Université Lumière Lyon 2
Ecole Doctorale Informatique et Mathématiques

Thèse pour obtenir le grade de
Docteur en Informatique

Présentée par

Soumaya Ben Hassine - Guetari

Évaluation et requêtage de données multisources : une approche guidée par la préférence et la qualité des données

Application aux campagnes marketing B2B dans les bases de données de prospection

Préparée conjointement au sein du Laboratoire ERIC – Université Lumière Lyon 2 et de la société A.I.D.

Sous la direction de

Pr. Jérôme Darmont Pr. Jean-Hugues Chauchat

Soutenue publiquement le 10 Octobre 2014 devant le jury

Mme. Laure BERTI-EQUILLE	Dr, IRD	(Rapporteur)
Mr. Pascal PONCELET	Pr, Université Montpellier2	(Rapporteur)
Mr. Jacky AKOKA	Pr, CNAM Paris	(Examinateur)
Mme. Zoubida KEDAD	MCF HDR, Université de Versailles Saint Quentin en Yvelines	(Examinateur)
Mme. Sylvie SERVIGNE	MCF, INSA Lyon	(Examinateur)
Mme. Virginie THION	MCF, Université Rennes 1	(Examinateur)
Mr. Olivier Coppet	Administrateur, GDE France	(Invité)
Mme. Brigitte LABOISSE	Directrice générale, GDE France	(Invité)
Mr. Jean-Hugues CHAUCHAT	Pr émérite, Université Lyon2	(Directeur)
Mr. Jérôme DARMONT	Pr, Université Lyon2	(Directeur)

Résumé

Avec l'avènement du traitement distribué et l'utilisation accrue des services web inter et intra organisationnels alimentée par la disponibilité des connexions réseaux à faibles coûts, les données multisources partagées ont de plus en plus envahi les systèmes d'informations. Ceci a induit, dans un premier temps, le changement de leurs architectures du centralisé au distribué en passant par le coopératif et le fédéré ; et dans un deuxième temps, une panoplie de problèmes d'exploitation allant du traitement des incohérences des données doubles à la synchronisation des données distribuées. C'est le cas des bases de prospection marketing où les données sont enrichies par des fichiers provenant de différents fournisseurs.

Nous nous intéressons au cadre particulier de construction de fichiers de prospection pour la réalisation de campagnes marketing B-to-B, tâche traitée manuellement par les experts métier. Nous visons alors à modéliser le raisonnement de brokers humains, afin d'optimiser et d'automatiser la sélection du « plan fichier » à partir d'un ensemble de données d'enrichissement multisources. L'optimisation en question s'exprimera en termes de gain (coût, qualité) des données sélectionnées, le coût se limitant à l'unique considération du prix d'utilisation de ces données.

Ce mémoire présente une triple contribution quant à la gestion des bases de données multisources. La première contribution concerne l'évaluation rigoureuse de la qualité des données multisources. La deuxième contribution porte sur la modélisation et l'agrégation préférentielle des critères d'évaluation qualité par l'intégrale de Choquet. La troisième contribution concerne BrokerACO, un prototype d'automatisation et d'optimisation du brokering multisources basé sur l'algorithme heuristique d'optimisation par les colonies de fourmis (ACO) et dont la Pareto-optimalité de la solution est assurée par l'utilisation de la fonction d'agrégation des préférences des utilisateurs définie dans la deuxième contribution. L'efficacité du prototype est montrée par l'analyse de campagnes marketing tests effectuées sur des données réelles de prospection.

Mots-Clefs : qualité des données, évaluation de la qualité, bases de données multisources, estimation de la préférence, intégrale de Choquet, optimisation de la sélection, algorithme de colonies de fourmis, algorithme évolutionnaires.

Abstract

In Business-to-Business (B-to-B) marketing campaigns, manufacturing “the highest volume of sales at the lowest cost” and achieving the best return on investment (ROI) score is a significant challenge. ROI performance depends on a set of subjective and objective factors such as dialogue strategy, invested budget, marketing technology and organisation, and above all data and, particularly, data quality. However, data issues in marketing databases are overwhelming, leading to insufficient target knowledge that handicaps B-to-B salespersons when interacting with prospects. B-to-B prospection data is indeed mainly structured through a set of independent, heterogeneous, separate and sometimes overlapping files that form a messy multisource prospect selection environment. Data quality thus appears as a crucial issue when dealing with prospection databases. Moreover, beyond data quality, the ROI metric mainly depends on campaigns costs. Given the vagueness of (direct and indirect) cost definition, we limit our focus to price considerations.

Price and quality thus define the fundamental constraints data marketers consider when designing a marketing campaign file, as they typically look for the "best-qualified selection at the lowest price". However, this goal is not always reachable and compromises often have to be defined. Compromise must first be modelled and formalized, and then deployed for multisource selection issues. In this thesis, we propose a preference-driven selection approach for multisource environments that aims at: 1) modelling and quantifying decision makers’ preferences, and 2) defining and optimizing a selection routine based on these preferences. Concretely, we first deal with the data marketer’s quality preference modelling by appraising multisource data using robust evaluation criteria (quality dimensions) that are rigorously summarized into a global quality score. Based on this global quality score and data price, we exploit in a second step a preference-based selection algorithm to return "the best qualified records bearing the lowest possible price". An optimisation algorithm, BrokerACO, is finally run to generate the best selection result.

Keywords: data quality, quality assessment, multisource databases, user preferences, Choquet integral, selection optimization, ant colony optimization, evolutionary algorithms

Table des matières

Table des figures

Liste des tableaux

Chapitre 1 : Introduction

« Tout le monde savait que c'était impossible à faire. Puis un jour quelqu'un est arrivé qui ne le savait pas, et il l'a fait. »
Winston Churchill

1. Contexte métier

Henri Dujardin, artisan dans le Puy de Dôme, est-il un des futurs clients de l'entreprise MaisonPhoning ? Henri ne le sait pas, mais son profil fait l'objet d'une réflexion marketing dans la salle de réunion de MaisonPhoning où analystes et commerciaux sont en train d'étudier la composition du plan de prospection de la prochaine campagne marketing de l'entreprise. En effet, MaisonPhoning se voit être en pleine phase de recrutement de nouvelle clientèle. Publicité et marketing direct sont alors ses principaux leviers d'action. Ainsi, si MaisonPhoning choisit le marketing direct, elle va d'abord analyser l'ensemble de ses clients, voire meilleurs clients, puis chercher les prospects de profils similaires.

Oui, mais où trouver ces cibles ? MaisonPhoning, comme la majorité des annonceurs Français[1], enrichit sa base de prospection moyennant l'achat ou la location de fichiers tiers proposés par des « courtiers d'adresses » (encore appelés *listbrokers* ou *brokers*) [2]. MaisonPhoning utilise, alors, les services de Marlène, courtier d'adresses de son métier, et lui expose sa cible (les artisans du département du Puy de Dôme) et le canal de communication (email, par exemple). Marlène contacte, alors, ses différents fournisseurs, calcule le nombre de prospects chez chacun ainsi que le taux de recouvrement des différents fichiers (contacts communs), puis propose « un plan fichier », à savoir une liste des prospects avec l'ensemble des sources qui les fournissent ainsi que le prix de leur utilisation (achat ou location). C'est là qu'intervient l'expertise du broker : Marlène « connaît » en effet les fichiers qui « marchent », qui génèrent des clients (appelés *leads*[3]) et est capable, de ce fait, de décider des sources de fichiers appropriées au contexte de la campagne, à savoir la cible et l'offre à prioriser.

---

[1] Annonceur : entreprise à l'origine d'une opération de communication publicitaire ou marketing qui vise à promouvoir ses produits ou sa marque.

[2] Courtier d'adresses : aussi appelée broker / list broker, l[es courtiers d'adresses sont des acteurs dont le but est de conseiller les annonceurs dans le domaine des campagnes de marketing direct. Ils assurent le conseil dans le domaine de la recherche et de la sélection de fichiers de marketing direct en établissant des plans fichiers [1].

[3] Lead : contact commercial obtenu à partir d'une campagne marketing. Internet est un canal très efficace de collecte de leads qualifiés, car il facilite la réactivité et le recueil de l'information à travers des formulaires [1].

C’est dans l’optique d’une automatisation de cette sélection que se situe notre prestation pour MaisonPhoning. Nous nous intéressons, en effet, à modéliser le raisonnement du broker humain, Marlène en l’occurrence, afin d’optimiser la sélection du « plan fichier » étant donné un ensemble de données d’enrichissement multisources. L’optimisation en question s’exprimera en termes de bénéfices coûts et qualité des données sélectionnées, le coût se limitant à l’unique considération du prix d’utilisation de ces données.

Dans cette thèse, nous abordons la problématique de la prospection marketing de deux points de vue :

1 un point de vue fonctionnel arborant la prospection multi-fournisseurs, les politiques de prix dégressives qui s’y appliquent ainsi que les contraintes de sélection minimale par fournisseur qui la conditionnent;
2 un point de vue technique s’intéressant à la gestion de données multisources et traitant, dans notre exemple, des problématiques de modélisation des préférences des brokers et celles de l’optimisation de la sélection dans un système d’information multisources.

Les sections suivantes détaillent ces deux aspects. Nous précisons, à chaque fois, les différentes contraintes que nous nous proposons de prendre en compte dans notre méthodologie de résolution de la sélection des données dans les bases de données multisources.

2. La prospection marketing de point de vue fonctionnel

Cette section a pour but de poser les piliers fonctionnels du marketing en général et de la prospection marketing en particulier. Nous décrivons, entre autres, l’objectif des opérations marketing, la composition des bases de données marketing et le déroulement d’une campagne de prospection classique. Ces notions et concepts sont, en effet, nécessaires pour cerner les contraintes techniques que nous nous proposons de traiter tout au long de ce manuscrit.

