] BECKER J., ROSEMANN M., UTHMANN C., « Guidelines of Business Process Modeling », *Business Process Management*, Lecture Notes in Computer Science, vol. 1806, p. 30-49, 2000.

[BER 07] BERTI-EQUILLE L., Quality Awareness for Data Managing and Mining, Habilitation à diriger des recherches, Université de Rennes 1, France, 2007.

[CAN 05] CANFORA G., GARCÍA F., PIATTINI M., RUIZ F., VISAGGIO C.A., « A family of experiments to validate metrics for software process models », *Journal of Systems and Software*, vol. 77, n° 2, p. 113-129, 2005.

[CIM 02] CIMATTI A., CLARKE E.M., GIUNCHIGLIA E., GIUNCHIGLIA F., PISTORE M., ROVERI M., SEBASTIANI R., TACCHELLA A., « NuSMV 2 : An OpenSource Tool for Symbolic Model Checking », *Proc. of the Intl. Conf. on Computer Aided Verification*, vol. 2404 de Lecture Notes in Computer Science, Springer, p. 359-364, 2002.

[DEM 86] DEMING W.E., « Out of the Crisis », *MIT Center for Advanced Engineering Study*, 1986.

[DEG 94] DEGENNE A., FORSÉ M., *Les réseaux sociaux. Une analyse structurale en sociologie*, Armand Colin, Paris, 1994.

[DEF 90] DE FEO J.A., BARNARD W., *JURAN Institute's Six Sigma Breakthrough and Beyond – Quality Performance Breakthrough Methods*, McGraw-Hill Professional, New York, 2005.

[DAV 98] DAVENPORT T.H., PRUSAK L., *Working knowledge : How organizations manage what they know*, Harvard Business School Press, Boston, MA, 1998.

[DAV 90] DAVENPORT T.H., SHORT J.E., « The New Industrial Engineering : Information Technology and Business Process Redesign », *Sloan Management Review*, p. 11-27, 1990.

[ENG 99] ENGLISH L.P., *Improving datawarehouse and business information quality : methods for reducing costs and increasing profits*, John Wiley & Sons, New York, 1999.

[ERI 00] ERIKSSON H., PENKER M., *Business Modeling with UML – Business Patterns at Work*, John Wiley & Sons, New York, 2000.

[GAR 05] GARTNER NOTE NUMBER, Business Process Management's Success Hinges on Business-Led Initiatives Michael James Melenovsky Source : G00129411, 26 juillet 2005.

[GEN 08] GENERO M., POELS G., PIATTINI M., « Defining and validating metrics for assessing the understandability of entity-relationship diagrams », *Data Knowl. Eng.*, vol. 64, n° 3, p. 534-557, 2008.

[GRI 10] GRIM-YEFSAH M., ROSENTHAL-SABROUX C., THION-GOASDOUÉ V., « Changing Provider in an Outsourced Information System Project : Good Practices for Knowledge transfer », *Proceedings of the International Conference on Knowledge Management and Information Sharing* (KMIS), 2011.

[GRI 11] GRIM-YEFSAH M., ROSENTHAL-SABROUX C., THION-GOASDOUE V., « Evaluation de la qualité d'un processus métier à l'aide d'informations issues de réseaux informels », *Revue des Sciences et Technologies de l'Information, série Ingénierie des Systèmes d'Information* (RSTI série ISI), vol. 15, n° 6, p. 63-83, Lavoisier, Paris, 2010.

[GEM 03] GEMINO A., WAND Y., « Evaluating modeling techniques based on models of learning », *Commun. ACM (CACM)*, vol. 46, n° 10, p. 79-84, 2003.

[HAS 09] HASSAN N.R., « Using Social Network Analysis to Measure IT-Enabled Business Process Performance », *Journal of Information Systems Management*, vol. 26, p. 61-76, 2009.

[HMR 08] HERAVIZADEH M., MENDLING J., ROSEMANN M., « Dimensions of Business Processes Quality (QoBP) », *Proceedings of Business Process Management Workshops*, p. 80-91, 2008.

[ISO 00] ISO 9001:2000 INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, « Systèmes de Management de la Qualité – Principes Essentiels et Vocabulaire », http://www.iso.org/, 2000.

[LAC 93] LACITY M.C., HIRSCHHEIM R.A., « The Information Systems Outsourcing Bandwagon », *Sloan Management Review*, vol. 35, n° 1, p. 73-86, 1993.

[LAU 06] LAUE R., GRUHN V., « Complexity Metrics for business Process Models », *Proceedings of the International Conference on Business Information Systems (BIS)*, p. 1-12, 2006.

[LAU 11] LAUE R., GADATSCH, A., « Measuring the Understandability of Business Process Models – Are We Asking the Right Questions ? », *Business Process Management Workshops*, p. 37-48, 2010.

[LEE 01] LEE J.-N., « The impact of knowledge sharing, organizational capability and partnership quality on IS outsourcing success », *Journal Information and Management*, vol. 38, n° 5, p. 323-335, 2001.

[LIN 94] LINDLAND O.I., SINDRE G., SØLVBERG A., « Understanding Quality in Conceptual Modeling », *IEEE Software*, p. 42-49, 1994.

[MAK 10] MAKNI L., KHLIF W., ZAABOUB HADDAR N., BEN-ABDALLAH H., « A Tool for Evaluationg the Quality of Business Process Models », *Proceedings of ISSS and BPSC*, p. 230-242, 2010.

[MEN 08] MENDLING J., STREMBECK M., « Influence factors of understanding business process models », *Proceedings of the International Conference on Business Information Systems (BIS)*, p. 142-153, 2008.

[MEN 10] MENDLING J., REIJERS H.A., VAN DER AALST W.M.P., « Seven process modeling guidelines (7PMG) », *Information & Software Technology,* vol. 52, n° 2, p. 127-136, 2010.

[MOO 05] MOODY D.L., « Theoretical and practical issues in evaluating the quality of conceptual models : current state and future directions », *Data Knowl. Eng.*, vol. 55, p. 243-276, 2005.

[MOR 07] MORLEY C., HUGUES J., LEBLANC B., HUGUES O., *Processus Métiers et systèmes d'information : Evaluation, modélisation, mise en œuvre*, 2e édition, Dunod, Paris, 2007.

[NON 91] NONAKA I., « The knowledge-creating company », *Harvard Business Review*, vol. 69, n° 6, p. 96-104, 1991.

[PAR 92] PARSONS J., WAND Y., « Guidelines for Evaluating Classes in Data Modeling », *Proc. of the International Conference on Information Systems (ICIS)*, p. 1-8, 1992.

[PER 06] PERALTA V., BOUZEGHOUB M., « Data Freshness Evaluation in Different Application Scenarios », *Revue Nouvelles Technologies de l'information (RNTI)*, vol. E5, p. 373-378, 2006.

[PIP 02] PIPINO L.L., LEE Y.W., WANG R.Y., « Data quality assessment », *Commun. ACM 45*, p. 211-218, 2002.

[R] R web site, http://www.r-project.org/.

[RED 97] REDMAN T.C., *Data Quality for the Information Age*, 1re édition, ArtechHouse, Norwood, MA, 1997.

[RED 98] REDMAN T.C., « The impact of poor data quality on the typical enterprise », *Communications of the ACM,* vol. 41, n° 2, p. 79-82, 1998.

[SAN 10] SÁNCHEZ-GONZÁLEZ L., GARCÍA F., MENDLING J., RUIZ F., PIATTINI M., « Prediction of Business Process Model Quality Based on Structural Metrics », *Proceeding of International Conference on the Entity-Relationship (ER)*, p. 458-463, 2010.

[SAS] SAS web site, http://www.sas.com/.

[SHE 80] SHEWHART W.A., *Economic Control of Quality of Manufactured Product/50th Anniversary Commemorative Issue*, American Society for Quality, 1980.

[VAN 07] VANDERFEESTEN I., CARDOSO J., REIJERS H.A., VAN DER AALST W., « Quality Metrics for Business Process Models », *Proceedings of BPM and Workflow Handbook*, p. 179-190, 2007.

[WAH 05] WAHL T., SINDRE G., « An Analytical Evaluation of BPMN Using a Semiotic Quality Framework », *Proceedings of CAiSE Workshops*, vol. 1, p. 533-544, 2005.

[WAN 00] WANG R., ZIAD M., LEE Y., *Data Quality*, Kluwer Academic Publishers, Boston, 2000.

[WAN 90] WAND Y., WEBER R., « An Ontological Model of an Information System », *IEEE Trans. Software Eng.*, vol. 16, n° 11, p. 1282-1292, 1990.

[WAS 94] WASSERMAN S., FAUST K., et Anonyme, « Social Network Analysis : Methods and Applications », *Structural analysis in the social sciences*, Cambridge University Press, New York, 1994.

[WHI 04] WHITE S., Business Process Modeling Notation (BPMN), Version 1.0, www.bpmi.org, 3 mai 2004.

[WOH 06] WOHED P., VAN DER AASLT W.M.P, DUMAS M., TER HOFSTEDE A.H.M., RUSSELL N., « On the Suitability of BPMN for Business Process Modelling », *Business Process Management,* p. 161-176, 2006.

Chapitre 8

L'excellence des données : valorisation et gouvernance

8.1. Introduction

Les entreprises obéissent à des impératifs mêlant efficacité immédiate et valorisation financière à long terme. Mais ces objectifs ne semblent pas totalement alignés avec la gestion des données de référence ou suivant le terme anglais couramment employé *master data management* (MDM).

Existe-t-il néanmoins une solution afin de transformer l'illusion MDM en rêve éveillé ? Cherchons la réponse du côté d'un concept émergent : l'excellence des données.

Alors que volume et dissémination des données ne cessent d'augmenter, il devient très difficile de gérer ces « montagnes », surtout d'un point de vue qualitatif.

Bonne ou mauvaise, la qualité des données demeure un facteur déterminant dans la course à l'Excellence et les entreprises doivent la considérer comme une arme à double tranchant.

8.1.1. *Contexte général*

Jamais le monde n'a changé aussi vite, en particulier dans le domaine des flux de données multicanaux. Le passage de l'ère industrielle à celle de l'information, les défis financiers et économiques sont déjà en train d'impacter notre comportement et,

---

Chapitre rédigé par Walid EL ABED.

de ce fait, nous réclamons plus d'intelligence dans les données afin d'orienter nos actions ou nos décisions. La survie d'une entreprise dépendra de son agilité, élément fondamental de sa pérennité et de sa croissance. Nous croyons vraiment que, sans gouvernance de données, l'agilité de l'entreprise ne peut pas être possible. Notre société consomme de plus en plus de données, leur qualité et leur gouvernance deviennent fondamentales afin de supporter la création de valeur et de permettre la croissance. Par la suite, il est impératif de définir les cadres de travail, les modèles, les méthodologies, les outils, et les plateformes technologiques afin d'assurer un bon niveau de cohérence et de confidentialité des données. Le partage de l'information et des données devient nécessaire afin de favoriser la collaboration et l'optimisation de la chaîne de valeur. Il est indispensable cependant de protéger les politiques de confidentialité au-delà du pare-feu de l'entreprise afin d'augmenter la confiance entre partenaires. De plus, avec une gestion des données valorisée, le coût total de la possession de données peut être particulièrement allégé.

Dans un monde dirigé par les données, il en va de la responsabilité des entreprises de disposer de tels cadres de travail, de préserver, de tester et de soutenir la valeur globale tout au long de la chaîne de l'économie.

Nous allons constater que ces entreprises vont continuer de gagner des affaires grâce à leurs compétiteurs, car elles sont focalisées davantage sur la capture de leur plein potentiel en maximisant un meilleur service au client et dans le même temps en optimisant l'exécution de leurs processus métier. Elles reconnaissent de concert, que le service au client et tous les processus métier dépendent de la façon dont les données se calquent à l'objectif. Ces entreprises, comme de plus en plus de sociétés, bénéficient du retour de leurs investissements dans des données d'entreprise de très haute qualité.

En comparaison, les organisations qui ne prêtent pas suffisamment attention à la qualité des données finissent avec des facturations retardées, des processus d'encaissement ralentis, des ventes inefficaces, des processus de production et des services non optimaux et des achats non maitrisés. S'ajoute à cela une confiance émoussée dans les rapports résultant de prises de décisions infructueuses. Le service client se trouve donc impacté et la confiance du client amoindrie.

Alors, que faire dans l'immédiat ? Il s'agit d'abord de :

– commencer par mesurer la conformité des données aux objectifs et aux exigences métier ;

– faire un audit de valorisation tangible de la qualité des données critiques supportant ces exigences métier ;

– lier les indicateurs qualité de données résultantes à la valeur et à l'impact métier, c'est-à-dire quels coûts et quels risques allez-vous engager et prendre à cause de cette mauvaise qualité des données ?

– calculer la valeur métier que vous pouvez générer avec une haute qualité des données. (Un résultat rapide et générateur de valeur sera la surprise garantie. Des mesures rapides peuvent être entreprises en se concentrant immédiatement sur les activités opérationnelles les plus critiques. En se basant sur une demande d'amélioration au cours des trois prochains mois, miser sur les succès rapides et changer la culture vers l'approche « gestion des données par la valeur et l'impact de leurs utilisations ») ;

– pérenniser l'élan ;

– lier toute initiative de données aux objectifs stratégiques, définir les exigences et les règles métier critiques supportant ces objectifs ;

– identifier les responsables métier de ces exigences ;

– identifier les données supportant les exigences et règles métier ;

– mesurer la conformité de ces données aux exigences et règles métier ;

– valoriser la qualité (valeur) et la non-qualité (impact et risque) par rapport aux transactions métier utilisatrices de ces données ;

– prioriser la correction des données non conformes par rapport à l'impact ;

– identifier les responsables des données défectueuses pour action.

Ainsi le patrimoine de données sera maximisé. Il s'agit de la véritable gouvernance des données par leur valeur et impact.

Gouverner les données n'est pas une initiative optionnelle, mais un réel programme de valeur ajoutée et de gestion de risques. C'est pourquoi nous croyons que seules les organisations dotées d'une approche de gouvernance des données efficace et intégrée dans la stratégie de l'entreprise, seront en mesure de transformer ces données en un réel avantage concurrentiel, leur apportant ainsi une valeur à court et à long terme tout en leur assurant la réussite et la pérennité.

8.1.2. Tempus fugit[1]... *le défi*

Les programmes de modernisation des métiers se concentrent généralement sur la standardisation des processus afin d'en obtenir des bénéfices mesurables et avec une efficacité industrialisable. Les technologies d'ERP (progiciels pour gestion des ressources de l'entreprise et d'automatisation des processus métier) remplissent les exigences de standardisation des processus et sont aujourd'hui devenues un point central pour la gestion des processus métier. Cependant, les systèmes ERP

1. Le temps s'enfuit...

n'empêchent pas la mauvaise qualité des données de pénétrer dans les systèmes et ne mesurent pas son impact sur l'efficacité d'un processus métier. Aujourd'hui, la plupart des organisations ont les mêmes systèmes ERP (SAP, Oracle, Dynamics AX...) configurés par les mêmes consultants. Par conséquent, la spécificité et la portée de l'avantage concurrentiel d'une organisation sont aujourd'hui définies par les individus et les données.

Le « master data management » (*MDM* ou gestion des données de référence) peut faire partie de la solution à plus long terme, cependant il ne livrera pas de valeur sur le court terme. Le défi est dans le paradoxe du temps, où la priorité du métier est d'exécuter les opérations et résoudre les problèmes critiques pour apporter une valeur immédiate tandis qu'un projet de MDM est une initiative stratégique axée sur la fourniture de valeur à plus long terme. Dès lors, comment pouvons-nous concilier le proche et le lointain ?

Seule une approche focalisée, non invasive, progressive et tenant compte des objectifs à court et à long terme sera acceptée et exécutée par les métiers. Notre objectif est de présenter une méthode exécutable favorisant une stratégie de gouvernance des données efficace, durable et offrant une valeur immédiate tout en soutenant la pérennité du processus d'amélioration.

8.1.3. *Les obstacles à surmonter*

Des données de haute qualité sont nécessaires pour exploiter pleinement le potentiel de l'entreprise et offrir tous les avantages des nouveaux systèmes (ERP, CRM, BI, conformité, MDM, *datawarehouse*, etc.). Inversement, des données de mauvaise qualité entraînent les transactions, les processus et les projets à l'échec, conduisant à une augmentation des coûts et des risques, réduisant la confiance dans les données de l'entreprise et entrainant éventuellement une perte d'affaires. Les nouveaux systèmes ERP sont souvent perçus à tort comme une solution à la qualité et à la gouvernance des données compte tenu de leur rôle central pour les capturer et les maintenir. Mais les systèmes ERP n'ont jamais été conçus pour gérer la qualité des données, d'autant plus que toutes les données de l'entreprise ne résident pas dans ces systèmes ERP. La responsabilité de la qualité des données repose souvent sur une équipe de gouvernance des données ou sur une équipe de gestion des données qui agit uniquement à titre consultatif, qui rapporte la plus part du temps à l'informatique et qui a très peu de pouvoir ou d'influence sur les budgets de l'entreprise. Le budget et les ressources de toute grande organisation passant sur un nouvel ERP sont généralement épuisés par l'acquisition et la mise en œuvre des applications : moins d'attention est de fait accordée aux processus de pérennisation de la qualité des données à travers l'entreprise.

De ce fait, les personnes en charge du budget comprennent de manière anecdotique que la qualité des données aboutit à l'efficacité des processus et à une confiance accrue dans l'entreprise. Il n'y a souvent pas de compréhension :

– de la justification du coût ou de la valeur des intendants aux données (*data stewards*) ;

– des processus ou des rôles axés sur la gestion des données en tant qu'actifs ;

– des impacts métier et du potentiel de création de valeur d'une équipe dédiée à la qualité des données ;

– de la justification du coût d'un système de gouvernance de données.

En conséquence, la plupart des organisations ont de la difficulté à construire des initiatives sur la qualité et la gouvernance des données, à mettre en œuvre et à gérer continûment des processus de qualité des données.

8.2. L'excellence des données entre en scène

Revenons à la question du CRM (gestion de la relation client). Le consommateur est le fil rouge du propos.

Si l'on affirme que : chaque métier ne se sent concerné par la qualité des données que lorsqu'elle impacte son processus et ses transactions : commandes, facturation, achats, crédits, prospection, etc. Comment faire pour que la qualité des données soit parlante aux yeux de chaque chef de service, lui même référent auprès du service consommateur ?

C'est dans cette optique que nous allons décrire la méthode de l'excellence des données (MED) qui se veut adaptée aux réalités de l'entreprise, à ses métiers et à ses priorités.

8.2.1. *MED : trois forces, trois étapes, un seul objectif*

Pour chacune des phases de la MED, on trouve un « trio de gouvernance » : les référents métiers, les gestionnaires de données et les informaticiens.

Leur coordination devient possible par l'intervention des *data stewards* afin de nourrir un objectif commun. Fruit d'une démarche méthodologique, MED s'appuiera sur les forces vives de l'entreprise.

En premier lieu, il s'agit d'aligner les impératifs métiers et les projets informatiques avec la gestion des données. Notre trio de gouvernance vise le même objectif, défini par les impératifs du métier. Il va de soi que leur action doit être reliée et supportée par les exigences d'excellence dans les métiers (EEM) que sont les pré-requis, les politiques, les lois, les règles métier, les bonnes pratiques et les standards. Mesurer et visualiser la valeur et l'impact des données constitue le point suivant : l'essentiel est de relier les chiffres de la qualité des données à une valeur pécuniaire. On observe d'abord la valorisation des processus et des transactions métier avec des données de haute qualité.

Puis, on reconnaîtra facilement les pertes et risques encourus avec des entrées de mauvaise qualité. Pour ce faire, on se base à la fois sur un indice qui montre la cohérence des données face aux règles métier et sur les « indicateurs clé de valeur » (ICV) qui représentent l'aspect tangible de la valeur ou de l'impact. L'originalité de cette méthode est qu'elle garantit une répartition claire des rôles des collaborateurs. Chacun peut ainsi visualiser la valeur et l'impact des données selon ses propres intérêts et responsabilités : zone géographique, services, processus, règles métier, etc. Par exemple : un directeur financier responsable de la facturation sur la zone Asie pourra visualiser la valeur et l'impact des données sur ce processus et pour sa zone.

La troisième étape tend à organiser et exécuter un processus pérennisant l'excellence des données, en fonction des différents points de vue et de l'organisation réelle de l'entreprise : les référents métier qui observent le système à travers le prisme des règles métier (gouvernance), les responsables des données qui réfléchissent en termes d'ERP, CRM, MDM (opérations) et les informaticiens qui agissent en tant que garants du support technique.

Au final, si 80 % des données génèrent de la valeur et une amélioration des processus métier, les 20 % qui impactent négativement ces processus définiront les prochaines priorités de travail.

8.2.2. *Le modèle de maturité de l'excellence des données*

Le modèle de maturité de l'excellence des données suit l'évolution d'une organisation depuis le début de la gouvernance des données, souvent décrite comme la phase « chaotique », jusqu'au stade le plus mature où les données sont utilisées comme un actif de base de l'entreprise et décrite comme la phase « prédictive ». Le modèle de maturité est souvent utilisé pour comprendre la meilleure approche dans les projets et initiatives visant à introduire progressivement les concepts de la méthode d'excellence des données, MED[2], proposée par la société « Global Data

2. La méthode MED est plus couramment connue sous le nom *data excellence framework* ou DEF.

Excellence Ltd. ». L'objectif de ce modèle est de positionner une entreprise en fonction de sa capacité à générer de la valeur à partir des données de l'entreprise. La figure 8.1 montre le modèle de maturité de l'excellence des données.

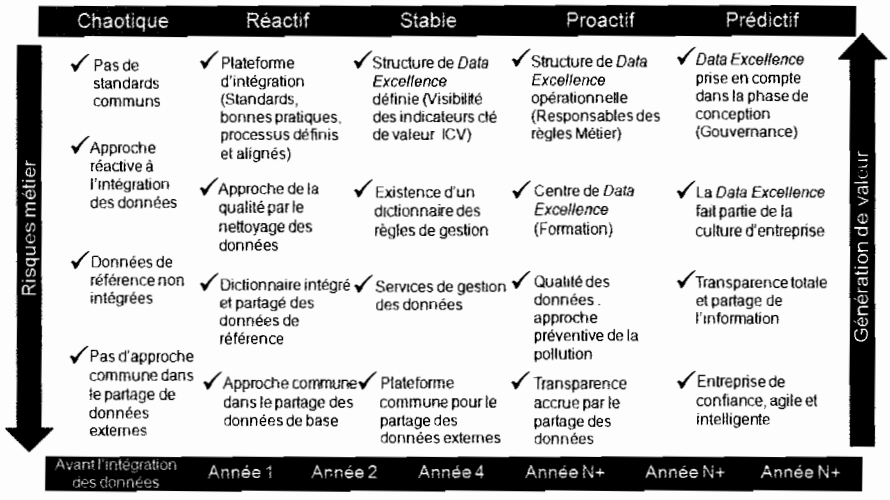

Figure 8.1. *Le modèle de maturité de l'excellence des données*

Le temps pris par les entreprises pour progresser au travers des étapes de manière incrémentale, met en évidence le mauvais alignement que l'on observe souvent entre les objectifs des programmes de gouvernance des données et les objectifs définis par les métiers. La méthode d'excellence des données supporte l'approche dite « proche et lointaine » qui peut alors livrer en parallèle, une valeur immédiate à la vision à long terme attendue par les métiers.

8.3. La méthode de l'excellence des données

MED décrit la méthodologie, les processus et les rôles nécessaires pour maximiser la valeur d'utilisation des données de l'entreprise et optimiser leur coût de gestion. Les processus métier utilisant des données de qualité conforme aux exigences métier et aux règles de gestion bénéficieront d'une efficacité optimale. La méthode prend en charge la création du changement de culture qui sera focalisé sur l'excellence des données, en motivant globalement les équipes et en soutenant la collaboration entre les intervenants. La méthode prend en considération le fait que la solution, bien que simple à énoncer, soit complexe et comprenne plusieurs dimensions. Il ya donc une focalisation forte sur la culture de l'entreprise comme clé d'une solution durable. Une différence fondamentale de MED par rapport à la

plupart des initiatives relatives aux données est que MED est axée sur la génération de valeur tandis que les autres sont conceptuelles ou ne sont axées que sur la réduction des coûts.

La notion de valeur est devenue un concept intangible dans la plupart des entreprises. Pourtant dans notre vie courante, la valeur est constamment présente dans nos prises de décision. Nous nous séparons difficilement d'un objet ou d'un vêtement car nous lui prêtons une valeur sentimentale, nous choisissons avec attention les biens que nous achetons en fonction de l'utilisation ou du plaisir qu'il va nous rapporter. Nous savons intuitivement que ce qui ne nous sert pas n'a aucune valeur, même si cela peut avoir un prix ou un coût. Il convient donc de faire la distinction entre prix, coût et valeur.

Le prix d'un objet est le montant affiché pour acquérir l'objet. Il est le plus souvent fixe, connu et égal (objectif) pour tous. On peut parfois le négocier en fonction des circonstances (lieu, quantité).

Le coût d'un objet est le prix de revient total de fabrication ou d'utilisation. L'exemple du coût d'utilisation d'une automobile est parlant : alors que le prix mesure la somme que l'on doit payer pour l'acquérir, le coût représente l'ensemble des dépenses attachées à son utilisation. Le coût est également objectif.

La valeur de l'objet est quant à elle totalement subjective : elle dépend de ce que l'on va faire avec l'objet. Une bouteille d'eau au supermarché du coin ne vaut pas la même chose qu'une bouteille d'eau dans le désert. Dans le premier cas, sa valeur est très proche de son prix affiché voire nulle si je n'en ai pas besoin, dans le dernier cas sa valeur est inestimable (celle de ma vie, car sans elle, ma vie ne vaut plus rien). La valeur est donc contextuelle et elle s'instancie lors de la transaction.

L'entreprise a depuis longtemps séparé la génération de valeur et la gestion des coûts, sans doute depuis la rationalisation du travail (Taylor). On a donc cantonné la gestion des coûts dans les fonctions de support à qui on demande de rationaliser et d'optimiser les processus donc, en gros, d'en fournir plus pour moins cher. Et on a confié la création de valeur au « business ». Cette séparation a déconnecté les moyens (processus) et la création de valeur et a « dévalorisé » en quelque sorte les fonctions de support.

MED est l'unique méthode disponible sur le marché permettant la gouvernance des données suivant le contexte et la valeur tangible de leurs utilisations. Néanmoins nombreuses sont les méthodes qui adressent les problématiques de gouvernance de qualité et de gestion des données de façon exhaustive mais indépendamment de l'usage et du contexte. Pour n'en citer que quelques-unes des plus connues, nous pouvons nous référer à la méthodologie de Larry English : « *Total Quality data*

Management » qui est devenue « *Total Information Quality Management* » [ENG 99]. Nous pouvons également citer l'approche pratique décrite par Danette McGilvray : « *Executing Data Quality Projects : Ten Steps to Quality Data and Trusted Information* » [MGI 09]. Quant à la gouvernance des données, la méthodologie la plus connue et celle de Gwen Thomas : « *The DGI Data Governance Framework* » [THO 11]. Nous pouvons aussi citer : « *The Non-Invasive Data Governance : Implementing Data Governance in a Non-Threatening Way* », de Robert S. Seiner [SEI 11]. D'autres tentatives intégrant la notion de valeur des données et sa gestion comme un actif d'entreprise ont été élaborées sans pour autant réussir à formaliser un modèle de valorisation organisée de la donnée permettant la systématisation ou la justification de sa gouvernance. En effet, ces tentatives se trouvent prisonnières de la valeur intangible ou difficile à déterminer des données alors que son prix et son coût de gestion sont plus faciles à appréhender. Parmi les ouvrages de référence, nous pouvons évoquer : « *Data Driven : Profiting from Your Most Important Business Asset* » de Thomas C. Redman [RED 08] et « *The Data Asset : How Smart Companies Govern Their Data for Business Success* » de Tony Fisher [FIS 09]. Enfin, certaines méthodologies visent à résoudre la problématique de gouvernance et de qualité des données par l'implémentation des modèles et des systèmes de gestion des données de référence (*MDM*). Nous citons à titre d'exemple l'ouvrage « *management des données d'entreprise – master data management et modélisation sémantique* » de Pierre Bonnet [BON 09]. L'éventail des ouvrages et des méthodes traitant de la problématique des données sont très nombreux et nous n'avons pas l'ambition de mener une étude comparative exhaustive ; c'est pourquoi nous nous sommes limités à n'en citer que les plus pertinents à nos yeux afin d'illustrer l'originalité de notre méthode MED. MED se caractérise par le postulat de base qui stipule d'une part, que la valeur des données est dynamiquement dérivée de la valeur générée par son usage et d'autre part, que la qualité de données doit être liée dynamiquement à cette valeur. Par conséquent, la détermination de la valeur et de la qualité requise des données se fait par l'instanciation de la relation entre des exigences métier des données participant à la génération de cette valeur et le résultat attendu de l'activité suivant le contexte et l'objectif du moment.

La méthode MED repose sur quatre piliers de valeur (figure 8.2) que nous pensons essentiels à la survie de toute organisation ou entreprise dans l'ère de l'information : agilité, confiance, intelligence et transparence. Nous sommes convaincus que ces caractéristiques sont des piliers fondamentaux de la valeur permettant d'assurer la pérennité des entreprises tout en soutenant la croissance économique.

Les quatre piliers (agilité, confiance, intelligence et transparence) soutiennent les impératifs business les plus courants que les dirigeants financent aujourd'hui. L'agilité est nécessaire pour réagir aux changements externes et internes et assurer rapidement une intégration réussie des transformations grâce à l'harmonisation des

processus, des fusions et acquisitions (F&A), des cessions et des réorganisations. La confiance est associée à l'intégrité des données : ainsi les étiquettes sur les denrées alimentaires doivent être exactes – sinon la confiance dans la marque est perdue. Si un produit financier promet un retour sur investissement incorrect, les acheteurs ne font plus confiance à la marque. L'intelligence à tous les niveaux de l'entreprise conduit à une exécution sans faille, à l'efficacité opérationnelle et à la consolidation financière précise basée sur «le juste à temps » de la qualité des données provenant des systèmes d'information et d'applications -globaux et opérationnels. Finalement, les avantages de la transparence ne sont apparus essentiels à la performance de l'organisation que récemment : elle est nécessaire pour accroître la visibilité et la collaboration au sein et en dehors de l'écosystème. La responsabilité sociale des entreprises sera rendue possible par la capacité de partager des données en interne dans l'entreprise et en externe avec des partenaires. Il en résultera de nouvelles façons de travailler et cela permettra d'abaisser davantage le coût d'utilisation des données.

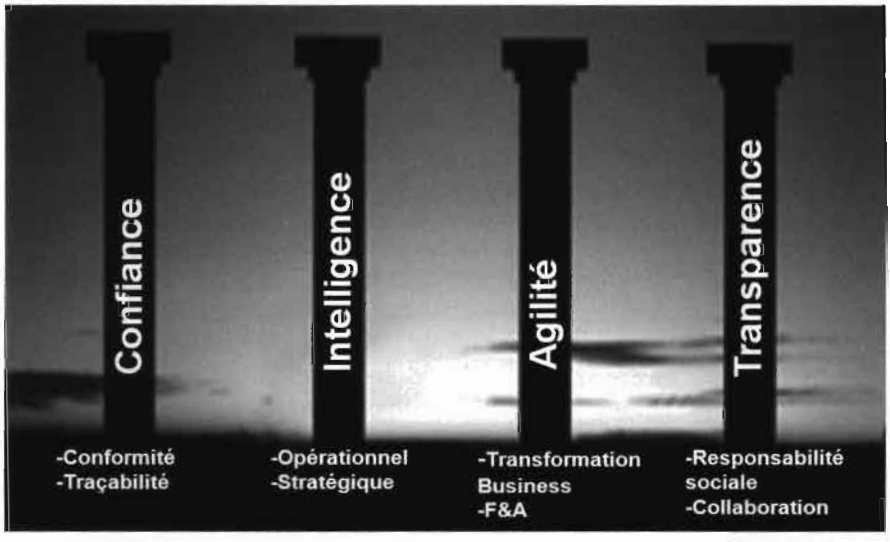

Figure 8.2. *Les quatre piliers de valeur*

Pour qu'une organisation puisse s'orienter vers la valorisation de ses données en tant qu'actif de l'entreprise, elle a besoin de changer sa culture et la façon dont les données sont gérées. Les données ne sont pas détenues par des individus, elles sont la propriété de l'entreprise pour soutenir ses objectifs. L'organisation a besoin de définir des rôles d'intendants des données (*Data steward*) qui prennent à leur compte la responsabilité des règles métier de l'entreprise et la valeur des données.

La figure 8.3 montre notre proposition visant à cultiver les données de l'entreprise comme un actif et illustre le changement de comportement et de mentalité qui sont nécessaires pour progresser vers la pérennité de l'excellence du métier.

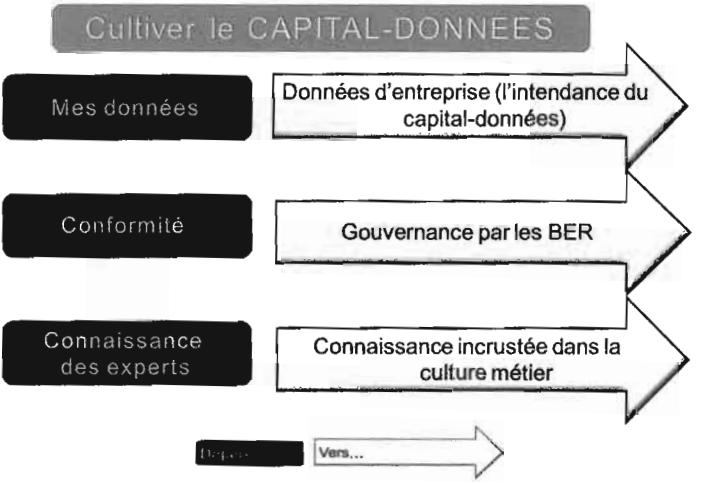

Figure 8.3. *La proposition de gestion du changement*

8.3.1. *Les dimensions de la qualité des données*

La méthode d'excellence des données MED définit sept dimensions pour mesurer la qualité des données. Nous pouvons imaginer d'autres dimensions de qualité néanmoins nous pensons que toute mesure de qualité peut être classée dans une de ces dimensions. La définition d'un nombre limité de dimensions simplifie la présentation et la gestion de la qualité des données et facilite sa mesure et son contrôle. La qualité des données peut être liée à un processus métier à travers l'identification et la mesure de la conformité des données avec les règles métier de base qui sont essentielles pour que ce processus réussisse. Le tableau 8.1 donne la définition des sept dimensions de la qualité des données retenues dans MED.

La spécificité des dimensions de qualité définies dans MED est que ces dimensions sont dynamiques car elles sont définies par les différentes collections de règles métier valides à un moment donné. Ceci permet d'augmenter les exigences de qualité progressivement en ajoutant des règles métier à la collection de base. Ainsi, le seuil acceptable de qualité des données peut être ajusté suivant le contexte et les exigences des processus et des transactions métier.

Dimension	Description
Spécificité (*Uniqueness*)	La dimension de la spécificité est la collection des règles métier qui permet l'identification d'une entité de façon déterministe, d'une relation ou d'une instance d'événement dans un contexte défini permettant d'exécuter un processus métier spécifique.
Complétude (*Completeness*)	La dimension de la complétude est la collection des règles métier qui garantit que toutes les données requises pour une exécution réussie d'un processus dans un domaine défini soient présentes dans la base de données.
Exactitude (*Accuracy*)	La dimension de l'exactitude est la collection des règles métier qui prouve que les données reflètent la réalité à travers un contexte et un processus défini.
Non-obsolescence	La dimension de la non-obsolescence est la collection de règles métier qui garantit que les données requises pour exécuter un processus défini dans un contexte spécifique soit actuelles et à jour.
Cohérence (*Consistency*)	La dimension de la cohérence est la collection des règles métier requises pour assurer que les valeurs des données sont fournies « bonnes dès la première fois » dans toutes les bases de données et systèmes pour l'exécution d'un processus métier spécifique dans un contexte défini.
Rapidité (*Timeliness*)	La dimension de la rapidité est la collection de règles métier qui garantit la livraison des données « bonnes dès la première fois » requises pour faciliter l'exécution des processus métier et remplir les accords de niveau de services.
Accessibilité (*Accessibility*)	La dimension de l'accessibilité est la collection des règles métier assurant que les personnes, les systèmes et les processus ont accès aux données selon leurs rôles et leurs responsabilités.

Tableau 8.1. *Les dimensions de la qualité des données*

8.3.2. *Exigences d'excellence dans les métiers (EEM)*

La méthode MED définit les exigences d'excellence dans les métiers (EEM) comme des pré-requis, des règles métier, des standards, des politiques ou des bonnes pratiques auxquels les processus, les transactions ou les données doivent se conformer afin que les objectifs du métier soient atteints sans erreurs et génèrent de la valeur. Pour chaque objet de données (client, fournisseur, banque, matériel, actif, emplacement, etc.) et pour chaque dimension de la qualité des données, un ensemble spécifique de règles métiers doit être identifié, documenté et géré. L'objet des données doit toujours être lié au contexte et aux processus métier lors de la définition d'une règle métier qui s'y rattache. Il est important d'adopter une approche pragmatique pour la gestion de la qualité des données et ainsi se concentrer sur un plus petit nombre d'exigences et règles métier critiques, plutôt que de chercher à atteindre le 100 % de qualité en conformant toutes les données de

l'organisation à l'ensemble exhaustif des exigences et des règles métier. Le 100 % qualité ne pourra jamais être atteint ou requis pour toutes les données de l'entreprise. Nous recommandons que le niveau optimal de la qualité des données soit impérativement ciblé afin d'optimiser la valeur métier et éviter les retards. L'ensemble des règles de gestion soutenant la qualité des données augmente au fil du temps dans le cadre du processus d'amélioration continue. La figure 8.4 illustre un exemple anecdotique de la manière dont une règle de gestion se traduit en impact et en valeur métier.