2.1. L’histoire du marketing

Remontons tout d’abord aux origines du terme « marketing » et définissons ses principaux fondements. Après la seconde guerre mondiale, une économie de production s’est mise en place : la demande était très largement supérieure à l’offre et tout ce qui se fabriquait s’achetait. L’urgence était de construire un appareil de production capable de répondre à des besoins en

croissance exponentielle. Le consommateur à séduire était alors une figure inconnue. Les années 60, définirent l'ère d'une économie de distribution où l'offre équilibre la demande et où s'instaure un début de diversification des produits et services proposés. C'est alors qu'apparut le *marketing*, avec ses stratégies de « push » et de « pull » (qui se définissent respectivement à amener les produits vers les consommateurs et inversement), comme une évolution des techniques de vente substituant la satisfaction des besoins des consommateurs aux vieilles techniques agressives de marchandage. Ensuite, au cours des années 70 et 80, le marketing fut renforcé avec l'apparition de la publicité et de la stratégie publicitaire qui se définissait à l'époque comme « toute forme de communication faite dans le cadre d'une activité commerciale, industrielle, artisanale ou libérale dans le but de promouvoir la fourniture de biens ou services, y compris les biens immeubles, les droits et les obligations » (directive 84/450/CEE, Septembre 1984) [2]. Progressivement, on assistait à une évolution dans les comportements du consommateur qui, depuis le début des années 90, s'est vu guidé par la tendance du *burrowing* (de l'anglais *burrow* qui signifie terrier) et du *cocooning* (faisant référence au cocon) l'encourageant, entre autres, à se faire livrer son repas à domicile, effectuer ses virements bancaires en ligne et se balader dans un supermarché virtuel pour y faire ses courses sans bouger de son canapé [Boisdevésy 96]. Ce changement dans l'attitude de consommation, alimenté par l'avènement du marketing relationnel [Berry al. 83] parmi les universitaires puis du « CRM » (*Customer Relationship Marketing*) chez les praticiens, stimula dans les années 1990 l'intérêt des chercheurs pour le marketing direct ou, plus récemment, le marketing interactif, d'autant plus que de nouvelles demandes avaient émergé, notamment en termes de services. Le marketing direct, ainsi historiquement « parent pauvre de la publicité en termes de recherche et d'investissements, a récemment gagné ses lettres de noblesse avec l'avènement du marketing relationnel et connaît un essor sans précédent par l'explosion des canaux de communication directs » notamment via Internet [Micheaux 07]. Son défi majeur concerne l'hyperpersonnalisation de la relation client tout en garantissant la même efficacité artisanale qu'autrefois. C'est alors que nous commencions à parler de marketing des bases de données, de marketing de l'animation et de marketing d'information, autant d'outils permettant de nourrir les bases de données en générant du contact, de la relation et de la recommandation de qualité sur un mode industriel [Boisdevésy 96].

Le processus de ciblage d'Henri Dujardin se place dans un contexte de marketing direct : Notre client, la société MaisonPhoning, a, en effet, défini une cible de prospection et souhaite instaurer une relation durable, à distance avec chacun de ses prospects. Plus particulièrement, comme MaisonPhoning s'adresse, dans ses campagnes de prospection, aux prospects entreprises, la

stratégie marketing utilisée est dans ce cas une stratégie de marketing direct B-to-B (B2B ou *Business-to-Business*)[4].

2.2. Composition d'une base de prospection marketing B-to-B

2.2.1. Le répertoire SIRENE : le socle de toute base de prospection

Généralement, une base de prospection B-to-B rassemble l'ensemble des entreprises prospects et leurs décideurs. En France, les annonceurs sont aidés par l'INSEE (Institut national de la statistique et des études économiques) [4]. L'INSEE collecte, produit, analyse et diffuse des informations sur l'économie et la société françaises. Elle met à disposition du public le répertoire SIRENE (Système Informatisé du Répertoire National des Entreprises et des Etablissements) qui recense l'ensemble des entreprises et établissements déclarés en France. Si l'information est exhaustive en termes de noms et d'adresses physiques (sauf certains périmètres tels que les associations), elle est très partielle pour les téléphones et inexistante pour les emails et les contacts nominatifs (noms des décideurs). Il est donc nécessaire d'enrichir ce socle de base. Cependant, un problème se pose quant à la mise en place d'une telle stratégie : les fichiers de prospection dont dispose généralement l'annonceur ne sont pas aussi « qualifiés » que les fichiers clients.

2.2.2. Les données d'enrichissement

Une solution consiste à enrichir cette base de prospection par des informations récoltées sur les prospects B-to-B (aussi bien les entreprises que leurs dirigeants, gérants ou décideurs) proposées par des fournisseurs spécialisés dans la vente ou la location de ce genre d'informations. Dans une autre solution, on s'appuie sur les réseaux sociaux professionnels (de type LinkedIn ou Viadeo) ou grand public avec possibilité d'usage professionnel (de type Facebook ou Twitter). Cette dernière approche, si elle a le mérite de « s'auto-alimenter » par

[4] B-to-B : Terme anglais désignant l'échange de biens ou services entre deux entités commerciales. Dans le monde du commerce électronique et d'internet, Business-to-Business (B2B ou B-to-B) est le nom donné à l'ensemble d'architectures techniques et logicielles informatiques permettant de mettre en relation des entreprises, dans un cadre de relations clients/fournisseurs. Ceci se fait notamment au travers de l'intranet de l'entreprise étendue, ou extranet. L'intranet de l'entreprise est connecté aux intranets des fournisseurs, des sous-traitants, des clients, des distributeurs et des partenaires. [3]

les mises à jour volontaires des membres, comporte à ce jour, de notre point de vue, une faille. En effet, outre les restrictions imposées par la CNIL[5] concernant l'utilisation des données des réseaux sociaux étant considérées des données à caractère personnel, de récentes études montrent l'écart important existant entre le taux d'abonnement de la Catégorie SocioProfessionnelle (catégorie CSP+) ci-ciblée par MaisonPhoning comparé à celui des cadres. En effet, une étude réalisée par A.I.D. en 2012[6] montre que le taux d'abonnement de la catégorie CSP+[7] aux réseaux sociaux est seulement de 4%. Ce taux est d'autant plus faible pour la sous-catégorie des « artisans » (dont Henri Dujardin fait partie), biaisée de toute évidence par la composante statistique « chef d'entreprises ». En l'occurrence, Henri Dujardin n'est pas abonné à LinkedIn ni à aucun réseau social ce qui justifie l'option d'enrichissement par fichiers de fournisseurs spécialisés choisie par MaisonPhoning.

Détaillons le détail de ces fichiers. D'une manière générale, l'enrichissement des bases de prospection se fait sur les attributs suivants :

- le téléphone, par un rapprochement avec un annuaire ou par questionnaire sur site Web. En France, il n'existe pas à ce jour d'annuaire exhaustif. Deux sources principales existent :
  - les Pages Jaunes ou Pages Blanches. Cet annuaire liste l'ensemble des abonnés ligne fixe hormis les personnes en liste rouge, ayant demandé explicitement à ne pas apparaitre,
  - l'annuaire universel : sur la base d'un accord explicite du titulaire, cet annuaire comporte les titulaires de lignes fixes ou mobiles, tout opérateur confondu ;
- les contacts et leurs adresses emails : les noms figurant dans les statuts des sociétés (direction générale, conseil de surveillance,..) font l'objet de déclarations aux CFE (Centres de Formalités des Entreprises) et sont souvent « vendus » par des fournisseurs de données qui captent l'information. Si pour les petites entreprises ces contacts sont souvent les décideurs exclusifs, dans les moyennes et grandes entreprises, ils ne représentent que la face visible de l'iceberg de décision ; d'où un marché pour les agences de marketing relationnel B-to-B qui qualifient ces décideurs par enquêtes ou encore celles souhaitant « rentabiliser » leur propre base de données client par la mise à

[5] CNIL : Commission Nationale de l'Informatique et des Libertés.

[6] Enquête réalisée par A.I.D. par Internet du 28 juin au 16 juillet 2012 sur une population de 10934 internautes Français redressés sur les critères démographiques suivants : sexe, âge et CSP.

[7] La catégorie CSP+ décrit à la fois les artisans, commerçants et chefs d'entreprises.

disposition en vente ou en location de contacts chez leurs clients. De telles entreprises sont les fournisseurs de notre courtière Marlène, des sociétés de vente par correspondance d'articles de bureaux, des sociétés organisatrices de séminaires et expositions ou encore des entreprises de ventes de matériel médical aux professionnels de la santé.

De toutes les données de qualification des prospects, l'adresse email est considérée comme la donnée la plus critique à obtenir car non captée dans les formulaires des CFE d'autant plus qu'elle est protégée par la loi LCEN (Loi pour la Confiance dans l'Economie Numérique) et nécessite une autorisation de la personne pour diffusion de la donnée à des partenaires (ce qu'on appelle en jargon métier *opt-in*). Cette donnée, très coûteuse lorsqu'elle est vendue par les fournisseurs de fichiers, est en général louée.

2.3. Déroulement classique d'une campagne de prospection

Une méthode classique consiste à analyser les fichiers client (appelés base client) afin d'établir les différents profils et segments des clients et de comparer les différences comportementales entre ces différentes classes. L'analyse comportementale permet ainsi de préciser le format (publicitaire ou non) du message de la campagne selon la segmentation ciblée.

Reprenons l'exemple de MaisonPhoning et détaillons le déroulement d'une de ses campagnes de prospection. Tout commence par une décision du département marketing concernant la commercialisation d'une offre. Soit l'offre suivante : une tablette avec une connexion 4G pour un prix de 200 euros. Une demande est alors faite pour un ciblage des architectes là où le réseau 4G est disponible, soit le département Ile de France. Les responsables marketing précisent, par ailleurs, le canal de ciblage. Il s'agit en l'occurrence d'une communication par email.

La cible marketing ainsi définie, l'équipe informatique sélectionne, à partir de sa base de prospection reposant sur le répertoire SIRENE de l'INSEE, les SIRETs[8] ciblés. C'est alors que MaisonPhoning demande à son broker, Marlène en l'occurrence, d'enrichir cette liste de SIRETs par les contacts dirigeants et leurs emails correspondants. Marlène va alors contacter l'ensemble de ses fournisseurs et sélectionner, parmi les sources les mieux réputées, les contacts et leurs emails respectifs. Le recours à ces courtiers pour le ciblage et la qualification des données de campagne est, cependant, sujet à une subjectivité assez restrictive. En effet, les

[8] SIRET : identifiant unique des établissements Français, défini par l'INSEE.

informations de qualification retenues lors de la sélection dépendent, d'une part, de la réputation de leurs fournisseurs (réputation acquise par les courtiers grâce à leur expérience) et, d'autre part, des taux de commission perçus par ces courtiers en fonction de la quantité de l'information achetée ou louée.

2.4. Les problèmes qualité des bases de prospection

2.4.1. La qualité, un facteur rarement considéré

Le nombre de fichiers de prospection B-to-B en France est passé de 500 en 2008 [5], à plusieurs milliers de fichiers [6] « représentant un potentiel de plusieurs millions d'adresses ». Cette diversité au niveau de la provenance des données de prospection induit, en dépit de la richesse et la complétude de la qualification qu'elle propose, une hétérogénéité dans la base de prospection ainsi formée ; une hétérogénéité qui se traduit, dans la plupart des cas, par des problèmes considérables de cohérence et, plus généralement, de qualité. Par ailleurs, la plupart des données marketing est plutôt proposée en location impliquant, outre les coûts d'acquisition, leur volatilité et, par conséquent, leur indisponibilité. Ainsi, trois facteurs principaux sont à prendre en compte lors de la réalisation d'une opération de ciblage des données marketing de prospection : la qualité, la volatilité et le coût.

Malheureusement, en 2009, lorsque nous commencions à étudier la qualité des bases de prospection, ces facteurs avaient rarement été considérés lors de la mise en place des campagnes. En effet, selon QAS [7] (éditeur de logiciels pour la gestion des adresses postales), la motivation des professionnels à améliorer la qualité de leurs données est moindre en B-to-B qu'en B-to-C (Business to Client) [8]. Aussi, une brève analyse des fichiers d'entreprise montre que :

- il n'est pas fourni d'indication sur la date de mise à jour des enregistrements, sur les enregistrements concernés par cette mise à jour, ni sur leurs provenances ;
- certains fournisseurs ont tendance à fournir des noms de contacts en double, voire en triple. Ainsi, un employé ayant deux fonctions différentes dans une entreprise figurera au minimum deux fois dans le fichier de prospection.

Pour résumer, l'état de l'art signale une exploitation de données imparfaite qui engendre des effets négatifs. Selon le magazine « Le journal du Net » en 2007, 67 % des entreprises emploient en moyenne trois méthodes de collecte des données et 62 % n'ont pas formalisé la récolte de données standards, au risque de se procurer des informations en double ou incomplètes (absence

d'adresse e-mail, mauvais libellé de l'adresse, contact obsolète…). En outre, 27 % des entreprises interrogées ne nettoient leurs données qu'une fois par an et 7 % ne l'ont jamais fait [8].

Ce constat montre la nécessité de l'intégration d'un processus d'évaluation puis d'amélioration de la qualité des fichiers de prospection afin d'assurer l'efficacité des campagnes marketing. Cela dit, l'évaluation qualité des fichiers de prospection n'est pas garante, à elle seule, de cette efficacité du retour sur investissement (ROI) des campagnes marketing. Une autre problématique taraude, en effet, l'esprit des marketeurs : le contrôle voire l'optimisation de la sélection de leur « plan fichier » en fonction des contraintes budgétaires. Un compromis entre la qualité et le coût d'exploitation devra alors être trouvé.

C'est dans ce contexte que se situe notre problématique métier. Nous nous proposons de développer un système d'aide à la prospection marketing ayant pour but d'améliorer cette stratégie de ciblage des prospects en lui ajoutant une dimension objective et contextuelle. Le ciblage que nous recommandons se basera, en effet, sur des critères évaluant la qualité de chaque donnée sélectionnée dans le fichier ainsi que leurs prix respectifs d'achat ou de location. Ce processus de sélection multisources est connu dans le contexte marketing sous le nom de *brokering* (courtage).

3. Contexte scientifique de la gestion des données multisources

Du contexte métier de prospection dans les bases volumineuses de marketing se distingue la problématique scientifique de la gestion des données multisources, qui impose des stratégies de résolution dépendant à la fois du contexte et des exigences des professionnels du métier. Ainsi, des architectures centralisées aux architectures distribuées en passant par les architectures coopératives et fédérées, une panoplie de solutions a été développée dans la littérature pour définir la configuration d'intégration adéquate selon le contexte métier. Nous distinguons, entre autres, les architectures physiques d'intégration utilisant les techniques de fusion et les architectures distribuées logiques de type LAV (*Local-As-View*) et GAV (*Global-As-View*). Dans le contexte particulier des campagnes marketing, la base de ciblage diffère en fonction de la campagne et les données de ciblage varient en fonction des populations (ou segments de populations) et produits désignés par les *marketers*. C'est pour cette raison que nous préconisons une stratégie d'intégration logique de type LAV. En effet, au sein de cette stratégie,

les différentes sources de données restent autonomes et l'intégration se fait à la volée lors de la génération de réponses aux requêtes des utilisateurs.

La gestion à la volée des données multisources, bien que garantissant l'autonomie des sources au sein d'une architecture distribuée, est contrainte par les situations d'incohérences causées par les données hétérogènes. Ces situations d'incohérence sont généralement gérées moyennant des règles métier et/ou des contraintes d'exactitude et de cohérence toutes exprimées lors de la définition de la requête de l'utilisateur. Elles ont pour but de générer la sélection la plus adéquate aux attentes de celui-ci.

Idéalement, le meilleur moyen d'atteindre cet objectif de satisfaction de l'utilisateur serait de l'impliquer dans la définition des contraintes de sélection. Dans ce contexte, des techniques de modélisation et de gestion des préférences ont été proposées par la littérature. De nouveaux langages de type SQLf [Bosc et al. 95] et PreferenceSQL [Kießling et al. 02], calqués sur le langage SQL, en ont été proposés pour inclure les préférences des utilisateurs dans les requêtes de sélection. D'autres approches plus interopérables se sont plutôt focalisées sur la modélisation et l'expression des préférences comme critère de sélection dans les requêtes SQL classiques. Ces préférences sont alors exprimées soit au niveau des données par des métadonnées qualifiant les attributs, soit au niveau du schéma global pour exprimer les préférences sur les associations d'attributs des différentes sources. Le souci avec ces variantes de représentations de la préférence se manifeste en cas de bipolarité, lorsque le moteur de requêtage se trouve contraint de choisir entre des doublons de données de critères de satisfaction incomparables. Dans la plupart des cas, ce genre de conflits est résolu manuellement par l'expert métier. Ainsi, soit il définit subjectivement des poids dénotant l'importance relative des différents critères de préférence en question, auquel cas le choix devient non rigoureux, soit il choisit la meilleure alternative parmi les alternatives sujettes à conflits à chaque fois qu'un cas conflictuel se présente, et dans ce cas, la résolution des conflits devient fastidieuse pour l'expert.

4. Problématiques de recherche

Notre problématique se déclinant dans le contexte industriel, l'approche de brokering que nous nous proposons de soutenir tout au long de ce mémoire se fixe deux objectifs principaux: la quantification rigoureuse des préférences des utilisateurs et l'utilisation de ces préférences pour l'optimisation de la sélection des données dans un environnement multi-fournisseurs ; le but étant d'améliorer le ciblage des prospects dans le contexte d'une campagne marketing B-to-B.

Dans un premier temps, les préférences de l'utilisateur sont exprimées pour évaluer les données de prospection multisources. Elles dénotent alors les importances et les synergies existant entre les différents critères d'évaluation qualité lesquelles sont utilisées pour calculer un score qualité global. Dans un deuxième temps, une autre série de préférences est utilisée pour la sélection optimisée des données de prospection. Elle exprime alors le compromis qualité/prix Pareto-optimal[9] que doit refléter le plan fichier.

L'approche d'évaluation de la qualité ainsi proposée analyse les dimensions et métriques qualité qualifiant les doublons conflictuels afin de gérer les incohérences causées par l'intégration des données multisources. En effet, chacune de ces données est décrite par un ensemble de critères qualité définis à la fois par des utilisateurs métier et des experts techniques. Cette double définition des attributs qualité est nécessaire pour garantir la pertinence de l'évaluation et sa contextualité. Cette approche devient cependant immédiatement problématique dès qu'il s'agit de confronter des données de critères incomparables ou difficilement comparables. Une solution consiste à agréger ces métriques en un score unique fédérateur et global.

Les métriques qualité étant des variables dépendantes et non commensurables, la fonction d'agrégation que nous proposerons doit tenir compte de ces deux contraintes. En effet, la dépendance entre les dimensions qualité a été prouvée par différentes études qui ont montré l'existence d'une synergie entre les dimensions qualité [Helfert et al. 09]. Ces dépendances peuvent être négatives (au cas où l'amélioration d'une dimension entraîne la dégradation d'une autre dimension) ou positives (au cas où l'amélioration d'une dimension engendre l'amélioration d'une autre dimension). De plus, la non-commensurabilité des scores assure la non-compensation des métriques entre elles.

Outre ces contraintes de calcul, la fonction d'agrégation doit intégrer une contrainte subjective cruciale : la préférence d'agrégation de l'expert humain, de manière à modéliser et quantifier une fonction de compromis Pareto-optimal. En effet, dans la pratique, en fonction du type de campagne, une métrique qualité voit son importance changer, impliquant un changement du score qualité. Prenons l'exemple de la donnée *numéro de téléphone* qualifiée par trois métriques : l'exactitude syntaxique, la fraîcheur (dénotant la récence de l'enregistrement) et la valeur ajoutée (calculée par rapport à la nature du numéro). Dans le cadre d'une campagne téléphonique, la métrique fraîcheur est plus importante que les deux autres métriques. En effet, le responsable marketing estime qu'un numéro fixe récent est plus fiable qu'un numéro mobile

[9] Une solution est dite "optimum de Pareto" si elle est possible et s'il n'existe aucune autre solution qui lui soit préférée au sens de Pareto. Une solution étant un ensemble d'individu, un état optimal au sens Pareto est un état où on ne peut améliorer la situation d'un individu sans détériorer celle d'un autre.

mais plus ancien. Par contre, dans le cas d'une campagne emailing ou courrier, le responsable marketing estime que les trois métriques se valent (en terme d'importance) pour évaluer la fiabilité de la donnée téléphonique.