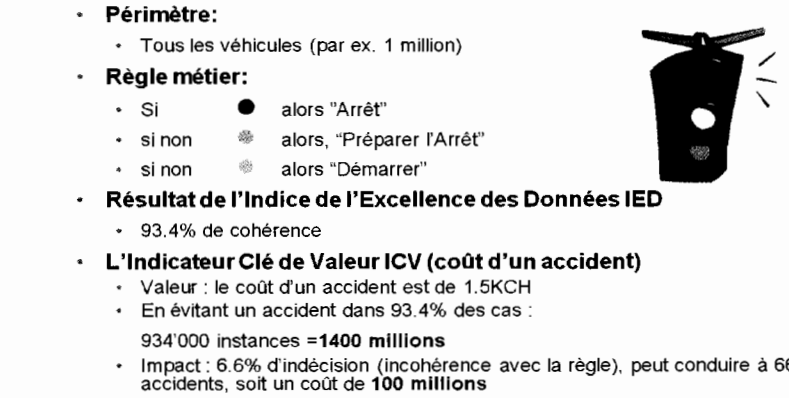

Figure 8.4. *Règle de gestion, Périmètre, IED et ICV*

Des exemples de quelques règles de gestion usuelles sont présentés :

– un enregistrement commande client doit disposer d'un code produits non-obsolète ;

– un enregistrement client doit avoir un « score de crédit » en cours pour qu'une commande soit transformée ;

– un dossier de client final doit avoir une date de naissance valide pour que cet enregistrement puisse être inclus dans les campagnes de marketing où l'âge est un critère impératif ;

– l'adresse de courriel doit être remplie pour toute une série de campagnes de marketing *via* Internet ;

– tous les enregistrements de banque doivent avoir le code ISO du pays dans les 5e et 6e rangs du code SWIFT pour sa validation ;

– le code de la devise doit être compatible avec le code pays afin de valider la cohérence.

Lorsque la méthode est appliquée, chaque règle de l'entreprise doit être attribuée à un propriétaire afin d'identifier l'individu responsable qui assurera que la règle est correcte et appliquée. Le tableau 8.2 montre quelques exemples de règles métier supplémentaires. Chacune des règles de gestion doit être liée aux transactions métier appropriées en vue d'évaluer et de quantifier une valeur métier tangible ou une valeur de risque.

Domaine	Règle métier	Indicateur clé de valeur (ICV)	
		Impact business	Valeur business
Ventes	Les prix doivent être consistent avec le type de client	Prix erronés fournis au client	Diminution des réclamations des clients
Marketing	Les adresses doivent être précises	Pertes dues aux frais postaux et au conditionnement	Améliorer l'efficacité du marketing
Finance - Conformité	Les clients doivent avoir une limite de crédit correcte	Non-conformité et risque dans l'octroi de crédits	Confiance dans le processus de contrôle de crédit
Finance	Les conditions de paiement doivent refléter celles du contrat	Les objectifs de délais de payement ne sont pas atteints	Amélioration de la gestion du cash
Achat	Les conditions de paiement du fournisseur doivent concorder avec le type de fournisseur	Les paiements ne sont pas conformes au contrat	Amélioration du processus business
Sécurité	Les informations sur les risques d'allergies possibles d'un produit sont complètes	Vie du consommateur	Confiance du consommateur

Tableau 8.2. *Exemple de règles métier*

8.3.3. *Indice d'excellence des données et indicateurs-clés de valeur*

L'indice d'excellence des données IED ou le « data excellence index » est une des principales réalisations attendues de la méthode. L'IED est une mesure du pourcentage d'enregistrements qui exécutent avec succès les règles métier ainsi qu'une liste d'enregistrements qui ont violé ces règles métier. L'IED est la base pour le calcul de l'impact métier et de la valeur générée.

L'indice est calculé de manière unique dans le sens où chaque enregistrement ne peut impacter l'index qu'une fois par violation des règles métier. Cela permet à l'IED de mesurer « la santé » de l'organisation d'un point de vue qualité des données en dépit du fait que les mesures couvrent l'ensemble objets de données (par exemple, l'IED mesure la santé de l'entreprise en termes de capacité à fonctionner globalement suivant les exigence du contexte, de la même façon qu'un

thermomètre mesure la santé de l'organisme en tenant compte des différentes parties qui le composent comme le bras, les jambes, la tête et le torse).

L'IED mesure la qualité du contenu des données, leur conformité aux règles de fonctionnement et aux règles d'accessibilité, ainsi que le temps pris pour leur livraison (rapidité) et les processus de leur gestion permettant l'exécution fluide et sans faille des processus métier et des transactions. L'accent est surtout mis sur la promotion d'une culture où chaque activité doit être accomplie conformément aux exigences du contexte et sans faille, « bon du premier coup ».

Les indicateurs-clés de valeur, ICV, connus sous le nom anglais *key value indicators* (KVI) sont un livrable clé de la méthode MED. Un ICV est la mesure de la valeur et de l'impact de l'IED sur les fonctions de l'entreprise :

– la valeur est liée à un ensemble d'éléments de données de l'indice d'excellence de données IED qui est conforme aux exigences et règles métier ;

– l'impact est lié à un ensemble d'éléments de données erronées de IED qui n'est pas conforme aux exigences et règles métier.

Un ICV fondamental représente la valeur clé du métier et l'impact découlant d'une seule exigence spécifique du métier. Un ou plusieurs ICV peuvent être liés à un IED spécifique. Nous pouvons identifier quatre macro-indicateurs-clés de valeur, ICV, pour dériver la valeur liée à l'activité opérationnelle de l'entreprise :

– le montant de la vente dérivé du « bon de commande » et de la transaction associée. Cet ICV sera utilisé pour mesurer la valeur et l'impact des données sur l'activité de la vente (de la commande à l'encaissement) ;

– le montant de dépense dérivé du « bon d'achat » et de la transaction associée. Cet ICV sera utilisé pour mesurer la valeur et l'impact des données sur l'activité des achats (de l'approvisionnement au règlement) ;

– le portefeuille produit ou service dérivé du « bon de production ou de service » et de la transaction associée. Cet ICV sera utilisé pour mesurer la valeur et l'impact des données sur l'activité de production ou de service (de la conception à la livraison) ;

– le capital humain (ressources humaines) dérivé du « bon de travail ou de mission » et de la transaction associée, des éléments de coût peuvent être associés à cet ICV tels que la fiche du salaire, les charges et le temps pris pour l'accomplissement du « bon de travail ou mission». La valeur maximale de l'ICV est atteinte lorsque toutes les activités de l'entreprise sont exécutées sans faille conformément aux exigences métier et au contexte. Dans ce cas nous dirons que l'indice de valeur

clé « bon du premier coup » est atteint à 100 %. Cet ICV sera utilisé pour mesurer la valeur et l'impact des données sur l'activité des ressources humaines (de l'embauche à la retraite).

Les quatre macro-indicateurs-clés de valeur, ICV, peuvent être décomposés en multiples micro-ICV relatifs à toute transaction intermédiaire importante dans un processus métier. Ces ICV peuvent être agrégés à chaque niveau de l'organisation de l'entreprise.

Nous pouvons considérer les transactions du planning prévisionnel pour chacun de ces ICV fondamentaux afin d'anticiper le calcul de la valeur et de l'impact permettant une gouvernance par anticipation.

Nous pouvons également identifier trois macro-ICV relatifs à l'activité stratégique de l'entreprise :

- la pertinence des rapports et des tableaux de bord de pilotage ;
- la pertinence de la consolidation financière ;
- la pertinence de la mise en conformité.

Le propos de ces trois ICV est de fiabiliser le support décisionnel et les rapports liés au risque opérationnel et financier (de l'élaboration des rapports à la définition et exécution de la stratégie).

Les caractéristiques et les principes des ICV sont les suivants :

- valeur tangible : un indicateur-clé de valeur doit posséder une valeur tangible pour être considéré comme un actif et être lié aux données par un indice d'excellence de données IED. La définition des objectifs et les mesures de leur réalisation sont des pré-requis pour une réelle gestion des performances. Il est important de savoir ce qui doit être réalisé, comment cela doit être fait et comment il est possible de savoir si l'on est sur la bonne route ;

- intendance : chaque indicateur doit avoir une gouvernance claire par rapport aux fonctions métier et aux niveaux de la structure d'organisation. L'intendant de la fonction doit engager sa responsabilité pour s'efforcer d'améliorer l'indicateur ;

- chaque indicateur doit être basé sur une exigences d'excellence dans les métiers bien définie ;

- adéquation : les indicateurs doivent refléter des facteurs métier réels et importants pour permettre la maîtrise, l'analyse ou l'évaluation des opérations du métier et être une d'une réelle aide à la décision ;

– simplicité et bon sens : les indicateurs doivent avoir une signification univoque. Leur impact sur le métier doit être clairement compréhensible sur la base du sens commun ;

– précision et non-ambiguïté : les valeurs mesurées ne doivent pas laisser de place à l'interprétation ou la négociation des résultats ;

– complétude : les indicateurs doivent décrire tous les aspects qui sont nécessaires pour atteindre les résultats attendus ;

– publication : le calcul des indicateurs et leur présentation aux personnes intéressées doivent être directs et ne doivent pas impliquer une interaction personnelle ;

– limités en nombre : l'ensemble complet des mesures devrait être visualisé en un coup d'œil lorsque l'on évalue les performances.

Les figures 8.5 et 8.6 illustrent deux exemples.

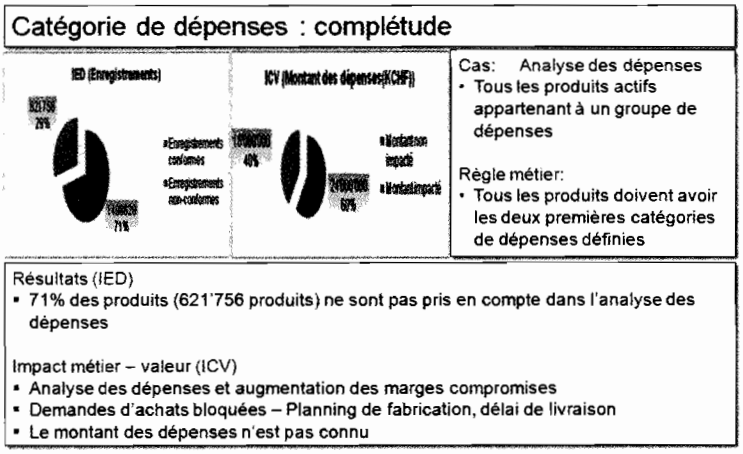

Figure 8.5. *Exemple d'ICV de la règle de gestion des analyses de dépenses*

8.3.4. *Le processus de l'excellence des données*

Une gouvernance des données réussie et durable ne peut pas être atteinte sans des processus communs à travers l'entreprise. Les pratiques et les méthodes communes promeuvent l'exécution des processus d'amélioration continue. Ces méthodologies s'axent sur les analyses des causes premières afin de résoudre l'origine du problème et améliorer les processus métier. Notre proposition est de se

concentrer principalement sur la résolution des problèmes mis en évidence par les ICV impactant la livraison de valeur à court terme, puis de passer à l'analyse des causes fondamentales pour améliorer les processus métier, et idéalement, de mettre en œuvre les processus de prévention de la pollution des données.

Le processus de l'excellence des données vise à aider l'entreprise à accélérer le changement de méthodologie afin d'atteindre l'excellence dans les données et dans les métiers Il se déroule en cinq étapes distinctes montrées dans la figure 8.7.

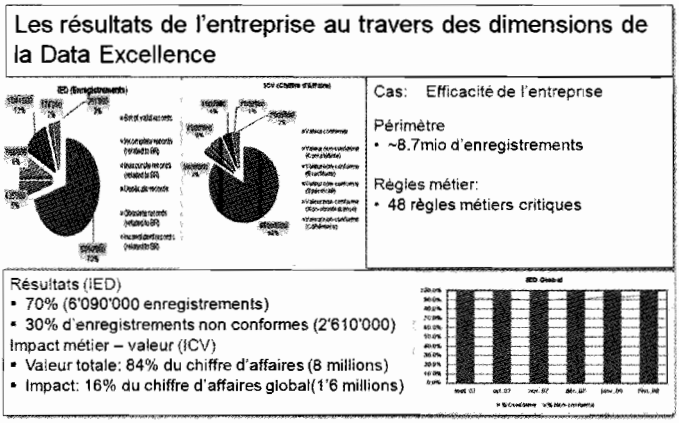

Figure 8.6. *Exemple des résultats de l'entreprise au travers des dimensions de l'excellence des données*

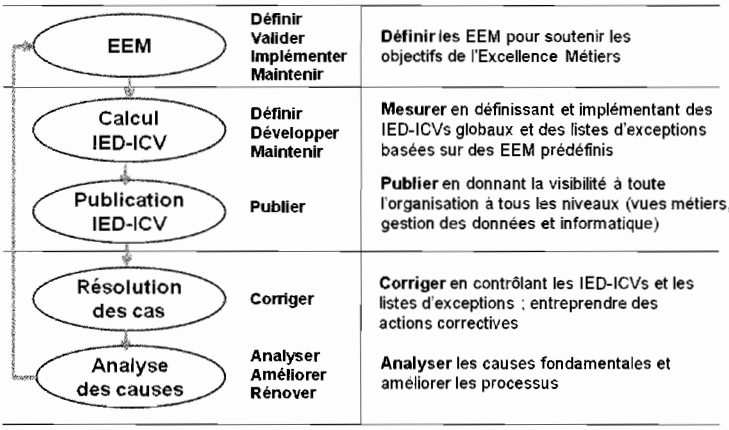

Figure 8.7. *Le processus de l'excellence des données*

8.3.5. *Le modèle de gouvernance de l'excellence des données*

Un des principes directeurs pour l'acceptation et l'application des programmes de gouvernance des données est d'éviter les augmentations substantielles des effectifs en tirant parti des ressources humaines, des outils et des infrastructures actuels. L'approche doit être progressive et non invasive, renforçant ainsi une culture de responsabilisation axée sur les priorités. Pour qu'une organisation s'oriente vers la valorisation de ses données en tant qu'actif de l'entreprise, elle a besoin de faire évoluer sa culture et de changer la façon dont les données sont gérées. Les données ne sont pas détenues par les individus, elles sont la propriété de l'organisation pour soutenir les objectifs de l'entreprise. L'organisation a besoin de définir des rôles d'intendance des données qui prennent la responsabilité des règles métier et des données qui s'y rapportent. Plus précisément, les délégués aux données sont des individus nommés à chaque niveau de l'organisation et de leur localisation. Ces individus prennent la responsabilité :

- des règles de gestion et des indicateurs-clés de valeur (ICV) ;
- des niveaux de qualité des données ;
- des corrections de données.

Les rôles suivants sont définis :

- l'intendant des données (*Data Steward*), individu responsable d'un ensemble de règles métier pour l'entreprise et qui actionne les processus de l'excellence des données ;
- le référent des données (*Data Accountable*), individu qui est le référent dans son domaine pour l'application des règles métier et les objectifs liés aux ICV ;
- le responsable des données (*Data Responsible*), individu qui est responsable de chaque donnée source individuelle liée aux règles métier et des objectifs liés aux IED.

Le modèle de gouvernance est soutenu par une équipe dédiée, le « centre d'excellence des données », qui a pour objectif la mise en place de la méthode d'excellence des données, de ses processus et de son organisation. Ce centre d'excellence sera chargé de faciliter l'exécution des processus de la méthode au sein de l'organisation, d'accompagner les intendants et de maintenir leurs connaissances et leur réseau.

Les figures 8.8 et 8.9 montrent le modèle de gouvernance de l'excellence des données et la répartition des rôles.

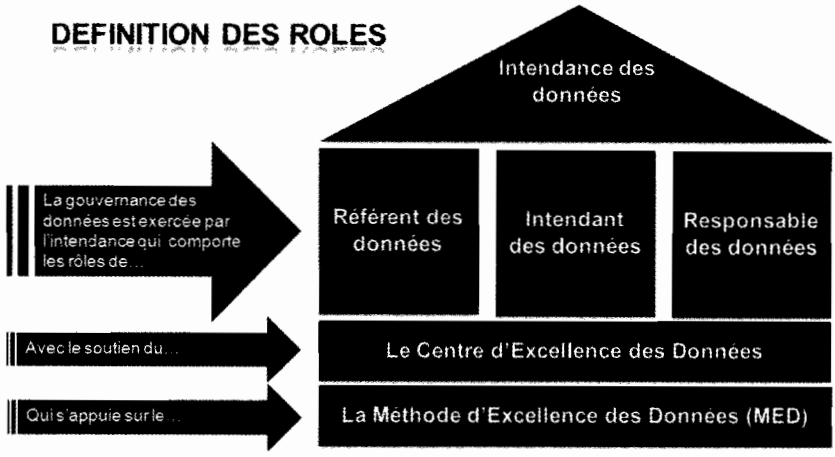

Figure 8.8. *Le modèle de gouvernance de l'excellence des données*

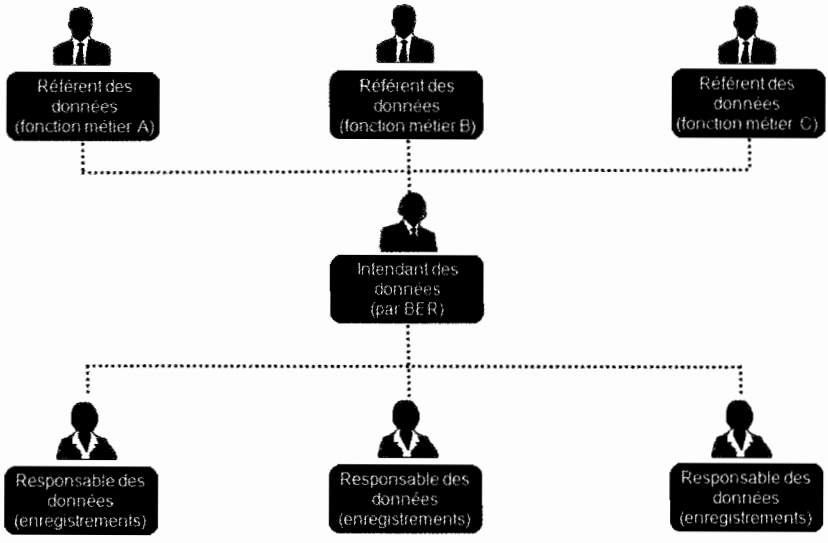

Figure 8.9. *La répartition des rôles*

8.3.6. *Les système de gestion de l'excellence des données*

Il est important d'avoir la vision de l'état final d'une plateforme technologique pour accompagner le voyage vers l'excellence des données, du chaos à la phase prédictive, tout en gardant à l'esprit que son implantation doit être introduite

progressivement pour soutenir l'exécution et la mise en œuvre de la stratégie de MED. L'ajout de chaque composant doit être aligné avec les objectifs métier qui sont présents dans la feuille de route globale du programme d'excellence des données. Par conséquent, nous ne devons pas attendre la mise en œuvre de la plateforme complète sous prétexte que l'organisation ne pourrait pas être prête ou qu'il serait difficile de justifier l'investissement.

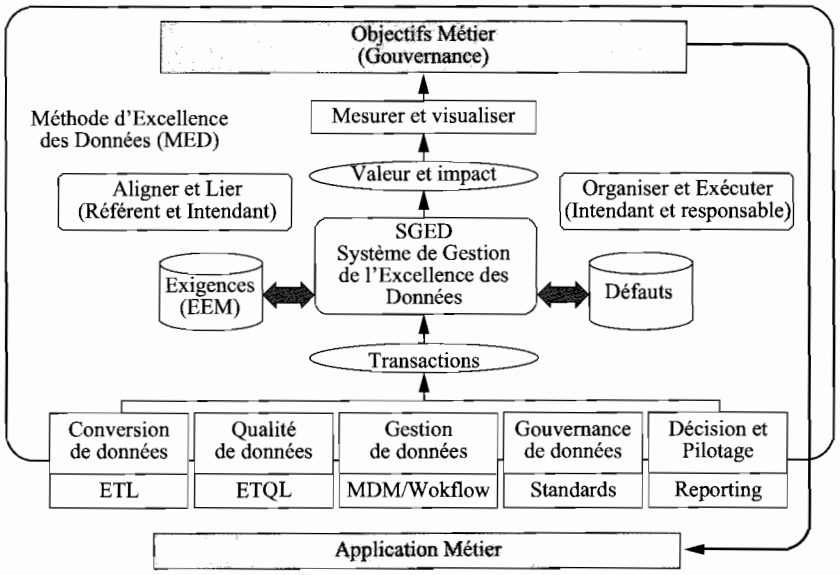

Figure 8.10. *Système de gestion de l'excellence des données*

Au fil du temps le système de gestion de l'excellence des données (SGED) ou « data excellence management system, (DEMS) » (figure 8.10) se verra étendu pour couvrir les différents processus dans le cycle de vie de la donnée. Le système accompagnera tous les projets de données afin de maximiser la valeur de la donnée et optimiser le temps d'exécution des projets – y compris les projets de nettoyage, de conversion, de migration, de standardisation, de MDM, de qualité ou de gouvernance des données.

8.4. Mise en œuvre de l'excellence des données

8.4.1. *Le contexte*

Le système d'information d'une entreprise est souvent complexe et caractérisé par une diversité d'applications et des réplications de données à différents niveaux.

La conséquence est un manque de vision globale des processus et des données, ainsi qu'une difficulté à définir qui est responsable de quoi.

Dans le cadre d'un tel système d'information, un certain nombre de processus ne sont généralement pas totalement maîtrisés. Le niveau de qualité des données entraine de nombreuses corrections *a posteriori* et les résultats obtenus ne sont pas toujours fiables. En conséquence, certains traitements présentent des délais d'exécution incompatibles avec la nécessité de prises de décision rapides et basées sur des données précises que la gestion implique aujourd'hui. Le manque de qualité des données, qui est souvent à l'origine de dysfonctionnements, est difficilement identifiable. D'où des actions correctives de masse coûteuses, alors qu'une intervention ciblée sur les données clés serait plus efficace.

Par ailleurs, l'impact d'un changement de processus ou de données, qu'il soit délibérément décidé ou imposé par un événement externe (par exemple une nouvelle législation), est difficilement mesurable.

Dans cette partie nous allons présenter une démarche progressive afin de simplifier la mise en route de l'excellence des données et d'en réduire le coût tout en générant de la valeur significative et contextuelle (en ligne avec les objectifs stratégiques de l'entreprise).

8.4.2. *Les objectifs de la mise en œuvre*

Les objectifs principaux de la mise en œuvre de l'excellence des données sont :

– mettre en place une structure afin que l'organisation, les processus fonctionnels et les traitements informatisés puissent assurer la qualité des données selon l'approche de la méthode d'excellence des données MED décrite précédemment ;

– définir et implémenter des règles métier pour améliorer et maintenir les différentes dimensions de la qualité à moyen et long terme ;

– définir explicitement les rôles et responsabilités vis-à-vis des processus et des règles métier et mettre en place l'organisation adéquate ;

– fournir des indicateurs-clés de valeur plus fiables et plus rapidement.

Au final, et par le biais d'un accroissement de la qualité des données, permettre à l'entreprise de gagner en : *intelligence, agilité, transparence, confiance.*

8.4.3. *Identification et description des processus et des acteurs*

Le but est de disposer à terme d'un catalogue des processus métier par domaine comprenant :

- les étapes de l'exécution avec une notion temporelle ;
- les activités détaillées de chaque étape ;
- les conditions de l'exécution ;
- les objets de données impactés ;
- l'identification des responsables, des exécutants et des bénéficiaires.

Démarche	Réunions et interviews. Examiner la documentation.
Participant(s)	Responsables métier.
Livrable	Catalogue des processus. Descriptions des processus, des systèmes associés et des objets de données.

Tableau 8.3. *Plan d'identification des exécutants et des bénéficiaires*

8.4.4. *Définition et maintenance des règles de gestion (métier)*

Le but est d'établir un référentiel des règles métier pour soutenir la qualité des données.

Démarche	Identifier les règles. Proposer les règles et les valider. Nommer le propriétaire. Attribuer des priorités de mise en œuvre. Publier les règles. Maintenir les règles. Ecrire les spécifications d'extraction des données.
Participant(s)	Responsables métier, informaticiens.
Livrable	Référentiel des règles métiers (document Excel ou base SGED). Spécification détaillée de l'extraction des données.

Tableau 8.4. *Plan de définition et de maintenance des règles métier*

La définition d'une règle et sa maintenance est déclenchée par une proposition, un nouveau besoin ou une nouvelle contrainte. Une nouvelle règle est soumise à une validation formelle par son propriétaire et par les équipes responsables des processus concernés pour confirmer sa raison d'être et attribuer sa priorité de mise en œuvre. Elle est ensuite stockée dans le référentiel des règles métier avec toutes les informations nécessaires à son implantation. Le référentiel, centralisé, est rendu visible à tous les acteurs de tous les processus.

Dans la mise en œuvre, une spécification détaillée doit être établie par le responsable des données pour préciser les modalités d'extraction des données.

8.4.5. *Définition et production des indicateurs-clés*

Le but est de définir les indicateurs, de développer les programmes nécessaires à leur production et de publier les résultats obtenus, ainsi que produire des listes d'anomalies.

La définition d'un indicateur doit se conformer à un certain nombre de règles :

- avoir un responsable nommé ;
- s'appuyer sur des règles métier définies et des données disponibles ;
- avoir un impact visible sur les affaires ;
- être précis, non ambigu, directement compréhensible ;
- être objectif et ne pas impliquer des interactions de personnes.

Dans un domaine donné, les ICV seront volontairement limités en nombre.

Démarche	Définir les ICV périmètre, critère d'agrégation des données, fréquence de la mesure, résultat attendu). Implémenter les ICV. Publier les ICV. Ecrire et tester les programmes d'extraction des données. Tester la production des ICV.
Participant(s)	Informaticiens, responsables métier, direction.
Livrable	Description des ICV. Programmes d'extraction des données. Programmes de calcul des indicateurs.

Tableau 8.5. *Plan de définition et production des indicateurs-clés de valeur*

8.4.6. *Mise en place du processus d'excellence des données et de correction des anomalies*

Le but est de s'assurer que les processus et les outils mis en place pour garantir la qualité des données sont bien suivis et que les anomalies détectées sont résolues.

Normalement, un propriétaire est déjà assigné pour chaque règle métier, un responsable nommé pour chaque ICV et un responsable des données concernées est désigné. Il s'agit donc de mettre en place un processus couvrant les tâches de détection, d'analyse, d'évaluation et de correction des anomalies. La tâche de correction ne couvre pas seulement la mise à jour de données dans un système par des données correctes, mais aussi des actions de changements et d'amélioration dans les processus métier.

Démarche	Définir et mettre en place une organisation d'excellence : intendant, référent et responsable des données. Surveiller et corriger les anomalies. Analyser les causes des anomalies et apporter des changements. Définir les actions correctives à court et à long terme. Si nécessaire, ajuster les droits d'accès dans les systèmes.
Participant(s)	Responsables métier, responsable des données.
Livrable	L'organisation d'excellence par domaine, processus et règle métier est opérationnelle. Listes d'anomalies.

Tableau 8.6. *Plan de mise en place du processus d'excellence des données et de correction des anomalies*

8.4.7. *Mise en œuvre d'un système de gestion de l'excellence des données SGED*

Cette tâche comporte d'une part l'installation technique du progiciel dans l'environnement informatique existant, et d'autre part la saisie des données assurant son fonctionnement dans le cadre du projet avec :

- les paramètres techniques ;
- les données de référence (valeurs dans les listes déroulantes par exemple) ;
- les règles métier ;
- les arbres de gouvernance et les indicateurs représentant les structures organisationnelles pour la visualisation et la gouvernance ;

– les paramètres pour le chargement des données extraites et les exécutions automatiques.

Démarche	Installer le progiciel. Configurer le logiciel, saisir les données de référence. Saisir les règles métier. Définir les ICV. Ecrire et installer les scripts de chargement des données.
Participant(s)	Informaticiens, responsables métier.
Livrable	Logiciel SGED installé et configuré. Règles métier et ICV introduits dans le logiciel. Les scripts de chargement des données sont opérationnels.

Tableau 8.7. *Plan de mise en œuvre d'un système de gestion de l'excellence des données*

8.4.8. *Les gains attendus*

8.4.8.1. *Diminution des délais (rapidité)*

La mise en œuvre de règles métier associées à une automatisation de l'extraction des données et à un outil logiciel de calcul des indicateurs accélérera la production des indicateurs-clés.

La diminution des délais permettra une visibilité plus rapide de l'impact et assurera que les mesures correctives accélèrent la réalisation de la valeur.

8.4.8.2. *Amélioration de la qualité des données fournies*

L'existence de règles métier et le respect de celles-ci assuré par la correction des anomalies détectées apporteront une plus grande qualité dans les données traitées selon les dimensions définies pour :

– la suppression des doublons ;

– le renseignement de données manquantes ;

– l'actualisation des données reflétant la situation réelle ;

– la mise en cohérence des données ;

– la mise à disposition des données disponibles au moment où elles sont requises.

8.4.8.3. *Sécurité dans les opérations (définition des rôles et responsabilités)*

Les règles métier peuvent inclure des notions de droits d'exécution, de délais ou de dépendance qui assurent que les opérations sont réalisées par les personnes autorisées dans des délais définis et selon une séquence précise.

De telles règles permettent à terme de fiabiliser le traitement des données dans son ensemble.

Par ailleurs, l'augmentation de la qualité des données mentionnée ci-dessus évite l'échec d'exécution de certains traitements par l'absence ou la mauvaise valeur d'une donnée.

8.4.8.4. *Renforcement de la confiance (interne et externe)*

L'amélioration de la qualité des données en termes de présence, de disponibilité, de précision et de cohérence aura un impact direct sur le niveau de confiance octroyé à l'entreprise tant par les collaborateurs internes que par les instances externes. Le niveau de confiance impacte à son tour l'image de l'entreprise.

8.4.8.5. *Augmentation de la visibilité des données dans l'entreprise*

Une focalisation sur la production de données de qualité et sur la motivation à fournir de telles données, soutenue par la publication des indicateurs de qualité et des ICV doit conduire à un changement durable de la culture d'entreprise où, comme énoncé précédemment, la donnée n'est pas la propriété d'un individu ou d'un groupe d'individu, mais de l'entreprise.

Une meilleure visibilité de l'état de l'entreprise amène à de meilleures prises de décision à tous les niveaux : changements d'organisation, allocation des ressources, nouvelles opportunités.

8.5. La démarche proposée

8.5.1. *Etape 1 : pilote*

Dans une première étape, nous proposons d'appliquer la méthode MED sur un périmètre restreint afin d'en démontrer la valeur et les résultats tangibles qui en découlent.

Ainsi, commencez par analyser deux processus types de votre entreprise. Pour chacun de ces processus, définissez quatre règles métier. Pour chaque règle, décrivez

et réalisez les programmes d'extraction des données. Sur la base de ces règles, mettez en place des indicateurs-clés faisant ressortir les différentes dimensions de la qualité des données. Parallèlement, sur la base des rôles et responsabilités, établissez et définissez le rôle du *steward* pour chacune des règles avec le processus de détection et de correction des erreurs. Enfin, mettez en place le système de data excellence SGED permettant le stockage des règles et des données extraites ainsi que la définition, le calcul et la publication des indicateurs-clés. Le bilan établi à la fin de cette étape permettra d'aborder l'étape d'extension du pilote en bénéficiant d'une première expérience.

Durée : de un à deux mois.

8.5.2. *Etape 2 : consolidation et industrialisation*

Il s'agit d'une étape de transition vers l'industrialisation de l'excellence par l'amélioration de la structure mise en place dans l'étape pilote et par l'implantation de nouvelles règles métier portant sur des processus qui auront été préalablement identifiés.

Mettez alors en activité des structures, des documents standards et des démarches à suivre permettant de créer et d'exploiter ces nouvelles règles et les ICV. Formalisez le processus de correction des données. Augmentez la capacité du logiciel SGED dans un environnement de production. La génération des indicateurs sera automatisée.

Durée : de trois à six mois selon le nombre de règles, leur complexité et la flexibilité de votre organisation.

8.5.3. *Etape 3 : régime de croisière*

Dans cette étape, passez à un fonctionnement en mode continu où la création, l'implantation et l'exploitation de nouvelles règles ainsi que la maintenance des règles existantes est un processus standard de votre entreprise.

Les processus, les documents standards et les structures permettant de créer et d'exploiter de nouvelles règles et ICV sont confirmés.

L'impact des règles existantes est analysé et, si nécessaire, des corrections sont apportées.

Durée : de trois à douze mois.

8.6. Conclusion

Nous avons présenté la méthode d'excellence des données MED axé sur la création de valeur pour les métiers à partir des données de l'entreprise. L'approche que nous avons décrite est essentiellement pragmatique et facile à utiliser. L'exécution chirurgicale de notre méthode permet d'obtenir des résultats plus rapides, facilite la collaboration entre les différentes fonctions et est alignée avec la vision à long terme pour pérenniser l'excellence dans les métiers. Notre approche met les transactions métier au centre de la qualité et de la gouvernance des données alors que les modèles de gouvernance traditionnels se focalisent sur la façon de gouverner les données de référence. La gouvernance des données, dans notre modèle, est un moyen d'offrir une valeur métier, elle ne dessert aucun objectif supérieur.

8.6.1. *Vole comme le papillon, pique comme l'abeille*

Il faudra désormais imaginer l'entreprise à travers les valeurs de transparence, confiance, intelligence et agilité : les piliers de l'excellence des données. Ceci n'est pas une initiative optionnelle, mais un réel programme de valeur ajoutée et de gestion de risques. Pour les sociétés, une prise de conscience s'impose : la mauvaise qualité des données entraîne les transactions, les processus et les projets à l'échec. La certitude demeure que la haute qualité des données est nécessaire pour révéler à la fois le potentiel de l'entreprise et les avantages des systèmes de gestion des données. De ce fait, l'excellence des données sera la fidèle compagne de la mise en place d'un référentiel afin de répartir l'effort dans le temps et prioriser les actions sur une base de valorisation continue des résultats.

8.6.2. *Prenez de la hauteur !*

Que dire de plus sinon qu'il est grand temps pour les entreprises de passer à l'action et de cultiver leurs données comme un actif. Dès aujourd'hui, et peu importe sa nature, un projet informatique de mise en commun des standards, d'un référentiel (MDM), de gouvernance des données ou de qualité des données, devra intégrer une vision supérieure : L'excellence des données !

8.7. Bibliographie

[BON 09] BONNET P., *Management des données d'entreprise – Master Data Management et modélisation sémantique*, Hermès, Paris, 2009.

[BON 10] BONNET P., *Enterprise Data Governance – Reference and Master Data Management – Semantic Modeling*, Wiley, New York, 2010.

[ELA 96] EL ABED W., Système d'interrogation de bases de données relationnelles en langage naturel, DEA, Université de Franche-Comté, Besançon, 1996.

[ELA 00] EL ABED W., Méta modèle sémantique et noyau informatique pour l'interrogation multilingue des bases de données en langue naturelle (théorie et application), Doctorat, Université de Franche-Comté, Besançon, 1996.

[ELA 08] EL ABED W., « Data Excellence Framework from Vision to Execution and Value Generation in Global Environment », *Information and Data Quality Conference*, San Antonio, Texas, 2008.

[ELA 08] EL ABED W., « Global Data Excellence Framework from Vision to Value Generation », *Data Management & Information Quality Conference Europe*, Londres, 2008.

[ELA 08] EL ABED W., « How are other organizations approaching the challenge of delivering quality global data ? How might PwC benefit from these approaches and potential solutions ? », *The Global Information Conference 2008 PriceWaterhouseCoopers (PWC)*, Rising above the Numbers, Estoril, Portugal, 2008.

[ELA 09] EL ABED W., « Data Governance : A Business Value-Driven Approach », *A White Paper*, USA, 2009.

[ELA 09] EL ABED W., *La Gouvernance Des Données : Une Approche De Valeur Conduite Par Les Métiers*, Papier Blanc, France, 2009.

[ELA 09] EL ABED W., « Mergers and Acquisitions : The Data Dimension », *A White Paper*, USA, 2009.

[ELA 09] EL ABED W., « The Data Excellence Framework », *The Hague, I-CHLAR International Conference on Hospitality & Leisure Applied Research*, Pays-Bas, 2009.

[ELA 09] EL ABED W., « The Data Excellence Framework to Improve Global Safety and Security», *ISMTCL Proceedings, Besançon, International Review Bulag, PUFC*, 2009.

[ELA 10] EL ABED W., « Comment créer un modèle de gouvernance des données fiable et pérenne », *Formation Agrée, Conférence Data Excellence Paris, La Maison des Polytechniciens*, Paris, France, 2010.

[ELA 11] EL ABED W., « Introduction to Data Excellence », *Tutorial, Data Governance Conference Europe 2011*, Londres, 2011.

[ELA 11] EL ABED W., « Linking Data Quality Metrics to Business Value and Risk », *Tutorial, Data Governance & Information Quality Conference (DGIQ)*, San Diego, Californie, USA, 2011.

[ELA 11] EL ABED W., « Master Data Management (MDM) : rêve ou illusion ? », *Article, ICT Journal*, Suisse, 2011.

[ENG 99] ENGLISH P.L., *Improving Data Warehouse and Business Information Quality*, Wiley, New York, 1999.

[ENG 09] ENGLISH P.L., *Information Quality Applied : Best Practices for Improving Business Information, Processes and Systems*, Wiley, New York, 2009.

Chapitre 9

Retour d'expérience sur un programme de gouvernance de données

9.1. Introduction

Ce chapitre constitue un témoignage sur la mise en place d'un programme de gouvernance de données au sein d'une compagnie d'assurance, de la décision d'entreprendre une action en faveur de la qualité des données aux premiers signes d'une amélioration de cette dernière.

9.2. Contexte

9.2.1. *La santé en Suisse*

En Suisse, les soins médicaux sont dispensés au sein d'établissements publiques, semi-privés et privés. Les hôpitaux universitaires, de ville ou de région, représentent le secteur public. Les établissements semi-privés prennent en charge la réadaptation tandis que les médecins exerçant en cabinet et les cliniques privées forment le secteur privé. Si le système de santé suisse n'est pas étatisé, une loi fédérale oblige chaque habitant du pays à s'assurer auprès d'une compagnie privée afin de couvrir ses frais de soins de base. Chacun peut étendre les prestations minimales garanties par la loi par le biais d'assurances complémentaires.

Chapitre rédigé par Yvan ZERMATTEN.

9.2.2. *Le marché*

L'entreprise Groupe Mutuel est une des trois plus grandes compagnies d'assurance « santé » de Suisse. Le domaine d'activité de l'entreprise est varié, il s'étend sur plusieurs segments du marché : assurance-maladie, assurance complémentaire (au sens des mutuelles françaises), assurance-accident, assurance-vie et prévoyance professionnelle. L'entreprise touche au social dans certains segments et dans d'autres se retrouve dans un monde de totale concurrence.

L'entreprise compte plus d'un million de clients pour un chiffre d'affaires annuel de l'ordre de quatre milliards de francs suisses.

9.2.3. *Le contexte légal*

Chacun des segments évoqués plus haut, est soumis à des lois et règlements particuliers. L'entreprise n'a pas le même degré de liberté d'action selon le domaine d'activité observé, l'assurance-maladie étant fortement règlementée, les autres domaines beaucoup moins.

Divers organismes d'état surveillent la bonne application des lois et des règlements.

9.2.4. *Les produits*

L'entreprise est donc soumise aux contraintes du marché et à de fortes contraintes légales. Il s'agit d'être attentif à disposer du produit adéquat au bon moment. Par conséquence de nombreux produits figurent au catalogue, la plus grande partie d'entre eux sont complexes et doivent pouvoir évoluer plus ou moins rapidement. D'autre part, le catalogue est en constante évolution pas tant de par les changements de loi mais surtout sous la pressions de la concurrence pour tous les domaines qui ne sont pas régis par une loi fédérale. La complexité des produits se manifeste par :

- les nombreuses variantes ;
- les nombreux tarifs (découpage en région) ;
- les rabais et remises (concurrence).

9.3. Les précurseurs

9.3.1. *La sensibilité*

Il existe sans conteste une sensibilité à la qualité dans l'entreprise. Que ce soit pour améliorer celle des processus métiers ou celle du contenu de son système d'information, divers projets ont été mis en place par le passé. Deux évènements marquants qui l'attestent ont été retenus ici, le premier étant réalisé par le métier et le second par le département informatique.

9.3.2. *La certification*

Une démarche menant à la certification ISO a été conduite par le métier. Classiquement, les différents processus métiers ont été analysés, décrits et documentés. Une organisation dédiée a été mise en place afin de supporter la démarche et assurer son exploitation au quotidien.

Le système fonctionne toujours à l'heure actuelle en mode « amélioration continue ».

La qualité du contenu est traitée de façon indirecte par le biais des contrôles effectués sur les processus eux-mêmes. Les dérangements affectant les processus ainsi que les données manipulées par ces derniers, sont analysés. Des responsables de processus sont chargés du traitement de listes de correction.

9.3.3. *Qualida*

De son côté, le département Informatique ne fut pas en reste. Un projet – du nom de « Qualida » – regroupant diverses actions d'amélioration de la qualité du contenu fut mis sur pied quelques mois après le démarrage du programme de certification ISO déjà évoqué. L'objectif de départ était de traiter les problèmes de qualité du contenu en travaillant uniquement sur les données.

Malheureusement, des ressources réduites furent affectées au projet ce qui eut pour conséquence de limiter fortement sa portée et l'ambition de départ. L'équipe de projet dut se concentrer sur quelques points ciblés et proposa la stratégie suivante : mettre en place des règles décrivant des contraintes sur les données. Une dizaine de contraintes furent identifiées et décrites. On les mesura ensuite. Les résultats obtenus furent publiés accompagnés des propositions de correction.

Le manque de moyens à disposition eut ici les conséquences les plus néfastes. Il ne fut pas possible de corriger les programmes à l'origine de certaines erreurs, seules

les actions sur les données furent entreprises : épuration des adresses et nettoyage des factures de prestations médicales obsolètes.

Le projet fut abandonné après cet épisode.

9.3.4. *Le bilan*

Point commun entre les deux approches, les deux démarches sont construites sur des règles qu'il s'agit de vérifier. Le traitement des résultats diverge ensuite fortement : une action continue d'amélioration d'un côté et une action opportuniste de l'autre.

La démarche ISO conduite par le métier a bien un impact indirect sur le contenu mais la qualité des données n'est pas au centre des préoccupations, c'est au processus que la majeure partie de l'attention est portée. De plus, la démarche ne couvre pas l'ensemble des départements de l'entreprise ; comme l'information est consommée par chacun d'eux, le risque de voir se dégrader une donnée fiable, est bien présent.

La démarche émanant de l'informatique se concentrait bien sur la qualité des données. Mais elle a produit uniquement des résultats partiels, une correction des erreurs très (trop) ciblée. Notez que le manque de moyens mis à disposition a conduit à ce résultat et provoqué l'application de mesures correctives de type « rustine ».

9.3.5. *L'enseignement à tirer*

Il est facile de détecter des erreurs dans un système. Alors quel est le problème ? La difficulté est dans la correction, dans la durée de celle-ci ainsi que dans les moyens parfois considérables à mettre en œuvre pour accomplir un travail sérieux.

9.4. Un nouvel essai

9.4.1. *La genèse*

Arrive le moment pour l'entreprise de remplacer son application centrale. Le projet est d'emblée considéré comme un projet d'entreprise car l'objectif est non seulement de remplacer l'application mais également de réfléchir aux processus et de les améliorer. Un projet de grande envergure donc, auquel des moyens considérables vont être alloués.

Le projet prévoit un volet de migration des données vers le nouveau système. L'approche prudente choisie implique de commencer par la conversion de petits volumes de clients.

La migration est considérée comme une activité « technique » sans valeur ajoutée, il s'agit ici de « transporter » des données d'un point de départ vers un point d'arrivée. Comme aucun analyste métier n'est rattaché à la migration, il n'existe pas de spécification.

Cependant, la qualité du contenu de la nouvelle application est un facteur de réussite du projet. Déterminer cette qualité devient donc rapidement une préoccupation majeure. Si les diverses campagnes de tests permettent de vérifier la bonne facture des fonctionnalités livrées dans l'application, qu'en est-il de l'interaction avec le contenu ? Qu'en est-il du contenu lui-même ?

De plus, à chaque ajout d'un nouveau volume de clients, il est nécessaire de déterminer le niveau de qualité de l'ensemble. Comment faire cela ?

9.4.2. *Répondre à ces interrogations*

L'équipe en charge des tests de fonctionnement va développer une batterie de contrôles s'appliquant spécifiquement aux données.

Mais ces tests empiriques vont rapidement montrer leurs limites. S'agissant principalement de contrôles de cohérence entre les données de la source et la cible, ces tests ne permettent pas de préciser la qualité du contenu. En effet, lors d'une conversion de cette ampleur, les données sources ne couvrent qu'environ un tiers des données cibles. Il est nécessaire de créer de l'information pour assurer la complétude des nouvelles structures de données et ainsi le bon fonctionnement de la nouvelle application.

Contrôler la cohérence ne suffit donc pas, ces tests « simplistes » ne rassurent pas le métier. Il faut donc faire autre chose. Un audit de la méthode de reprise est alors confié à un partenaire externe, l'entreprise Global Data Excellence (GDE).

9.4.3. *Le cadre méthodologique*

Le cadre méthodologique proposé par Global Data Excellence (GDE) et inspiré des travaux du Dr Walid el Abed (voir chapitre huit du présent ouvrage : « l'excellence des données : valorisation et gouvernance ») offre une démarche structurée s'appuyant sur la vérification de règles de gestion, le calcul de l'impact des erreurs détectées ainsi que la recherche des causes originelles (*root causes*).

D'emblée, la démarche va plaire car elle permet de mesurer la qualité des données de la nouvelle application dans leur ensemble : celles provenant de l'ancien système et celles créés lors de la conversion.

Apparemment, un autre point fort de la méthode est la facilité avec laquelle il est possible d'établir la criticité des contenus et de définir la priorité des corrections en fonction de l'impact des erreurs détectées. Les divergences d'opinion sur ce qu'il faut faire et à quel moment sont ainsi éliminées. Pour terminer, le fait de disposer d'une méthode et de la suivre a un côté si rassurant à ce point d'avancement du projet, qu'il renforce encore la conviction du management qu'il faut saisir cette opportunité. La proposition de GDE va être ainsi rapidement acceptée et sa mise en place décidée.

Pourquoi la démarche de GDE a-t-elle été considérée comme pertinente ? La réponse est en fait très simple. Outre le fait d'être présentée à un moment critique du projet de migration, une étude détaillée du cadre méthodologique fit ressortir les avantages suivants :

– le cadre méthodologique agit comme un « accélérateur » sur le projet ;

– le cadre méthodologique est fourni avec un échéancier précisant les différentes étapes ;

– les règles sont acceptées par tous et sont publiques, d'où une autorégulation ;

– les règles servent de critères de mesure ;

– les résultats sont factuels, l'analyse des causes est simple et rapide ;

– une démarche d'intégration apparaît en filigrane ce qui a pour effet de limiter les inefficacités ;

– la démarche est non invasive pour le métier, il n'est pas nécessaire de recourir à de nouvelles ressources.

Nous allons maintenant voir dans quelle mesure toutes ces promesses ont été tenues.

9.5. Au commencement

Une fois prise la décision d'entreprendre une nouvelle action en faveur de la qualité des données, le projet a débuté classiquement par la mise en place d'une équipe de gouvernance de données.

A ce stade, seul le noyau dur de l'équipe est constitué, le réseau de gouvernance sera créé bien plus tard. Il est composé de quatre personnes, possédant de nombreuses années d'expérience dans leur domaine respectif. Le commanditaire du projet est un membre de la direction opérationnelle, une partie du personnel est fournie par le département Finance tandis que la direction des systèmes d'information (DSI) fournit le reste du personnel ainsi que les équipements et les logiciels.

L'équipe se voit confier la mission de mettre en place d'une démarche de gouvernance de données au sein de l'entreprise, en prenant en charge tous les aspects du projet.

Chaotique	**Réactif**	**Stable**	**Proactif**	**Prédictif**
Pas de standards communs Approche réactive et opportuniste de la qualité du contenu	Passage de l'approche « Nettoyage des données » à l'approche « Qualité des données » Les règles de gestion sont capturées dans un référentiel Des listes d'erreurs sont disponibles	La structure de Data Excellence est en place pour un nombre réduit de règles (stewardship[1] et data responsible) Extension de la démarche dans l'entreprise	Approche de la qualité des données au travers de la prévention de la pollution	Data Excellence dès le design Data Excellence fait partie de la culture d'entreprise
Avant	Année 1	Année 2	Année 4	Année + n

Tableau 9.1. *La progression*

9.5.1. *L'apprentissage*

La première étape a consisté bien évidement en la formation de l'équipe de base à la démarche de gouvernance de données et aux détails du cadre méthodologique.

Nous avons pris le pari de nous concentrer sur l'apprentissage du cadre méthodologique de GDE et avons travaillé intensivement (trois à quatre jours par mois) avec un accompagnateur (*coach*) durant un semestre. Les interventions de

1. *Stewardship* : fonction d'un (data) steward ; *(data) steward* : régisseur ou intendant ou coordinateur des données correspond à la fonction récurrente de prendre soin de la donnée. Il intervient au point d'entrée, au point de transformation, au point de consommation. Il gère opérationnellement la qualité de la donnée et supporte la mise en œuvre des règles de gouvernance (voir chapitre dix du présent ouvrage : « Rôle et responsabilités des acteurs de la gouvernance des données : de la théorie à la pratique ».

l'accompagnateur se sont espacées ensuite (un jour par mois) à mesure que la compréhension du concept de gouvernance augmentait pour cesser un peu après la fin de la première année. Depuis lors, le *coach* intervient sur la sollicitation de l'équipe.

9.5.2. *Choisir les étapes*

La phase d'apprentissage passée, nous avons adapté l'échéancier proposé par le cadre méthodologique à la situation de l'entreprise. Celle-ci pouvant, sans hésitation, être qualifiée de chaotique, nous avons proposé une démarche stratégique qui nous permettrait de faire évoluer l'entreprise vers un état réactif puis stable. Le passage à un mode de fonctionnement proactif et plus tard prédictif n'avait pas été formalisé à ce stade, il représentait un objectif si lointain qu'il nous semblait utopique de tenter de le traduire en actes concrets.

9.5.3. *Le calendrier*

Après quelques semaines, il a été possible de préciser l'objectif global en le détaillant en un ensemble d'étapes. Une proposition de planification grossière va ainsi être faite sur cet ensemble, elle contiendra cependant les grandes étapes de notre action. Il n'était bien sûr pas possible de faire plus à ce stade pour un projet d'une telle ampleur.

Succès rapides	**Plus**	**Encore plus**	**Toujours plus**	**Pour finir**
5 règles Mesures Listes *ad hoc*	Règles + n Impacte Mesurer la migration des filiales	Règles + n Industrialisation des mesures Publication *via* un Intranet	Règles + n Automatisation Diffusion des résultats à plusieurs domaines métiers et à différents niveaux du management	Gouvernance des données de l'entreprise
Année 1 (< six mois)	Année 1	Année 2	Année 3	Plus tard

Tableau 9.2. *La planification*

9.5.4. *La première règle*

Le calendrier établi, il a fallu s'atteler à l'écriture de la première règle de gouvernance. Comme le métier n'était pas encore impliqué à ce moment du projet, il revenait aux membres du noyau dur de l'équipe d'effectuer ce travail. Peut-être trouverez-vous cette approche cavalière, mais n'oubliez pas que la moitié de l'équipe provenait du métier de l'assurance. Nous pouvions donc procéder de la sorte sans aucune arrière-pensée.

Il a d'abord fallu élire un domaine métier susceptible de faire l'objet de l'attention de l'équipe. Le choix se portera sur le processus de facturation des primes d'assurance, domaine dans lequel il nous paraissait évident de travailler en priorité. Pour cette première règle, nous nous sommes intéressés à la personne payant la prime en nous disant qu'il était important que chaque contrat d'assurance soit assorti, durant toute sa durée, d'un payeur de prime.

Ecrire cette première règle va s'avérer difficile. En effet, n'ayant pas de définition commune des objets métiers, il a fallu trouver un compromis acceptable par tout le monde sur ce qu'était un payeur de prime. Un autre point de friction a été la question de la portée de la mesure : quels contrats fallait-il observer ? Si le contrat était actif, fallait-il observer l'histoire de ce dernier et y rechercher d'éventuelles erreurs ? Le résultat de ce premier exercice d'écriture ne figure pas ici, il sort du cadre de cet ouvrage, mais sachez que nous avons tout de même fini par y parvenir après de très vives discussions.

Paradoxalement, implémenter la première mesure de qualité ne nous a pas créé de difficultés particulières. Développé en langage SQL, le traitement envoyait simplement le résultat dans un tableur. Le résultat se présentait sous la forme d'une liste d'erreurs contenant tous les contrats qui ne satisfaisaient pas à la règle. Un comptage primaire des contrats corrects fournissait une valeur de comparaison et nous permettait de calculer le premier ratio.

9.6. Ce qui a été fait

9.6.1. *Les actions accomplies*

Dès la mi-parcours de la phase d'apprentissage du cadre méthodologique, l'équipe a travaillé à la description de quelques règles de gouvernance et au développement des premiers outils de gestion de ces dernières. Il s'agissait simplement de pourvoir disposer d'un référentiel commun, ainsi que de quelques outils permettant de saisir une règle, de la décrire, de la modifier, de la partager et de

suivre son évolution. Simultanément, un outillage rudimentaire avait été mis en place afin de récolter les premières mesures de qualité. Il sera amélioré par la suite.

Durant les deux années qui vont suivre, l'équipe va se concentrer sur l'extension des contrôles, le travail avec le réseau de gouvernance ainsi que l'amélioration des outils à disposition afin d'atteindre les objectifs assignés par la direction.

9.6.2. *L'outillage*

De la phase initiale à aujourd'hui, pas moins de trois plates-formes différentes vont être développées. Cela peut paraître beaucoup et faire penser que le projet a souffert de quelques égarements. Mais il n'en est rien. Il s'agit plutôt d'un processus de maturation tout-à-fait normal. Pour accompagner ce mûrissement, nous avons donc prudemment choisi de développer nos propres outils avant de faire l'acquisition d'un produit spécifique.

Comme déjà évoqué dans la section précédente, nous avons commencé par construire le référentiel et les outils de gestion de ce dernier. Ce référentiel se composait d'un ensemble de tables placées dans une base de données ainsi que d'une interface de gestion développé sous MS Access. Ce premier référentiel nous servira durant deux ans avant d'être remplacé dans la plate-forme actuelle par le nouvel outil. Immédiatement après, nous sommes passés à la mise en place de la première plate-forme de gouvernance.

Cette plate-forme se composait d'outils fabriqués à l'aide de scripts, en langage SQL, et générait des fichiers plats. Le résultat obtenu est une liste d'erreurs distribuée et traitée de manière empirique. Chaque règle faisait l'objet d'un traitement particulier à base de plusieurs requêtes qui devaient être exécutées dans un ordre bien précis. Il n'y avait pas de script commun aux règles, ce qui conduisit rapidement à des difficultés de gestion dès que le nombre de règles devint trop important, au-delà de quinze règles cette approche était impossible à gérer. Nous avons travaillé avec cette technologie durant une dizaine de mois.

Entre temps, la seconde plate-forme sera développée. Il s'agissait cette fois d'un travail beaucoup plus abouti comprenant :

- un référentiel de règles métier amélioré ;
- un système de paramétrage et de pilotage des différents moteurs ;
- un moteur d'extraction ;
- un moteur de calcul ;
- un intranet pour la diffusion des résultats et la communication.

Cette seconde plate-forme a été développée à l'aide du générateur de code de Talend pour ce qui est des fonctionnalités permettant le pilotage, l'extraction et le calcul. L'intranet était bâti sur l'outil Confluence, un simple Wiki équipé de quelques extensions.

La palette des outils s'était certes enrichie, mais il manquait encore de nombreuses fonctionnalités et l'ensemble était trop rigide. La présentation des résultats n'était pas dynamique, mais figée dans un mode convenu avec le consommateur des résultats qu'il n'était pas possible de modifier. Il n'était pas possible de naviguer entre les éléments de l'organisation, il n'y avait pas de *drill-down* (exploration par niveau de détail croissant) et les possibilités de filtrage étaient limitées à la plus simple expression seuls quelques tris étant proposés.

Les avantages de cette version étaient un point d'entrée unique pour tous les acteurs, un outil de communication, des résultats mieux présentés parmi lesquels il était possible de naviguer de manière sommaire.

Mais la maintenance de la plate-forme était très lourde donc coûteuse. Surtout en ce qui concernait l'intranet, modifier les pages était difficile, Confluence n'étant qu'un Wiki. L'évolution de la plate-forme n'était ainsi pas garantie. De plus, il n'était pas franchement aisé de calculer des agrégats, ils étaient pré-calculés à l'aide de Talend, stockés dans le système de paramétrage pour être ensuite assemblés à l'aide de l'outil Confluence. Dans ces conditions, il valait mieux ne pas vouloir les modifier. Il fallait donc trouver autre chose.

La troisième plate-forme sera alors développée. Construite autour de l'outil DEMS de GDE, elle intègre un référentiel, un moteur de calcul sophistiqué ainsi que les outils de communication. DEMS fournit la majeure partie des fonctionnalités de la plate-forme :

- le référentiel centralisé ;
- des arbres de navigation permettant de parcourir les résultats selon l'organisation de l'entreprise ;
- un moteur de calcul produisant les résultats de toutes les dimensions et tous les agrégats ;
- un outil de *reporting* Web.

Nous recourons toujours à Talend pour ce qui est de l'extraction des données sources et l'alimentation du DEMS. D'autre part, le lancement du traitement est confié à un ordonnanceur, le logiciel UC4 choisi par l'entreprise, par souci d'intégration dans le flot des activités courantes.

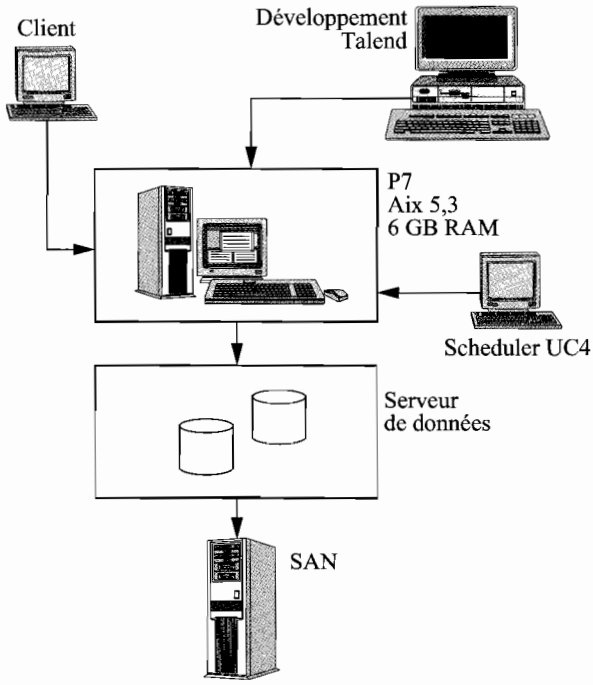

Figure 9.1. *Infrastructure matérielle et logicielle*

Bâtie sur un serveur 64 bits tournant sous AIX, la plate-forme nous permet de réaliser dans un laps de temps raisonnable l'intégralité de la chaîne de traitement de la gouvernance. Ce serveur abrite le serveur d'application faisant fonctionner le DEMS, le code Java généré à l'aide de Talend y est également déployé.

Cette nouvelle plate-forme a été mise en production à la fin du premier semestre 2011 et le portage des règles métiers existantes vient de se terminer. D'ici à quelques semaines, cette nouvelle plate-forme aura atteint le niveau de stabilité requit, nous pourrons alors complètement abandonner les deux anciens systèmes.

9.6.3. *Le traitement actuel*

Les différentes étapes du traitement à l'aide de DEMS sont les suivantes :

– création de la DAL (*Data Abstraction Layer*), permet de s'affranchir des modifications structures de données ;

– chargement des extracteurs, limite les données traitées à un ensemble plus réduit ;

– recherche des enregistrements bons ou mauvais, ceux qui satisfont ou non à une règle de gestion ;

– génération et transfert des fichiers XML (enregistrements bon ou mauvais), interfaçage avec l'outil DEMS ;

– chargement des enregistrements bons ou mauvais dans l'outil DEMS :

– calcul des clichés de la gouvernance et de la gestion des défauts ;

– publication des résultats au réseau.

Environ 40 heures sont nécessaires pour effectuer toutes ces étapes durant lesquelles 50 millions de lignes sont manipulées. Ce volume de données se répartit entre les différentes DAL et extracteurs ainsi que les informations utilisées par le DEMS pour le calcul des statistiques. Il faut encore ajouter à ces volumes les 60 fichiers XML permettant l'interfaçage avec le DEMS, ceux-ci représentent un volume de 6 GB. Ces chiffres sont en augmentation constante en raison de l'accroissement du nombre de règles métiers ainsi que de la campagne de migration sur la nouvelle plate-forme. Pour mémoire, cette dernière héberge 15 % des clients de l'entreprise au moment d'écrire ces lignes.

Le traitement sera optimisé au fil du temps afin de garantir des performances optimales malgré l'augmentation des volumes de données.

9.6.4. *Le développement d'une nouvelle règle*

Nous avons industrialisé notre activité de développement grâce à la mise en place d'un modèle. Cette approche nous permet de développer une nouvelle règle, de l'analyse au test unitaire, en trois jours.

Prendre en charge une nouvelle règle se résume maintenant en deux étapes. La première consiste en une analyse qui inclut l'étude du besoin, la faisabilité, l'étude des sources de données, l'étude du résultat attendu, l'étude des contraintes ainsi que les coûts. La seconde étape couvre le développement lui-même qui se fait par clonage d'un modèle d'implémentation. Une fois ce travail effectué, la nouvelle règle passe tout-à-fait classiquement par toute une série de tests : unitaire, intégration et validation. Cette batterie de tests passés avec succès, la nouvelle règle est déployée sur la plate-forme de production et devient disponible pour notre réseau.

9.6.5. *Les résultats*

Pendant longtemps, nous n'avons eu que des listes d'erreurs à proposer. Ces listes se présentaient sous la forme de fichiers Excel. La plate-forme initiale ne permettait pas de faire autre chose.

Ensuite les listes ont été complétées par la mise à disposition de résultats avec un historique permettant de faire apparaître des tendances d'une mesure à l'autre, d'une semaine à l'autre pour nous. Quelques résultats agrégés étaient également disponibles, mais ils n'étaient pas très complexes.

Aujourd'hui les résultats se présentent sous la forme de tableaux et graphiques affichant les indices de qualité, les listes d'erreurs ainsi que les impacts de la non-conformité des données, et cela, en tenant compte de l'organisation de l'entreprise. La figure 9.2 montre un extrait de ces résultats. L'évolution du nombre d'enregistrements ne satisfaisant pas à une règle de gestion est affichée dans cet exemple.

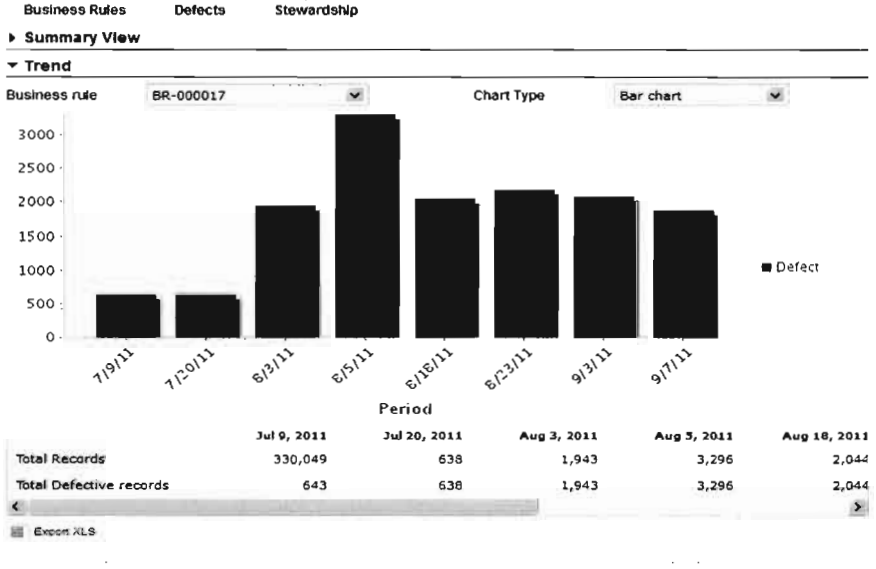

Figure 9.2. *Les résultats*

9.6.6. *Le réseau*

La mise en place du réseau est certainement la tâche la moins avancée actuellement. Notre réseau est composé, à ce stade, de deux *stewards* et d'un responsable de données. Deux domaines métiers sont ainsi couverts : la gestion des

partenaires et celle des contrats. C'est peu mais les contrôles sont judicieusement placés et de nouvelles personnes devraient nous rejoindre bientôt.

9.6.7. *Les coûts*

Dans la phase initiale du projet, nous n'avons pas jugé nécessaire de faire l'acquisition d'une application particulière. Nous avons fait le choix de commencer par investir dans la formation et indirectement donc dans l'humain.

Nous n'avons pas investi d'argent non plus dans les outils de la première plate-forme, il s'agissait uniquement d'un éditeur de requêtes à disposition dans l'entreprise.

Pour la seconde plate-forme, nous avons utilisé un outil « open source » pour construire les moteurs d'extraction et de calcul ainsi que le pilotage de ces derniers. Quant à l'intranet, il s'appuyait sur un Wiki déjà présent dans l'entreprise lui aussi. Il fut donc possible de limiter sérieusement les investissements en matériel durant les premières phases du projet.

Mais à mesure que le nombre de règles augmentait, nous avons dû investir dans du matériel plus sérieux afin de garantir des temps de traitement acceptables, un processus robuste ainsi qu'un meilleur visuel.

Charges et investissements	**Ressources**	**Coûts**
Formation	Consultant (coaching)	195 500,00
Développement	Consultant + spécialistes	1 176 000,00
Activité réseau	Gestionnaires	44 400,00
Plates-formes (I, II, III)	Licences (Talend, DEMS, etc.)	175 500,00
	Serveurs	30 000,00
		sFr 1 620 900,00

Tableau 9.3. *Les coûts*

Le tableau 9.3 montre clairement que le principal poste de dépenses est constitué des charges de personnel, le matériel représentant, jusqu'à présent, environ 10 % des coûts du projet.

9.6.8. *Les délais*

Comment parler de délais dans un projet tel que celui-ci. Pour mesurer la progression, je vais me limiter à la comparaison entre les objectifs et la liste des actions accomplies. Cette comparaison montre que les objectifs assignés ont été couverts et pour la plupart, dans les délais impartis.

Des activités de pure gouvernance et le contrôle de la qualité des migrations sont demandés à l'équipe, ce qui n'est pas sans influence sur le calendrier originel mais dans l'ensemble les délais sont respectés.

Après trois ans de travail, trente règles sont actives. La mesure des indicateurs est industrialisée et automatisée. Les résultats sont diffusés à plusieurs domaines métiers ainsi qu'à différents niveaux du management.

Arrivé à ce stade, il est possible que la diffusion de la gouvernance à d'autres domaines métiers prenne un peu plus de temps qu'initialement prévu et déborde sur la quatrième année du projet.

9.6.9. *Les indicateurs de succès*

Quels sont les changements intervenus ? Tout d'abord, à ce jour, il y a trente règles suivies par le réseau. Ensuite, bien que modeste, le réseau fonctionne. Pour terminer, la qualité de la migration peut être quantifiée et son évolution suivie d'une exécution à l'autre.

Ces trois affirmations indiquent clairement que le projet a rencontré le succès jusqu'à présent. Et cela, en suivant le tableau des objectifs établi dans la phase initiale.

Mais le fait le plus marquant est sans doute d'avoir permis à l'entreprise de passer d'un état initial chaotique à un état stable avec un cadre méthodologique en place pour un nombre de règles déjà conséquent. Une anecdote illustre parfaitement cette évolution : lorsque nous avons livré notre premier résultat fiable des contrôles sur les règles de facturation, le directeur financier s'est précipité sur le système de facturation afin de vérifier l'exactitude de notre rapport. Je n'aurais pas écrit ces lignes si le rapport n'avait pas été correct, mais quoi qu'il en soit, ce responsable consulte aujourd'hui les résultats de nos traitements en toute sérénité. L'entreprise a donc bien évolué.

9.7. Ce qui reste à faire

9.7.1. *La plate-forme*

Il est plus que probable que nous tenions maintenant la plate-forme adaptée. Si cette dernière est riche de fonctions, il faut cependant la faire évoluer afin d'absorber la charge des mois et années à venir. Mais le modèle de développement ne devrait pas changer ces prochains mois. Il va sans dire que nous suivrons les évolutions de l'application DEMS, elle va demeurer au centre de notre infrastructure.

9.7.2. *Les traitements de la gouvernance*

Les développements vont pouvoir s'intensifier. Le passage à la plate-forme DEMS nous a permis de simplifier les différentes étapes de nos traitements. Nous pouvons ainsi nous contenter d'appliquer la démarche expliquée au paragraphe décrivant la méthode de développement des règles pour augmenter le nombre de contrôles et étendre la gouvernance à tous les métiers de l'entreprise.

9.7.3. *Le réseau*

C'est à coup sûr la tâche qui va nous demander le plus d'attention dans les mois à venir. Le réseau doit être étendu, « en hauteur et en largeur ». Un pas vers la direction opérationnelle va se faire prochainement mais il faut aller plus haut.

L'extension horizontale doit également se poursuivre afin d'amener plus de *stewards* et de responsables de données à intégrer le réseau.

9.7.4. *La gouvernance de données*

Il va falloir consolider la gouvernance de données en fiabilisant les indicateurs de qualité. Une autre avancée significative sera la publication d'un premier indicateur-clé de valeur ou *KVI*. Ce point est d'une grande importance pour l'entreprise, car seul le calcul de l'impact peut faire ressortir des problèmes cachés par de bons indicateurs de qualité. Ainsi, la crédibilité de la gouvernance sera augmentée. Ce projet doit devenir une préoccupation de l'entreprise, ne pas rester dans le giron de l'informatique. Il va falloir évangéliser un plus grand nombre de personnes pour atteindre cet objectif et pérenniser ainsi la démarche de gouvernance.

9.8. Bilan

– Bien qu'il soit entré dans une phase stratégique et technique particulièrement aboutie tout récemment, le projet doit durer encore plusieurs années. Le bilan ne peut donc être qu'intermédiaire.

– Le point le plus important est que la gouvernance de données fonctionne aujourd'hui. Certes la couverture des domaines métiers est encore modeste, mais la qualité de l'information des domaines observés est contrôlée et améliorée. Il y a donc dans ce fait une reconnaissance du bien-fondé de la démarche et du travail de l'équipe.

– Cependant l'adhésion au projet n'est pas complète, car la perception de l'apport de la démarche n'est pas uniforme. Si l'encadrement réagit favorablement aux indicateurs de qualité, les responsables intermédiaires susceptibles d'œuvrer comme *steward*, considèrent la gouvernance comme un moyen de résoudre des problématiques opérationnelles. Il reste donc beaucoup à faire en termes de communication afin de renforcer la compréhension de notre activité et espérer ainsi étendre notre réseau.

– Le projet vit sa troisième année, mais l'utilisation de KVI n'a pas encore été industrialisée. Cette situation peut paraître paradoxale étant donné que la capacité à déterminer le coût des défauts de qualité était un des arguments ayant fortement contribué à l'acceptation du projet. Il y a cependant une explication à cette ambiguïté. Au début de la démarche, nous avions valorisé de nombreuses règles métier, or nos indicateurs se sont avérés imprécis. Nous étions capables de donner l'impact potentiel d'une erreur, mais le montant choisi pour valoriser l'erreur a été rapidement sujet à contradiction. Ensuite, le métier n'est pas demandeur de valorisation, du moins les interlocuteurs actuels. Ils veulent des listes d'erreurs et s'appuient sur les indices de qualité (rapport entre les bons et les mauvais enregistrements) pour suivre l'évolution de la qualité d'une semaine à l'autre. Le métier prend un risque en agissant de la sorte. De sérieux problèmes peuvent se cacher derrière un indice de qualité élevé, que seul le KVI, peut faire apparaître. Enfin, nous avons commis une erreur. Nous avons choisi de lier le KVI à un flux financier uniquement et cela dans le but de faire taire les détracteurs apparus au début de la démarche. En mesurant les erreurs contenues dans le processus de facturation des primes par exemple, il nous semblait facile de calculer le coût de celles-ci. Le montant de la facture devait nous fournir cette information. Or, les flux financiers sont complexes et concernent plusieurs départements de l'entreprise. Nous n'avons pas cherché à mieux mettre en évidence cette notion d'indicateur-clé de valeur (KVI) car nous ne possédons pas, après tout ce temps, suffisamment de règles qui misent bout-à-bout, couvriraient le flux financier de la facturation des primes.

Aujourd'hui, nous nous préparons à revenir sur cette question en mettant en avant le fait que l'indicateur-clé de valeur peut se calculer sur autre chose qu'un flux financier : un client, une adresse, un contrat. En résumé tout ce qui représente de la valeur pour une entreprise.

9.8.1. *Le budget*

Investissez d'abord dans la formation. Ce choix peut surprendre mais il était dicté par le bon sens. En effet mieux vaut comprendre les principes d'une bonne gouvernance de données plutôt que de commencer par faire l'acquisition d'un logiciel à un million d'euros et de se rendre compte par la suite que ce produit nous enferme dans un moule, réduit notre autonomie et coûte la même somme en consultants et support.

9.8.2. *Le calendrier*

Une planification grossière, pilotée par les objectifs suffit. Une évidence dans le cadre d'un projet s'étalant sur plusieurs années. Les objectifs sont un fil rouge, la planification ne peut que tendre vers cette limite et nous aider ainsi à atteindre les objectifs même au prix de quelques détours.

9.8.3. *Le personnel*

Une petite équipe fortement motivée fait merveille. Mais il faut bien que ces membres aient les qualités requises pour travailler sur un projet de ce genre qui sont : esprit de synthèse, sens de la communication, ténacité, résistance et ouverture d'esprit.

9.8.4. *L'architecture des données*

Disposez d'un ensemble de données spécifiques à la gouvernance, isolées des données transactionnelles. L'ensemble peut être une copie de la production mais il est mieux de lui faire subir une transformation en le chargeant dans une couche d'abstraction.

Construisez une *DAL* (*data abstraction layer*). Elle représente un découplage entre les données utilisées pour les mesures de qualité et les données de production et permet de s'affranchir des changements intervenant sur les structures de données utilisées par les diverses applications de l'entreprise.

Généralisez l'emploi d'extracteurs de données. Situés au-dessus de la DAL, ils permettent de facilement filtrer les données afin de les faire correspondre à la portée de chaque mesure. De plus, les extracteurs sont réutilisables dans plusieurs mesures.

9.8.5. *Le développement*

Industrialisez le développement le plus vite possible. Nous avons appliqué ce principe dès la deuxième plate-forme et l'avons encore optimisé lors de la mise en place de l'infrastructure actuelle.