La sélection multisources (brokering) optimisée se basera aussi sur l'agrégation préférentielle où un compromis entre la qualité globale des données et leur coût (qui est réduit à leur prix d'achat dans ce projet) doit être estimé. La phase de brokering est alors assimilée à une problématique décisionnelle d'optimisation multiobjectifs où la décision se définit par le choix des données à intégrer à la sélection. Ainsi, l'approche d'optimisation préconisée se base sur la fonction de compromis que nous définissons pour déterminer la sélection la plus conforme aux attentes de l'utilisateur.

5. Contributions et organisation du mémoire

Etant données les problématiques fonctionnelles et techniques décrites précédemment, nous présentons dans ce mémoire une triple contribution quant à la gestion des bases de données multisources.

1 La première contribution concerne l'évaluation rigoureuse de la qualité des données multisources. L'approche utilisée est inspirée du dénominateur commun des différentes méthodologies d'évaluation qualité. Elle suppose, entre autres, l'analyse des données multisources, l'analyse des besoins en qualité de données étant donné le contexte de réalisation de campagnes marketing, la définition des dimensions adéquates, le coût de la non-qualité des données ainsi que les bénéfices perçus suite à l'instauration d'un tel processus d'évaluation qualité.

2 La deuxième contribution porte sur la modélisation et l'agrégation préférentielle des critères d'évaluation qualité. Les préférences d'agrégation des utilisateurs sont définies selon l'attribut et le contexte de la sélection (la nature de la campagne marketing), puis exprimées sous forme de relations de pré-ordre entre les doublons conflictuels. Elles seront ensuite estimées par une fonction de scoring floue : l'intégrale de Choquet, qui, en plus de quantifier ces relations de pré-ordre par des poids sur les critères et groupes de critères, permet d'intégrer les synergies entre les attributs.

3 La troisième contribution concerne l'automatisation du brokering multisources et son optimisation pour que la sélection réalisée corresponde aux attentes du décideur marketing. Nous utilisons l'algorithme heuristique d'optimisation par les colonies de fourmis (ACO)

et la Pareto-optimalité de la solution est assurée par l'utilisation de la fonction d'agrégation des préférences des utilisateurs définie dans la deuxième contribution.

Par conséquent, le plan du mémoire s'énonce comme suit. Le deuxième chapitre décrit l'état de l'art de la gestion des données multisources et des approches de brokering déjà utilisées dans la littérature. Le troisième chapitre détaille notre proposition d'évaluation qualité où nous décrivons les principales dimensions et métriques définies à la fois par les experts métiers et techniques, critères que nous nous proposons d'agréger selon les préférences de l'utilisateur final (responsable marketing, en l'occurrence). Le quatrième chapitre développe le principe du brokering automatique optimisé et propose une estimation du compromis qualité/prix en fonction des préférences du décideur. Le cinquième chapitre expose le déploiement de notre approche de brokering, approche guidée par la préférence de l'utilisateur final dans le cadre de réalisation d'une campagne marketing B-to-B. Les expériences en question portent sur la conception de plan fichier optimisé pour la réalisation de campagnes emailing au sein d'une entreprise de télécommunication française de type MaisonPhoning dont les fichiers de qualifications de l'email forment notre environnement de prospection multisources. Enfin, nous dressons, dans le denier chapitre, le bilan de ce projet et discutons de ces perspectives scientifiques et fonctionnelles.

Chapitre 2 : De l'intégration de données multisources

« Lois d'existence scolaire - Elève: J'existe car je suis évalué. - Enseignant: J'existe car j'évalue. - Directeur d'école: J'existe car j'ordonne d'évaluer. - Ministère de l'éducation: Rien n'existe hormis l'évaluation. »
Abbé Ernest

Nous dressons dans ce chapitre un état de l'art des différentes approches de gestion des données multisources proposées dans la littérature. Nous définirons les grandes approches d'intégration de données multisources en distinguant les techniques de l'intégration logique et celles de l'intégration physique. Nous montrons que ces approches se reposent, pour la plupart, sur la qualité des données qui est utilisée comme critère influant d'intégration. Nous décrivons alors les méthodologies d'évaluation de la qualité des données en détaillant celles qui sont les plus répandues dans la littérature. Nous mettons en valeur notamment le rôle de la qualité des données dans la résolution des situations conflictuelles impliquées par la présence de doublons dans les données multisources.

Par ailleurs, notre objectif étant la sélection optimisée du fichier de ciblage parmi les données multisources de la base de prospection, nous décrivons les approches de brokering proposées par la littérature et insistons sur le rôle de la qualité des données et l'importance de la prise en compte des préférences des utilisateurs dans l'efficacité de cette tâche de sélection à la volée de données multisources.

1. Gestion des données multisources

Avec l'avènement du traitement distribué et l'utilisation abondante des services Web inter et intra organisationnels, les données multisources partagées sont de plus en plus présentes dans les systèmes d'informations, modifiant leurs architectures vers des structures délocalisées, plus décentralisées. C'est alors qu'apparurent les systèmes d'information coopératifs, pair-à-pair (de l'anglais *peer-to-peer*) ou encore fédérés, entrainant une panoplie de problèmes d'exploitation allant du traitement des incohérences des données doubles à la synchronisation des données distribuées, sans oublier 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 80 pour permettre le développement de mécanismes de négociation pour la gestion des incohérences au sein des systèmes coopératifs. Ainsi, une multitude de stratégies d'intégration et de fusion des données a été proposée ; leur objectif commun étant de trouver un mécanisme de communication fiable entre les sources de données afin de mieux interpréter les informations qui y circulent. Les solutions proposées varient 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 classons les architectures de gestion des données multisources en deux groupes distincts : celles qui visent une intégration logique préservant l'autonomie des sources locales moyennant des vues locales ou globales et celles qui visent 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 multisources [Bleiholder et al. 08] (Figure 2.1).

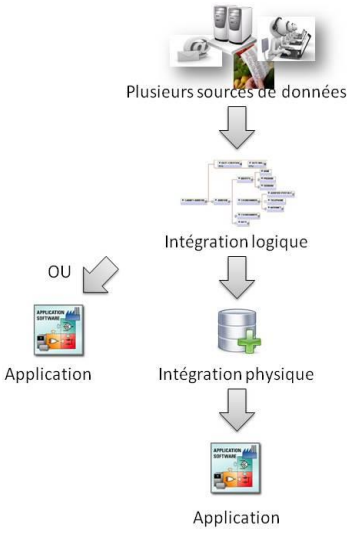


Figure 2.1 Processus général d'intégration des données [Bleiholder et al. 08]

Nous soulignons dans la suite le rôle de la qualité des données dans la mise en place de ces architectures d'intégration.

1.1. Approche logique

1.1.1 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 2.2 [Hacid et al. 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, ce modèle d'intégration théorique prend des désignations et des formes différentes. Par exemple, nous distinguons les bases de données fédérées, les systèmes coopératifs, les bases de données interopérables et les bases de données distribuées. Ces architectures distribuées diffèrent dans la façon avec laquelle les « intégrateurs » construisent leur schéma global d'interfaçage entre les sources et la couche applicative [Tari et al. 98].

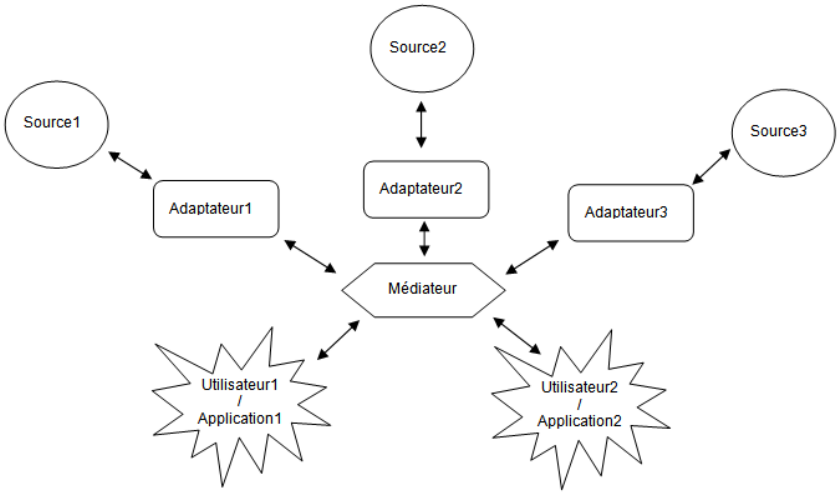


Figure 2.2 Schéma simplifié d'une architecture de médiation [Hacid et al. 04]

Dans ce qui suit, nous détaillons le principe des architectures distribuées les plus répandues dans la littérature tout en soulignant les différences qui existent entre elles.

i- Architecture fédérée

Une base de données fédérée est un ensemble de systèmes d'informations coopératifs autonomes et éventuellement hétérogènes. Elle se définit par trois caractéristiques principales : la répartition des données sur plusieurs bases de données (locales ou physiquement distribuées),

l'hétérogénéité (qui se manifeste soit au niveau du système d'information de chacune des entités de la base fédérée, soit au niveau des langages de requêtage, au niveau des modèles des données ou encore au niveau de la sémantique des données), et l'autonomie et l'indépendance des entités organisationnelles (à savoir l'autonomie au niveau de la modélisation, l'autonomie au niveau de la communication avec les autres entités de la base fédérée et l'autonomie au niveau de l'exécution des opérations ou transactions commandées par les utilisateurs locaux) [Sheth et al. 90].

Les composantes de cette base fédérée sont définies comme des sources structurées auxquelles on accède via des schémas d'export et des schémas de fédération (Figure 2.3). Nous distinguons alors trois types d'architectures : les systèmes d'information faiblement couplés, les bases de données fortement couplées et les systèmes d'information à base de médiateurs. L'analyse détaillée de ces différentes variantes est basée sur les travaux de [Busse et al. 99], qui comportent une étude détaillée des systèmes fédérés et de leurs différentes variantes.