Pour industrialiser le développement nous sommes passés non seulement par la codification de bonnes pratiques et par la formation des personnes, mais aussi par l'utilisation d'outils spécifiques et par la création d'un modèle d'implémentation.

Si un outil spécifique est disponible, n'hésitez pas à l'utiliser le plus rapidement possible. L'achat d'une application de gouvernance est un investissement qui sera très rapidement rentabilisé. L'introduction de l'outil DEMS a coïncidé avec la mise en place de la troisième plate-forme de gouvernance. Il faut rendre hommage ici à la branche « informatique » de l'équipe qui a fourni un travail considérable, mais si nous avions disposé de cet outil plus tôt, nous aurions pu, au minimum, éviter la construction d'une plate-forme.

Le modèle d'implémentation permet de diminuer fortement le temps de développement d'une nouvelle règle étant donné qu'il suffit de cloner le modèle et de se concentrer sur les parties spécifiques de chaque règle, comme les extracteurs et la recherche des « bons ou mauvais » enregistrements. De plus, le modèle empêche la mise en place de solutions ésotériques, farfelues risquant de mettre en péril la productivité de l'équipe. Si la créativité du développeur est ainsi limitée, la maintenance est simplifiée.

L'industrialisation nous garantit également des exécutions robustes des traitements, ils traversent une batterie complète de tests avant d'arriver en production. Il n'y a donc pas de risque de subir un plantage inopiné de toute la chaîne lors de l'ajout d'une nouvelle mesure.

9.8.6. *A retenir*

Dans la phase initiale, il vaut mieux éviter le développement d'un trop grand nombre de règles au détriment d'autres activités. Le métier risque ne pas suffisamment s'impliquer, il est ensuite plus difficile de regagner son attention.

Mieux vaut donc travailler avec un petit nombre de règles mais en allant au bout de la démarche avec chacune d'elle.

Pendant longtemps la gouvernance n'a pas été comprise par la faute d'une communication déficiente. Elle peine toujours à s'imposer auprès de la direction opérationnelle. Un projet d'entreprise nécessite que l'on porte une attention toute particulière à la communication, sinon l'échec est programmé. D'autre part, les différents acteurs risquent de ne pas percevoir la gouvernance comme une activité stratégique mais plutôt comme un outil permettant de corriger des carences opérationnelles.

Une conséquence directe des erreurs de communication est la suivante : la gouvernance finit par produire uniquement des listes d'erreurs. Il ne faut pas se contenter de servir uniquement les *stewards* mais surtout répondre à la préoccupation de la direction opérationnelle sensible elle aux indices de qualité, ce qui saura favoriser le projet.

Si vous travaillez avec un *coach*, faites en sorte qu'il soit actif sur le projet le plus longtemps possible. Ayant souvent une forte personnalité, ce *coach* peut apporter une aide précieuse dans de nombreux domaines comme la promotion du projet auprès de la direction, la correction de la communication, etc.

Les évolutions technologiques doivent être limitées. Nous avons utilisé trois plates-formes, c'est certainement une de trop, car à chaque évolution, il y a un gros travail de portage consommateur de ressources et de temps.

N'hésitez pas à mettre en place une architecture de données solide incluant une couche d'abstraction. Nous n'avons pas mis en place cette couche dans la deuxième plate-forme, cette omission nous a coûté plusieurs semaines de travail de maintenance lorsque les structures de données ont évolué.

Un vieil adage dit que : « L'union fait la force ». Deux précédentes initiatives d'amélioration de la qualité ont échoué ou n'ont que partiellement réussi, la mise en commun des forces du métier et celles de l'informatique a contribué à la réussite de la démarche en cours.

Prenez garde à l'interprétation du KVI, il s'agit d'un sujet délicat. Le KVI ne doit pas forcément être établi sur un flux financier. Il peut être calculé sur tout ce qui représente de la valeur pour une entreprise et il est clair que le contexte dans lequel l'entreprise évolue influence sur le sens à donner à cet indicateur-clé.

Chapitre 10

Les rôles et responsabilités des acteurs de la gouvernance des données : de la théorie à la pratique

10.1. Introduction

Dans ce chapitre, nous présentons le témoignage de cinq sociétés qui ont mis en place ou sont en cours de mise en place d'une cellule de gouvernance. Après un rappel rapide des rôles décrits dans les méthodes d'organisation de la gouvernance, et une présentation de nos cinq témoins, nous avons voulu décrire les rôles et responsabilités des acteurs de la gouvernance à travers les expériences communes de ces témoins : quelles missions, quelles compétences sont nécessaires pour gouverner la donnée ? Tâche qui s'est révélée difficile, car sous un même terme, nous avons observé des missions différentes et des responsabilités différentes. Nous avons pris le parti d'illustrer systématiquement notre propos : pourquoi a-t-on besoin de ce rôle ? A quel moment de vie de la donnée intervient-il ?

Après une présentation des différents acteurs et des rôles selon la vie de la donnée, nous positionnons les rôles couverts par la cellule de gouvernance *versus* les rôles non assurés par la cellule mais par des interlocuteurs externes à la cellule. Enfin, nous indiquons pour chaque rôle les facteurs clés de succès qui ressortent des différentes expériences. Les tableaux 10.7 à 10.11 en annexe section 10.9, récapitulent les « rôles théoriques/rôles observés » pour permettre au lecteur d'avoir une synthèse des correspondances entre les termes employés.

Chapitre rédigé par Evelyne ROSSIN, Nathalie BARTHÉLÉMY et Brigitte LABOISSE.

10.2. Rôles et responsabilités dans la littérature sur la gouvernance de données

Dans la littérature [LOS 09, REG 08, BON 10, BER07, DATA], les rôles généralement décrits sont les suivants.

Le parrain (ou sponsor) : il met à disposition les moyens, pilote les alignements stratégiques et s'assure du maintien de cette stratégie en mode opérationnel. Il garantit l'adoption de l'entreprise à une politique de gouvernance de la donnée. Le parrain négocie des engagements sur la qualité de données avec les fournisseurs externes.

Le propriétaire de la donnée (ou Data owner *ou* Business Data Owner *ou BDO)* : en tant que référent métier, il va indiquer les règles et les exigences. On parle de multipropriété car on s'aperçoit que, souvent, un objet n'appartient pas exclusivement à un métier.

Le régisseur ou intendant ou coordinateur (ou *data steward*) : correspond à l'action récurrente de prendre soin de la donnée. Il intervient au point d'entrée, au point de transformation, au point de consommation. Il gère opérationnellement la qualité de la donnée et supporte la mise en œuvre des règles de gouvernance.

Les rôles de parrain, propriétaire de la donnée et régisseur sont décrits dans la majorité des ouvrages sur la gouvernance de données. Les rôles ci-dessous sont moins cités.

L'analyste (ou data analyst) : modélise les données. Il supervise le projet, il représente le référent métier. Il met en pratique les procédures sémantiques de la modélisation de données.

L'architecte de données (ou data architect) : il a un rôle transversal dans l'organisation, supervise l'évolution des modèles. Il valide la cohérence des modèles. Il s'assure que les objets ne comportent ni redondance ni incohérence d'un métier à l'autre.

Le responsable des coûts de la donnée (ou data cost accountant) : il suit les coûts de modélisation de la donnée. Il est en charge d'établir les lignes de bilan concernant le budget en particulier des données de référence.

Gardien de la donnée (ou data custodian) : supervise la sécurité du transport et du stockage de la donnée. Si le contenu est important pour lui, il se focalise sur l'infrastructure et les activités pour garder la donnée protégée et disponible aux utilisateurs. Il travaille avec le *data steward* pour résoudre les problèmes, faire évoluer les systèmes et implémenter les transformations de données.

10.3. Présentation des cas pratiques et des contributeurs

Nous présentons cinq témoignages de sociétés qui ont implémenté une cellule de gouvernance de données ou sont en cours d'implémentation.

Michelin	MICHELIN Une meilleure façon d'avancer
EDF DOAAT	EDF
GDF SUEZ Branche Global Gaz & GNL	GDF SUEZ
Bouygues Telecom	Bouygues Telecom
CTI – Centre des Technologies de l'Information – Etat de Genève	RÉPUBLIQUE ET CANTON DE GENÈVE

Tableau 10.1. *Les contributeurs*

Ces sociétés ont un niveau de maturité différent dans la gouvernance de données, mais une approche similaire des équipes dédiées à la qualité de données et à la gouvernance de données. Ci-après, nous présentons les fiches d'identité de ces cellules contributrices. Pour des raisons de confidentialité, le nom des sociétés n'est pas repris sur les fiches d'identité qui présentent les sociétés dans un ordre aléatoire.

Nos commentaires : on observe un rattachement soit au métier, soit à la DSI. Les cellules rattachées au métier étant reconnues « plus légitimes » dans leur compréhension des règles métier et des niveaux d'exigence. *A contrario*, celles rattachées à la DSI ont plus facilement accès à des ressources de développement ou de mise en œuvre d'outils. Il faut noter l'influence de facteurs tels que la culture de la société, d'un parrainage plus ou moins convaincu, plus ou moins influent.

Contributeur n° 1	
Nom de la cellule	Mission Management de l'Information
Nombre de personnes en équivalent temps plein	5
Profil	Conseils système d'information
Formation en qualité	Non
Positionnement dans la hiérarchie de l'entreprise	4
Domaine de rattachement	DSI[1] métier
Date de création	2010
Missions	Gestion des métadonnées (objets métiers). Modèle conceptuel des données, description des cas d'utilisation ; glossaire. Travaille en collaboration avec le service architecture. Définition et mise en œuvre d'une politique de données et d'une gouvernance de données (processus, rôles, responsabilités, instances) Audit du respect des contraintes réglementaires sur les données. Gestion opérationnelle de la qualité des données (rôle de régisseur en place sur un type de données), avec animation d'un réseau de référents données Métiers et d'un directoire données de référence.
Type de données manipulées	Marché, météo (fournisseurs prochaine étape)
Communication	Lettre d'information, Flash infos, espace collaboratif, présentations dans les réunions de service métier, formations, journal interne de la filière système d'information
Indicateurs publiés	Animation d'un espace collaboratif qualité des données ; partage de connaissances et animation plan d'actions amélioration qualité
Budget de la cellule	Budget propre reventilé aux métiers

Tableau 10.2. *Cellule du contributeur 1*

1. DSI : Direction des systèmes d'information.

Contributeur n° 2	
Nom de la cellule	Observatoire statistique
Nombre de personnes en équivalent temps plein	5
Profil	Analyste conseil en management du système d'information – Chef de projet métier/système d'information – Pilote de processus d'administration de données
Formation en qualité	Non
Positionnement dans la hiérarchie de l'entreprise	NC
Domaine de rattachement	Métier
Date de création	2010
Missions	Promotion d'une politique de la donnée et fédération des acteurs-actions autour de règles, actions, organisations de maitrise des données basée sur cinq principes (patrimoine, référentiel, unicité du point de sortie, traçabilité, efficacité d'accès aux données). Mise en place de processus et d'outils d'acquisition-contrôle- mise à disposition de données (de référence, historiques, décisionnelles)
Type de données manipulées	Production, commercialisation, finance, données d'analyses métier, indicateurs
Communication	Communication via les comités de pilotage projets, Commissions, rencontres managériales, conférences « grand public » au sein de l'entité couverte, et par essaimage au fur et à mesure de l'implication des acteurs dans la mise en œuvre des services et des pratiques visant la maitrise des données pour la performance des métiers
Indicateurs publiés	Tableaux de bord qualité de données hétérogènes en cours d'harmonisation et de structuration (suivi de l'amélioration continue)
Budget de la cellule	Budget alloué aux trois projets opérationnels

Tableau 10.3. *Cellule du contributeur 2*

Contributeur n° 3	
Nom de la cellule	AGD – Administration et gouvernance des données
Nombre de personnes en équivalent temps plein	6
Profil	Equipe mixte composée d'architectes système d'information et métier marketing
Formation en qualité	Formation outil qualité de données
Positionnement dans la hiérarchie de l'entreprise	4
Domaine de rattachement	Informatique décisionnelle métier
Date de création	2008
Missions	Gestion partagée des métadonnées utilisateurs, règles métier. Améliorer la qualité intrinsèque des données par des actions ciblées. Garantir le bon niveau de qualité des données et du bon usage des référentiels par des actions de pilotage et de contrôle. Améliorer la structuration des données et de la gestion des référentiels par une démarche d'entreprise.
Type de données manipulées	Huit référentiels définis comme prioritaires en 2011. Notamment clients (particulier et entreprises), distributeurs, offres, produits.
Communication	Plan de communication. Mise à disposition d'un dictionnaire à travers un espace collaboratif. JPO (journées portes ouvertes).
Indicateurs publiés	Mise en place de tableaux de bord qualité de données
Budget de la cellule	Budget d'étude et de fonctionnement

Tableau 10.4. *Cellule du contributeur 3*

Contributeur n° 4	
Nom de la cellule	Cellule MDM – master data management
Nombre de personnes en équivalent temps plein	6
Profil	Architectes système d'information, profil technique et administration des données
Formation en qualité	Formation outil qualité de données
Positionnement dans la hiérarchie de l'entreprise	4
Domaine de rattachement	DSI
Date de création	2008
Missions	Gestion partagée des métadonnées utilisateurs, règles métier. Concevoir, évoluer, déployer le modèle de données entreprise (données référentielles et transactionnelles). Mise en œuvre de la gouvernance des données et notamment des données de référence et de la qualité des données. Valider l'architecture d'applications dans le cadre de la gouvernance de l'architecture d'entreprise. Création des cartographies des données de référence et concepts afin de s'intégrer dans les plans d'actions par domaines. Promouvoir les méthodes et outils de qualité de donnée. Intégration des plans d'action référentiels dans les plans d'action d'architecture fonctionnelle.
Type de données manipulées	Données référentielles (organisations, zones géographiques, classification achats, clients, fournisseurs, etc.)
Communication	Gouvernance des données avec des réunions mensuelles avec les Data Owners des métiers. Dictionnaire sur Intranet groupe et espace collaboratif.
Indicateurs publiés	Mise en place d'indicateurs pour certains référentiels.
Budget de la cellule	Budget d'étude et de fonctionnement, ainsi qu'un budget outils

Tableau 10.5. *Cellule du contributeur 4*

Contributeur n° 5	
Nom de la cellule	Centre d'expertise qualité de données au sein du service gestion des données et de l'information
Nombre de personnes en équivalent temps plein	5
Profil	Administrateur de bases de données, Informatique décisionnelle, profil architecte
Formation en qualité ou gouvernance de données des personnes de la cellule	Oui, formation à la gouvernance de données par une société spécialisée
Positionnement dans la hiérarchie de l'entreprise	3
Domaine de rattachement	DSI
Date de création	2010
Missions	Promouvoir la qualité des données et sensibiliser les équipes métiers. Support auprès de la Maitrise d'ouvrage pour définir les règles, exigences et gouvernance. Auditer, mesurer, corriger et surveiller l'évolution de la qualité. Gérer les règles métiers et les indicateurs. Organiser la gouvernance qualité.
Type de données manipulées	Données référentielles, données personnes, adresses, très variées.
Communication	NC
Indicateurs publiés	Des indicateurs liés à certains référentiels font l'objet de publication. D'autres sont en cours de définition.
Budget de la cellule	Pas de budget spécifique, budget de la DSI.

Tableau 10.6. *Cellule du contributeur 5*

En termes de profil des membres de la cellule, on observe une majorité d'architectes Système d'Information, accompagnés (ou pas) de personnes métier. La formation spécifique en qualité et gouvernance des données est rarement suivie, ce qui s'explique parfaitement par le peu de cursus disponibles à ce jour. L'échange à travers des conférences est privilégié par la majorité des cellules.

10.4. Une modélisation des rôles pour se comprendre

Comprendre comment chaque entreprise gouverne ses données, comparer, trouver les points communs, est un exercice périlleux lorsque le vocabulaire employé par les entreprises ne couvre pas les mêmes gestes pour les mêmes objectifs. Les termes de « *data steward* », « propriétaire de la donnée », ou « architecte de données », s'ils sont présentés de manière générique dans les méthodes d'organisation vues plus haut, ont des interprétations très différentes dans chacune des sociétés témoins. Qui plus est, l'organisation de chaque entreprise induit sur chaque individu un découpage de gestes et d'objectifs qui peuvent être similaires en partie aux gestes et objectifs de professionnels d'autres entreprises, mais rarement égaux !

C'est pourquoi, il apparait nécessaire pour faire émerger les « gestes » que tous font, de bien les préciser concrètement, de les regrouper lorsqu'ils répondent à un même objectif et de trouver un vocabulaire commun, standard et évocateur, pour nommer les rôles ainsi reconstitués.

C'est ce modèle de rôles que nous explicitons dans la section suivante, déconnecté des organisations formelles des entreprises (organigrammes), au vu du retour d'expérience de cinq entreprises.

Ces rôles peuvent être joués pour des enjeux d'entreprise différents ou de priorités différentes (par exemple, la sécurité des hommes et de l'environnement, l'image de marque, la réglementation, la performance financière, etc.), dans des contextes de sensibilités différentes (plus facile à mobiliser sur les erreurs que sur les opportunités, par exemple) et de maturités différentes (rôles formels ou informels, à mailles très localisées ou couvrant toute l'entreprise, etc.). Ces différences induisent des démarches particularisées composées, par exemple, d'approches par les euros, les clients ou le référentiel de données, d'approches par les risques ou les principes (selon les enjeux et la sensibilité de l'entreprise), ou bien encore d'approches par la sensibilisation et la règle (selon la maturité de l'entreprise).

Aussi certains rôles prennent une importance et un contenu dans un contexte, qu'ils n'auront pas dans un autre. Le modèle, ci-dessous présenté, se veut « maximaliste ». Le lecteur en fonction de son contexte, pondèrera chaque geste, en termes d'enjeux, de collaborations nécessaires à sa réalisation, de compétences et de charges de travail.

Attention au piège d'assimiler un rôle à une personne ! Plusieurs personnes peuvent être sur un même rôle, ou inversement une personne peut tenir plusieurs rôles.

10.4.1. *Les rôles pour maîtriser les données*

Pour faire comprendre le pourquoi de ces rôles, c'est-à-dire le pourquoi de la mise en œuvre de gestes pour disposer, à minima, de données dans des délais et avec un niveau de confiance « acceptable », nous reconstituons le parcours d'une donnée de sa « naissance » jusqu'à sa mise sous « assurance qualité ».

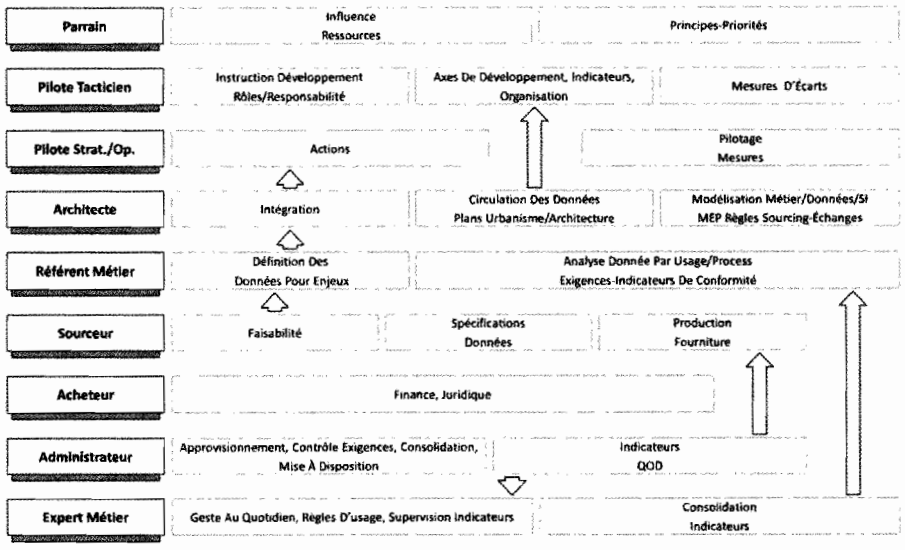

Figure 10.1. *Rôles et Missions des acteurs*

La figure 10.1 fournit une vue d'ensemble des différents rôles de la création de la donnée à sa mise sous contrôle. Dans un souci pédagogique, nous montrons uniquement les interactions majeures entre rôles.

La donnée n'existe pas ex-nihilo. Elle a été demandée, à un moment donné, par un acteur pour être consommée dans une activité ou un processus. La donnée a d'abord suivi un parcours d'émergence avant d'être créée : définition pour une bonne interprétation, évaluation de son intérêt pour les enjeux d'entreprise, établissement de sa circulation et de son partage. Au centre de cette production et de cette livraison, des gestes regroupés sur trois rôles sont détaillés ci-après.

10.4.2. *Le référent métier*

Par domaine métier, le référent métier définit, pour la performance du processus ou de l'activité métier qu'il représente, le type de données souhaitées, leurs usages et les exigences qualité afférentes, la traduction en indicateurs et seuils, leur niveau d'accessibilité par les utilisateurs/consommateurs. Il a le pouvoir d'ajuster processus et donnée pour obtenir le meilleur résultat possible. Il connait ce qui est « réaliste » dans son organisation en matière de changement et établit la définition des données et de leurs usages en conséquence.

Il approuve le contrat de fourniture, et le dispositif de gouvernance.

Il définit les seuils d'alerte sur les indicateurs et mène l'analyse et les actions structurelles si les valeurs des indicateurs sont en deçà.

Le référent donne les règles métier et, dans une approche de valorisation, est capable d'indiquer la valeur de la donnée et les impacts financiers de la non qualité.

10.4.3. *Le sourceur*

Le sourceur contribue à définir précisément la donnée (ses caractéristiques et attributs) selon la capacité de son organisation à en disposer, soit en la produisant, soit en réadaptant des données déjà existantes dans son organisation. Son objectif est de livrer au référent métier des données qui répondent au plus près aux usages et aux exigences qualité demandés par le référent métier.

Il définit le niveau d'accessibilité des données qu'il va mettre à disposition. En effet, elles peuvent revêtir un caractère « à ne pas divulguer » pour les enjeux de l'organisation du sourceur.

Il désigne les acteurs-fournisseurs des valeurs.

Il contractualise au nom de sa direction la fourniture des données, ainsi que le dispositif de suivi de la qualité de fourniture à l'organisation du référent métier.

10.4.4. *L'acheteur (côté référent métier)*

L'acheteur assure les gestes de passage de contrat avec l'organisation du sourceur en co-rédigeant le contrat ou en vérifiant les termes du contrat (fourniture, traçabilité de la qualité, niveau d'accès aux données livrées, etc.).

Il assure les gestes d'achats (commandes, etc.).

EXEMPLE 1.– Dans le cadre de projets (nouvelle activité, nouvel outil, par exemple), un ou plusieurs agents sont désignés pour leur bonne connaissance du ou des métiers et des organisations concernées par le projet, pour définir quels seront les futurs gestes métiers, et quelles informations devront être collectées pour les rendre possibles. Ils jouent le rôle de référent métier.

Ils se tournent vers l'acteur interne ou externe de l'entreprise, qui à leurs yeux, pourrait disposer de ces informations (le sourceur), et voient avec cet acteur, s'il accepte de les leur mettre à disposition. Le comité de pilotage projet entérine le choix de cette mise à disposition aux conditions du sourceur. Le travail commun peut alors se concrétiser par la rédaction de dictionnaires pour définir les données, de cahiers des charges décrivant comment les données seront produites (cahier des charges internes au sourceur) et comment les valeurs seront fournies aux organisations des référents métiers (dossier d'urbanisme et d'architecture d'échange des données).

Cette mise à disposition peut revêtir des conditions : facturation de la mise à disposition, modalités de fourniture (quand, à qui et sous quelles formes et supports), avec quel niveau de qualité intrinsèque (complétude des valeurs, exactitude, etc.), avec quel niveau de diffusion ou de confidentialité, avec quelles tolérances d'échec avant de réviser le service de mise à disposition, etc. Dans les organisations encore peu sensibles à l'enjeu commun sur les données de l'entreprise, les sourceurs ne s'obligent qu'à fournir les données qu'eux-mêmes utilisent, et à leur propre niveau d'exigence.

Comment garantir l'obtention de cette valeur, au quotidien, par l'usage attendu ?

Au quotidien, ce sont les valeurs des données mises à disposition, qui sont fournies par l'organisation d'acteurs désignée par le sourceur, aux consommateurs désignés par le référent métier.

Les fournisseurs de valeurs peuvent être très nombreux, tout comme ceux qui réceptionnent et utilisent les données. Comment assurer alors, que cette fourniture arrive aux bonnes personnes selon les conditions définies ?

Ceci est le rôle de l'administrateur.

10.4.5. *L'administrateur*

L'administrateur acquiert les données/valeurs des acteurs désignés par le sourceur. Il contrôle si elles respectent les critères définis par le référent métier

(la qualité plus particulièrement) et l'architecte (voir section 10.4.7) et il assure la remise en conformité le cas échéant.

Il met à disposition les valeurs aux consommateurs selon les modalités de fourniture contractualisées.

Pour suivre et piloter la qualité de la mise à disposition selon les exigences requises par les usages attendus, il définit des indicateurs sur la mise à disposition des données.

Ce rôle peut prendre des formes d'organisation très différentes, tout particulièrement en nombre d'agents, et en rattachement de ces agents à un service ou à un autre. En effet, les gestes factuels peuvent aller de la simple habilitation d'une personne à accéder à une base de données, à la conception et au pilotage d'un processus sous assurance qualité, dont les briques fonctionnelles sont définies ci-dessus et dont les contributeurs apportent une connaissance fonctionnelle des données spécifique à chaque métier.

L'administrateur est le rôle qui tire le mieux profit de la technologie informatique ; l'outillage est une aide précieuse pour modéliser les données (voir rôle architecte) et vérifier ensuite la cohérence entre les terminologies des données injectées dans les bases, ou encore pour réaliser automatiquement des tests sur la qualité des valeurs.

Cependant, est-ce suffisant d'assurer la conformité de la mise à disposition des données, afin qu'elles remplissent la mission pour laquelle il a été jugé initialement intéressant de les créer ? Non, bien sûr ! Il faut également qu'elles soient utilisées comme le référent métier l'avait défini.

Ceci est le rôle de l'expert métier.

10.4.6. *L'expert métier*

L'expert métier définit les procédures métier (ou coordonne leur définition) et les règles de gestion associées aux données.

Il supervise les indicateurs de la qualité de fourniture et de mise à disposition, et commandite les actions au jour le jour de remise en conformité, relevant de sa responsabilité métier.

Il est en lien étroit avec le référent métier, et l'équipe de l'administrateur, pour la mise en œuvre des corrections ou des adaptations.

EXEMPLE 2.– Qui ne s'est jamais demandé s'il pouvait avoir confiance dans les données qu'il utilisait et n'a pas cherché par plusieurs voies à confronter ses données ?

Dès que plusieurs personnes se posent les mêmes questions sur une donnée, des bases de données collectives émergent, sensées être le lieu de mise à disposition unique des données pour l'entité de ces personnes. Un acteur est chargé de vérifier qu'elle est bien approvisionnée, que les données sont rafraichies ; ce sont les prémices du rôle d'administrateur. Il est en lien direct avec les informaticiens, chargés d'assurer la continuité des flux techniques d'approvisionnement de la base en données. Parfois, des utilisateurs viennent se plaindre pour des anomalies liées à une gestion non conforme ; par exemple, une donnée a été transformée dans le cadre d'une activité métier, et n'a pas été replacée à l'endroit convenu dans la base avec les autres utilisateurs, au bon moment ou de la bonne façon. Les autres utilisateurs ne peuvent plus travailler.

Le rôle d'expert fonctionnel métier, dans sa dimension « expert », émerge à cette occasion. Il analyse les gestes métier qui ont conduit à cette anomalie, en comprend la raison, et révise la procédure de bon usage collectif ou la fait évoluer, et bien sûr, il la repartage avec ses collègues.

Comment être sûr de se fournir à la source légitime des données ? Comment assurer la cohérence entre les valeurs de données utilisées par plusieurs processus ou plusieurs activités de l'entreprise ? Comment assurer que tous les utilisateurs interprètent de la même façon une donnée ?

Aller chercher la donnée auprès de qui l'a déjà, n'est pas forcément la bonne stratégie pour disposer de données aux valeurs exactes. C'est comme prendre les rumeurs pour la vérité ! Il est nécessaire de revenir aux sources (comme pour les journalistes) pour se garantir au mieux de l'exactitude et de la fraicheur des valeurs. Les sources sont les organisations dont les missions institutionnelles légitiment le savoir-faire dans la production de la donnée recherchée. Estimez-vous plus fiable de savoir combien de clients a l'entreprise en interrogeant le service commercial ou le service technique ?

D'autre part, créer des données qui ressemblent à des données déjà existantes (définition et caractéristiques similaires mais avec des périmètres techniques différents, ou encore dénominations identiques pour des définitions différentes, etc.), induit fatalement le risque de confusion lors de leurs futures utilisations, et donc un usage inapproprié de leurs valeurs. A moins que leurs similitudes et leurs différences ne soient clairement identifiées et que les liens qui doivent, compte tenu des choix de modélisation, exister entre leurs valeurs, ne fassent l'objet de contrôles réguliers.

Désigner les sources fiables pour une demande de donnée, définir les relations entre valeurs de données, telles sont les missions de l'architecte.

10.4.7. *L'architecte*

L'architecte modélise les données pour les intégrer les unes aux autres, et définit le flux de fourniture des données aux consommateurs.

Plus précisément, à l'occasion d'une demande métier.

Il analyse avec le référent métier les données dont il a besoin (les éléments qui composent la donnée, ce qui fait que le référent la considèrera factuellement de qualité). Il établit les relations avec les données existantes produites et consommées dans son organisation. Il décrit techniquement les données contractualisées avec le sourceur et définit les exigences techniques sur les échanges entre son organisation et celle du sourceur, ainsi que les applications de mise à disposition de l'utilisateur. Il bâtit et entretient au fil des demandes un schéma d'intégration des données entre les processus métier et un schéma de flux des données pour répondre à des questions du type : d'où viennent les données, comment entrent-elles dans son organisation, par quels processus, activités et acteurs de son organisation passent-elles ?

Il règle les flux de données (pouvoir législateur). Cette activité l'amène à être le rôle le plus pertinent pour :

– décider de la source de la donnée pertinente pour son organisation et répondre aux questions telles que : quel entrepôt, quel acteur la détient dans un rôle de sourceur ou de sourceur délégué au sein de l'organisation ?

– définir les règles qui régissent les relations entre les attributs des données (formule de conversion par exemple) et dont le respect sera vérifié au quotidien par les administrateurs ;

– décider de l'emplacement des données de référence (par exemple, canal de diffusion officiel, etc.) ;

– choisir les architectures et progiciels de gestion de données.

EXEMPLE 3.– L'architecte est généralement un informaticien dans les entreprises peu sensibles à l'enjeu de maîtrise des données.

Dans le cadre des projets, il définit les « tuyaux techniques » entre les deux applications que lui désigne le demandeur, sans regard à l'ensemble du parcours de la donnée dans le Système d'Information. La circulation des informations, résultante de ce type d'activités d'architecture réduites à la mise en œuvre technique d'un flux

inter applicatif, ne suit pas l'ordre prédéfini des activités de l'entreprise, contributrices au produit final.

L'architecte qui, avec l'expérience, connait les tenants et aboutissants des fils qu'il tire, montre facilement qu'une donnée supposée être la même pour deux consommateurs distincts, peut prendre des valeurs différentes à l'usage (donnée fournie par deux applications différentes par exemple). C'est le début de la prise de conscience par le management de l'enjeu « qualité de la donnée ». L'architecte qui modélise globalement les flux entre processus et les activités métiers, et non pas seulement entre applications, offre alors aux métiers un puissant outil d'aide à la décision, qui permet de sortir son activité du domaine purement informatique et de la faire entrer dans le domaine de l'urbanisme fonctionnel.

Un architecte curieux des contenus des flux et non pas seulement de leurs contenants techniques, est une aide précieuse dans la recherche de performance !

Comment faire travailler ensemble tous les acteurs porteurs de ces rôles ?

Nous venons de définir six rôles. Des acteurs à divers degrés de la hiérarchie, appartenant à des entités organisationnelles différentes, collaborent sur un même rôle et correspondent avec les acteurs des autres rôles. Or, pour être efficace, leur travail doit converger vers les mêmes objectifs, et de façon organisée. Les objectifs prennent leur sens dans les enjeux métier de leur organisation. En effet, quel responsable mobiliserait des ressources s'il n'y avait pas un gain ou un risque à la clé ?

Quels sont ces objectifs et comment émergent ils ? Quels sont les engagements que les acteurs métiers et les acteurs de la gouvernance des données sont prêts à passer les uns avec les autres pour partager et faire circuler les données ? Comment fait-on respecter ces engagements ?

C'est le rôle des pilotes, de faire émerger les opportunités de gain ou les risques, et de montrer en quoi, et comment les aborder dans le contexte de l'organisation, pour les traiter à l'avantage de l'entreprise. Ils assurent également les actions décidées, avec les moyens et les ressources alloués.

10.4.8. *Les pilotes*

Les pilotes sont de deux types : le pilote stratégique ou opérationnel et le pilote tacticien.

Le *Pilote stratégique ou opérationnel* est un entrepreneur (selon les entreprises, il peut s'agir de chefs de projet, de chefs de processus ou de managers métiers). Il décline la stratégie de l'organisation en axes opérationnels pour chaque métier et

pour les données de son périmètre. Il responsabilise les acteurs couvrant les rôles ci-dessus. Il assure la communication stratégique ou opérationnelle. Les directions métiers sont associées à sa prise de décision.

Comment faire en sorte que ces pilotes ne partent pas dans des directions différentes, c'est-à-dire dans des initiatives qui, à une maille supérieure de leur périmètre, nuiraient ou n'optimiseraient pas les initiatives en cours ailleurs dans l'organisation ? Cette convergence est portée par le tacticien.

Pilote tacticien identifie, harmonise c'est-à-dire met en cohérence, facilite et fédère les initiatives locales et partielles qui existent à plusieurs endroits de l'entreprise, sur la base de priorités ou de principes sur lesquels l'ensemble des managers et les pilotes stratégiques et opérationnels ont convergé. C'est le pivot de la gouvernance des données.

Il pilote la gestion et l'amélioration de la qualité des données, au niveau stratégique.

Il définit et fait partager des règles ou des pratiques de maîtrise des données aux pilotes stratégiques et opérationnels, aux instances de gouvernance métier et SI. Ces règles sont déclinées de priorités et de principes approuvés.

Il met en œuvre des états des lieux et des indicateurs pour identifier les lacunes et les faiblesses sur la mise en œuvre des priorités ou des principes. De cette observation, il instruit des axes de développement.

Les pilotes stratégiques et opérationnels auront la charge d'évaluer l'intérêt de faire, ou de ne pas faire, en vue de décider la mise en œuvre. Par exemple, dans une entreprise qui prend conscience des enjeux de la Donnée et nomme un pilote tacticien pour éclairer le sujet, l'un des premiers axes de développement proposé peut être l'identification d'acteurs pour porter les responsabilités de gouvernance des données sur leur périmètre métier ou organisationnel.

EXEMPLE 4.– Le pilotage s'établit sur un périmètre global de données, s'il n'existe pas encore de priorisation entre données.

Le pilote tacticien sensibilise les responsables métiers sur la base :

– des risques ou du manque de visibilité sur les risques de non maîtrise des données (qualité des valeurs et définitions partagées des données, notamment) ;

– de la capacité à résorber ces risques, en agissant sur un périmètre de fonctionnement plus large que celui de leur entité (apport de l'architecte fonctionnel).

Il communique sur les actions et les acteurs existants qui aident à consolider la qualité des données.

Il détecte les lacunes en termes de maîtrise des données par les organisations existantes (projet, processus, administration d'ensemble de données relevant du même métier (commercial, production, etc.), etc.), d'où la mise au point d'un tableau de bord et de préconisations.

Il conseille les métiers sur des pratiques pour bien administrer leurs données (conventions avec fournisseurs, dictionnaire de données, exigences qualité, traçage des données, thèmes à aborder en revue de processus métier, etc.). L'objectif est de passer de démarches locales de corrections de valeurs de données à une démarche collective qui vise la prévention.

Le rôle de pilote stratégique et opérationnel prend forme à travers des projets dans un premier temps, puis est pérennisé par la responsabilité de maintenir les procédures établies. Cette pérennisation est difficile à acquérir sans lien objectivé entre qualité des données et résultats métiers.

Les projets peuvent être aussi bien portés sur la mise en place de nouvelles activités métiers que sur la consolidation de modes d'acquisition, partage et diffusion de certaines données à fort enjeux pour l'organisation.

Sur quels principes et avec quelles régulations de moyens, faire converger les actions de maîtrise des données vers la performance de l'entreprise ?

Toute démarche organisée, c'est-à-dire qui dispose, à minima, de pilotes pour éclairer les enjeux et les choix, débute par une volonté au faîte de l'organisation. Faire émerger cette volonté est le rôle d'un acteur qui appartient au comité de direction et détient un budget : le parrain (ou sponsor).

10.4.9. *Le parrain (ou sponsor)*

Le parrain définit et porte des principes, en lien avec la gestion du patrimoine de données, auprès de ses pairs de la Direction. Il convainc du bon sens de ces principes pour les priorités de l'organisation, sur laquelle il exerce une influence permanente.

Il délègue la gouvernance des données au pilote tacticien, mais demeure le stratège et continue de porter les principes dans toutes ses décisions et actions de promotion.

EXEMPLE 5.– Le parrain émerge souvent des structures Système d'Information, car ces structures sont les premières à détecter les anomalies concrètes que pose une non-maîtrise collective métier-SI des données.

Le parrain initie une clarification des causes profondes des anomalies (de pilotage et d'organisation) et de leurs conséquences, et en tire des principes (par exemple : « une personne externe à l'entreprise ne doit pas recevoir de données incohérentes des différents canaux de diffusion de l'entreprise », « les décisions susceptibles d'être auditées doivent pouvoir être expliquées sur la base des informations utilisées dont l'intégrité doit être vérifiée », « les données qui sont stables et servent à plusieurs processus ou activités doivent être gérées par une administration unique »).