Les bases de données faiblement couplées ne nécessitent pas un schéma fédéré, mais plutôt un langage de requêtage des bases de données multisources. L'avantage de cette architecture réside dans la conservation de l'indépendance et de l'autonomie des différentes bases locales. Par contre, 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. De plus, dans une architecture fédérée faiblement couplée, l'utilisateur est responsable de l'intégration des données et doit, de ce fait, gérer tous les problèmes de conflits que cette tâche implique.

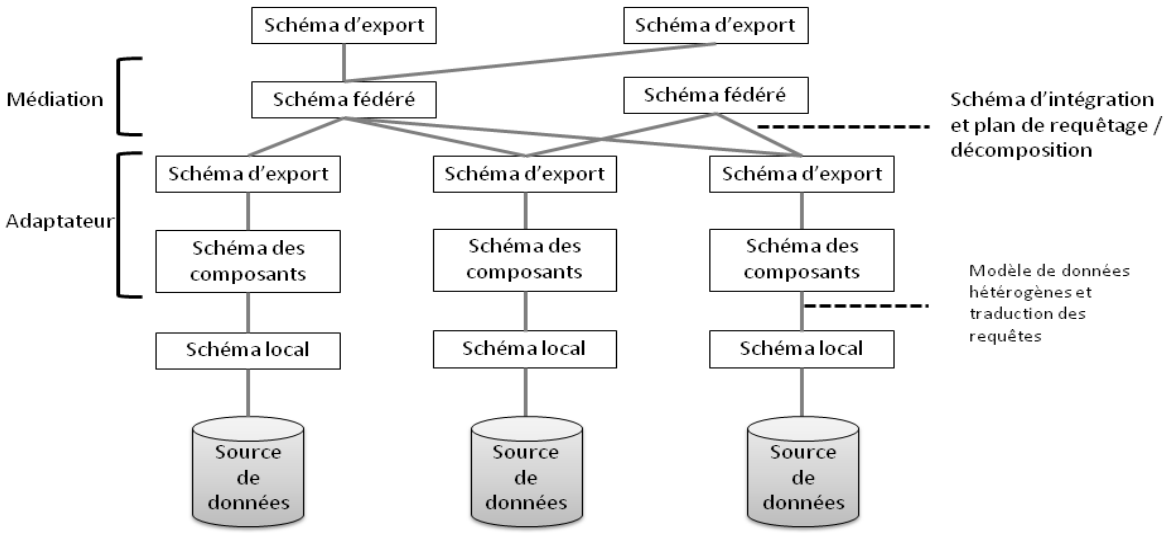


Figure 2.3 Architecture des bases de données fédérées [Busse et al. 99]

Les architectures fortement couplées offrent, quant à elles, 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 structure, ainsi que les contraintes transactionnelles relatives à la gestion des accès multiples concurrents.

Enfin, dans les systèmes d'information à base de médiateurs (Figure 2.4), introduits par [Wiederhold 93], un médiateur gère l'accès entre l'utilisateur et chaque source locale. Ils se distinguent des autres systèmes fédérés par la façon avec laquelle le schéma global est défini. En effet, il est conçu de manière descendante (*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 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'architectures, 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 classiques (car il ne nécessite pas un schéma global figé complet et minimaliste).

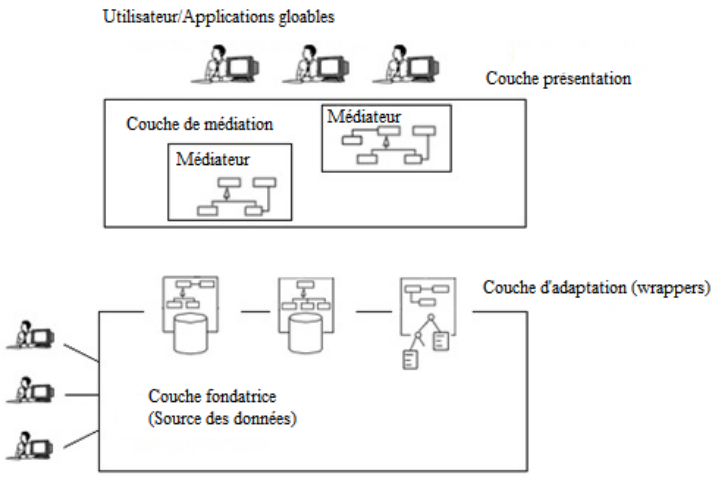


Figure 2.4 Architecture d'un système d'informations à base de médiateurs [Busse et al. 99]

ii- Systèmes coopératifs

Les systèmes coopératifs (encore appelés systèmes collaboratifs) se définissent comme des systèmes organisationnels différents, interconnectés, autonomes, géographiquement répartis, interopérables et partageant des objectifs communs [Mecella et al. 02]. Du point de vue technique, alors que les systèmes fédérés se gèrent par un langage de requête adapté, les systèmes coopératifs nécessitent des programmes de communication sophistiqués appelés agents informatiques, qui forment la couche coopérative. Ces systèmes partagent donc des services plutôt que des données [Mecella et al. 02, Tari et al. 98]. Les agents assurent une autonomie flexible au niveau des sources ainsi que la sécurité requise et les mécanismes appropriés pour la compréhension sémantique des processus et services offerts par les sources locales. Dans ce contexte, [Tari et al. 98] proposent une architecture à base d'agents pour la gestion de ce genre de systèmes. Une telle architecture est composée (Figure 2.5):

- 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 nécessaires à l'exécution des requêtes des utilisateurs,
- d'agents d'encapsulation qui se trouvent au niveau des sources locales et offrent un environnement coopératif de partage d'informations et de gestion de la sécurité des services avec différents niveaux d'autonomie.

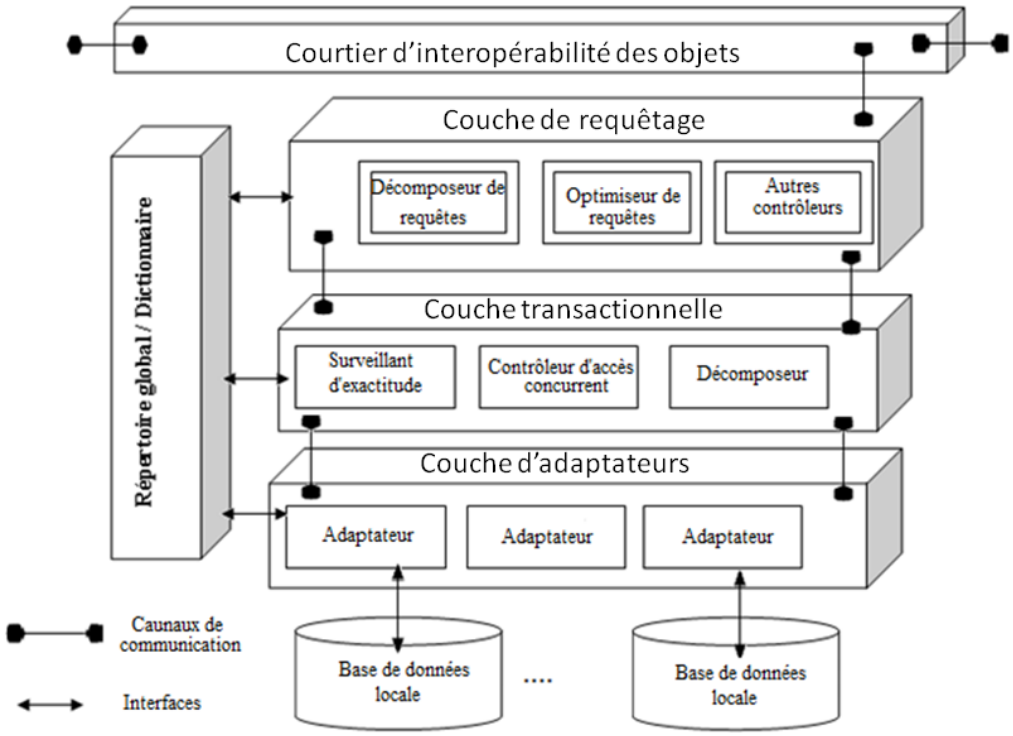


Figure 2.5 Architecture de coopération [Tari et al. 98]

iii- Architectures fédérées distribuées dynamiques

Outre les bases fédérées et les bases coopératives, de nouvelles formes de bases de données multisources, les bases de données fédérées distribuées et dynamiques, ont vu le jour [Bent et al. 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 ignorent leur localisation dans le réseau. Concrètement, cet accès *ad hoc* est réalisé grâce au mécanisme « stocke localement, interroge partout » (de l'anglais *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.

iv- Architectures pair à pair

Les bases de données pair à pair (*peer-to-peer*) sont caractérisées par la préservation de l'autonomie et de 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 [Calvanese et al. 05]. Les bases de données pair à pair sont, en effet, fondamentalement différentes des bases de données distribuées traditionnelles. Une première différence réside dans le fait que les nœuds composant le réseau peuvent s'y joindre et s'en déconnecter à tout moment. La base ainsi formée a une structure dynamique et flexible, d'autant plus que les nœuds la formant ne sont pas identifiés par avance (sauf dans le cas des réseaux sécurisés). De plus, les schémas de données sont différents d'un nœud à l'autre, ce qui affecte considérablement la fiabilité et la complétude des réponses aux requêtes des utilisateurs [Yong 04].

1.1.2. Techniques d'intégration logique

Nous avons défini le principe des architectures distribuées les plus répandues dans la littérature. Nous avons montré qu'elles diffèrent uniquement par la manière avec laquelle elles définissent leurs schémas d'intégration. Nous étudions à présent les techniques proposées pour la construction du schéma d'intégration. En effet, 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 (elle part du schéma global et descend vers les sources). Par exemple, l'attribut *argent* du schéma local se définit comme la somme des deux attributs *pièces* et *billets du schéma global*;
- 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, l'attribut *argent* du schéma global se définit comme la somme des deux attributs *pièces* et *billets* du schéma local.