Il les partage avec la direction et donne les moyens de leur promotion en missionnant des pilotes et des projets.

10.5. La cellule de gouvernance ou le moteur du dispositif

La cellule de gouvernance des données est l'organe qui va permettre à l'ensemble du dispositif de gestion de la qualité des données de fonctionner. C'est le moteur, le cœur du dispositif. Cette cellule n'a pas pour objectif de remplir l'ensemble des rôles décrits ci-dessus mais d'en assurer un certain nombre et de coordonner les autres. La figure 10.2 illustre un résumé des rôles couverts par la cellule de gouvernance.

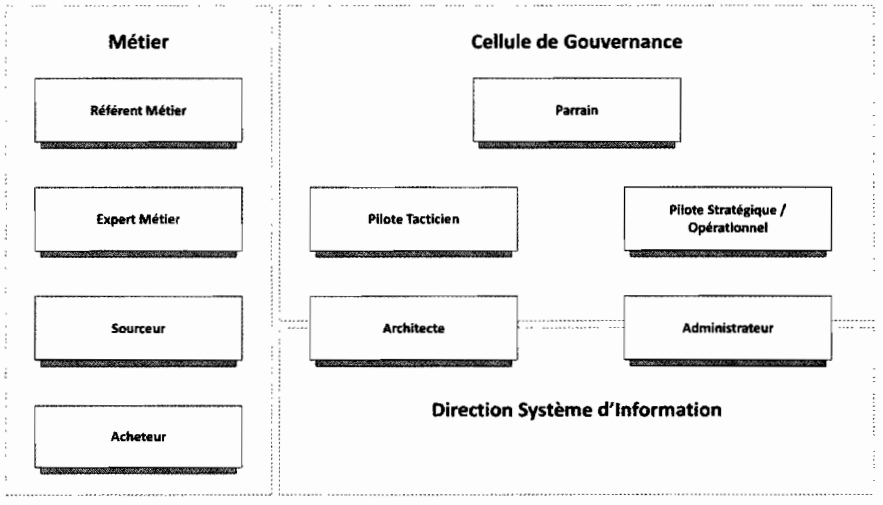

Figure 10.2. *Cellule gouvernance des données*

Elle a un rôle transversal multiforme : relier les différents acteurs et rôles, faciliter leur dialogue, garantir une prise en compte des besoins et contraintes de tous les profils d'utilisateurs, organiser les arbitrages et irriguer l'organisation avec les bonnes pratiques.

10.5.1. *Le cœur*

Le parrain, qui prend la décision de créer la cellule et lui attribue un budget, fait partie « symboliquement » de la cellule. Il en garantit l'ancrage à un niveau hiérarchique suffisant pour que la cellule ait un véritable pouvoir d'action et d'influence.

Le parrain mandate le pilote tacticien pour définir la stratégie, la politique et les grands axes de la démarche qualité de données, et en diriger la mise en œuvre. Le pilote est la tête de la cellule et porte la responsabilité du bon fonctionnement du dispositif sur tout son périmètre d'action.

REMARQUE.– Le périmètre de la cellule couvre la gouvernance des données de référence, données stables et partagées. Ces données sont celles pour lesquelles le besoin d'une gouvernance transverse se fait le plus naturellement sentir. Se concentrer sur les données de référence permet de prouver l'apport de la démarche avant de les étendre aux autres types de données (POC = *Proof of Concept*). Le périmètre de la cellule est structuré en domaines : par donnée, processus métier utilisant les données, ou métier.

Chaque domaine est sous la responsabilité d'un *pilote stratégique ou opérationnel*. Celui-ci décline la stratégie et la politique portées par le pilote de la cellule en axes opérationnels sur son domaine et les met en œuvre, avec l'implication des métiers manipulant les données.

Cette implication des métiers doit être forte et constante sous peine d'isoler la cellule et de la couper des processus métier et du reste de l'organisation. La cellule doit être en constante relation avec les processus pour garantir que les besoins auxquels elle répond sont bien des besoins pérennes et clés pour le métier, ce qui assoit sa légitimité dans l'organisation.

10.5.2. *Le relais avec le métier*

Cette implication est portée par les *référents métiers*. Il s'agit d'interlocuteurs métiers opérationnels qui ne font pas partie à proprement parler de la cellule, mais vont en être le relais, la courroie de transmission entre elle et les métiers.

Ils font remonter vers la cellule les besoins de la structure organisationnelle qu'ils représentent.

Ils apportent leur connaissance des usages de la donnée effectués dans leur structure, ainsi que l'expertise développée par leur métier sur cette donnée.

Ils diffusent dans leur équipe les règles et les bonnes pratiques, la politique de données élaborée par la cellule.

Ils valident pour leur structure les décisions concernant la gouvernance des données (et relaient vers leur management les besoins en décision), comme la validation de la stratégie de la gouvernance et sa déclinaison en axes opérationnels.

Ce sont les *experts métiers* qui, mobilisés par les référents en réponse aux demandes de pilotes opérationnels ou stratégiques, vont apporter leur expertise pointue sur des thèmes bien précis liés à la donnée.

10.5.3. *Les partenaires*

D'autres rôles et types d'acteurs sont impliqués dans la gouvernance des données et peuvent, ou non, selon les choix d'organisation faire partie de la cellule de gouvernance.

C'est d'abord le cas des *administrateurs*, qui de par leur rôle ont développé une connaissance fine de la donnée et de certains de ses usages, connaissance complétant celle des référents métiers et des experts métiers. Ils ont un profil « informatique » s'ils sont administrateurs techniques des applications (outils informatiques) manipulant les données sous gouvernance, et ne font pas partie à proprement parler de la cellule ; ils ont un profil dit « fonctionnel » s'ils sont régisseurs (*data stewards*), et font dans ce cas partie de la cellule de gouvernance.

Un autre type de partenaire clé est *l'architecte système d'information.* Il est le relais de la cellule de gouvernance vers les équipes informatiques. L'architecte intervient :

- sur la modélisation des données et des flux ;
- sur le respect, par les projets informatiques, des règles de normalisation ;
- sur la prise en compte dans les démarches « schéma directeur » et « gestion du portefeuille de projets » des besoins liés à la gestion des données, comme la mise en place de référentiels de données ou d'outils d'échange de données.

L'appartenance d'un architecte fonctionnel à la cellule de gouvernance est identifiée comme facteur clé de succès (voir section 10.6).

Deux autres rôles apportent un soutien précieux à la cellule de gouvernance :

– le *sourceur* : qu'il soit interne ou externe, il est un maillon essentiel de la qualité de donnée et doit être associé aux travaux de la cellule ;

– l'*acheteur* : qui va optimiser l'achat des données externes, négocier pour que des critères de qualité de données soient précisés dans les contrats avec les fournisseurs et apporter son expertise sur la dimension juridique liée aux données (contraintes à respecter).

10.6. Facteurs-clés de succès

Nous avons identifié à partir de nos différentes expériences, les principaux facteurs clé de succès associés à chacun des rôles impliqués dans la gouvernance de données.

Le parrain doit être situé assez haut dans la hiérarchie pour permettre à la cellule de gouvernance d'avoir un pouvoir de décision à effet transverse ; il doit posséder assez de charisme et être lui-même convaincu de l'importance et de l'utilité de la cellule pour relayer les messages de celle-ci et les renforcer. Le parrain ne doit pas seulement prendre la décision de création de la cellule de gouvernance, il doit aussi porter cette décision à chaque occasion.

Le pilote tacticien doit posséder une capacité à communiquer et convaincre pour bien vendre la gouvernance et faire en sorte qu'elle soit bien prise en compte dans les processus métiers ; il doit bien comprendre les enjeux du métier pour toujours aligner la stratégie et les activités de la cellule sur les besoins métier. Le pilote tacticien doit enfin animer et fédérer sans cesse les différentes parties prenantes de la manipulation des données, pour garantir et pérenniser la transversalité de la gouvernance des données, et l'implication des acteurs métiers tels que les référents et les experts.

Le pilote opérationnel et stratégique, pour bien mettre en œuvre les axes opérationnels de la gouvernance, doit maîtriser, en plus du domaine sur lequel il intervient, les principes et méthodes de la qualité de données : dimensions qualité, outils, expérience de gestion de projet qualité de données.

L'administrateur, qu'il soit de profil informatique (administrateur de base de données ou d'application informatique) ou de profil fonctionnel (intendant des données), doit posséder une sensibilité à la donnée et à tout ce qu'implique une bonne gestion de celle-ci.

Un architecte fonctionnel (ou urbaniste), prenant en charge la modélisation des données et portant les règles à respecter sur ces données dans les projets de systèmes d'informations constitue un facteur clé de succès pour la gouvernance des données.

Le référent métier, pour assurer son rôle clé de relais entre la cellule et l'équipe au nom de laquelle il doit « parler », doit être réellement représentatif de cette équipe : représentatif par la connaissance qu'il a de l'usage des données au sein de la totalité de l'équipe et par l'écoute qu'il suscite au sein de cette même équipe et auprès de son responsable. Sinon, les messages de la gouvernance ne seront pas relayés, ni les réels besoins de son équipe, ce qui coupe la cellule du métier. Le référent doit se voir accorder par son responsable suffisamment de disponibilité et de reconnaissance pour pérenniser le rôle. Il doit enfin savoir mener des actions de conduite du changement au sein de la structure qu'il représente.

L'expert métier doit être convaincu par l'utilité de la gouvernance des données pour l'amélioration des activités métier, dont les siennes, et par l'importance de l'apport de son expertise pour le bon fonctionnement du dispositif de gouvernance. Un expert se sentant un devoir d'alerte sur les évènements et incidents impactant les données constitue un maillon essentiel de la chaine de gestion de la qualité des données.

Un *sourceur* avec un réel souci de la satisfaction client est un facteur clé de succès.

L'acheteur enfin, peut apporter un réel plus s'il connait bien l'offre des sourceurs et a une sensibilité donnée lui permettant d'intégrer cette dimension dans ses négociations.

10.7. Comment gérer la gouvernance ?

10.7.1. *Exemple de la branche Global Gas & GNL de GDF SUEZ*

La cellule de gouvernance des données de référence est située au sein de la DSI de cette branche, donc une DSI métier. Ceci a permis de porter la dimension transverse de la cellule et une neutralité garantissant la prise en compte de tous les besoins des différents métiers de la branche.

La cellule comporte un pilote, des administrateurs fonctionnels (*data steward*, qui gèrent au jour le jour la qualité des données), un architecte fonctionnel. Elle est rattachée au directeur adjoint en charge de la maîtrise d'ouvrage du SI. Elle a pour partenaires privilégiés, dans le dispositif de gouvernance, des référents métiers qui mobilisent des experts métiers en fonction des besoins. Elle a de nombreux échanges avec les fournisseurs de données, qui sont un des maillons de la qualité des données, et avec les acheteurs, pour mieux contractualiser la relation avec ces fournisseurs.

Pour faire « fonctionner » ensemble tous ces acteurs situés en différents endroits de l'organisation, un processus transverse « gestion des données de référence » a formalisé les rôles, responsabilités de chacun, ainsi que les activités à mener et les indicateurs de pilotage et de performance associés. Le pilote de la cellule de gouvernance est aussi le pilote du processus (garant de sa mise en œuvre efficace).

Les référents métiers sont associés dans un réseau, à vocation opérationnelle, qui se réunit tous les un à deux mois. Tous les trimestres, les responsables de ces référents, eux-mêmes rassemblés dans un réseau, se réunissent pour valider et prendre les décisions transverses liées à la gouvernance de données. Réseau des référents et réseau des responsables traitent à la fois du contenant (les outils de gestion des données) et du contenu (les données elles-mêmes). L'animation de ces deux réseaux est une mission essentielle du pilote de la cellule de gouvernance.

Les liens avec les instances informatiques, sont portés en collaboration avec l'architecte fonctionnel : comités de pilotage des projets impliquant les données, démarches « schémas directeur » et « gestion du portefeuille de projets » et normalisation des échanges.

10.7.2. *Exemple de Bouygues Telecom*

Le service administration et gouvernance des données est rattaché à la direction études et connaissance client, direction métier donc. Sa priorité est néanmoins sur l'amélioration et la gouvernance des données opérationnelles, ce qui en fera bénéficier indirectement les fonctions de l'informatique décisionnelle.

Le service n'est pas directement en charge des données ou des référentiels, mais de la gouvernance des données. Sous le parrainage d'un membre du comité de direction générale (CoDG), la structure s'appuie :

– en interne, sur des chargés de projet multi-compétences (rôles expert métier, pilotes, chef de projet, chargé d'étude) ;

– en externe sur des responsables de référentiel situés dans les directions métiers, nommés par le coDG. Ceux-là ont le rôle de référent métier, l'administration et le pilotage du référentiel étant généralement délégué au sein de sa direction à un responsable opérationnel (rôles administrateur et pilote opérationnel).

Le pilotage s'effectue à plusieurs niveaux, sous forme :

– d'un comité exécutif validant les orientations et l'implication de la stratégie données de l'entreprise ;

– d'un comité de pilotage des référentiels, permettant l'alignement et l'amélioration des processus, méthodes et gouvernance de chaque référentiel ;

– de suivis individualisés entre la structure de gouvernance et les responsables de référentiels et leurs délégués.

Les évolutions et projets d'amélioration de données s'inscrivent dans la démarche de pilotage projet de l'entreprise comme toute évolution fonctionnelle. Il n'y a pas de pilotage projet indépendant de la vue « donnée » de l'entreprise. Le tableau 10.7 donné en annexe, section 10.9, représente une synthèse des acteurs faisant partie de la cellule de gouvernance ou interlocuteurs de la cellule.

10.8. Conclusion

Nous reprenons les rôles présentés dans la littérature, ceux définis chez nos témoins et nous avons essayé de faire « une table d'équivalence » : on voit par exemple que le rôle théorique du parrain est chez nos témoins réparti entre celui qui est identifié comme parrain, avec une délégation forte aux pilotes tacticiens et stratégique et opérationnel. Pour la mise en pratique, il ne faut pas oublier comme indiqué plus haut que « un rôle n'égale pas une personne » et en réalité, une personne peut avoir plusieurs rôles ou assurer un certain rôle à hauteur de x % de son activité.

Un facteur clé de succès est certainement la sensibilisation, l'intérêt, l'adhésion que pourront susciter la cellule de gouvernance qui permettront d' « enrôler » les bonnes personnes (au sens reconnues comme porteuses des valeurs de l'entreprise quel que soit leur niveau hiérarchique). L'important est de donner l'exemple : en démarrant avec les rôles que nous décrivons clairement identifiés, cette démonstration des personnalités reconnues dans leur domaine sera le meilleur moyen de provoquer un vote général pour une donnée de qualité.

10.9. Annexes

Théorie		Pratique	
Il met à disposition les moyens, pilote les alignements stratégiques et s'assure du maintien de cette stratégie en mode opérationnel. Il garantit l'adoption par l'entreprise d'une donnée de haute qualité mesurable. Le sponsor négocie avec les fournisseurs externes des engagements sur la qualité de données.	Parrain (Sponsor)	Parrain	Il mandate le ou les pilotes de la cellule de gouvernance. C'est sa caution qui va permettre à la cellule d'exister. Il met à disposition les moyens.
		Pilote tacticien	Identifie, harmonise (met en cohérence) et fédère les initiatives locales et partielles qui existent à plusieurs endroits de l'entreprise (on part rarement de rien). Pilote la gestion et l'amélioration de la qualité des données, au niveau stratégique. Met en œuvre un suivi des indicateurs.
		Pilote stratégique ou opérationnel	Décline la stratégie en axes opérationnels pour chaque métier pour les données de son périmètre. Responsabilise les acteurs. Assure la communication stratégique ou opérationnelle. Les directions métiers sont associées à la prise de décision.
C'est celui qui va indiquer les règles, les exigences. On parle de multipropriété car on s'aperçoit que souvent un objet n'appartient pas exclusivement à un métier.	Propriétaire de la donnée (Data owner, Business Data owner, BDO)	Expert métier	Cette notion de « propriétaire des données » est répartie entre référent, expert métier et sourceur mais c'est bien le même rôle.
		Référent métier	Définit les procédures métiers (ou coordonne leur définition), et les règles associées aux données. Supervise les indicateurs qualité de fourniture (ou de mise à disposition) et commandite les actions au jour le jour de remise en conformité, relevant de sa responsabilité métier. En lien étroit avec le référent données métier, les fournisseurs, l'équipe de « l'administrateur », pour la mise en œuvre des corrections.

Théorie		Pratique	
			Par domaine métier : - « définit » le type de données souhaité, leurs usages, les exigences de qualité pour l'usage de ces données, les attendus qualité (indicateurs), le niveau d'accessibilité ; - approuve le contrat de fourniture, et le dispositif de gouvernance ; - définit les seuils d'alerte sur les indicateurs et mène l'analyse et les actions structurelles si les valeurs des indicateurs sont en deçà.
Correspond à l'action récurrente de prendre soin de la donnée. Le steward intervient au point d'entrée de l'application, au point de transformation, au point d'utilisation. Il gère la qualité de la donnée et supporte la mise en œuvre des règles de gouvernance.	Régisseur ou intendant ou coordinateur (Data steward)	Adminis-trateur	On parle d'administrateurs de la donnée avec des data stewards qui ont une connaissance fonctionnelle de la donnée.
Celui qui modélise les données. En pratique, ce rôle est souvent délégué à la DSI car les compétences sont rarement présentes au sein du métier.	Analyste (Data analyst)	Architecte	Analyse avec le référent métier les données dont il a besoin et recherche une source (modélisation). Définit et maintient le schéma d'intégration des données entre les processus métier. Conseille puis décide la source de la donnée pertinente et l'emplacement des données maitres. Décrit techniquement les données. Définit les exigences techniques sur les échanges avec les sources et les applications de mise à disposition de l'utilisateur (choix des architectures et progiciels MDM).

Théorie		Pratique	
Il a un rôle transverse dans l'organisation, supervise l'évolution des modèles. Il valide la cohérence des modèles. Il s'assure que les objets ne comportent ni redondance ni incohérence d'un métier à l'autre.	Architecte de données (Data architect)	Architecte	Voir ci-dessus. Dans la pratique, les rôles d'analyste ou architecte de données sont confondus.
Il suit les coûts de modélisation de la donnée. Il est en charge d'établir les lignes de bilan concernant le budget en particulier des données de référence.	Responsable des coûts de la donnée (data cost accountant)	Pilote stratégique ou opérationnel	Rôle pris en charge par le pilote stratégique ou opérationnel.
Supervise la sécurité du transport et du stockage de la donnée. Si le contenu est important pour lui, il se focalise sur l'infrastructure et les activités pour garder la donnée protégée et disponible aux utilisateurs. Il travaille avec le data steward pour résoudre les problèmes, faire évoluer les systèmes et implémenter les transformations de données.	Gardien de la donnée (Data custodian	Administra-teur	Responsable de la mise à jour des données partagées et de leurs valeurs : modes opératoires, processus de collecte, de contrôle et validation des données en entrée, de mise à disposition des données aux usagers ou aux applicatifs. Responsable du contrôle qualité des données en entrée et sortie ; de la consolidation des valeurs des indicateurs. Responsable de la mise en œuvre des corrections. Responsable des habilitations dans l'application (l'aspect sécurité des données peut être géré par une direction « sécurité ») en fonction de la structure de la société, l'administrateur peut être un responsable qui délègue.

Théorie		Pratique	
		Sourceur	Contribue à définir la donnée (ses caractéristiques, ses attributs) selon sa capacité à la mettre à disposition du référent métier, pour répondre aux usages et exigences qualité demandés et définit son niveau d'accessibilité. Désigne les fournisseurs des valeurs. Contractualise au nom de sa Direction la fourniture des données et le dispositif de suivi de la qualité de fourniture.
		Acheteur	Contractualise la fourniture, la traçabilité de la qualité, le niveau d'accès aux données livrées.

Tableau 10.7. Synthèse des rôles et responsabilités en théorie et en pratique

10.10. Bibliographie

[LOS 09] LOSHIN D., *Master Data Management*, Morgan Kaufmann Publishers, New York, 2009.

[REG 08] REGNIER-PECASTAING F., GABASSI M., FINET J., *MDM : Enjeux et méthodes de la gestion des données*, Dunod, Paris, 2008.

[BON 10] BONNET P., *Enterprise Data governance : Reference & master data management semantic modeling*, ISTE, Londres, 2010.

[BER 07] BERSON A., DUBOV L., *Master Data Management and Customer Data Integration for a Global Enterprise*, McGraw-Hill Osborne Media, New York, 2007.

[DATA] DATA GOVERNANCE BLOG, http://www.primedataconsulting.com/blogPDC/.

Chapitre 11

La valeur de la qualité en gouvernance des données dans la chaîne logistique

11.1. Introduction

11.1.1. *Pourquoi donner une valeur à la qualité des données ?*

Les données sont essentielles au fonctionnement d'une entreprise moderne. La qualité des données est un facteur de compétitivité vital, voire de survie face à une pression accrue de régulation (CNIL[1], SOX[2], Bâle III[3], etc.). S'il est aisé de comprendre intuitivement l'importance de la qualité des données qui soutiennent les processus informatisés, rendre cette importance tangible par l'intermédiaire de critères économiques quantifiables demeure une gageure.

La littérature propose le plus souvent des estimations quantitatives, fondées sur des critères statistiques et économétriques, par exemple : « les problèmes de qualité de données client coûtent aux entreprises américaines le montant stupéfiant de 611 milliards de dollars par an. »[4] D'autres estimations sont basées sur des appréciations comptables : « Les coûts de la non qualité des données, [...], peuvent atteindre 10 à 25 % du revenu ou du budget total d'une organisation. »[5]

Chapitre rédigé par Thierry DÉLEZ et Nicole BUSSAT.

1. Commission nationale de l'informatique et des libertés, voir [CNI 05].
2. Sarbanes-Oxley Act, régissant la publication des résultats des entreprises cotées en bourse aux Etats-Unis, voir [SOX 02].
3. Accord global de régulation de l'activité bancaire, voir [BAS 10].
4. [ECK 02, p. 5].
5. [ENG 99, p. 12].

D'autres approchent la non-qualité par l'estimation des coûts unitaires des erreurs de données : « même à 10$ par erreur, une estimation conservatrice, la compagnie [d'assurances] est exposée à plus de dix millions de dollars par année aux risques de la non-qualité des données. »[6] Ces chiffres représentent souvent un argument promotionnel pour justifier l'acquisition d'outils complexes et coûteux de gestion de la qualité des données, de services de consultance ou d'investissements internes.

Dans le contexte de l'entreprise, de tels indicateurs génériques n'offrent qu'une information partielle et peu pratique que le responsable des données ne peut exploiter pour justifier les ressources et les investissements nécessaires à l'accomplissement de sa mission. Alors que le coût de la gestion des données est explicite (consommation de ressources, outils et services par les processus), les bénéfices générés demeurent opaques.

Les coûts de la gestion des données augmentent constamment avec les exigences croissantes d'intégration entre processus, acteurs et systèmes, multipliant exponentiellement le nombre des interfaces et par conséquence la vulnérabilité à la qualité des données. De plus, les problèmes de qualité, autrefois contenus dans l'entreprise, sont désormais exposés à la vue des clients et des fournisseurs. Ces problèmes sont exacerbés dans les entreprises opérant dans des cadres législatifs hautement demandeurs en information (par exemple, dans les domaines agroalimentaire, pharmaceutique, chimique, financier, etc.).

Les entreprises doivent donc justifier les demandes d'investissement accrues en matière de qualité de données sur la base de prévention de risques (pertes commerciales, prévention de problèmes légaux, réduction d'exposition à une image négative, etc.), dont la quantification est essentiellement fondée sur des méthodes probabilistes et statistiques. De plus, il demeure impossible de déterminer *a posteriori* si les bénéfices ont été effectivement réalisés, les risques évités demeurant intangibles. Par contre, les situations d'échec sont cruellement évidentes par la matérialisation des risques.

La gestion des données est souvent confiée au service informatique, qui privilégie une approche technique, souvent orientée sur le contrôle des métadonnées, à forte composante technologique, et supportée par des processus souvent inadéquats, hérités des méthodes du support applicatif. Cependant, les investissements technologiques consentis en matière de gestion de qualité de l'information ne produisent pas les effets escomptés.

6. [ECK 02, p. 6].

Ainsi, la justification des investissements est-elle souvent fondée sur des expériences négatives de l'entreprise, afin d'éviter le renouvellement de problèmes passés. Cette approche, au demeurant compréhensible, ne permet pas la systématisation de processus préventifs et prédictifs. De même, elle aboutit rarement à des améliorations structurelles fondamentales, faute de pouvoir générer un consensus entre le métier et les services informatiques sur les priorités et les moyens à mettre en place.

Cette situation est directement imputable au manque de capacité des gestionnaires de données de déterminer la validité économique (*business case*) de la qualité des données et d'en démontrer la tangibilité dans la durée. Comme tout cadre responsable, il devrait être à même de répondre à des questions managériales de base, telles que :

- quel est le coût de la non-qualité des données ?
- quelle est la valeur supportée par les données de qualité ?
- quel est le retour sur investissement en matière de qualité des données ?
- quelle est la valeur générée par x pourcent de qualité supplémentaire (ou à l'inverse, la valeur menacée par une érosion de la qualité des données ?)
- quel est le niveau de qualité de données optimal en matière de coûts/bénéfices ?

Ces questions peuvent s'inscrire dans un contexte immédiat (photographie à un moment donné) ou dans une vision historique afin de déterminer l'efficacité des investissements et des dépenses consentis.

Il faut donc munir le gestionnaire de données d'outils et méthodes de valorisation simples et efficaces, permettant de *démontrer que la qualité des données génère une valeur tangible pour l'entreprise*[7]. Cette valeur doit être tangible (en relation directe avec l'activité économique réelle), financièrement crédible (les valeurs générées ou menacées sont directement conformes avec la situation comptable) et immédiatement utilisable (les informations retournées déterminent les priorités d'action et justifient le coût des mesures de correction).

Ces outils ne remplacent pas ceux utilisés dans un contexte opérationnel de gestion des données. Ils les complémentent en fournissant une vision purement managériale destinée à fournir des instruments économiques de gestion.

7. Nous considérerons que la valeur de la qualité des données est liée à leur utilité pour les processus de l'entreprise, l'utilité étant déterminée par la contribution totale à la réalisation des objectifs économiques. Cette définition exclut d'emblée toute approche marginaliste.

11.1.2. *Valorisation de la qualité des données*

On distinguera deux sortes de données dans le système de valorisation présenté dans ce chapitre[8] :

– les *données de base* ou de *référence*[9], sont le plus souvent la représentation dans les systèmes d'information d'entités du monde réel, telles que clients, produits, etc. Elles sont généralement relativement statiques. Leur rôle consiste à permettre l'exécution des processus de l'entreprise par le biais des transactions ;

– les *données transactionnelles* représentent les opérations de la chaîne logistique, telles que commandes clients, commandes fournisseurs, ordres de fabrication, bulletins de livraison, factures, etc. Ces opérations sont le plus souvent directement liées aux processus de l'activité économique, dont ils sont les *porteurs de valeur*.

Certaines doctrines assimilent les données à des actifs de l'entreprise[10]. Cette perception est à la fois erronée et perverse lorsqu'elle est prise au pied de la lettre, car elle empêche de considérer la valorisation des données de manière efficace.

En effet, *les données de base n'ont aucune valeur réelle intrinsèque*. Par exemple, la perte de données d'un produit affectera la capacité de l'entreprise à exécuter les processus qui s'y rapportent (ventes, production, expéditions, etc.) mais les inventaires physiques dans les entrepôts demeureront invariants, tant en quantité qu'en valeur. De même, la suppression de données de base inutilisées ne change en rien la valeur du bilan, alors que la cession ou l'élimination d'un actif physique sera comptabilisée. On peut comparer la relation entre les données et les entités qu'elles représentent à celle d'une carte avec son territoire.

L'utilité des données de base est déterminée par leur rôle dans les processus transactionnels de l'entreprise (en termes imagés, ce à quoi sert la carte). De bonnes données permettent une exécution optimale des transactions. A l'opposé, des données de base de mauvaise qualité compromettent la capacité opérationnelle de la société.

---

8. Le modèle présenté dans ce chapitre se focalise exclusivement sur les systèmes transactionnels et exclut par définition tous les autres types de données, tels que données produit, bases de données informationnelles ou statistiques, données de configuration, etc.
9. *Master Data* en anglais.
10. Les données produit sont certainement les seules qui peuvent s'appliquer à cette classification.

Ainsi, toute tentative de donner une valeur aux données de base (hors coûts d'acquisition, de création et de maintenance), indépendamment de leur usage, est vouée par avance à l'échec. Par conséquence, *la valeur de la qualité des données de base est déterminée par leur utilité transactionnelle.*

Les données transactionnelles représentant l'activité économique de l'entreprise sont généralement valorisables de manière directe. Par exemple, la valeur d'une commande client correspond à la somme de la valeur représentée par chaque ligne (produit des quantités et prix des produits ou services).

De plus, la qualité des données de base conditionne directement la capacité de l'entreprise à exécuter ses transactions. Il existe un lien de causalité entre la qualité des données de base et la valeur transactionnelle supportée ou menacée. Ainsi, la qualité de l'information contribue directement à la réalisation de la valeur transactionnelle, *a contrario*, la non-qualité empêche la génération de valeur.

Cela implique que la valeur de la qualité (ou de la non-qualité) d'un objet de données spécifique est variable et fluctue en fonction de l'activité de l'entreprise représentée par la situation transactionnelle, qui dépend de l'activité économique. La qualité d'une donnée inutilisée n'a donc par définition aucune valeur.

La valeur de la qualité des données de base est déterminée par leur contribution aux transactions qu'elles supportent. Elles obéissent donc à une logique de valeur variable.

Le principe de base est intuitif mais sa mise en opérations nécessite quelques précautions. En effet, il convient de bien comprendre l'usage des données au sein des transactions pour établir correctement la valeur, par exemple :

– une donnée produit peut être utilisée à la fois dans un contexte prospectif (prévisions de vente) et opérationnel (ventes actuelles), entre lesquels il existe des recouvrements ;

– une transaction peut utiliser plusieurs objets de données de natures différentes : une commande client contient des données client, dont la qualité affecte la totalité de la commande, et des données produit dont la qualité est associée aux lignes de commande. Cependant, le cumul de la valeur de la qualité des données client et produit ne peut excéder celle de la commande ;

– les transactions sont liées de manière non-séquentielle : une commande client reste ouverte tant qu'elle n'est pas honorée. Il convient donc d'éviter de cumuler la valeur de la commande, de l'ordre de production associé et du bulletin de livraison, la valeur totale étant celle de la commande ;

– la qualité d'une donnée à un moment donné peut affecter des transactions non encore exécutées. Cependant, il est nécessaire de déterminer l'impact de la non-qualité de manière prospective afin de prendre les mesures de prévention nécessaires.

Ces exemples illustrent la complexité de l'intrication entre données, transactions et processus dans l'entreprise. La méthode de valorisation, pour demeurer crédible, doit tenir compte de cette complexité afin de fournir une information claire et utile. Cette méthode illustre la manière de déterminer la valeur de la qualité ou de la non-qualité d'une manière pratique et efficace.

11.1.3. *Caractéristiques de la méthode de valorisation de la qualité des données*

Le but recherché est de fournir un instrument de gouvernance, efficace, aisé à mettre en œuvre, crédible et robuste. Il n'est donc nullement question de fournir une vision scientifique, holistique ou absolue de la valorisation des données. La méthode se focalisera donc exclusivement sur les domaines et activités de l'entreprise qui ont une relation directe avec une valeur économique, en laissant de côté les activités non-directement valorisables (telles que la maintenance des données de configuration des systèmes, de la sécurité informatique, des e-mails du personnel, etc.).

La méthode démontre la valeur de la qualité et la valeur menacée par la non-qualité des données en conformité avec la réalité économique de l'entreprise. Elle est applicable en mode instantané (permettant de mesurer la valeur de la qualité à un moment t) et en mode prospectif (permettant de prévoir l'impact de la qualité actuelle des données sur les activités futures de l'entreprise).

La méthode s'appuie sur des principes pragmatiques et indiscutables :

– crédibilité : la valeur de la qualité des données doit reposer sur une base méthodologique rigoureuse, conforme avec la situation financière et comptable de l'entreprise ;

– simplicité : la méthode de valorisation doit être aisément compréhensible par chaque métier impliqué et aisée à mettre en œuvre ;

– représentativité : la méthode se focalise sur les données les plus importantes pour les processus principaux de l'entreprise (maîtrise du niveau de détail) ;

– utilité : les résultats sont directement utilisables pour la gouvernance et l'amélioration continue.

11.2. Approche générale

11.2.1. *Aperçu de la méthode*

La qualité des données détermine la capacité à réaliser la valeur des transactions courantes ou à venir. Ainsi, la valeur de la qualité des données doit être considérée selon une approche contributive : seules les données de qualité permettent de réaliser la valeur de la transaction. Le calcul de la valeur de la qualité s'exprime de manière binaire : la valeur d'une transaction est soit confirmée, soit menacée par la qualité des données qu'elle contient. Afin d'aboutir à ce résultat, il convient de déterminer :

– le contexte de valorisation de chaque transaction, déterminant son objectif de gouvernance ;

– la qualité objective des données associée à chaque transaction par l'intermédiaire de mesures de conformité avec les règles de gestion ;

– la répartition de la valeur des transactions sur les données de base en fonction de la segmentation des transactions et des dimensions d'impact.

Le *contexte* définit des groupes de transactions distincts, satisfaisant des objectifs de gouvernance particuliers. En général, ils correspondent aux horizons de gouvernance de l'entreprise, de l'opérationnel à la planification à long terme comme illustré par la figure 11.1.

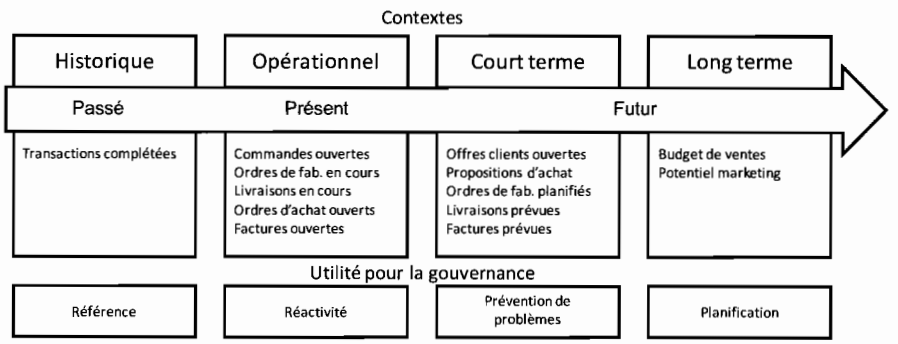

Figure 11.1. *Exemple de contextes et transactions associées*

Chaque contexte doit être considéré séparément pour éviter de mêler des natures différentes de valeurs, en adéquation avec les principes comptables. En effet, il ne serait pas très rigoureux de mêler valeur prospective (budgétée) avec valeur

effective (réalisée). Les objets de données (clients, fournisseurs, produits, matières premières) apparaissant généralement dans plusieurs contextes, cette distinction est incontournable. Chaque contexte contient des transactions (ou sous-type de transactions) clairement identifiables, dont le champ d'application est très clairement déterminé et normalement propre à un seul contexte. Le contexte nécessite également la détermination d'une *valeur de référence* qui fournira la base de calcul pour la valorisation de toutes les transactions chaînées au sein d'un contexte.

La qualité objective des données de référence est établie en en mesurant leur conformité avec les règles de gestion objectives qui les gouvernent. Ceci implique que ces règles existent et sont associées de manière précise et non ambigüe aux transactions et que l'impact de leur non-respect est clairement identifié. La méthodologie sera détaillée ci-dessous.

Enfin, la méthode détermine la manière dont la valeur des transactions est *répartie sur les données de référence*. En effet, les transactions peuvent recourir à plusieurs données de référence et être segmentées de diverses manières. Une information client critique, absente peut menacer la totalité de la valeur d'une commande. En revanche, une erreur dans un produit ne menace que la valeur de la ligne de commande qui s'y réfère. De même, tous les champs d'un objet de données n'ont pas une importance égale : certaines données essentielles sont potentiellement destructrices de valeur, d'autres peuvent engendrer des nuisances moins sévères : délais, risques, etc. La détermination de la valeur de la qualité ou de la non-qualité requiert une analyse d'impact de chaque élément de données sur la transaction. Ainsi, la valeur d'une transaction sera répartie sur les objets de données de référence selon une dimension structurelle relative à la segmentation de la transaction et une dimension d'impact déterminant les catégories de risques liées à la non-qualité. Les dimensions d'impact représentent l'effet de la non-qualité sur une transaction : destruction de valeur si la transaction est totalement compromise, délais dans l'exécution, risques légaux, etc. Ces dimensions d'impact sont par définition clairement distinctes, ce qui permet de considérer simultanément plusieurs types de risques pour le même objet de données sans poser de problématiques particulières de cumul de valeur.

La méthode de valorisation de la qualité des données, illustrée par la figure 11.2, répartit la *valeur de référence* sur les *instances d'objets de données*[11], identifiant ainsi leur *valeur contributive* (valeur *potentielle* que l'objet de données supporte

11. L'instance d'un objet de données identifie une occurrence particulière de celui-ci. Par exemple, l'objet « *produit fini* » contient l'ensemble des produits finis de l'entreprise. L'instance « *produit X* » identifie un produit déterminé : X.

dans la transaction). La *qualité de l'objet de données* détermine si la valeur contributive est *confirmée* par la qualité ou *menacée* par la non-qualité, et *qualifie* cette valeur par *dimensions d'impact* (critique, délais, risques, etc.). La valeur est ensuite consolidée au niveau de l'objet de données, de la transaction ou du contexte en fonction de leur nature (confirmée ou menacée) et de leur qualification.