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 possède 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. Par contre, 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 l'ajout d'une nouvelle source remet en question le schéma global du médiateur et nécessite sa mise à jour. Aussi, dans une architecture LAV, le schéma de médiation doit référencer l'ensemble des attributs de toutes les relations même si certains ne sont pas nécessaires pour l'intégration. De la même manière dans une architecture GAV, le schéma de médiation doit contenir toutes les relations ou ensemble de relations ce qui rend le schéma de médiation dépendent des relations entre les sources locales.

Pour pallier les inconvénients des architectures LAV et GAV, de nouveaux schémas d'intégration ont été proposés. En 1999, l'approche GLAV (Global-Local-As-View) [Friedman et al. 99], considérée comme une extension de l'approche LAV, résout le problème d'exhaustivité qu'impose cette dernière. En effet, elle se base sur le principe de récursivité pour permettre une définition flexible des schémas de médiation indépendamment des détails particuliers des sources et abrégeant de cette manière la structure du schéma de médiation.

Quatre ans plus tard, l'approche BAV (*Both As View*) a été définie comme étant la meilleure architecture d'intégration [Dey 04] combinant à la fois les avantages des approches LAV et GAV [Mc. Brien et al. 03]. Elle se base sur un schéma d'intégration hybride et utilise un schéma de transformation de séquences réversible permettant de passer facilement d'une vue du schéma local au schéma global et vice-versa. Dans cette architecture, les relations sémantiques entre les sources des données ainsi que les relations spécifiées au sein du schéma de médiation sont exprimées via le modèle HDM (Hypergraph-based Data Model), un modèle de données bas niveau.

Toujours dans le cadre des modèles hybrides, Rizopoulos propose en 2010 un schéma de *mapping* (mise en correspondance ou association des enregistrements, appelée aussi *dédoublonnage* en jargon métier) qui se base sur l'étude et l'analyse de 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 par des scores de classification de 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,

l'objectif étant d'exploiter cette approche d'association dans le moteur de requêtage lors de la formulation des réponses aux requêtes des utilisateurs [Rizopoulos 10].

1.2. Approche physique

1.2.1. 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'information 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 et conservent leur aspect multisources.

L'architecture générale d'une approche d'intégration physique est résumée dans la Figure 2.6.

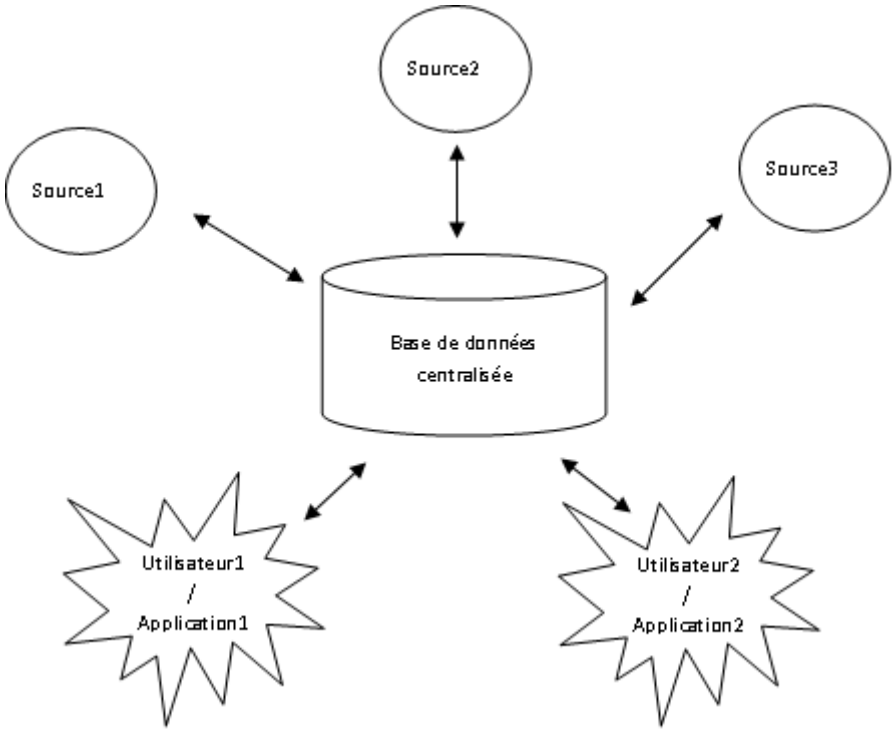


Figure 2.6 Processus d'intégration physique

Avant de spécifier les différents aspects fonctionnels et techniques de l'approche physique de l'intégration des données, commençons par distinguer la différence entre fusion de données et intégration de données. La fusion des données est définie comme une étape du processus général de l'intégration physique qui intervient après le *mapping* des schémas et la détection des doublons [Bleiholder et al. 98]. Ainsi, l'intégration des données est un processus basé sur les étapes suivantes.

- ***Mise en forme des données*** : Cette étape définit un schéma global à partir de l'ensemble des schémas locaux hétérogènes. Deux approches sont alors utilisées : l'intégration des schémas et le *mapping* (mise en correspondance) des schémas. L'intégration des schémas nécessite une bonne connaissance des sources de données et consiste à analyser les schémas locaux afin de générer, dans la mesure du possible, un nouveau schéma qui soit à la fois complet et minimal, correct et compréhensible, tout en respectant les schémas initiaux des sources. Le *mapping* des schémas consiste quant à lui en la mise en correspondance d'un ensemble de schémas locaux étant donné un schéma cible préalablement défini, indépendamment des structures locales. Ainsi, la différence entre ces deux approches consiste dans la manière avec laquelle le schéma global est établi. En effet, dans l'approche de *mapping*, contrairement à l'approche d'intégration, il se construit par rapport à un schéma cible qui n'est pas l'union de tous les schémas locaux de la base. Cela dit, ces deux techniques sont utilisées pour accomplir un même objectif : construire un schéma global commun aux schémas locaux, approche utilisée lors de l'intégration logique de type GAV.
- ***Détection des doublons***, couplage des données (*record linkage*) ou réconciliation des références (*reference reconciliation*) : L'objectif de cette étape est d'identifier les représentations multiples du même objet réel. Il s'agit, en effet, d'une étape primordiale prémisse de la fusion des données. Cette tâche est requise aussi dans l'intégration logique, notamment dans les approches de type LAV et GAV.
- ***Intégration des données*** : L'objectif est d'avoir un schéma à la fois complet et concis. Ainsi, dans [Bleiholder et al. 08], l'intégration des données vise 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 ajoutant, au niveau du système central, de nouvelles sources (de nouveaux objets, de nouveaux attributs décrivant les objets). La *minimalité*, quant à elle, est assurée par l'élimination des 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.

1.2.2. Techniques d'intégration physique : gestion des conflits dans le cadre de la fusion

i- Résolution par l'étude de la provenance

[Wang et al. 01] définissent un modèle de gestion des conflits, induit par la fusion des données hétérogènes et basé sur l'étude de la provenance des données. Ce modèle est un des premiers modèles offrant une solution à base de métadonnées pour la gestion de l'hétérogénéité des systèmes multisources. En effet, son principe se base sur l'analyse de la crédibilité des données et se repose sur deux métadonnées principales : la connaissance des sources de données et la connaissance des sources intermédiaires de données. Ces deux métadonnées permettent d'avoir une idée de la crédibilité des données, indicateur utilisé comme pilier de détermination de l'exactitude des données conflictuelles.

Toujours dans le contexte de l'étude de provenance, [Dong et al. 09a] proposent 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 multisources étant donné un ensemble de sources web. En effet, ce modèle analyse la confiance que l'on peut accorder aux données intégrées, d'une part, en évaluant l'exactitude de leurs sources ainsi que les dépendances qui peuvent exister entre ces différentes sources ; et d'autre part, en utilisant des modèles Bayésiens afin d'estimer la probabilité d'exactitude des données. Pour ce faire, les auteurs s'intéressent à l'analyse des dépendances entre les sources. Cela prend en effet tout son sens lorsque deux sources partagent de fausses informations. Des modèles Bayésiens sont alors utilisés pour estimer la probabilité que deux sources soient dépendantes. Ensuite, grâce à cette 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 de « copies indirectes » (*indirect copying*) entre sources.

La provenance des données pour la gestion des problèmes d'intégration des données multisources est de plus en plus utilisée récemment, surtout avec l'avènement des réseaux sociaux et des données publiques ouvertes (*open data*). En effet, ces nouvelles sources de données sont utilisées pour l'enrichissement des données comme les données clients et les données prospects des bases de données marketing. Cette utilisation induit des problèmes de qualité des données qui déteint, si l'on reste sur le même exemple, sur l'efficacité des campagnes marketing. Une telle problématique est alors principalement résolue par l'analyse

des sources des données. Dans ce contexte, [Talburt 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.

D'autres approches utilisent des méthodes de *scoring* et les appliquent au contexte de la gestion des conflits des données non-structurées (notamment les données Web) privilégiant une source (un site) sur une autre, étant donné le nombre de visites impliquant l'ordre d'apparition dans le navigateur [Wu et al. 11].

ii- Résolution par mise en correspondance des données

Outre l'étude de la provenance des données, des modèles de *mapping* (mise en correspondance) ont été proposés pour la résolution des conflits dus à la fusion des données multisources. Ainsi, Kolovos, Paige et Polack définissent un modèle basé sur le langage EML (*Epsilon Merging Language*) [Kolovos et al. 06]. Il s'agit d'un modèle à base de règles établissant la mise en correspondance des données en question étant donné un processus à quatre étapes :

1 la comparaison des composants, basée sur le langage de comparaison des modèles ECL (*Epsilon Comparison Language*) utilisant des règles d'association (*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 qu'elles peuvent transformer ainsi que l'incidence de cette transformation dans le modèle résultant. Ceci s'effectue grâce à deux langages : le langage de transformation des modèles ETL (*Epsilon model Transformation Language*) et le langage de transformation des textes EGL (*Epsilon Generation Language*) ;
4 la réconciliation qui consiste à résoudre les conflits éventuels induits par le *mapping*.