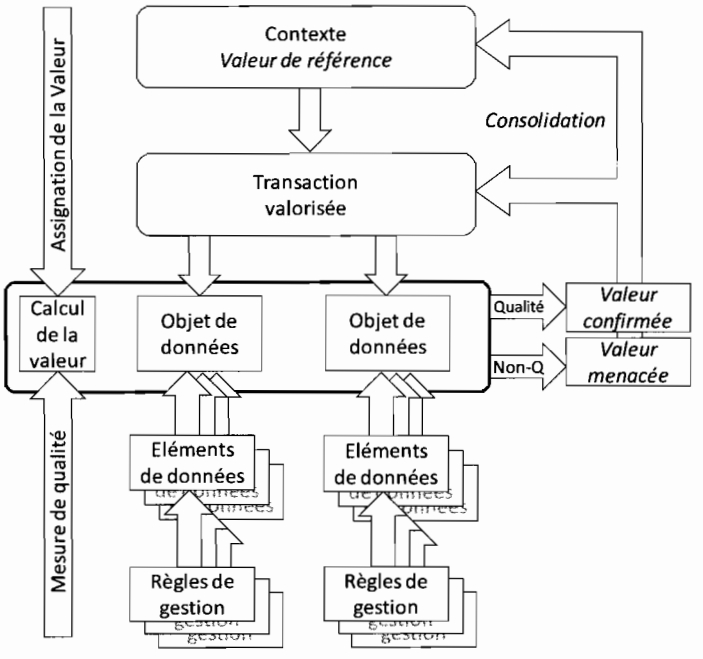

Figure 11.2. *Méthode de valorisation de la qualité des données*

Par définition, les données qui ne peuvent être associées à au moins une transaction valorisée, quel que soit le contexte ou la dimension d'impact, n'ont aucune valeur déterminable. Ceci inclut les données de base sans utilisation actuelle ou prévue, celles qui n'ont jamais été utilisées et sans prévision d'utilisation et celles utilisées exclusivement dans des transactions non valorisables. Ceci règle le sort des données obsolètes, dont la valeur est intrinsèquement nulle[12].

12. Il est de bonne guerre de valoriser les données obsolète à zéro mais de les considérer dans le calcul du coût de maintenance sur une base moyenne. En effet, cette méthode fournit un motif économique pour les éliminer.

11.2.2. *Pré-requis de la méthode*

L'approche requiert deux pré-requis essentiels :

– une connaissance précise des processus principaux de la chaîne logistique de l'entreprise et des transactions qui les supportent afin de déterminer le contexte de gouvernance des données ;

– une base objective de mesure de la qualité des données fondées sur des règles de gestion.

11.2.2.1. *Mesure de la qualité des données*

La qualité des données s'entend par *la mesure de la conformité entre les données de référence et les règles objectives qui les gouvernent.* Cette notion peut sembler étrange ou restrictive au premier abord pour les gestionnaires familiarisés avec les notions de dimensions de qualité des données. Il s'agit cependant exactement des mêmes concepts, mais encapsulés dans une structure compréhensible pour les métiers[13].

Les règles de gestion fixent les conditions qui déterminent si un élément de données particulier, dans un contexte d'application spécifié, satisfait aux critères de qualité recherchés. Il peut s'agir de critères d'unicité, de complétude, de cohérence, etc. Chaque règle est associée à des attributs particuliers : responsabilité (qui est responsable de sa formulation, de son application et suivi, de sa mise en œuvre), applicabilité (tous les produits finis, tous les clients européens, etc.), impact de la non-qualité sur les transactions (empêche l'exécution, retarde l'exécution, compromet l'image de l'entreprise, génère un risque légal, etc.).

Les règles de gestion sont énoncées dans le langage naturel du métier, afin que son interprétation soit aussi claire et non ambiguë que possible. Plusieurs règles différentes de diverses natures peuvent s'appliquer à un élément de données si nécessaire. Dans la mesure où la qualité est déterminée par l'absence de défaut, le nombre de règles appliquées à un objet de données et ses éléments n'est pas vraiment un facteur important, si ce n'est la complexité inhérente au suivi d'une pléthore de règles trop élémentaires.

Les règles de gestion déterminent le périmètre de valorisation de la qualité : un objet de données ou un de ses éléments associé à aucune règle de gestion (en général ou dans un contexte particulier) n'est pas pertinent d'un point de vue de la

13. La typologie rigoureuse de la qualité des données est souvent trop complexe pour être abordée à un niveau managérial, voir par exemple le chapitre trois *Data Quality*, section *Data Quality Rules*, de Adelman, Moss et Abai [ADE 05], p. 51-57 qui est très précis sur le plan formel mais incompréhensible pour la plupart des non-initiés.

gouvernance, car personne n'est en mesure de déterminer la qualité objective et donc de prendre les mesures de corrections adéquates. Ainsi, l'extension du périmètre de la gouvernance ne s'exprime que par le biais de règles de gestion supplémentaires, déterminant clairement les objectifs et les responsabilités de chacune. Ce modèle exclut donc toute vision absolue ou pure de la qualité pour l'ancrer fermement dans un contexte de gestion et de contrôle clairement délimité, gage de bonnes pratiques managériales.

La méthodologie supportant les règles de gestion de la qualité des données doit reposer sur un cadre formel produisant des indicateurs fiables, crédibles et alignés avec les besoins métiers[14]. La méthodologie doit en particulier garantir que la qualité d'une donnée dans son contexte d'application spécifique, ne soit représentée que par un indicateur unique de manière à permettre la lisibilité des résultats et la détermination claire de la valeur.

Le résultat de l'analyse de la qualité est une information binaire : une instance spécifique d'un objet de données ne peut être que correcte ou incorrecte. Il est important de préciser que la mesure de la qualité doit impérativement être déterministe (au niveau de l'instance de l'objet et du champ de données) et non pas statistique, ce qui impose le recours à de règles de gestion des données préétablies. C'est par cette détermination objective, crédible et précise de la qualité que le modèle de permettra d'identifier les opérations mises en danger par la non-qualité.

Les données étant toujours utilisées dans un contexte temporel déterminé, *la qualité des données n'est pertinente qu'au moment où les informations sont utilisées*. Il serait par exemple inutile et même contre-productif de maintenir des données obsolètes à n'importe quel standard de qualité. De même, une anticipation trop importante sur les besoins de données peut consommer des ressources plus utiles dans un contexte plus immédiat.

Ainsi, les règles de gestion utilisées pour la valorisation[15] déterminent la qualité des données en fonction de critères d'applicabilité rigoureux incluant la dimension de qualité, le domaine d'impact (la transaction) et la dimension d'impact (risque sur la valeur).

11.2.2.2. *Connaissance des processus de la chaîne logistique*

L'entreprise doit également disposer d'une bonne connaissance de ses processus, du moins par l'intermédiaire de ses flux transactionnels afin d'établir la cartographie

14. Le présent article peut être considéré comme une extension naturelle de la méthodologie intitulée *Global Data Excellence Framework* proposée par [ELA 09], la seule qui soit, à notre connaissance, à la fois rigoureuse et pratique.
15. Il n'est pas nécessaire de valoriser toutes les règles de qualité de données.

des flux de valeur. En effet, par convention, on considère qu'une commande client est ouverte tant qu'elle n'est pas honorée (livrée). Ainsi, la valorisation additive des commandes, des productions sur commande et des livraisons pourrait, par double comptage, retourner une valeur totale supérieure à la valeur réelle engagée dans les processus. De même, les approvisionnements conditionnés par les commandes s'inscrivent soit dans un contexte de valeur clairement dissocié du flux commercial, soit en tant que composant intégré à la commande.

Cette condition détermine la manière dont la valeur doit être traitée, en fonction des objectifs de gouvernance recherchés. Une société très orientée sur les ventes optera pour ramener la totalité de la valeur sur la commande du client, considérant que les transactions en aval (ordres de fabrications, livraisons, facturations, etc.) y sont totalement subordonnées. Les résultats de la détermination de la valeur de la qualité des données seront considérés sous l'angle de leur contribution à la réussite ou à l'entrave de l'exécution des commandes.

Une organisation centrée sur des objectifs de production utilisera les ordres de fabrication (valorisés au prix de revient ou au prix de vente au choix) comme base. Une société de distribution pourra recourir aux ordres de livraison, etc. Chaque société ou unité se référera à son contexte et ses objectifs particuliers pour déterminer la manière dont la valeur sera traitée.

L'enchaînement des transactions est important, car il détermine la manière dont la valeur se transmet dans les différentes étapes de la chaîne logistique. Ce point sera détaillé dans la section 11.3 traitant de l'implémentation de la méthode.

Une bonne connaissance des processus est incontournable pour bâtir un modèle de valorisation crédible, ne serait-ce que dans un contexte étroit. Cette condition est d'ailleurs nécessaire pour l'établissement de bonnes règles de gestion des données.

11.3. Implémentation de la valorisation de la qualité des données

11.3.1. *Aperçu général de la méthode*

La méthode d'implémentation requiert la mise en place de quatre constituants, essentiels au calcul de la valeur de la qualité et de la non-qualité.

La *chaîne de valorisation*, qui s'inscrit dans un contexte de gouvernance bien déterminé (opérationnel, prévisionnel, etc.) et qui représente un groupe logique de transactions uni par une même *valeur de référence*.

Les *transactions valorisées,* qui quantifient la valeur de chaque chaîne de valorisation et dont la structure conditionnera la structure de calcul de la valeur contributive des objets de données.

Les *objets de données*, dont la qualité déterminera si la valeur contributive est confirmée ou menacée. Ils conditionneront à leur tour la sélection des règles de gestion pertinentes dans le contexte de valorisation.

Les règles de gestion, qui établissent les critères formels de qualité des données et déterminent l'impact de la non-qualité.

Une fois les constituants en place, la méthode précise la manière dont le calcul de la valeur de la qualité ou de la non qualité des données est effectué.

11.3.2. *Détermination du contexte et des chaînes de valorisation*

La chaîne de valorisation détermine :

– un contexte de gouvernance particulier ;

– une nature de valeur, définie par une source déterminée ;

– un ensemble de transactions partageant la même nature de valeur, liée à une même source ;

– une séquence de transactions qui permet de déterminer la manière dont la valeur est transmise dans les processus de l'entreprise et autorisant une vision prédictive.

La figure 11.3, « Exemple de chaîne de valorisation de gestion des commandes », illustre un cas typique. Elle s'inscrit dans un contexte opérationnel, dont la source de valeur est constituée par la commande client et dont la nature est le prix de vente des produits finis. L'exécution de chaque commande nécessite un ordre de fabrication (on suppose une production sur ordre), une livraison et une facturation.

La valeur est transmise de transaction à transaction. Ainsi l'ensemble des ordres de fabrication liés à une commande spécifique contient la valeur de la commande associée.

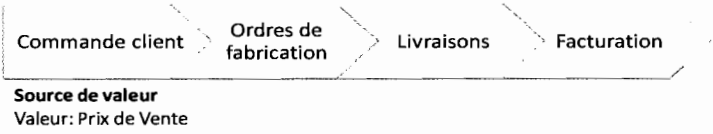

Figure 11.3. *Exemple de chaîne de valorisation de gestion des commandes*

La chaîne de valorisation autorise une vision prédictive. Sachant que toute commande doit être livrée, la mesure de la qualité des données au moment de la commande permet d'anticiper les problèmes que la non-qualité posera aux transactions de production, de livraison et de facturation, en quantifiant la valeur potentiellement impactée. Par exemple, l'absence de codification EAN[16] d'un produit empêchera sa livraison. La valeur menacée par ce manque correspond à la somme de toutes les lignes de commandes relatives à ce produit. Similairement, la facturation requiert qu'un code TVA valide soit associé au client. Au moment de la commande, il est possible de déterminer quel sera l'impact de la violation de cette règle, en déterminant la valeur associée de manière directe.

Un autre cas est illustré par la figure 11.4 « Exemple de chaîne de valorisation de production sur prévisions », qui représente le flux d'approvisionnement d'une entreprise produisant sur la base de prévisions de vente. La source de valeur est constituée par les prévisions de vente de chaque produit, fondées sur le prix de vente estimé. Chaque prévision nécessite un plan de production, lequel requiert une planification et une exécution des approvisionnements. Au moment de l'établissement des prévisions de vente, des défauts mesurés dans les données de base impacteront les transactions associées. Ainsi, un produit dont la composition ou la recette de fabrication est absente lors des prévisions de vente, ne pourra être produit et ses approvisionnements ne pourront être planifiés ou exécutés. De même, une matière première dont la source n'est pas déterminée ne pourra pas être commandée. En déterminant les produits finis dans lesquels cette matière première est nécessaire, on détermine la part des ventes prévisionnelles menacée par cette non-qualité.

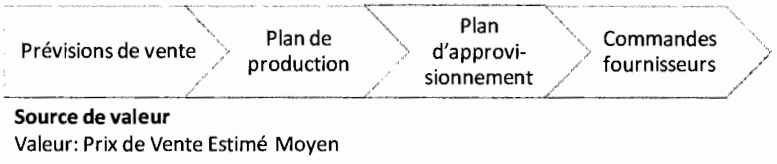

Figure 11.4. *Exemple de chaîne de valorisation de production sur prévisions*

Les chaînes de valorisation sont en principe indépendantes, en particulier lorsque les contextes représentés ne sont pas miscibles. Cependant, il est possible parfois de consolider plusieurs chaînes de valorisation lorsque leurs natures de valeur et leurs

16. Numéro d'article européen (*european article number)*, codification standardisée de produits utilisée dans les codes-barres.

contextes sont compatibles. Par exemple, les modes de production *make to stock*[17] (MTS) et *make to order*[18] (MTO) devraient en principe être considérés selon des chaînes de valorisation spécifiques. Cependant, le MTO est associé à un flux de vente sur stock, qui se réfère à une commande client, tout comme un MTO. Ceci ouvre la possibilité de constituer une vision consolidée afin de proposer une représentation plus unifiée de la valeur de la qualité des données issue des commandes clients. Cette consolidation n'est pas exempte de risques car elle peut potentiellement multiplier la valeur représentée par double comptage, lorsque la qualité d'une donnée spécifique impacte plus d'une chaîne de valorisation. Le gestionnaire de données devra prêter une attention particulière à ces impacts croisés et les traiter de manière appropriée.

Le choix de la source et de la nature des valeurs correspondent à la problématique de gouvernance représentée et à la culture de l'entité. Une entreprise focalisée sur les ventes optera pour une valorisation au prix de vente, tandis qu'une unité de production privilégiera par exemple le prix de revient.

La construction des chaînes de valorisation doit respecter certains principes :

– *une chaîne de valorisation se rapporte à une seule valeur de référence, déterminable dans toutes ses transactions*. Une violation de cette règle mettrait en danger les objectifs de gouvernance recherchés, en particulier une vision de valeur prédictive fiable ;

– *une chaîne de valorisation appartient à un seul contexte (prévisionnel, opérationnel...)*. Chaque contexte est considéré de manière indépendante afin de ne pas mêler des valeurs de natures différentes, même si elles apparaissent similaires ;

– *la valeur de référence est constante*. La réalité peut être différente : quantité délivrée légèrement différente de la quantité commandée, prix ou conditions déterminés ultérieurement, etc. Il convient de se rappeler que la méthode de valorisation poursuit un objectif de gouvernance, qui ne requiert pas la rigueur d'une science exacte ;

– *la séquence des étapes de la chaîne de valorisation est claire et stable*. Une chaîne de valorisation dans laquelle la séquence dépend d'informations contextuelles ne permettrait pas de gérer la valeur prédictive et systématique. Il vaut mieux éviter les « processus à option », quitte à faire l'impasse sur certaines étapes ou phases de la chaîne logistique, ou à créer des chaînes subordonnées traitées de manière indépendantes.

17. Production sur stock, usuellement basé sur des prévisions ou sur une gestion par niveaux d'inventaire minimum.
18. Production sur commande.

Les chaînes de valorisation représentent des points de vue différents sur la qualité des données et leur impact sur les processus de l'entreprise. Cette approche différenciée est justifiée par les différents besoins de gouvernance et par la perception relative de la valeur. De même que la valeur subjective d'un verre d'eau fraîche variera selon que l'on se trouve dans un désert brûlant ou dans un bar climatisé, la valeur de la qualité d'une donnée particulière dépend du contexte de son utilisation. Les chaînes de valorisation permettent de supporter ces besoins contextuels.

11.3.3. *Détermination des transactions valorisées*

Les transactions valorisées sont les éléments constituants de la chaîne de valorisation. A partir des attributs relatifs à la chaîne de valorisation (contexte et définition de la valeur de référence), c'est par la structure des transactions en lien avec les objets de données que vont pouvoir être définies la valeur de transaction ainsi que la part de la valeur contributive.

Les transactions valorisées quantifient la valeur au sein d'une chaîne de valorisation. Afin d'éviter des confusions, on désignera comme « instance de transaction » : l'ensemble des occurrences particulières et uniques d'une transaction, cette dernière étant considérée comme leur catégorie générique. Une transaction (par exemple une commande client) requiert toujours les mêmes objets de données (un client et des produits). Une instance de transaction (la commande *x*) utilisera des instances d'objets de données particulières (le client « Marianne » et les produits « blanc », « bleu » et « rouge »). Les relations entre transactions et objets de données sont régies par la manière dont les premières sont structurées.

Les transactions reposent sur des objets de données (par exemple, le client, le produit), eux-mêmes composés d'éléments de données (l'adresse du client, le code EAN du produit, etc.). Les éléments de données sont généralement associés à un contexte fonctionnel, comme les ventes, les achats, la planification, etc., groupés dans des segments et des vues dans les systèmes d'information. Les règles de gestion des données sont toujours applicables au niveau de l'élément de données. Cependant, la valeur de la qualité ou de la non-qualité n'a guère de sens à ce niveau de granularité et sera donc consolidée au niveau de l'objet de données. On considère qu'une instance d'un objet de données (le produit « bleu ») est de qualité lorsque la totalité de ses constituants satisfait aux règles de gestion associées.

La structure des transactions définit quelles sont les données pertinentes pour son exécution. Ainsi, une commande client nécessitera des informations relatives au client (identification, adresse, typologie, limite de crédit, etc.) et aux produits (identification, spécifications, prix, etc.).

Une transaction peut recourir à de nombreux objets de données. Il convient cependant de se rappeler que la détermination de la valeur de la qualité des données suit un objectif de gouvernance et de pilotage. Il serait contre-productif de construire un modèle complet mais difficile à mettre en œuvre car incluant une pléthore d'objets et d'éléments de données. En effet, une bonne gouvernance requiert davantage une information lisible, pertinente et raisonnablement complète qu'une précision absolue mais inutilisable du fait de sa complexité.

Le gestionnaire de données se focalisera donc sur la capture des données essentielles, nécessaires à la réalisation de la valeur d'une transaction, qui seront représentatives de la situation qualitative des données, les détails ou les règles non essentielles étant traitées à un niveau purement opérationnel. De manière générale, on considèrera les données de base suivantes :

– *informations métier critiques,* sans lesquelles une transaction ne peut fonctionner correctement. Ces informations sont issues de la connaissance métier et ne peuvent être issues de règles de calcul ou de détermination. Les exemples typiques sont la désignation des produits, les identifiants des clients (raison sociale, adresse), la typologie technique des matières, etc. ;

– *données déterminantes de processus.* Ces informations conditionnent les variantes de processus ou d'exécution transactionnelle. On peut citer par exemple la typologie des matières (produit fini, matière première), le mode d'approvisionnement (sur niveau de stock, sur prévision ou sur commande), etc. ;

– *données ayant un impact direct ou significatif sur la valeur.* L'absence ou la non-qualité de ces informations affecte la valeur, de manière directe ou indirecte. Cette notion d'impact sera décrite ultérieurement.

Certains éléments de données peuvent affecter plusieurs transactions. Le gestionnaire peut choisir une approche simplifiée en associant chaque élément de données pertinent à une seule transaction par contexte. Il peut également choisir de montrer une vision plus transverse en mettant en évidence la multiplicité des impacts, selon ses objectifs.

11.3.4. *Détermination de la valeur contributive des données*

Cette étape détermine comment les données contribuent à la valeur des transactions, par l'association de la valeur des transactions aux objets de données. La valeur des transactions est conditionnée par la structure de ces dernières qui peuvent être *atomiques* (relatives à un seul objet de données) ou *segmentées* (contenant plus d'un objet de données). La segmentation requiert la détermination du mode de distribution de la valeur de la transaction sur les données. La figure 11.5

présente une commande qui se compose d'un en-tête contenant l'identification du client et de plusieurs lignes de commande qui contiennent les éléments produit, prix et quantité. Les lignes de commande sont les constituants de la valeur.

La logique de répartition de la valeur s'effectue selon une approche simple. On part du principe que la valeur d'une transaction est confirmée lorsque la totalité des données de base qui la constitue est de qualité (satisfait aux règles de gestion). Chaque violation de qualité supprime de la valeur. Ainsi, l'analyse de l'impact de la non-qualité détermine la clé de répartition de la valeur. Dans une commande, la non-qualité des données d'un client (par exemple l'absence d'adresse) affecte la totalité de la valeur de la transaction. *A contrario*, la non-qualité des données d'un produit fini n'affectera que la valeur de la ligne de commande qui le contient.

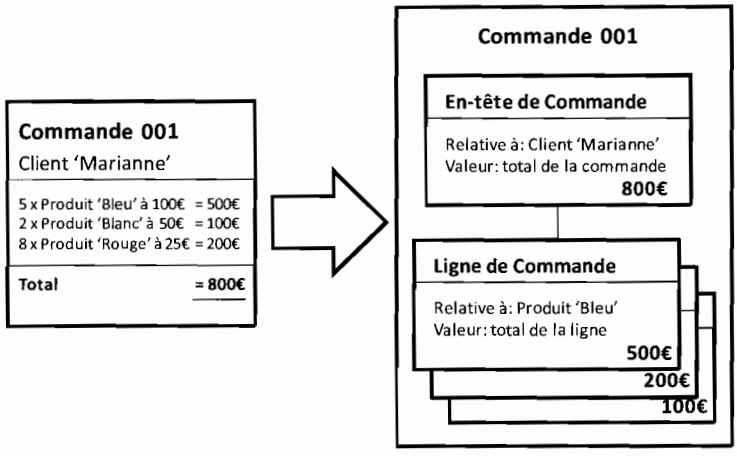

Figure 11.5. *Exemple de segmentation de la transaction « commande »*

Cette analyse d'impact nécessite donc de déterminer la structure des transactions afin d'y associer les données de manière pertinente. On distinguera les structures transactionnelles suivantes :

– *tomique* : la transaction repose sur un seul objet de données. L'objet de données reçoit l'intégralité de la valeur de la transaction ;

– *hiérarchique* : la transaction contient plusieurs niveaux (par exemple une commande, un bulletin de livraison ou une facture). La non-qualité d'une donnée affecte le niveau qui le concerne et la totalité des niveaux inférieurs. La figure 11.5 illustre la manière dont une commande est segmentée entre son en-tête, généralement relatif aux informations du client, et les lignes de commandes dont les données principales sont représentées par le produit, le prix et la quantité ;

– *concurrent* : la transaction contient plusieurs objets d'égale importance (par exemple, un ordre de fabrication requérant un produit, une formule et une recette). La non-qualité d'un objet de données affecte l'ensemble de la transaction, donc chaque objet est valorisé à hauteur de la totalité de la valeur de la transaction.

Des modèles de segmentation hybrides sont possibles : par exemple un niveau de segmentation hiérarchique peut contenir plusieurs objets de données concurrents. La seule contrainte réside dans la capacité de construire la méthode de détermination de la valeur d'un objet de données.

Cependant il convient une fois de plus de se rappeler que l'objectif suivi est la gouvernance dont les bonnes pratiques privilégient la flexibilité et la rapidité sur la complexité. Le gestionnaire devra trouver un juste équilibre en fonction de ses besoins de gouvernance. Il s'assurera de l'utilité de chaque élément introduit dans le modèle afin de ne pas diluer les informations réellement pertinentes. En règle générale, il est préférable de débuter avec une base simple quitte à en augmenter la sophistication ultérieurement, par exemple en se limitant à deux niveaux hiérarchiques par transaction. Cette approche permet d'attaquer les problèmes de qualité de données de manière incrémentale et de guider progressivement l'entreprise vers une culture de qualité des données.

11.3.5. *Intégrer la dimension qualitative*

Les règles de gestion permettent de déterminer si des éléments de données sont de qualité ou non et de déterminer l'impact de la non-qualité sur la réalisation de la valeur des transactions. On appellera *règle valorisée* une règle de gestion des données utilisées dans un contexte de calcul de valeur afin de la distinguer des règles « ordinaires » utilisées dans un contexte de gestion opérationnel de la qualité des données.

La relation entre les règles valorisées et les données est directe :

– chaque règle de gestion supporte un seul élément de données[19] ;

– chaque élément de données dont on veut valoriser la qualité est supporté par au moins une règle. La figure 11.6 illustre par un exemple simplifié, la manière dont les règles de gestion sont liées à la valeur des transactions.

19. Un élément de données est généralement équivalent à *champ de données*. Parfois l'élément est une petite collection homogène de champs (par exemple une adresse qui nécessite plusieurs lignes, chacune étant contenue dans un champ, mais qui constitue une seule information du point de vue du métier).

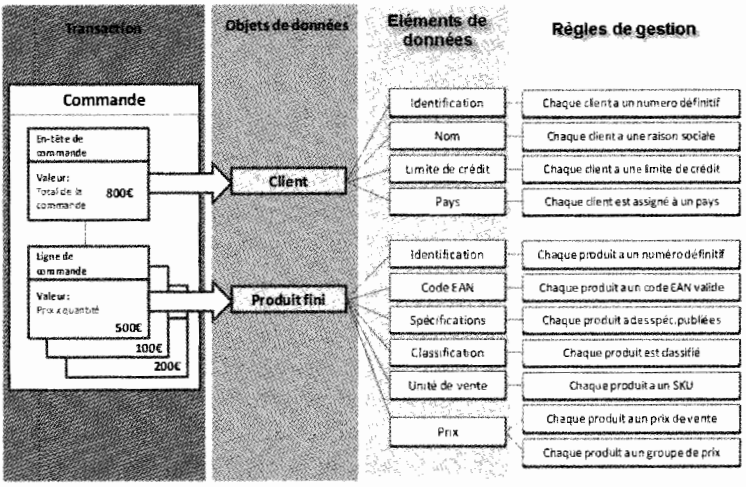

Figure 11.6. *Association des règles de gestion aux données*

Les règles valorisées contiennent des attributs spécifiques par rapport aux règles ordinaires :

– le périmètre d'application, déterminant le contexte et la ou les transactions affectées ;

– la nature de l'impact du non-respect de la règle sur la valeur.

Il n'est pas nécessaire de transformer toutes les règles de gestion des données en règles valorisées, tout comme il n'est pas nécessaire de valoriser toutes les données. En effet, rien n'impose d'associer une valeur à chaque règle, mais il est important de sélectionner celles qui sont utiles à la gouvernance.

L'impact de la non-qualité sur la valeur est un attribut impératif pour la valorisation de la qualité des données. En effet, on ne peut mêler règles critiques (dans lesquelles une erreur de données provoque une interruption irrémédiable du processus), règles avec impact sur les délais d'exécution et règles non-bloquantes mais provoquant un impact différé (par exemple, insatisfaction de la clientèle). Chaque nature d'impact représente donc une dimension distincte dans les rapports de valeur de qualité des données.

L'impact est très important dans l'établissement de la crédibilité du modèle. Par exemple, l'absence de source de matière première empêche tout approvisionnement, car la commande au fournisseur devient impossible. L'inexistence d'un code produit empêche sa vente. Ces cas mettent en danger la capacité de l'entreprise à exécuter ses

processus. Une sous-estimation du délai d'approvisionnement d'une matière première retardera la production, les livraisons, la facturation et donc le recouvrement. Une surestimation de la même information n'impactera pas la production, mais augmentera les inventaires. De même, une erreur d'épellation du nom d'un client aura un impact probable sur la satisfaction du client mais sans impact économique.

Il serait peu crédible de mêler des données critiques dont la non-qualité détruit la capacité à réaliser la valeur, avec les données dont la non-qualité génère des effets indirects ou intangibles. Il est donc impératif de savoir lier directement la notion de qualité des données objective (niveau de respect des règles) avec l'impact de la non-qualité sur la marche des affaires.

Le gestionnaire de données établira une classification d'impacts tout en en maîtrisant la prolifération qui s'avérerait contre-productive en complexifiant les rapports. En principe, une chaîne d'approvisionnement devrait se satisfaire de la classification générique suivante :

- données critiques pour l'exécution des processus, dont l'absence ou la non-qualité empêcherait l'exécution des processus et la réalisation de la valeur ;
- données engendrant des dérangements (délais, attentes, déconnections) ;
- données engendrant des risques (risques légaux, financiers, industriels) ;
- données impactant la réputation ou le service au client (plaintes, rejets).

Une règle ne peut contenir qu'une seule nature d'impact, sous peine de rendre le modèle de valorisation inutilement complexe. Il vaut mieux, le cas échéant, dupliquer une règle pour représenter des impacts multiples, tout en s'interrogeant sur la pertinence d'une telle mesure dans un contexte de gouvernance.

Les règles à impact critique méritent un traitement particulier. En effet, le non-respect de règles techniques critiques empêche l'exécution des transactions, ce qui par corollaire prévient toute valorisation. *A contrario*, si la transaction est enregistrée, cela signifie automatiquement que les éventuels problèmes critiques ont été déjà réglés : la règle retournera donc systématiquement une qualité parfaite. Dans un contexte de gouvernance, cette information a autant d'utilité que d'ouvrir son parapluie après l'orage. *Les règles techniques critiques n'ont de sens que dans un objectif de prévention.* On peut déterminer au moment de la commande que le traitement des expéditions ne pourra être effectué, faute d'adresse de livraison. Il convient donc de clairement établir à quel moment du processus la détection de ces règles est la plus pertinente.

Il existe aussi des règles métier critiques, qui autorisent l'exécution d'une partie ou de la totalité de la transaction mais mettraient en danger la réalisation finale de la

valeur. C'est le cas lorsque les spécifications de production ne correspondent pas aux demandes clients. La commande, la production et la livraison sont possibles, mais la valeur ne sera jamais réalisée car le client en refusera la réception.

Dans tous les cas de règles critiques, il importe de déceler les violations de la manière la plus anticipée possible. Ce principe impose *d'assigner les règles de gestion aux transactions en fonction des objectifs de gouvernance recherchés* et non pas en fonction de critères purement techniques.

11.3.6. *Calcul de la valeur de la qualité et de la non-qualité des données*

Une fois les composants de la méthode déterminés, le calcul de la valeur de la qualité ou de la non-qualité des données est effectué en quatre étapes :

– *étape 1* : détermination de la valeur contributive des données dans les transactions ;

– *étape 2* : détermination de la valeur confirmée ou menacée au niveau des éléments de données ;

– *étape 3* : consolidation de la valeur confirmée ou menacée au niveau de la transaction ;

– *étape 4* : consolidation finale et présentation des résultats.

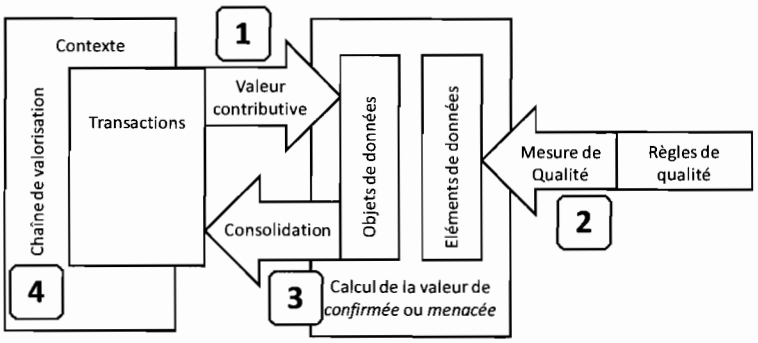

Figure 11.7. *Calcul de la valeur*

La figure 11.7 illustre ces étapes de manière schématique. Le but consiste à déterminer la proportion de la valeur contributive respectivement *confirmée* par la qualité des données ou *menacée* par la non-qualité. Comme la valeur est détruite par la non-qualité, on adopte une approche consistant à déduire la valeur menacée de la valeur contributive. Le solde restant constitue la valeur confirmée.

11.3.6.1. *Etape 1 : détermination de la valeur contributive*

La valeur contributive représente la contribution de chaque objet de données dans une transaction valorisée. Elle est déterminée par les sources et natures de valeur et par la segmentation des transactions. On considérera l'exemple de la chaîne de valorisation représentée par la figure 11.3, dont la source de valeur est déterminée par les commandes clients. La valeur contributive de chaque client est la somme de la valeur totale de chacune de ses commandes. La valeur contributive de chaque matériel est la somme des montants de toutes les lignes de commande qui les contiennent.

Le tableau 11.1 illustre l'exemple de trois commandes concernant deux clients, à des états d'avancement divers. Les chaînes de valorisation permettent de donner une visibilité sur les transactions suivantes. Ainsi, la commande 001 en est à la livraison, tandis que la commande 003 vient d'être produite. Les transactions complétées ne sont pas considérées dans le processus de valorisation.

Client	**Produit**	**Source de valeur**			**Statut transactions**		
		Identifiant commande	***En-tête (client)***	***Ligne (produit)***	***Production***	***Livraison***	***Facturation***
Marianne	Bleu	001	200	200	Complet	En cours	Prévu
	Blanc	002	500	500	Prévu	Prévu	Prévu
Paul	Rouge	003	750	150	Complet	Prévu	Prévu
	Bleu			600	Complet	Prévu	Prévu

Tableau 11.1. *Exemple de calcul de valeur contributive*

La valeur contributive est déterminée selon les principes décrits dans la section 11.3.4 : la valeur de chaque transaction est assignée aux clients (dont la valeur contributive est représentée par la totalité de la commande) et aux produits (dont la valeur contributive est déterminée par la valeur des lignes de commande). On constate à ce stade que la valeur contributive totale (clients et produits) n'est plus représentative de l'activité réelle de l'entreprise.

A l'issue de cette étape, on dispose de la valeur contributive de chaque objet ainsi que du périmètre utile, à savoir la liste des instances de transactions et d'objets de données pertinentes au moment de la valorisation.

11.3.6.2. *Etape 2 : détermination de la valeur confirmée ou menacée*

La valeur est dite *confirmée* si la qualité de l'objet de données ne présente aucun défaut. Elle est au contraire *menacée* dès qu'un seul élément de données n'est pas conforme à au moins une des règles de gestion qui y est associée. De manière générale, la valeur contributive est égale à la somme des valeurs confirmées et menacées, pour une dimension de risque et un contexte spécifique.

La détermination des valeurs confirmées ou menacées s'effectue par la mesure de la conformité par rapport aux règles valorisées. Elle s'effectue de manière indépendante pour chaque chaîne de valorisation et nature d'impact.

Le tableau 11.2 montre le résultat de l'exécution de trois règles relatives aux produits et deux règles associées aux clients, selon deux dimensions d'impact : critique (rendant l'exécution de la transaction impossible) ou délai (retardant l'exécution de l'opération) pour une transaction spécifique. Ainsi, la non-conformité de la règle P001 empêche l'exécution de toute production, tandis que le non-respect de la règle P002 ne causera que des délais dans l'exécution des livraisons (ce qui n'est nécessairement moins grave).

Règle	**Objet**	**Identif.**	**Impact**		
			Transaction	***Nature***	***Résultat***
P001	Produit	Bleu	Production	Critique	Passe
P002	Produit	Bleu	Livraison	Délai	Passe
P003	Produit	Bleu	Facturation	Critique	Echoue
P001	Produit	Blanc	Production	Critique	Passe
P002	Produit	Blanc	Livraison	Délai	Passe
P003	Produit	Blanc	Facturation	Critique	Passe
P001	Produit	Rouge	Production	Critique	Passe
P002	Produit	Rouge	Livraison	Délai	Echoue
P003	Produit	Rouge	Facturation	Critique	Passe
C001	Client	Marianne	Facturation	Délai	Passe
C002	Client	Marianne	Livraison	Critique	Passe
C001	Client	Paul	Facturation	Délai	Echoue
C002	Client	Paul	Livraison	Critique	Passe

Tableau 11.2. *Exemple de résultat de règles de gestion*

La valeur des instances d'objet qui ne sont pas conformes aux règles est menacée. Le tableau 11.3 illustre la manière dont la valeur confirmée est détruite en fonction du résultat des règles de gestion. On remarquera qu'il n'existe dans cet exemple aucune règle pour l'objet « client » relative à la production.

Impact	Client	Produit	Source de valeur			Valeur confirmée (en-tête)		Valeur confirmée (ligne)		
			Identifiant commande	*En-tête (client)*	*Ligne (produit)*	*Livraison*	*Facturation*	*Production*	*Livraison*	*Facturation*
Critique	Marianne	Bleu	001	200	200	200	200	-	200	~~200~~
		Blanc	002	500	500	500	500	500	500	500
	Paul	Rouge	003	750	150	-	-	-	150	150
		Bleu			600			-	600	~~600~~
Délai	Marianne	Bleu	001	200	200	200	200	-	200	200
		Blanc	002	500	500	500	500	500	500	500
	Paul	Rouge	003	750	150	750	~~750~~	150	~~150~~	150
		Bleu			600			600	600	600

Tableau 11.3. *Exemple de détermination de la valeur confirmée*

Cette étape permet déjà de fournir une information de valorisation pertinente au niveau de l'objet de données. Ainsi, le défaut critique du produit « Bleu » sur la règle P003, applicable à la transaction de facturation (par exemple, l'absence de code douanier) impactera 55 % du chiffre d'affaires en cours de l'entreprise. Il convient cependant de noter que les résultats de valeur confirmée ou menacée ne sont pas immédiatement consolidables. L'exemple de la commande 003 est à ce titre représentatif : la somme de la valeur menacée par les défauts du client « Paul » et du produit « Blanc » excède le montant de la commande. La consolidation s'effectuera dans l'étape suivante.