Le modèle sur lequel se basent les langages ETL et EGL se distingue par son extensibilité 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, il permet de gérer l'hétérogénéité des modèles au sein d'un même méta-

modèle, ce qui n'est pas le cas de toutes les approches proposées dans ce 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 le décideur peut se trouver confronté, si nous nous référons à l'exemple d'une migration des données vers un système unique, au *mapping* de centaines d'entités qui proviendraient de l'ancien système d'informations ou bien d'autres services dont il maîtrise très peu les flux.

Pour pallier ce problème décisionnel de la gestion des conflits, des approches proposent l'automatisation de ce processus. En effet, [Bleiholder et al. 08] les regroupe dans les catégories suivantes.

- Dans les stratégies d'ignorance des conflits (*conflict ignorance*), aucune décision n'est prise. Dans cette catégorie, nous trouvons l'approche « *ignorer* » (*pass it on*) où la décision est laissée à l'utilisateur. Une autre approche consiste à « 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.
- Parmi les stratégies d'évitement des conflits (*conflict avoidance*), nous distinguons 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. Nous citons également 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 le prix, la fiabilité, le volume des données fournies et d'autres critères qualité.
- Parmi les stratégies de résolution des conflits et d'intégration des informations, nous retrouvons 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 stratégies, la donnée imputée au système intégré résultant peut être différente des données sujets

au 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 d'une entreprise » 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 de manière systématique. Par exemple, une valeur de 54, qui résoudrait 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 [Talburt 11] ou par des méthodes plus sophistiquées utilisant les modèles probabilistes [Dong et al. 09b].

1.3. Discussion

Nous avons présenté les technologies et méthodologies proposées dans la littérature pour l'intégration logique et physique de données disparates. Ces approches sont multiples et leur utilisation en pratique dépend de la stratégie d'intégration privilégiée par les experts métier.

En effet, nous distinguons les approches logiques et les approches physiques d'intégration. Le Tableau 2.1 résume les caractéristiques et les avantages des deux approches.

Critères de comparaison	Approche logique	Approche physique
Autonomie	Conserve l'aspect autonome et distribué des sources via l'utilisation de schémas locaux. La communication est, dans la plupart des cas, gérée entre des couches de médiation entre les schémas locaux et un schéma fédérateur global.	Perte au niveau de l'autonomie des bases de données locales.

Critères de comparaison	Approche logique	Approche physique
Lien entre schémas locaux et schéma global	Le lien entre le schéma global et les schémas locaux s'exprime à travers des vues locales et globales (schémas de *mapping*).	Les données sont physiquement stockées (fusionnées) au sein d'une même base. Un grand effort est effectué au niveau de la fusion des données, moyennant des approches probabilistes et des méthodes empruntées au domaine de la qualité des données.
Couplage	Les niveaux de couplage (liaison) sont variés selon l'architecture logique adoptée.	

Tableau 2.1. Comparaison des approches logiques et des approches physiques d'intégration

De même, au sein d'une même approche, les architectures diffèrent. Nous distinguons les approches LAV et GAV dont le résumé des avantages et inconvénients est donné dans le Tableau 2.2.

Critères de comparaison	Approches LAV	Approches GAV
Principe de l'approche	Utilisation de vues locales sur le schéma global	Utilisation de vues globales en fonction de schémas locaux
Communication entre schéma global et schémas locaux	Chaque source propose la vue avec laquelle elle communique avec la couche médiatrice	Le schéma de médiation propose une vue globale à l'ensemble des sources

Critères de comparaison	Approches LAV	Approches GAV
Extensibilité	Ajout facile de nouvelles sources	Le rajout d'une source remet en question le schéma global de médiation
Requêtage	Réécriture assez fastidieuse des requêtes.	Réécriture assez simple et intuitive des requêtes

Tableau 2.2. Comparaison des approches LAV et GAV

Dans les approches d'intégration physiques, les divergences concernent principalement la méthodologie de résolution des incohérences et des conflits causés par les doublons dans les données multisources. Nous remarquons alors que la plupart de ces techniques utilise les méthodes basées sur des métriques d'évaluation de la qualité des données. Ainsi, l'approche basée sur l'étude de la provenance des données s'appuie, entre autres, sur l'analyse des dimensions de crédibilité, d'exactitude et de fiabilité pour évaluer la qualité des sources à intégrer. Par ailleurs, les méthodes de fusion basées sur l'étude des instances s'intéressent à l'évaluation de la qualité des données intrinsèques privilégiant ainsi les données récentes, cohérentes et syntaxiquement exactes. De plus, la qualité des données intervient dans l'évaluation des résultats des requêtes de médiation en analysant les dimensions de fraîcheur, de délai et de coût. Elle revêt, de ce fait, d'une importance capitale dans notre problématique d'optimisation de la sélection multisources.

2. Qualité des données

L'intérêt pour la qualité des données s'est développé lors de la dernière décennie avec l'importance accrue de la valeur et de la donnée, en particulier dans le processus de prise de décision stratégique. Ce constat s'est traduit par la croissance des activités centrées sur la qualité, qui a évolué de 73 % en 2007 [9] à 98 % en 2012 [10]. En particulier, le besoin en qualité s'est fait ressentir dans les organisations gérant au sein de leurs systèmes d'information des bases de données multisources où des indicateurs qualité sont sollicitées pour gérer, entre autres, les problèmes de conflits générés par la présence de doublons ou triplets.

Nous nous intéressons dans cette section à la description des fondements de la qualité des données. Nous nous focalisons, en particulier, sur la définition du rôle de la qualité des données

dans la gestion des problèmes d'intégration des données multisources à travers une revue des méthodologies qualité des données utilisées dans ce contexte. Auparavant, nous rappelons la terminologie utilisée.

Nous notons, en effet, une diversité dans l'appellation des critères d'évaluation de la qualité des données que [Wang et al. 96] regroupent en quatre désignations différentes : les catégories qualité, les dimensions qualité, les attributs qualité et les métriques qualité. La catégorie qualité est assimilée à un concept qualité regroupant quatre classes possibles : l'exactitude des données, la pertinence des données, la représentativité des données et l'accessibilité des données. La dimension qualité est ensuite définie comme un groupe d'indicateurs qualité qui appartiennent à une même catégorie incluant, par exemple, l'accessibilité et la sécurité d'accès dans la catégorie de l'accessibilité. Ces dimensions sont divisées en groupes d'attributs qualité décrivant les indicateurs qualité d'une même dimension. Ainsi, plusieurs indicateurs peuvent qualifier la dimension d'*accessibilité*, telles la rapidité d'accès (*speed of access*), la disponibilité (*availability*) ou encore l'actualité (*up-to-date*). Enfin, pour quantifier ces différents indicateurs, des fonctions et modèles mathématiques connus sous le nom de métriques qualité sont utilisés.

Dans le cadre de cet état de l'art, nous nous limitons à deux appellations des critères qualité : les dimensions et les métriques. Nous proposons de définir les dimensions les plus fréquentes dans la littérature et les plus utilisées dans les problématiques d'évaluation qualité, ainsi que les métriques correspondantes. Dans certains cas, nous utilisons à la fois la désignation anglaise et française de la dimension ou métrique qualité afin d'écarter toute confusion que la traduction pourrait induire.

2.1. Dimensions qualité

Au début des années 80, l'étude de la qualité des données se limitait à l'unique considération de la dimension d'exactitude (*accuracy*) pour évaluer la qualité d'une entité donnée [Olson 03], et s'appliquait principalement à l'évaluation du processus de saisie des données [Frank 07]. De nos jours, nous comptons quelques centaines de dimensions qualité dont l'usage dépasse la simple évaluation de l'exactitude et de la justesse des données pour décrire des critères plus larges d'accessibilité, de complétude, de fraîcheur et bien d'autres encore, et dont la définition et la quantification varient selon le contexte d'évaluation qualité en question. Ainsi, des dimensions telles que la fraîcheur (*freshness*), la récence (*recency*), l'exactitude (*accuracy*), la cohérence (*consistency*), la précision (*precision*) ou encore la complétude (*completeness*) ont vu le jour. Certaines, comme l'exactitude et la précision, sont des appellations différentes qui

réfèrent plus ou moins à la même caractéristique, à savoir la justesse de la valeur en question. Ce dernier constat se justifie par l'absence de standardisation quant à la définition des critères qualité de données.

Nous nous intéressons dans ce paragraphe à la description des approches de définition des dimensions qualité. Nous décrivons ensuite les dimensions les plus répandues et nous présentons enfin quelques-unes de leurs classifications.

2.1.1. Approches de définition des dimensions qualité des données

i- Approche intuitive

Dans cette approche, les dimensions qualité sont intuitivement choisies par les experts métier et varient, de ce fait, d'un domaine à un autre. Ainsi, en comptabilité et audit des données, la *fiabilité* est le critère le plus souvent considéré en évaluation de la qualité des données. Pour les systèmes d'information, on s'intéresse plutôt à l'étude de la *satisfaction de l'utilisateur* et la *qualité de l'information.*

Ceci nous pousse à ouvrir une parenthèse à propos de la distinction entre donnée et information. En effet, on appelle donnée toute représentation d'une information sous une forme conventionnelle destinée à faciliter son traitement. Il s'agit d'un élément brut qui n'a pas été interprété, c'est-à-dire mis en contexte. Au contraire, l'information se définit comme une donnée interprétée et, donc, comme tout renseignement ou tout élément de connaissance susceptible d'être représenté sous une forme adaptée à une communication, un enregistrement ou un traitement. Ainsi, la dimension *qualité de l'information* ne comprend pas un champ ou un attribut en particulier, mais plutôt l'ensemble des attributs conférant à son interprétation.

ii- Approche théorique

Cette approche vise à analyser la fiabilité de la donnée depuis sa création dans le système d'informations (SI). En particulier, elle vise à étudier la non-conformité de la représentation des données dans le système d'informations par rapport à leur représentation réelle (RR). Les dimensions qualité qui sont définies permettent de guider cette analyse.