Il est utile de noter que les transactions passées ne sont pas considérées : la qualité des données étant mesurée à un moment t, on ne peut tirer de conclusion quant à leur impact sur un événement antérieur. Il ne s'agit pas d'une règle absolue cependant. Un risque de qualité impactant un aspect légal ou d'image de l'entreprise peut éclater longtemps après les faits.

Dans tous les cas, un risque critique est par définition écarté de toute transaction passée, sans quoi celle-ci n'aurait pas pu être complétée en premier lieu.

L'application rétroactive de ce type de règle peut être cependant utile afin de déterminer le moment de leur violation.

11.3.6.3. *Etape 3 : consolidation de la valeur confirmée ou menacée sur les transactions*

Dans cette étape, on consolide la valeur confirmée ou menacée au niveau de la transaction pour permettre la cartographie des risques de qualité de données et de leur effet sur la valeur au sein d'une chaine de valorisation.

Cette détermination s'effectue en utilisant la logique inverse de la détermination de la valeur contributive (étape 1), mais en traitant séparément valeur confirmée et menacée.

On tient compte de la segmentation de chaque transaction tout en respectant le principe de qualité totale : la non-qualité est toujours prioritaire sur la qualité. Ainsi, la valeur menacée par la non-conformité des données du client « Paul » pour la facturation détruit l'intégralité de la valeur de la commande 003 dans la dimension d'impact « délai », quel que soit la qualité des produits qui y sont contenus.

Le résultat est présentable par exemple sous une forme similaire à celle du tableau 11.4. A l'issue de cette étape, la somme de la valeur confirmée et menacée est équivalente à la valeur de référence (ou de base).

Cette information présente clairement l'exposition de l'entreprise à la non-qualité, en termes purement valorisés, par catégorie de risques, en indiquant sans ambigüité les processus les plus exposés (en l'occurrence, la facturation). Le gestionnaire de données n'aura aucune difficulté à faire renforcer l'application des règles C001 et P003 qui concentrent l'essentiel des risques.

Il convient de noter la puissance de la valorisation par rapport à l'approche purement qualitative : trois exceptions ont été identifiées sur les treize mesures de qualité effectuées à l'étape deux, représentant un niveau de qualité de 77 %. En lecture plus fine, les règles critiques retournent une qualité de 87 % (une exception sur huit mesures). En se faisant l'avocat du diable, on peut également prétendre qu'il n'existe qu'un seul problème de données critiques. Est-il donc nécessaire de payer un gestionnaire pour ne traiter qu'un seul cas ?

Cependant, l'analyse de valeur indique que 55 % de la facturation est menacée par des risques critiques, indiquant une concentration des problèmes de qualité sur les données de base les plus importantes et sensibles, justifiant la mise en place de mesures de correction. Ce genre d'argumentation devient difficilement contestable.

Impact	Com-mande	Valeur de base	Production			Livraison			Facturation		
			Q	*NQ*	*%NQ*	*Q*	*NQ*	*%NQ*	*Q*	*NQ*	*%NQ*
Critique	001	200	-	-	0 %	200	0	0 %	0	200	100 %
	002	500	500	0	0 %	500	0	0 %	500	0	0 %
	003	750	-	-	0 %	750	0	0 %	150	600	80 %
	Total	1450	500	0	0 %	1450	0	0 %	650	800	55 %
Délai	001	200	-	-	0 %	200	0	0	200	0	0 %
	002	500	500	0	0 %	0	500	100 %	500	0	0 %
	003	750	-	-	0 %	750	0	0 %	0	750	100 %
	Total	1450	500	0	0 %	950	500	0 %	700	750	52 %

Q : valeur de la qualité, NQ : valeur de la non-qualité, %NQ : proportion menacée

Tableau 11.4. *Exemple de consolidation de valeur confirmée et menacée par transaction*

11.3.6.4. *Etape 4 : consolidation finale et présentation*

Les données produites dans les étapes deux et trois peuvent être exploitées de multiples manières. Quel que soit le but de gouvernance recherché, il convient de s'assurer que la consolidation s'effectue en respectant les principes mentionnés ci-dessus. La meilleure méthode consiste à vérifier que la somme des valeurs confirmées et menacées est toujours consistante avec la source de valeur.

La consolidation doit également respecter les règles de contexte. Lorsque des rapports consolident des valeurs issues de chaînes de valorisation différentes, par exemple entre vente sur stock et production sur commande, il convient de déterminer un dénominateur commun de valeur sur laquelle l'ensemble des résultats seront comparés. Il vaut mieux présenter deux rapports distincts et corrects que de tenter l'impossible en consolidant des résultats qui, une fois additionnés, ne représentent plus la réalité objective de l'entreprise, la crédibilité du modèle et du gestionnaire de données en dépend.

Certains types de consolidation sont aisés à produire, à comprendre et à utiliser dans un objectif de gouvernance. Le tableau 11.5 présente quelques exemples et leur utilité pour la gouvernance.

Type de consolidation	Description	Utilité pour la gouvernance
Par objet de données	Quantification comparative de l'impact de la non-qualité par objet de données	Définir les priorités par objet de données
Par transaction active au sein d'une chaîne de valorisation	Localiser et quantifier l'impact instantané de la non-qualité des données sur les processus	Identifier les vulnérabilités des processus face à la qualité des données, immédiatement et à court terme
Par règle de gestion	Quantification des risques par règle de gestion et mesure de l'impact sur l'entreprise	Définir les priorités par objet de données
Par chaîne de valorisation	Rapport de synthèse sur la qualité globale des données	Base pour la détermination des indicateurs de performance opérationnels

Tableau 11.5. *Quelques types de consolidation*

La consolidation par objet de données permet de déterminer la manière dont chaque objet de données contribue et menace la valeur dans les transactions actives de l'entreprise. Cette consolidation s'effectue par chaîne de valorisation pour éviter d'avoir à traiter les recoupements de valeurs. Le rapport indique quelles sont les données de l'entreprise les plus impactées par la non-qualité en termes de valeur contributive. Il est utilisé pour déterminer les priorités en matière d'amélioration des processus de gestion et de nettoyage des données par objet. Il est impératif de ne pas totaliser les colonnes de valeur car cette information n'a aucune signification réelle.

Gestion des commandes						
Objet de données	*Valeur contributive*	*Valeur de qualité*	*Valeur de non qualité*			
			Totale	*Critique*	*Délais*	*Finance*
Produit fini	2 140	1 240 (58 %)	900 (42 %)	700 (33 %)	200 (9 %)	0 (0 %)
Client	1 790	1 450 (81 %)	340 (19 %)	0 (0 %)	340 (19 %)	340 (19 %)
Recette	200	200 (100 %)	0 (0 %)	0 (0 %)	0 (0 %)	0 (0 %)

Tableau 11.6. *Exemple de consolidation par objet de données*

La consolidation par transaction active (déjà évoqué ci-dessus avec l'exemple du tableau 11.4) permet de déterminer la valeur instantanée de la qualité ou de la non-qualité sur les transactions. La consolidation s'effectue en sommant la valeur réelle et menacée par catégorie de risque, par type de transaction au sein d'une

chaîne de valorisation. Cette consolidation permet de construire des rapports illustrant les catégories d'impact et leur localisation dans les processus de la chaîne logistique. Elle permet de déterminer la vulnérabilité des transactions face à la qualité des données et d'identifier les priorités adéquates de résolution au niveau des règles de gestion.

La consolidation par règle de qualité représente la valeur détruite par la non-qualité, rapportée aux règles de gestion des données. Elle met en relation directe le niveau de qualité issue de la mesure de la règle avec la valeur mise en danger, permettant de déterminer si la gestion des données se focalise sur les bonnes priorités. Le tableau 11.7 illustre la manière dont cette information peut être exploitée. Si la structure de responsabilité associée aux règles de gestion est suffisamment claire, ce genre de rapport devient très mobilisateur, tant au niveau de la gouvernance des règles qu'à celui des équipes chargées de leur application.

Règle de gestion		**Trans. impactées**		**Non-qual. (%)**	**Valeur par type d'impact % valeur contributive totale**		
ID	***Description***	***Transaction***	***CdV***		***Critique***	***Délai***	***Fin.***
M001	Tout produit fini a un code EAN	Commande	GdC	5 %	700 (33 %)		
M012	Tout produit fini a une durée de production	Production	GdC	3 %		200 (9 %)	
C011	Tout client a des termes de paiement	Facture Client	GdC	7 %			340 (19 %)
C012	Tout client a une adresse de facturation vérifiée	Facture Client	GdC	7 %		340 (19 %)	

CdV : Chaîne de valorisation

GdC : Gestion des commandes

Tableau 11.7. *Exemple de consolidation par règle de qualité*

La consolidation par chaîne de valorisation, telle qu'illustrée dans le tableau 11.8, donne une information rapide sur la qualité des données et la valeur supportée ou menacée avec une vision processus. Son objectif est de fournir un indicateur de haut niveau qui intéressera plus particulièrement une direction métier peu soucieuse de détails, en complément des outils de suivi des processus. Il est également possible de décliner la non-qualité par dimension d'impact pour autant que l'audience soit

sensible à ces distinctions. Ce rapport constitue souvent un bon point d'entrée pour des discussions plus détaillées sur la qualité des données.

Domaine	**Valeur de référence**	**Valeur générée par la qualité**		**Valeur menacée par la non-qualité**	
		Valeur	***%***	***Valeur***	***%***
Gestion des commandes	2 140 €	900 €	42 %	1 240 €	58 %
Gestion des approvisionnements	4 850 €	3 900 €	80 %	950 €	20 %
Gestion des offres	1 200 €	1 000 €	83 %	200 €	17 %
Total		71 %		29 %	

Tableau 11.8. *Consolidation par chaîne de valorisation*

Ces rapports de consolidations devraient être assortis d'éléments quantitatifs déterminant le nombre d'instances de transactions et d'objets de données afin de donner une indication du volume sous gestion. Finalement, l'historique ce ces rapports devrait être conservé afin d'inscrire la valorisation des données dans un contexte temporel, en s'assurant de garder la trace de l'évolution du périmètre : l'ajout de nouvelles règles de gestion cause généralement une chute des résultats et peut engendrer un effet contre-productif. Il s'agit d'ailleurs d'un des rares risques managériaux de la méthode.

11.4. Conclusion

La méthode de valorisation de la qualité des données permet de remplir les objectifs visés par la gouvernance, à savoir produire une information fiable et crédible sur la valeur de la qualité des données et l'impact de la non-qualité. Elle est simple à expliquer et produit une information aisément compréhensible.

En matière d'implémentation, cette méthode requiert de disposer d'un cadre de gestion de qualité de données utilisant des règles de gestion formalisées. Les auteurs se sont reposés sur la méthodologie *Global Data Excellence framework* proposée par [ELA 09], dont l'extension à la valorisation ne requiert que peu d'efforts et d'investissements. La seule difficulté technique à résoudre consiste à déterminer la logique de distribution de la valeur contributive entre transactions et objets de données.

Le risque principal réside, comme dans tout outil de gouvernance, dans les attentes qu'un tel outil peut générer. Des rapports simples peuvent susciter une

demande pour davantage de périmètre, de détails ou de précision. *Trop d'information tue l'information*[20] : les bons outils de gouvernance doivent rester aisés à comprendre et à gérer. Il est possible d'augmenter le nombre de règles de gestion intégrés à l'outil sans nécessairement ajouter de dimensions supplémentaires aux rapports produits, qui en réduirait la lisibilité et donc l'utilité.

Le message de la gouvernance des données est limpide : la non-qualité des données étrangle les entreprises de manière insidieuse mais néanmoins réelle. Valoriser la qualité et surtout la non-qualité rend visible les risques encourus et permet d'y remédier en se focalisant sur les points essentiels. C'est le but de tout gestionnaire de données. Cette méthode n'a d'autre objectif que de l'y aider.

11.5. Bibliographie

[ADE 05] ADELMAN S., MOSS L., ABAI M., *Data Strategy*, Addison-Wesley Professional, Upper Saddle River, NJ, 2005.

[BAS 10] BASEL COMMITTEE ON BANKING SUPERVISION, *Basel III : A global regulatory framework for more resilient banks and banking systems*, Bank for International Settlements, Basel, 2010.

[CHU 04] CHU M., *Blissful Data : Wisdom and Strategies for Providing Data That's Meaningful, Useful, and Accessible for All Employees*, AMACOM, New York, 2004.

[CNI 05] Décret n° 2005-1309 du 20 octobre 2005 (France).

[ECK 02] ECKERSON W., *Data Quality and the Bottom Line : Achieving Business Success through a Commitment to High Quality Data*, The Data Warehouse Institute (TDWI), Chatsworth, 2002.

[ELA 09] EL-ABED W., « Data Governance : A Business Value-Driven Approach », *White Paper,* USA, novembre 2009.

[ELA 09] EL-ABED W., « Data Governance : La Gouvernance des Données : Une Approche de Valeur Conduite par les Métiers », *Papier Blanc,* France, novembre 2009.

[ELA 09] EL-ABED W., « The Data Excellence Framework to Improve Global Safety and Security », *ISMTCL Proceedings, International Review Bulag (2009) 277,* p. 94-99, septembre 2009.

[ELA 11] EL-ABED W., « Linking Data Quality Metrics to Business Value and Risk, Tutorial», *Data Governance & Information Quality Conference (DGIQ),* San Diego, Californie, USA, 27 juin 2011.

[ENG 02] ENGLISH L., « Mistakes to Avoid for DW Data Quality », http://www.information-management.com, 1er juin 2002.

20. Citation attribuée à Noël Mamère.

[ENG 99] ENGLISH L., *Improving Data Warehouse and Business Information Quality*, Wiley, New York, 1999.

[FER 08] FERNANDEZ A., *Les nouveaux tableaux de bord des managers*, Editions d'Organisation, Paris, 2008.

[KAR 07] KAREL R., KIRBY J., EVELSON B., MOORE C., BARNETT J., *Data Governance, What Works and What Doesn't*, Forrester Research, Cambridge, USA, 2007.

[REG 08] RÉGNIER-PÉCASTAING F., GABASSI M., FINET J., *MDM, Enjeux et méthodes de la gestion des données*, Dunod, Paris, 2008.

[SOX 02] Public Law 107 – 204 – Sarbanes-Oxley Act of 2002, 107th Congress (USA), 2002.

Chapitre 12

La gouvernance des données : apports de l'ingénierie des données dirigée par les modèles

12.1. Préambule

La modélisation est au développement informatique ce qu'est le solfège à la musique :

– pour le compositeur un moyen d'exprimer, dans ses créations, toutes les nuances de ses sentiments ;

– pour le chef d'orchestre la définition d'une œuvre à diriger subtilement pour en dégager l'harmonie ;

– pour l'instrumentiste ou soliste la description de ce qu'il doit interpréter avec toute sa sensibilité ;

– pour les initiés un moyen de communication pour se comprendre dans le temps et l'espace ;

– pour les musiciens amateurs un formalisme dont ils pensent pouvoir se passer aisément ;

– pour les profanes une *terra incognita* à explorer pour ses richesses insoupçonnées.

Chapitre rédigé par Vincent CISELET, Jean HENRARD, Jean-Marc HICK, Frumence MAYALA, Dominique ORBAN et Didier ROLAND.

12.2. Introduction

Enseignée dans toutes les écoles et les instituts de formation, recommandée en tant que bonne pratique par de très nombreux experts, la modélisation informatique est peu utilisée dans les organisations. Il existe, sans doute, de multiples raisons à ce paradoxe. Deux raisons sont évidentes : d'une part, l'aspect hermétique du langage qui semble vouloir réserver la modélisation aux seuls spécialistes et d'autre part, le peu de résultat concret produit au regard de l'investissement que la modélisation nécessite.

Cependant, depuis quelques années cette situation évolue sous l'effet de deux courants convergents :

– la nécessité pour les organisations de se doter de méthodes et d'outils face à leur difficulté de garder la maîtrise de leurs applications dont la complexité et le nombre ne font qu'augmenter ;

– la présence grandissante de solutions performantes, résultats de projets dans lesquels le code source n'est plus considéré comme l'élément central d'un logiciel, mais comme un élément dérivé de la modélisation.

La modélisation des données s'inscrit dans cette tendance. C'est ainsi que depuis plus de sept ans la société REVER industrialise et commercialise les résultats des recherches et développements menés depuis vingt cinq ans au sein du laboratoire d'ingénierie des bases de données (LIBD) de l'université de Namur (Belgique)[1].

Si la modélisation des données peut sembler pour certains n'être qu'anecdotique par rapport à l'enjeu global des entreprises, c'est à la fois perdre de vue que les données sont au cœur des applications informatiques et sous-estimer le rôle vital des données dans le fonctionnement des organisations.

Par ailleurs, les coûts très importants que les données engendrent, les valeurs patrimoniales et financières qu'elles représentent décuplent l'obligation pour les organisations de dépasser la simple gestion des données pour entrer dans la gouvernance des données.

A l'image de la gouvernance des entreprises et de toutes les autres formes de gouvernance, la gouvernance des données se doit de définir les règles dans lesquelles s'exercent les activités de gestion des données, de veiller au respect de ces règles et à leur mise en application, et d'en assurer les évolutions et les évaluations.

1. Voir présentation générale http://www.fundp.ac.be/info et publications http://www.fundp.ac.be/en/precise/page_view/publications.html.

C'est dans cette perspective qu'a été rédigé ce document qui s'adresse aux responsables, aux praticiens et à toutes personnes intéressées par la gouvernance des données. Il montre qu'en considérant les données comme un écosystème, en proposant des fonctionnalités innovantes telles que la cogénération, la coévolution et la comparaison d'écosystèmes, une démarche d'ingénierie des données dirigée par les modèles (IDDM) contribue aux objectifs de la gouvernance des données. En particulier, il illustre qu'outre le maintien permanent de la cohérence de l'écosystème, l'approche IDDM réalise le lien entre :

– les exigences stratégiques de la gouvernance exprimées par le métier à savoir :

- définir des systèmes d'information (SI) (création de bases de données) ;

- évaluer les SI existants (qualité des données, qualité des bases de données, risques, etc.) ;

- faire évoluer les SI (maintenances évolutives, migrations de bases de données, etc.) ;

- utiliser et réutiliser les données existantes (migration ou intégration de données, échanges, extractions, etc.) ;

– les méthodes à appliquer choisies par le service informatique ;

– les outils opérationnels indispensables aux intervenants techniques pour la réalisation des projets.

Les succès incontestables rencontrés projet après projet par l'utilisation des solutions (méthodes supportées par des outils) exposées dans ce document démontrent leur pertinence et leur efficacité. Adoptées par nombre de grandes organisations et d'intégrateurs, elles sont utilisées dans une grande diversité de projets, d'environnements techniques et organisationnels. Ces solutions offrent à leurs utilisateurs :

– des résultats de très haute qualité professionnelle ;

– une réduction drastique des risques techniques des projets grâce à la maîtrise de tous les composants de l'écosystème ;

– une réduction très importante de la durée et des charges de travail des projets provenant d'une très forte automatisation des processus ;

– le maintien permanent du lien entre le « métier » (ou maîtrise d'ouvrage – MOA) et la réalisation (ou maîtrise d'œuvre – MOE) garant d'évolutions sereines et de la pérennité des investissements.

12.3. Les concepts

12.3.1. *Ingénierie dirigée par les modèles*

Aujourd'hui, dans de nombreux domaines des sciences et techniques, l'utilisation de modèles est quotidienne et a montré son utilité et son efficacité en tant qu'outil :

– de description et de compréhension des systèmes ;

– de communication entre toutes les personnes concernées par une problématique spécifique ;

– d'abstraction permettant de raisonner indépendamment des contraintes techniques ;

– de prédiction permettant d'identifier *a priori* les impacts de modifications ou d'évolutions ;

– de simulation et d'action sur la réalité.

L'ingénierie dirigée par les modèles (IDM ou MDE en anglais pour *model driven engineering*) s'inscrit dans cette approche et peut être définie comme une forme d'ingénierie générative, qui se singularise par une démarche dans laquelle tout ou partie d'une application informatique est générée à partir de modèles.

Cette démarche correspond à un paradigme dans lequel le code source n'est plus considéré comme l'élément central d'un logiciel, mais comme un élément dérivé d'éléments de modélisation. Cette approche prend toute son importance dans le cadre des architectures logicielles et matérielles dirigées par les modèles utilisant des standards actuels.

De telles architectures s'intègrent tout naturellement dans un processus de développement à base de modèles s'assurant, à chaque niveau de modélisation, que les modèles obtenus et réutilisés ont les qualités requises. Cette démarche met le modèle au centre des préoccupations des analystes et des concepteurs.

Si le nom peut paraître nouveau, le processus, lui, ne l'est pas : les activités de modélisation sont le pain quotidien des développeurs depuis toujours. Cependant dans la plupart des cas, les modèles et les solutions restent implicites, ou tout au moins informels et sont appliqués manuellement. Ce que propose l'approche IDM est simplement de formaliser et de mécaniser le processus que les ingénieurs expérimentés suivent à la main. En d'autres termes, l'IDM est la simple transposition dans le domaine informatique de la démarche classique de l'ingénieur qui, avant de réaliser une pièce mécanique, en dresse le plan.

Pour que la démarche IDM soit utile et efficace, il faut bien sûr que les modèles et les processus soient rendus explicites et suffisamment précis pour être interprétés ou transformés par des machines. Dans ce cadre, les processus peuvent alors être vus comme un ensemble de transformations partiellement ordonné des modèles, chacune des transformations prenant un modèle en entrée et produisant un modèle en sortie, jusqu'à obtention d'artéfacts exécutables. Ainsi, lorsque l'on doit dériver une nouvelle solution, qu'elle soit une simple évolution d'un existant ou une nouvelle variante, on peut se contenter de rejouer la plus grande partie du processus en changeant simplement quelques détails ici et là dans le modèle.

L'ingénierie des données dirigée par les modèles (IDDM) dont il est question dans la suite de ce chapitre est tout simplement une démarche IDM appliquée aux écosystèmes des données.

12.3.2. *Ecosystème des données*

Les données permanentes des systèmes informatiques sont stockées dans des bases de données. Cette définition est générique et ne préjuge en aucun cas du type de système de gestion utilisé pour le stockage des données : ce système peut être composé de fichiers plats, de fichiers XML, de systèmes de gestion de bases de données (SGBD) de type hiérarchique, réseau, relationnel, ou de toute combinaison de ces containers.

Quel que soit le système utilisé, les données (pour être plus précis : leurs valeurs) sont stockées selon une structure définie pour pouvoir être traitées par des programmes.

Par ailleurs, les données sont comparables aux pièces d'un puzzle : prise isolément, chacune des pièces répond à des règles précises de hauteur, de flexibilité, etc., et, ensemble, chacune des pièces doit s'ajuster à ses voisines tant dans ses formes que dans ses couleurs. Il en est de même pour les données : isolément, elles doivent répondre à des règles précises (format, longueur, etc.) ; ensemble, elles ont des liens entre elles qui assurent la cohérence de l'information (par exemple : dans une base de données de soins de santé, les soins prénataux ne peuvent être dispensés qu'à des personnes de sexe féminin). Ces règles sont nommées règles de gestion ou plus simplement règles données.

Enfin, les données stockées sont accédées par des programmes pour être utilisées, manipulées, modifiées afin d'atteindre un résultat défini.

S'il est commun de considérer que les structures, les valeurs et les règles de gestion font partie de l'écosystème d'une base de données, il est plus rare d'y inclure

les accès aux données qui se trouvent dans les programmes. Dans une démarche d'IDDM qui se veut complète, il est cependant obligatoire de les inclure pour la bonne et simple raison que les règles de gestion ne sont pas localisées uniquement dans le système de stockage des données mais se répartissent de manière non homogène dans les systèmes de stockage et dans les programmes.

12.3.3. *Gouvernance et ingénierie des données*

Comme toute forme de gouvernance, la gouvernance des données nécessite trois niveaux d'intervention :

- stratégique ;
- tactique ;
- opérationnel.

12.3.3.1. *Niveau stratégique*

Le niveau stratégique définit la vision de l'organisation pour la gestion de ses données, indique les règles générales à appliquer, en évalue l'efficacité et la pertinence. Ce cadre de gouvernance permet aux utilisateurs de préciser leurs exigences en termes :

- de définition de systèmes d'information ;
- d'évaluation des systèmes existants ;
- d'évolution des systèmes mis en place ;
- de réutilisation des données disponibles.

	DÉFINIR	ÉVALUER	ÉVOLUER	REUTILISER
EXIGENCES « MÉTIER »	nouveaux programmes, nouvelles applications, réécritures d'applications,...	qualité des données, qualité des bases de données, adéquation des applications aux besoins,...	changements fonctionnels, organisationnels, fusion de systèmes d'informations, fusion d'applications...	migration et/ou intégration de données, fusion de bases de données, jeux de tests, archivage, échanges de données, ...

Figure 12.1. *Niveau stratégique*

12.3.3.2. *Niveau tactique*

Le niveau tactique définit les solutions (méthodes et outils) applicables aux écosystèmes de données. Ces solutions doivent permettre aux intervenants

techniques de répondre aux exigences imposées par le niveau stratégique et plus particulièrement :

- de développer des systèmes d'informations ;
- de comprendre et de mesurer les systèmes existants ;
- de modifier et de moderniser les systèmes mis en place ;
- d'exporter et d'importer des données.

EXIGENCES « MÉTIER »	DÉFINIR	ÉVALUER		ÉVOLUER		RÉUTILISER	
MÉTHODES & OUTILS	DÉVELOPPER	COMPRENDRE	MESURER	MODIFIER	MODERNISER	EXPORTER	IMPORTER

Figure 12.2. *Niveau tactique*

12.3.3.3. *Niveau opérationnel*

Le niveau opérationnel est l'application aux écosystèmes de données des solutions retenues au niveau tactique.

EXIGENCES « MÉTIER »		DÉFINIR	ÉVALUER		ÉVOLUER		RÉUTILISER	
MÉTHODES & OUTILS		DÉVELOPPER	COMPRENDRE	MESURER	MODIFIER	MODERNISER	EXPORTER	IMPORTER
BASE DE DONNÉES	STRUCTURES	√	√	√	√	√		
	VALEURS	√	√	√	√	√	√	√
	RÈGLES	√	√	√	√	√		
PROGRAMMES	ACCÈS	√	√	√	√	√		
	traitements & écrans							

ÉCOSYSTÈME

Figure 12.3. *Niveau opérationnel*

12.3.3.4. *Rôle de l'IDDM*

Dans ce cadre et pour la réalisation des projets, la démarche d'IDDM est le lien entre les exigences stratégiques, les méthodes et les outils opérationnels :

- les modèles définissent et formalisent les exigences métiers ;
- les transformations de modèle assurent le passage des exigences aux fonctionnalités ;

– les générateurs produisent les outils nécessaires pour l'exécution des fonctions.

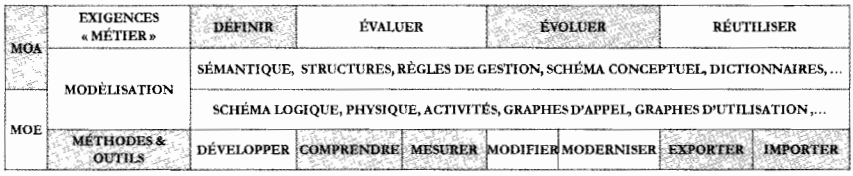

Figure 12.4. *Rôle de l'IDDM dans la gouvernance des données*

La suite de ce document décrit les démarches d'IDDM pour répondre à chacune des exigences métiers. Ces descriptions sont illustrées par des exemples provenant de projets réels réalisés au moyen de la plateforme de modélisation DB-MAIN, complétée par des solutions techniques intégrant des analyseurs et des générateurs de code source. Les exemples utilisés dans ce document proviennent de projets différents réalisés pour différents clients ayant bien entendu des environnements technologiques très diversifiés. Dès lors, ce qui pourrait paraître à première vue comme des exemples erratiques montre au contraire la généricité de la démarche d'IDDM.

Par ailleurs, pour des raisons de compréhension, le vocabulaire utilisé est celui du paradigme entité-association dans la mesure où les représentations utilisées dans les figures sont principalement exprimées dans ce paradigme. Pour rappel, la plateforme de modélisation DB-MAIN a été développée par l'université de Namur et est disponible gratuitement[2].

12.4. Développer

L'expression de nouveaux besoins utilisateurs se traduit *in fine* par les développements de nouveaux programmes ou par le développement de l'entièreté d'une nouvelle application ou encore par la réécriture d'une application existante. Dans toutes ces circonstances, l'objectif de l'IDDM est de mettre à la disposition des développeurs, quelle que soit leur fonction – analyste fonctionnel, programmeur, administrateur de base de données – des outils qui leur permettent d'accélérer la réalisation des nouveaux écosystèmes.

2. Voir : http://www.db-main.eu.

Le but poursuivi par les méthodes et outils est de permettre en partant de sa définition de co-générer l'écosystème des données : création des structures de la base données et des méthodes d'accès aux données.

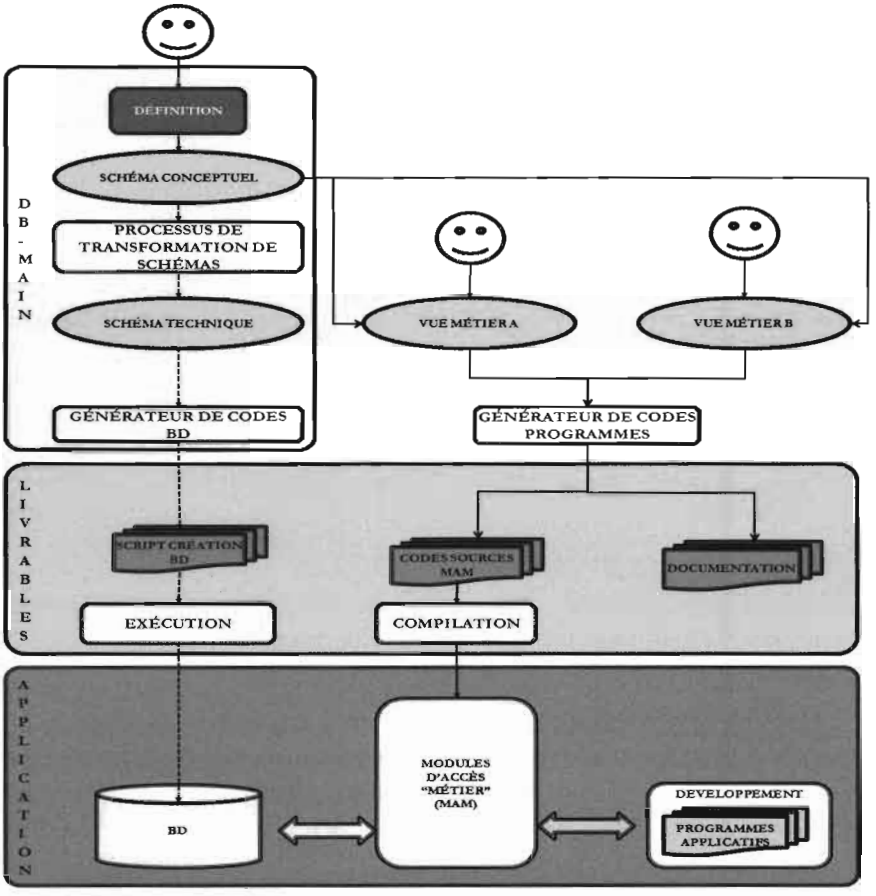

Figure 12.5. *Processus de cogénération*

Le processus de cogénération (voir figure 12.5) peut être synthétisé comme suit :

– il convient d'abord de définir un schéma conceptuel du SI. Ce schéma est une définition métier du SI à implémenter, indépendante de toute technologie. A partir du schéma conceptuel un processus automatique de transformation va produire un schéma technique qu'un générateur de code va utiliser pour produit le script de création de la base de données (BD) ;

– par ailleurs, les utilisateurs ou développeurs décrivent une ou plusieurs vues métiers. Ces vues sont des sous-ensembles du schéma conceptuel donnant des perceptions du S.I. correspondantes à des besoins métiers. Pour chacune des vues définies, un processus de génération permet de produire :

- le code source d'une couche d'intermédiation (*middleware* : les modules d'accès métier ou MAM). Cette couche contient l'ensemble des méthodes d'accès pour la gestion des données, les MAM offrant des accès à la BD selon la logique de la vue métier dont ils sont issus ;

- la documentation technique de la couche d'intermédiation destinée aux développeurs afin qu'ils puissent utiliser les MAM.

Ces fonctionnalités de base sont complétées par un générateur d'interfaces homme-machine permettant aux utilisateurs d'accéder directement aux données suivant la logique qu'ils ont définie dans la vue métier[3].

La cogénération supportée par les outils adéquats présente plusieurs intérêts majeurs :

– elle se base exclusivement sur des descriptions fonctionnelles qui sont traduites en termes techniques de manière cohérente et homogène par des automates ;

– l'approche proposée isole les processus de gestion des données des processus de traitements offrant ainsi une architecture agile permettant de faire évoluer les différentes couches techniques avec une certaine indépendance ;

– la méthodologie n'induit, ni n'impose, aucune architecture technique. Seul le générateur de MAM est dépendant du choix de l'architecture pour générer du code source. Ce dernier, lui, doit être conforme à l'architecture choisie et aux orientations stratégiques définies (centralisée, décentralisée, etc.) ;

– il s'agit d'une approche applicative, les vues métier n'étant pas cantonnées à une seule base de données. Les MAM accèdent à plusieurs bases, éventuellement implémentées dans différents SGBD ;

– la génération de code source peut se faire dans différents langage de programmation (JAVA, C, COBOL, etc.).

3. Voir un exemple http://www.rever.eu/DISTRIBUTION/DB-MAIN/DEASY-datasheet-fr.pdf et le produit, http://www.db-main.eu/?q=fr/node/220.

12.5. Evaluer

Pouvoir évaluer une application, mesurer la qualité des données, estimer les risques d'adaptation d'une application à des nouveaux besoins sont des fonctions indispensables pour une bonne gouvernance des données.

Cette exigence se heurte à la réalité des applications. En effet, les programmes s'accumulent au cours du temps, formant un agglomérat dont la complexité est renforcée par les évolutions technologiques, les maintenances évolutives et correctives. Cet accroissement de la complexité au cours du temps est accompagné en parallèle d'une perte de la connaissance de l'application : documentation incomplète, dépassée et les intervenants, ayant la connaissance, appelés à d'autres fonctions ou tâches.

Cette complexité des programmes engendre une opacité encore plus grande pour les données qui sont au cœur des applications. Dans ce contexte, le premier objectif de l'IDDM est de comprendre l'application afin d'en avoir une connaissance suffisamment détaillée pour ensuite pouvoir mesurer la qualité des données, la qualité des bases de données, les risques, etc.

12.5.1. *Comprendre*

La méthodologie de rétro-ingénierie explicitée ci-dessous a pour objectif de reconstruire la définition de l'écosystème du S.I. quelle que soit la diversité des structures de stockage et de gestion des données qui le compose. Cette définition passe par la reconstruction des différents niveaux de modèle. Il va de soi que si un modèle complet ou partiel est déjà disponible, il n'est pas obligatoire d'exécuter toutes les étapes du processus décrit en section 12.5.1.1. Par ailleurs, la précision (la granularité) des modèles est dépendante des étapes réalisées et il convient en fonction des besoins du projet de choisir le niveau de précision adéquat.

12.5.1.1. *Méthode*

Pour atteindre l'objectif défini, la méthode proposée consiste à analyser deux catégories d'éléments disponibles :

– les éléments techniques : les scripts de création de la BD, les codes sources de l'ensemble des processus applicatifs (procédures de base de données, *triggers*, programmes, JCL, scripts, etc.) et enfin les données elles-mêmes ;

– les éléments non techniques tels que les documentations existantes, la connaissance des développeurs et des utilisateurs.

L'analyse des éléments techniques se déroule en plusieurs étapes successives qui améliorent et valident au fur et à mesure de leur déroulement la précision et la qualité du modèle.

1. La première étape permet de reconstruire le modèle physique de la BD par simple analyse des scripts de création ou interrogation directe de la BD (pour la plupart des SGBD relationnels).

2. La deuxième étape permet de compléter le modèle précédent par des éléments déclarés explicitement dans les programmes et non déclarés dans la BD. Ainsi, à titre d'exemple, il est fréquent de trouver dans les BD des colonnes de plusieurs centaines de caractères dont la (ou les) description(s) est (sont) définie(s) dans les programmes. Cette étape est obligatoire lorsqu'il s'agit d'applications fonctionnant avec des fichiers plats.

3. La troisième étape permet de produire le modèle logique de l'application. Ce dernier est construit principalement en enrichissant les résultats de l'étape précédente par la découverte des règles de gestion se trouvant dans les programmes.

4. La quatrième étape consiste à valider les résultats obtenus par l'analyse des données. En effet, la non-conformité des valeurs des données aux règles définies dans le modèle a pour conséquence de s'interroger sur l'origine de l'écart : valeur erronée, règle erronée ou règle incomplète ? En outre, cette étape permet d'enrichir le modèle sur la base des valeurs analysées : colonnes inutilisées, valeurs par défaut, etc.

5. La dernière étape consiste à abstraire les résultats techniques pour produire un modèle conceptuel indépendant des technologies. Ce modèle conceptuel brut peut être complété par l'apport des connaissances provenant de la documentation et des expertises disponibles. Cet apport permet alors d'obtenir un schéma conceptuel dont la sémantique est enrichie et tend à exprimer au mieux la perception du SI du point de vue des utilisateurs.

Le processus décrit ci-dessus est automatisé à plus de 90 %. Les tâches manuelles sont les validations des résultats de chacune des étapes ainsi que l'enrichissement du modèle conceptuel brut au moyen de la documentation et des expertises humaines. Il convient également de souligner que la méthodologie proposée est totalement générique et est utilisable pour tous les types de SGBD, de langages et de systèmes d'exploitation.

12.5.1.2. *Exemples*

Il n'est évidemment pas possible ici de décrire de manière exhaustive tous les résultats du processus de rétro-ingénierie qui vient d'être décrit. Ce document présente essentiellement la capacité des automates à reconstruire le modèle des données. Outre la reconstruction des modèles, la rétro-ingénierie produit des

résultats complémentaires très utiles pour la compréhension d'un SI tels que la cartographie applicative, l'évaluation des risques techniques des SI, les analyses d'impacts, etc. Ces résultats ne sont toutefois pas décrits ici[4].

Les figures 12.6 à 12.9 illustrent chacune des étapes du processus de rétro-ingénierie montrant l'évolution de la précision du modèle. L'exemple est un sous-ensemble d'une base de données IDS2 (environnement BULL GCOS8) devant migrer vers un environnement IBM, Z/OS et DB2 : à titre indicatif la base de données complète comprend 255 types d'enregistrement et l'application environ 1,5 million de lignes de codes COBOL).

L'analyse du code de création de la base fait apparaître les types d'entité (type d'enregistrementIDS2) et les types d'association déclarés (figure 12.6). La notion d'*owner-member* propre aux bases réseaux indique le sens du type d'association (par exemple, pour une occurrence de IDENT1 il est possible d'avoir plusieurs INSTITUT et réciproquement, une occurrence d'INSTITUT ne peut dépendre que d'un et un seul IDENT1).

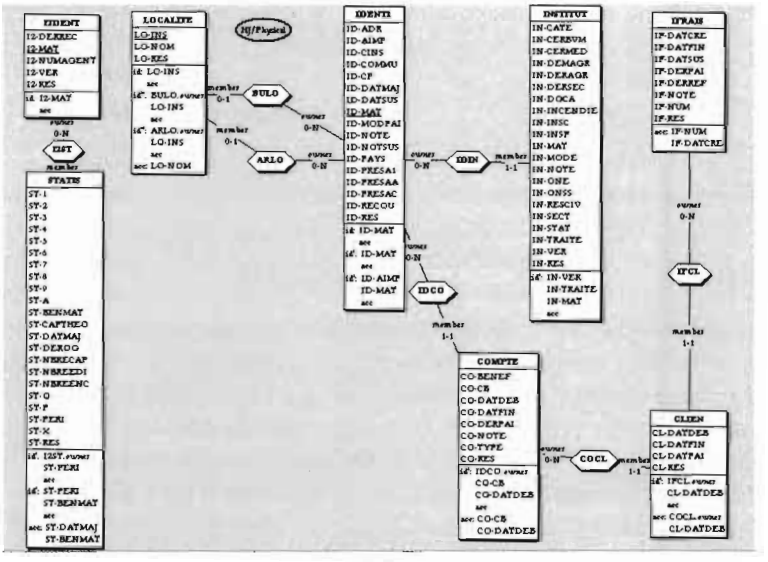

Figure 12.6. *Résultats de l'analyse du script de création de la BD*

4. Pour une description plus approfondie illustrée par des exemples détaillés voir http://www.rever.eu/white-papers/methodesIDDM-FR.pdf. Une description des outils mis en œuvre est fournie dans le document http://www.rever.eu/white-papers/solutionstechniques IDDM-FR.pdf.

L'analyse des codes sources des programmes permet de compléter le schéma physique notamment par l'ajout des décompositions des structures telles qu'elles sont utilisées par les programmes (figure 12.7).

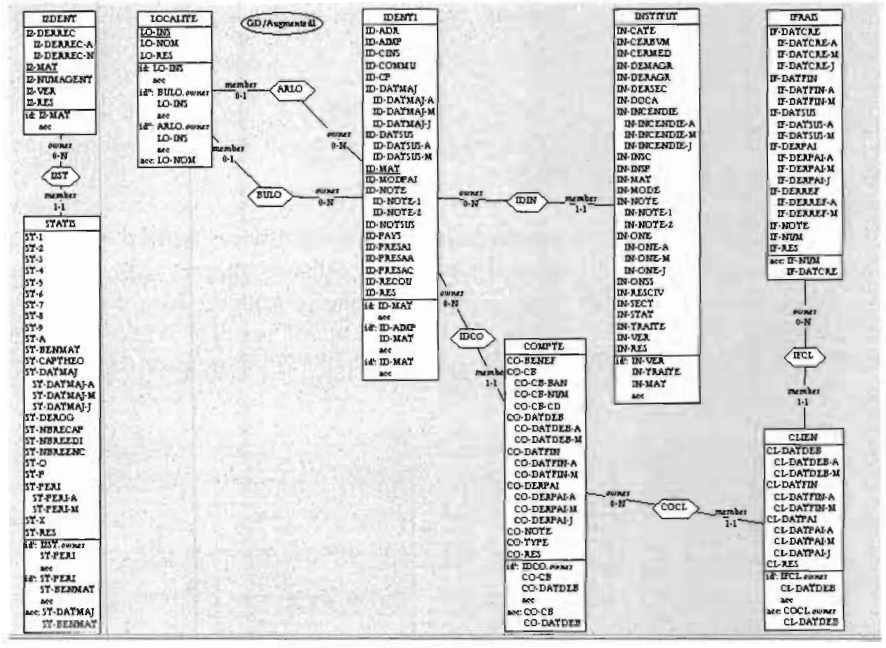

Figure 12.7. *Enrichissement par ajout de structures*

Une analyse plus détaillée des programmes permet de compléter le schéma par des types d'association complémentaires utilisés par les programmes. Ces types d'association entre types d'entité sont des règles données qui sont gérées exclusivement par les programmes. On notera sur l'exemple que les deux grappes de types d'entité qui semblaient être indépendantes l'une de l'autre sont en fait associées par l'utilisation qu'en font les programmes (figure 12.8).

Afin de s'assurer de l'exactitude du modèle reconstruit par l'analyse des codes sources, un processus automatique confronte les règles définies par le modèle aux données. Les règles ainsi contrôlées concernent tant le respect des formats que les types d'association ou les contraintes de dépendances entre attributs types (figure 12.9).

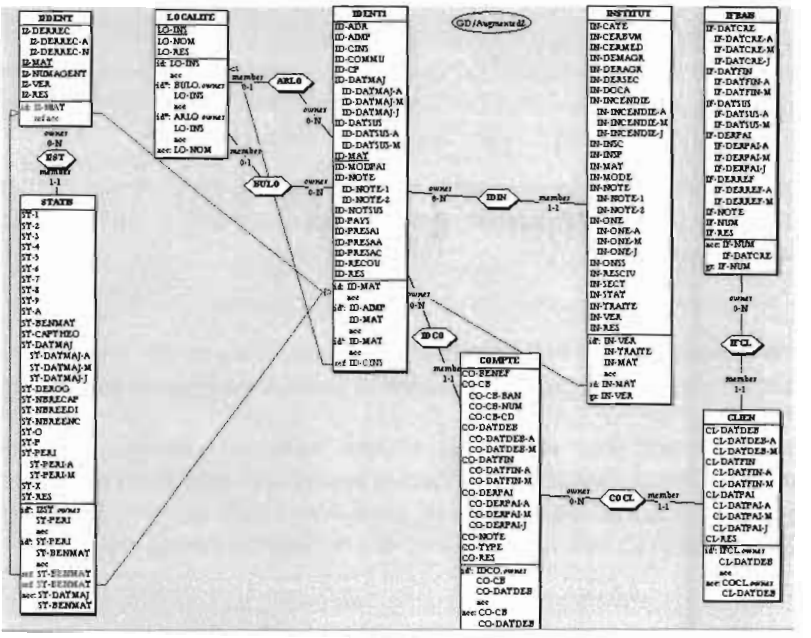

Figure 12.8. *Enrichissement par ajout de règles données*

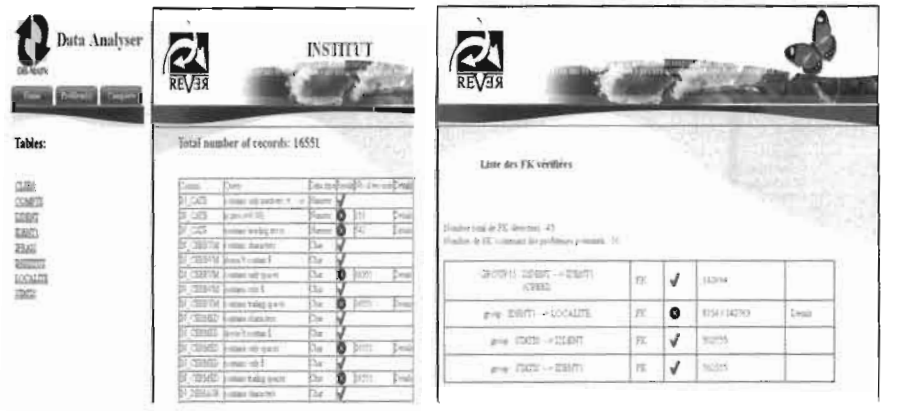

Figure 12.9. *Validation du modèle*

Pour produire le modèle conceptuel (figure 12.10), les actions suivantes ont été réalisées :

– l'analyse des données a montré une équivalence des clés entre IDENT1 et I2DENT, donc ils ont été fusionnés ;

– un enregistrement qui était l'implémentation d'une relation N-N est devenu un type d'association après suppression des attributs redondants ;

– dans INSTITUT, la contrainte rajoutée mettait en évidence le fait que l'attribut IN_MAT était redondant avec le type d'association IDIN et donc, cet attribut a été supprimé ;

– deux des contraintes ajoutées ont été transformées en types d'association ;

– les décompositions des dates en années, mois, jours ont été supprimées, ainsi que la décomposition des comptes bancaires et la décomposition de notes en lignes.

On notera que pour obtenir un schéma lisible, il convient également de renommer les types d'entité, les attributs et les types d'association afin d'avoir un vocabulaire significatif. Ce travail a bien été réalisé dans le cadre du projet, mais pour des raisons évidentes de confidentialité son résultat n'est pas présenté ici.

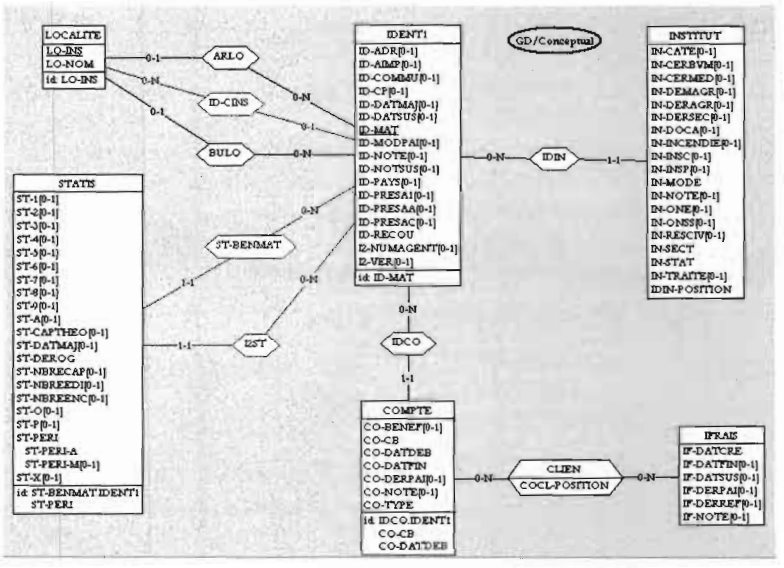

Figure 12.10. *Modèle conceptuel*

12.5.2. *Mesurer*

La connaissance détaillée d'une application permet d'apprécier la qualité des éléments de l'application. En fonction des besoins, les évaluations peuvent aller de

la simple mesure d'un ou de plusieurs composants à la mise en place d'une solution complète pour leur gestion. A titre d'exemple, la connaissance détaillée du modèle des données d'une application, permet en partie de mesurer son adéquation aux besoins informationnels des utilisateurs.

Il va de soi qu'il convient d'adapter les méthodes en fonction du type de composants que l'on souhaite mesurer voire améliorer. Quelques exemples de mesures possibles sont fournis ci-dessous. On notera toutefois, que dans la mesure où l'IDDM se limite aux accès des données, elle ne peut prétendre à une évaluation de la qualité des programmes : ce point relève d'un autre domaine de l'informatique.

12.5.2.1. *Qualité des données*

Au cours de la rétro-ingénierie, les processus utilisés à l'étape quatre génèrent automatiquement les outils permettant d'évaluer la conformité des données aux règles décrites dans le modèle. Cette évaluation est effectuée sur chacune des valeurs contenues dans la BD par rapport à l'ensemble des règles que cette valeur doit respecter. Toutes les non-conformités sont identifiées et rapportées en indiquant pour chacune d'entre elles, quelle est la règle qui n'est pas respectée, quelles sont les valeurs qui ne respectent pas la règle et quels sont les programmes concernés par ces données.

Pour mesurer la qualité des données, le principe est de construire un référentiel qui contient l'ensemble des règles de gestion. Outre les règles provenant du modèle, ce référentiel peut intégrer d'autres règles (conformités à des réglementations, règles internes, etc.) fournies directement par des intervenants. A partir de ce référentiel, des processus automatisés génèrent les requêtes permettant de confronter les données aux règles. Ce processus permet d'identifier :

- les données erronées susceptibles de faire l'objet de corrections ;
- les règles de gestion incomplètes ou inexactes et qui doivent être précisées.

La connaissance des dépendances données-programmes renforce la démarche de prévention des erreurs en particulier par l'amélioration des contrôles effectués dans les programmes.

Dans le même ordre d'idée il est possible d'adjoindre des modules complémentaires :

- d'historisation des résultats des contrôles permettant ainsi une mise en perspective des améliorations de la qualité des données et d'une évaluation du ratio des coûts des efforts d'amélioration de la qualité rapportés aux résultats obtenus ;

– d'évaluation de l'impact des données erronées en terme métier en se basant sur des règles d'évaluation métier par exemple, l'absence d'un code postal dans une adresse empêche l'expédition d'une facture ce qui représente un impact financier de x % du montant de la facture.

Il va de soi que les mesures effectuées par les systèmes décrits ici n'ont pas la prétention de résoudre l'entièreté des questions concernant la qualité des données. En effet, les données contenues dans les bases de données informatisées se doivent d'être le plus proche possible de la réalité du domaine qu'elles décrivent, en particulier elles doivent être exactes, fiables et à jour.

Quels qu'ils soient, les systèmes techniques de contrôle ne peuvent prétendre atteindre ce résultat qui reste une responsabilité humaine et organisationnelle : tout au plus peut-on espérer des systèmes techniques le contrôle d'une certaine cohérence des données.

En d'autres termes, il n'est pas possible pour un système technique de vérifier que madame X a bien trois enfants en revanche, il est possible de mettre en évidence qu'une base de données signale qu'elle en a trois alors qu'une autre n'en indique qu'un seul.

12.5.2.2. *Qualité des bases de données*

Au-delà de la qualité des données, l'IDDM permet d'apprécier la qualité de la base de données. Les critères à prendre en compte peuvent être très divers : nombre de colonnes par tables, nombre d'identifiants, redondance des attributs, règles de gestion définies dans le SGBD, utilisation des données par les programmes, etc. Il va de soi que ces appréciations viennent compléter les informations fournies par les SGBD en matière de performance, de taux d'utilisation, etc.

L'évaluation de la qualité de la base de données est utile notamment pour :

– évaluer la complexité des évolutions des applications : degré de dépendances des types d'entité, redondances des informations, dépendances des éléments, etc. ;

– compléter les évaluations de la qualité des programmes pour fournir une mesure de la qualité d'une application.

Des travaux plus approfondis en la matière sont actuellement en cours : ils concernent notamment la détection automatique de constructions complexes susceptibles d'être sources d'erreurs.

12.6. Evoluer

Aussi riches qu'ils soient, les systèmes d'information ne sont et ne seront jamais qu'un discours sur la réalité. A ce titre ils sont irrémédiablement condamnés à être complétés et enrichis afin de coller au plus près à la réalité qu'ils prétendent décrire. Par ailleurs, la réalité est changeante et évolue au cours du temps. L'exigence de l'évolution des systèmes d'information est donc une nécessité vitale de la gouvernance des données.

Du point de vue des analystes métier, les évolutions possibles d'un SI peuvent être classées en trois catégories :

– soit il s'agit de répondre à des changements fonctionnels liés à des changements organisationnels ou réglementaires, ce qui implique des modifications dans la base de données d'une ou plusieurs applications existantes ;

– soit il s'agit de changements techniques et en particulier un changement de SGBD tout en conservant l'application existante : il s'agit d'une modernisation de l'application ;

– soit enfin il s'agit de remplacer l'application existante par une nouvelle. Dans ce cas, il convient d'exporter les données d'une application existante pour les importer dans une autre base de données : ce type d'évolution est traité dans la section 12.7 concernant les exportations et importations des données.

12.6.1. *Modifier ou moderniser*

Si d'un point de vue fonctionnel et technique les deux types d'évolution pris en compte dans ce chapitre n'ont pas les mêmes objectifs, force est de constater que du point de vue méthodologique les méthodes utilisées dans les deux cas sont très proches.

Que ce soit pour modifier une base de données ou pour la moderniser (changements de SGBD) les processus à exécuter sont semblables et synthétisés dans la figure 12.11 : les différences dans la méthodologie sont mises en italique. La méthode prévoit cinq phases principales.

Comprendre. C'est une des spécificités fortes de la méthode proposée : il faut une connaissance suffisante de l'écosystème source. A l'image d'un GPS qui ne peut pas calculer un chemin sans se situer, la connaissance du point de départ d'un projet est indispensable à sa réussite.

Définir. Cette phase consiste à définir dans les modèles, les évolutions qui doivent être réalisées pour atteindre les résultats finaux espérés : changements des structures et des règles de gestion. Concrètement cela revient à définir le modèle conceptuel de la base cible. Dans le cas d'une modernisation, le modèle conceptuel de la base cible peut être identique ou différent de celui de la base source.

Contrôler. A ce stade, les situations source et cible étant connues, il convient d'identifier les obstacles qui pourraient empêcher le passage de la source à la cible. Ces obstacles, qui de fait sont les risques techniques des projets, proviennent de trois origines distinctes qui doivent faire l'objet de contrôles *a priori* :

– *contrôle des compatibilités :* il s'agit d'identifier les incompatibilités entre source et cible, c'est-à-dire toute évolution qui ne peut pas être réalisée par un simple transfert de valeurs. Les types d'incompatibilités dépendent bien entendu du type de projet, l'essentiel étant de pouvoir les détecter. A titre d'exemple, et sans que la liste ci-dessous ne soit exhaustive :

- dans le cas des évolutions de bases de données, les incompatibilités proviennent essentiellement de la non-conformité des données aux nouvelles règles de gestion définies pour la base cible : par exemple si une colonne a évolué pour devenir une clé étrangère il est nécessaire de vérifier que les valeurs des données soient conformes à cette règle ;

- dans le cas des changements de SGBD, les incompatibilités proviennent généralement de fonctionnalités supportées par le SGBD source et non supportées en tant que telles par le SGBD cible : par exemple dans la plupart des SGBD anciens ou *legacy systèmes* (IMS, IDS2, IDMS, etc.) l'ordre de présentation des lignes d'une table est défini au moment de l'écriture, alors que dans un SGBD relationnel cet ordre est défini au moment de la lecture. Cet ordre pouvant avoir de l'importance pour les programmes applicatifs, il convient dans ce cas de définir les clés de tri adéquates qui permettront le bon fonctionnement des programmes ;

– *contrôle de la qualité des données* : quelles que soient les règles d'évolution à appliquer, il convient de vérifier que les données source sont conformes aux règles du SI cible. Ce contrôle s'effectue en vérifiant la conformité des données par rapport au modèle cible ;

– *contrôle de la propagation des modifications* : la première étape de la méthode a pour résultat une connaissance approfondie des trois niveaux de dépendances : dépendances données-données (le modèle des données), dépendances données-programmes (graphe des programmes utilisant une donnée), programmes-programmes (graphe des appels entre programmes). Cette connaissance permet d'identifier pour chacune des modifications d'un élément de l'écosystème les répercutions sur les autres éléments.

Executer. Les activités à déployer au cours de cette phase sont :

– la génération automatique des outils nécessaires pour la création de l'écosystème cible :

- les structures de la base cible ;

- dans le cas de modernisation, les accès à la base cible ;

- les composants (programmes) de la migration des données ;

– la génération automatique des outils nécessaires à la migration des données depuis la base source vers la base cible :

- les programmes de déchargement en intégrant les transformations de valeur ou changements de format nécessaires ;

- les scripts de rechargement de la base cible ;

- l'exécution des programmes de déchargement pour produire des fichiers plats au format des tables de la base cible directement chargeables dans la base au moyen des scripts et des utilitaires standard du SGBD ;

– dans le cas de la modernisation, l'adaptation des programmes applicatifs pour qu'ils utilisent les accès à la base cible qui ont été générés.

On notera que les adaptations des programmes applicatifs pour qu'ils prennent en compte les évolutions fonctionnelles sont en dehors de la portée des automates : elles restent donc à réaliser par des intervenants.

Valider. Cette phase a pour but de garantir que la migration des données depuis le SI source vers le SI cible n'a pas perturbé la cohérence de l'écosystème. La démarche utilisée consiste à déduire des modèles sources et cibles et de leurs correspondances, un modèle commun à partir duquel il est possible de générer des programmes qui permettent de valider la migration sous deux angles :

– *validation de l'exhaustivité* : les programmes générés pour chacun des environnements sources et cibles effectuent un comptage du nombre d'occurrences physiques dans les bases et un contrôle fonctionnel (*checksum*) de chacun des attributs. Un rapport publie les résultats de ces comptages en mettant en exergue les éventuelles différences ;

– *validation de la cohérence* : les programmes générés extraient les données de chacune des bases source et cible. Les résultats des extractions sont comparés et un rapport publie les éventuelles différences de valeurs et d'occurrences.

On notera que cette méthode est utilisable quels que soient les types de SGBD et les structures des bases sources et cibles et qu'elle permet de valider une migration de données indépendamment de toute application.

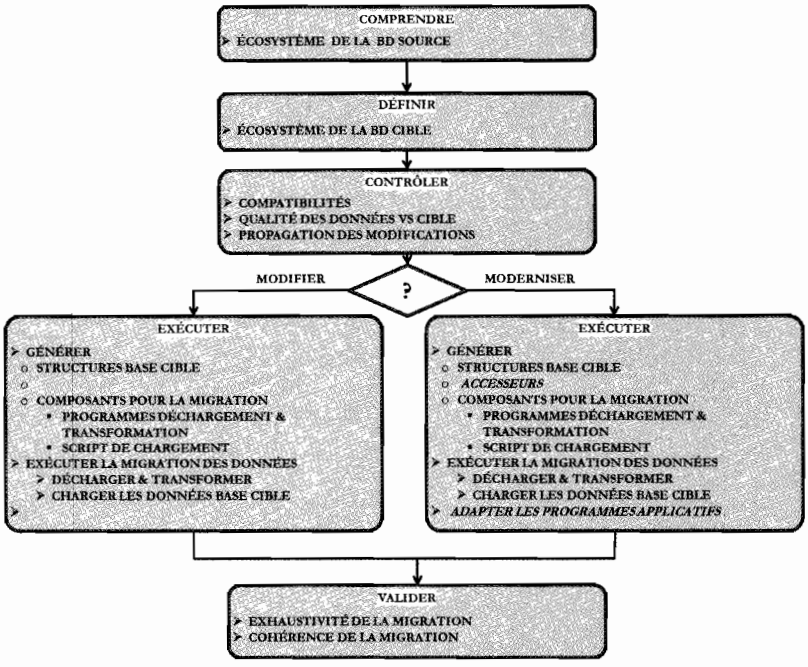

Figure 12.11. *Evolution ou modernisation d'un écosystème BD -structures, données, règle de gestion, accès*

12.6.2. *Coévolution*

La coévolution n'est qu'un cas particulier de la méthodologie de modification décrite ci-dessus : la seule différence est l'ajout dans la phase d'exécution du générateur de MAM pour la nouvelle définition de la base.

Ce processus part d'un modèle conceptuel existant et comprend six étapes successives :

1. la première consiste à définir le nouveau schéma conceptuel en partant du schéma conceptuel existant. Une fois que le nouveau schéma conceptuel est défini, un transformateur de schémas produit un nouveau schéma technique et une nouvelle vue métier ;

2. la deuxième étape a pour objectif de s'assurer que les données contenues dans la base de données sont compatibles avec les contraintes et règles définies pour la nouvelle version de la base. Un rapport indique les données non conformes : il est alors nécessaire soit de corriger ou d'adapter la ou les donnée(s), soit de redéfinir la ou les contrainte(s) ;

3. la troisième étape produit, en fonction des choix effectués, soit un script d'adaptation de la base de données technique existante soit un script de création d'une nouvelle base de données. Le choix entre l'un ou l'autre des scripts est souvent lié à l'importance des modifications effectuées ;

4. la quatrième étape a pour objectif d'adapter ou de migrer les données contenues dans la base existante vers la nouvelle base. Cette étape génère également les processus de validation qui permettent de garantir l'exhaustivité de la migration et le maintien de la cohérence des données ;

5. la cinquième étape permet de générer les modules d'accès métier (MAM) pour la nouvelle version de la base de données. Elle fournit également la liste des MAM existants qui sont impactés par les évolutions ;

6. enfin, la sixième et dernière étape a pour objectif de fournir aux développeurs, la liste des programmes ou modules applicatifs impactés par les modifications et qu'il convient d'adapter.

12.7. Réutiliser

Les données sont les matières premières pour la fabrication de l'information. Elles doivent donc pouvoir être utilisées dans d'autres contextes que ceux dans lesquels elles sont crées et gérées : il faut pouvoir les exporter c'est-à-dire les extraire pour les échanger, les archiver, ou plus simplement produire des échantillons, éventuellement rendus anonymes, à des fins de tests ou de formation.

Si les données doivent pouvoir être exportées, elles doivent pouvoir également être importées. Dans ce cas il faut pouvoir, sans qu'elles ne perdent leur cohérence et signification, les transformer, les agréger, les désagréger, les intégrer dans d'autres SI dont les structures et les règles de gestion sont fixées au préalable.

12.7.1. *Exporter*

La nécessité d'exporter tout ou partie des données d'une application est très fréquente : échanges de données avec des partenaires (clients, fournisseurs, etc.),

archivage de données inactives, mise à disposition auprès d'intervenants du métier de sous-ensembles à des fins de contrôles, de tests, de formation, etc.

L'objectif est de permettre d'extraire d'une base de données des sous-ensembles cohérents de données selon une logique métier.

Pour atteindre l'objectif, la méthode suivie consiste à :

– définir dans le modèle conceptuel, le ou les sous-ensemble(s) fonctionnel(s) souhaité(s). Ces sous-ensembles définissent la liste des types d'entité et des attributs à extraire ;

– définir les critères qui permettent de réduire la volumétrie des données au sein des sous-ensembles fonctionnels créés à l'étape précédente ;

– définir pour chacun des sous-ensembles fonctionnels les modèles cibles à produire : ces modèles cibles permettent en fait de définir les formats résultats de l'extraction (fichiers plats, base de données, fichier XML, etc.) ;

– générer le code des programmes de sélection : l'objectif de ces programmes est d'obtenir les clés des types d'entité racines des sous-ensembles fonctionnels ;

– générer le code des programmes d'extraction dont le but est d'extraire les données des bases de production pour toutes les occurrences de clés sélectionnées.

12.7.2. *Jeux de tests*

Pour la fabrication de jeux de tests, les outils d'extraction sont complétés par :

– l'utilisation des dépendances données-programmes pour déterminer automatiquement le sous-ensemble fonctionnel nécessaire et suffisant pour les programmes à tester ;

– des outils permettant :

- de générer des données à partir de règles prédéfinies. Cette génération permet la génération de lignes dans les tables et la génération de valeurs dans les colonnes de chacune des lignes ;

- d'anonymiser les données à des fins de confidentialité. Ce processus assure le maintien de la cohérence des données lorsque des identifiants sont modifiés par l'anonymisation ;

- de comparer les valeurs des données avant et après test, apportant une aide efficace au dépouillement ;

- des outils permettant d'évaluer la couverture technique des tests. Le but recherché ici est de déterminer le taux de couverture du jeu de données ayant servi aux tests : si le jeu de données n'a permis de tester que 30 % d'un programme, il est probable qu'il soit nécessaire de fournir un nouveau jeu de données.

12.7.3. *Epuration*

De la même manière que pour les jeux de tests, les outils d'extraction peuvent êtres complétés par des outils qui suppriment de la base les données qui ont été extraites.

12.7.4. *Importer*

Si les données doivent pouvoir être exportées, elles doivent pouvoir également être importées. Dans ce cas il faut pouvoir, sans qu'elles ne perdent leur cohérence et signification, les transformer, les agréger, les désagréger, les intégrer dans d'autres SI dont les structures et les règles de gestion sont fixées au préalable.

La méthodologie utilisée pour importer des données est proche de celle utilisée pour les évolutions des bases de données on y retrouve les mêmes cinq mots-clés.

Comprendre. Il s'agit de s'assurer que l'ensemble des éléments constitutifs des écosystèmes source et cible sont identifiés et connus avec un degré de précision suffisant pour atteindre les objectifs du projet.

Définir. Dans les processus d'importation, le modèle cible est déjà défini. Dès lors les définitions attendues dans cette phase sont les correspondances sources-cibles. Il va de soi que ces correspondances sont d'abord d'ordre sémantique puis dans un second temps d'ordre technique.

Contrôler. Une fois que les correspondances sont définies, il convient :

– d'identifier les incompatibilités qui proviennent :

- des différences dans les règles de gestion (y compris celles gérées par les programmes) ;

- des différences dans les définitions des structures ;

– de valider les règles de transformation qui vont permettre de résoudre les incompatibilités détectées. La diversité des situations et des règles de transformation ne permet pas d'automatiser cette phase. Les règles de transformations spécifiques doivent donc faire l'objet de développements appropriés. Au préalable, un prototy-

page des règles de transformation est réalisé afin de s'assurer de leur validité avant d'effectuer les développements proprement dits. Chacune des règles de transformation est testée sur la totalité des données qui la concerne ;

– à l'issue du prototypage de chacune des règles de transformation, il est nécessaire de contrôler la qualité des données transformées par rapport au modèle cible, afin de s'assurer de leur conformité aux règles du système cible. Ce contrôle permet, de fait, de valider les règles de transformation.

Executer. Il faut d'abord développer, sur la base des prototypes réalisés, les modules spécifiques qui assurent les transformations complexes. Ensuite, un générateur de codes sources permet de produire les programmes de déchargement en y intégrant les modules développés et les transformations standard telles que les changements de type ou les changements de format. Les scripts de chargement pour la base cible sont également générés. Enfin, l'exécution des programmes et des scripts générés permet d'effectuer la migration.

Valider. Cette phase permet de garantir que la migration des données depuis le SI source vers le SI cible a maintenu la cohérence des données existantes dans l'écosystème source : en d'autres mots, il faut s'assurer que les commandes de monsieur X n'ont pas été attribuées à madame Y. La démarche utilisée consiste à déduire des modèles sources et cibles, et des correspondances qui ont été établies, un modèle commun (souvent au format XML) à partir duquel il est possible de générer des programmes qui permettent de valider la migration :

– *validation de l'exhaustivité* : les programmes générés pour chacun des environnements sources et cibles effectuent un comptage du nombre d'occurrences physiques dans les bases et un contrôle fonctionnel de chacun des attributs. Un rapport au format HTML publie les résultats de ces comptages en mettant en exergue les éventuelles différences ;

– *validation de la cohérence* : les programmes générés extraient de chacune des bases, les données au format XML. Les fichiers résultats des extractions sont comparés et un rapport publie les éventuelles différences de valeurs et d'occurrence.

12.8. Conclusions et perspectives

Comme on le voit, l'approche IDDM offre un ensemble de solutions couvrant une très large part des besoins de la gouvernance des données en termes tactiques et opérationnels. Par ailleurs, l'industrialisation de ces solutions permet de se libérer de plus en plus des aspects purement technologiques.

La pratique de la démarche d'IDDM amène à constater qu'elle raccourcit la distance entre métier et informatique. Rien de plus normal somme toute : que peut-on espérer de mieux qu'une formalisation (les modèles) pour définir les exigences et utiliser cette formalisation pour la réalisation technique ?

A ce titre, elle est l'approche idéale pour s'assurer de la conformité d'une application informatique aux besoins métier, en particulier dans le cas de réception d'application telle qu'elle a été utilisée dans le cadre d'un projet important pour une administration publique. La méthode illustrée (figure 12.12) se base sur :

– une démarche de modélisation en amont de l'application informatique réalisée au moyen de l'outil d'ingénierie des exigences « *objectiver* » de la société RESPECT-IT[5]. « *Objectiver* » a une approche centrée sur les objectifs métier et permet de préciser au moyen de modèles :

- les besoins métier (modèle des exigences) permettant de générer des cahiers des charges ;

- les activités métier (modèle des opérations) mises en œuvre pour répondre aux exigences ;

– une démarche de rétro-ingénierie de l'application livrée. Outre le modèle des données la rétro-ingénierie a produit le modèle des traitements et le modèle des fonctions.

La confrontation (manuelle) du modèle des opérations et du modèle des fonctions a permis de vérifier la similitude des modèles (ce qui signifie dans ce cas que l'application est conforme aux spécifications).

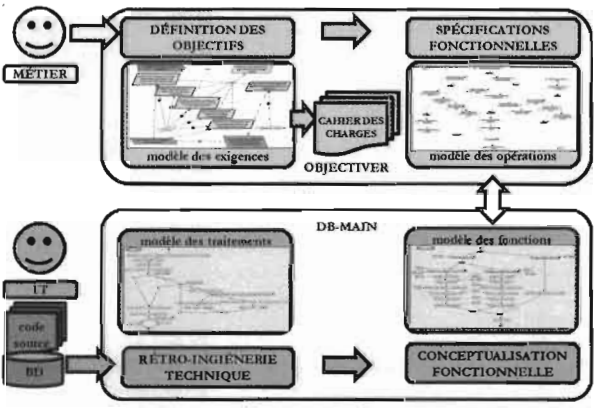

Figure 12.12. *Comparaison des modèles exigences-traitements*

5. voir http://www.objectiver.com/index.php?id=4&L=1.

En outre, un tableau des correspondances exigences-traitements (figure 12.13) a été établi et enrichi par un certain nombre d'informations permettant d'établir des priorités de tests et de prononcer des réceptions provisoires.

L'enrichissement du tableau s'est fait à partir de trois sources d'informations complémentaires :

– la criticité de l'exigence pour le métier ;

– l'évaluation des risques BD fournis par des outils de mesure des dépendances données-programmes ;

– l'évaluation de la complexité des programmes fournis par des outils de mesure de la qualité des programmes.

MODÈLE DES EXIGENCES	MODÈLE DES OPÉRATIONS	CRITICITÉ « MÉTIER »	MODÈLE DES FONCTIONS	MODÈLE DES TRAITEMENTS	COMPLEXITÉ BD	COMPLEXITÉ PROGRAMMES
CONDITION SUR L'ABSENCE DE RÉCLAMATIONS EN COURS POUR LES DETTES NON SCINDÉES ET VÉRIFIÉES	SÉLECTIONNER DETTES NON SCINDÉES	HAUTE	CHERCHER TOUTES LES DETTES	GETNN242MBYNN	1	3
			CHERCHER LES DETTES NON SCINDÉES	GETRAPB, GETRAPBBYARTICLE, GETRAPBDETAIL	12	9
	FILTRER LES DETTES CALCULÉES	HAUTE	VÉRIFIER LA DATE D'ÉCHÉANCE	GETRKOH	1	1
			CHERCHER LES PAIEMENTS	GETALLPAIEMENT, GETPAIEMENT, GETDETAILPAIEMENT, GET TOTALPAIEMENT	2	1
	FILTRER SUR L'ABSENCE DE RÉCLAMATION EN COURS	BASSE	CHERCHER UNE PROCÉDURE DE CONTESTATION ACTIVE	GETRSPEC, GETRBZW	2	3
		dépend des stratégies « métiers »			« risque » données	« risque » programmes

Figure 12.13. *Correspondances exigences-traitements*

Index

Impression & reliure sepec - France
Numéro d'impression : 01703120890 - Dépôt légal : septembre 2012

IMPRIM'VERT®

PEFC 10-31-1470 / Certifié PEFC / Ce produit est issu de forêts gérées durablement et de sources contrôlées. / pefc-france.org

La bonne qualité des données est aujourd'hui la clé de voûte de toute organisation. La gestion et l'amélioration de cette qualité sont des tâches coûteuses et difficiles, mais néanmoins incontournables.

Cet ouvrage propose une étude des différents outils et démarches qui assistent les spécialistes de la qualité et de la gouvernance des données. A travers les expériences de la communauté francophone animée par l'association ExQI (Excellence Qualité, Information), il présente, avec pédagogie et pragmatisme, un panorama des concepts-clés de la gestion de la qualité des données et leurs déclinaisons dans les entreprises (*Business Intelligence*, *Data Quality Management*, *Key Performance Indicator*, *Model Driven Engineering*, *Master Data Management*, etc.). Des solutions théoriques et techniques performantes sont détaillées et de nombreux retours d'expérience permettent d'illustrer les bonnes pratiques à adopter.

Mêlant contributions industrielles et académiques, cet ouvrage est un outil de référence en langue française sur la qualité et la gouvernance des données en entreprise.

La coordinatrice

Laure Berti-Equille est directrice de recherche en informatique à l'Institut de Recherche pour le Développement. Depuis plus de quinze ans, elle s'est spécialisée dans l'évaluation de la qualité des données en proposant des stratégies de détection et de correction des anomalies dans les bases de données.

hermes science publications
www.hermes-science.com

ExQi

978-2-7462-2510-7
9 782746 225107