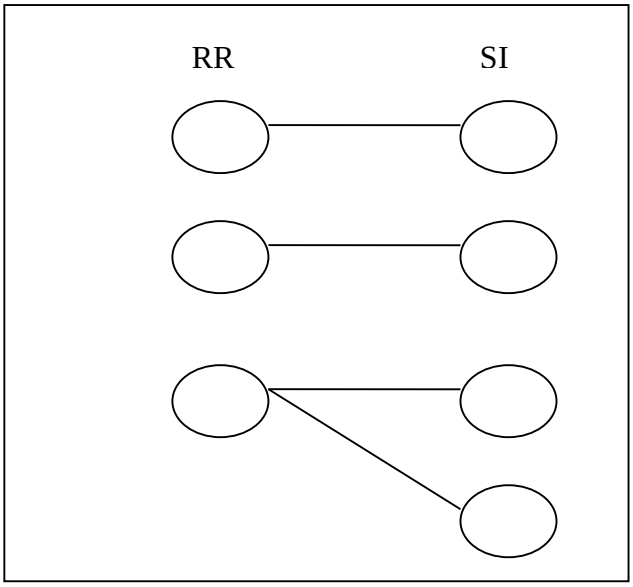


Figure 2.7 Représentation du monde réel [Wand et al. 96]

Toute déviation de cette bonne représentation génère des défaillances. Nous distinguons ainsi :

- les déficiences de représentation, à savoir l'incomplétude, l'ambiguïté et la non significativité de la représentation. Ceci est exprimé par les dimensions suivantes : incomplétude, ambiguïté et représentativité, dimensions modélisées dans Figure 2.8 ;
- les déficiences opérationnelles qui nous permettent de distinguer les dimensions suivantes : exactitude (*accuracy*), fiabilité (*reliability*), convenance temporelle (*timeliness*)[10], complétude (*completeness*) et cohérence (*consistency*).

[10] Dans la suite du manuscrit, nous l'appellation anglaise, plus répandue et couramment utilisée dans les articles francophones.

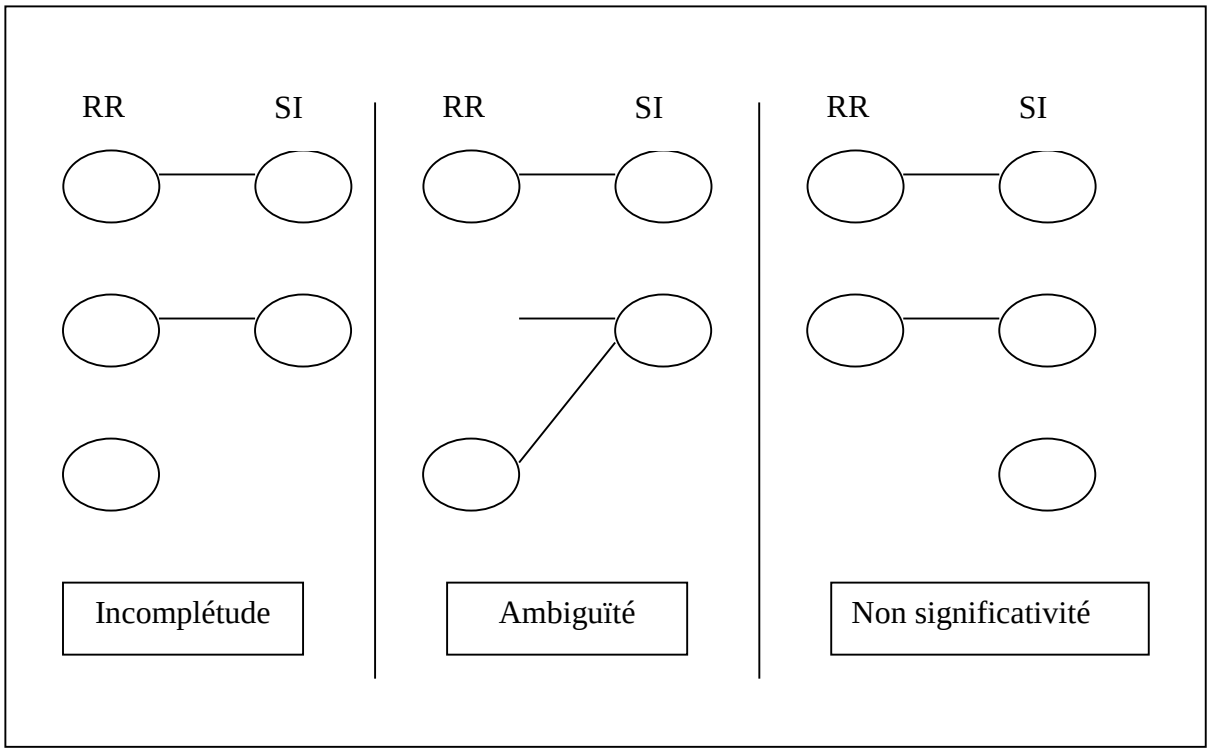


Figure 2.8 Les déficiences de représentation selon [Wand et al. 96]

iii- Approche empirique

Cette approche est utilisée quand les critères qualité sont définis par les utilisateurs. Elle évalue, ainsi, la qualité de leurs données selon leurs propres critères et exigences et nous permet, de ce fait, d'obtenir un ensemble de dimensions qualité pouvant intéresser les utilisateurs des données conformément à la règle *fitness for use* [Wang et al. 96].

2.1.2. Dimensions qualité les plus courantes

Une première revue de la littérature montre que nous pouvons grossièrement regrouper les dimensions qualité en quatre groupes génériques: les dimensions d'exactitude, les dimensions temporelles, les dimensions de complétude et les dimensions de confiance.

i- Dimensions d'exactitude

Le terme exactitude est souvent utilisé pour décrire plusieurs aspects reliés à la donnée intrinsèque [Peralta 06] : l'exactitude, la précision, l'absence d'ambiguïté, etc. Nous distinguons, dans ce qui suit, la définition de ces différentes dimensions.

- Exactitude (*Accuracy*) : Plusieurs définitions ont été proposées afin de qualifier la dimension *exactitude*. Ainsi, Redman la définit comme le taux de conformité d'un

ensemble de valeurs de données à un ensemble de données de référence (conventionnellement correct) [Redman 96]. Dans le même sens, [Batini et al. 06] l'identifient par la proximité d'une valeur *v'* à une valeur *v* considérée comme la représentation correcte d'une entité réelle que *v'* tente de représenter. Ils distinguent alors :

- l'exactitude syntaxique, définie par l'admissibilité de la valeur en question dans le domaine de définition de l'attribut. Par exemple, supposons que *v*, valeur correcte de l'entité réelle, soit *John*. Si v', la valeur de *v* représentée dans la base, est *Jack*, *v'* est considérée correcte, car admissible dans le domaine de définition qui est dans notre cas le domaine des prénoms ;
- l'exactitude sémantique, définie par la divergence par rapport à la valeur réelle. En reprenant l'exemple ci-dessus, *v'* serait ainsi considérée inexacte.

- Précision (*Precision*) : Elle est principalement utilisée pour définir l'exactitude des données géométriques. La norme ISO 19113 :2002 (revisitée par la norme ISO 19157 :2013) [11] distingue ainsi la précision géométrique relative à la résolution spatiale définie par :
  - la précision de position vérifiant que l'objet est plus ou moins positionné sur la carte,
  - la précision de forme vérifiant que la forme de l'objet est plus ou moins juste sur la carte ;
  - la précision sémantique définie par la différence entre la valeur d'un attribut du jeu de données et sa valeur dans le monde nominal.
- Cohérence (*Consistency*) : C'est l'absence d'ambiguïté. Elle est aussi définie, dans le contexte des bases de données, par la satisfaction des contraintes d'intégrité [Redman 96] et par la cohérence de représentation de la même entité réelle dans des tables différentes (contrainte d'intégrité référentielle de Codd) [Pipino et al. 02].

Nous remarquons que ces définitions qualifient les dimensions d'exactitude intrinsèquement, indépendamment des sources depuis lesquelles elles proviennent. D'autres définitions par contre les relient directement à la fiabilité de la source. Dans ce cas, les dimensions d'exactitude sont les métriques qui évaluent la réputation et la crédibilité des sources [Strong et al. 97].

ii- Dimensions temporelles

Une multitude d'indicateurs a été proposée dans la littérature afin de quantifier les dimensions d'actualité et de récence.

- La *Timeliness* : Il s'agit de la dimension temporelle la plus utilisée dans la littérature. Wand et Wang la définissent comme la propriété *intrinsèque* d'une donnée qui désigne principalement le caractère courant ou à jour des données au moment de leur diffusion [Wand et al. 96]. Insistant sur le côté contextuel, Leo Pipino la présente comme le degré d'actualité des données respectivement aux tâches dans lesquelles elles sont utilisées [Pipino et al. 02]. Par ailleurs, tout comme l'exactitude, certains chercheurs évaluent la *timeliness* en ramenant les données à leurs sources. Elle mesure alors la fréquence de mise à jour ou encore la fréquence de création de nouvelles données dans la source en question [Naumann et al. 99].
- La fraîcheur (*Freshness*) : Il s'agit du terme le plus générique décrivant l'aspect temporel des données en calculant (parfois en estimant) l'âge des données [Akoka et al. 07].
- L'actualité (*Currency*) : Souvent confondue avec la fraîcheur, cette dimension définit le degré d'ancienneté des données respectivement à leurs sources [Segev et Weiping 90] et mesure leur obsolescence et la rapidité de leurs mises à jour [Scannapieco et al. 05]. Elle décrit aussi leur degré d'ancienneté par le délai entre la date d'extraction des données à partir des sources et sa livraison aux utilisateurs [Akoka et al. 07].
- La volatilité (*Volatility*) : Cette dimension définit la durée pendant laquelle les données demeurent valides et est utilisée de ce fait comme un facteur d'évaluation de la *timeliness* [Pipino et al. 02]. Cependant, elle décrit aussi la fréquence avec laquelle les données varient dans le temps [Scannapieco et al. 05].
- La récence (*Recency*) : Cette dimension décrit l'âge des données et définit la fraîcheur [Peralta 06].

iii- Dimensions de complétude

La complétude est une dimension plutôt définie dans le contexte des modèles relationnels. Ainsi, [Pipino et al. 02] distinguent trois niveaux différents de complétude.

- La complétude au niveau du schéma décrit le degré avec lequel toutes les entités et attributs sont représentés.

- La complétude au niveau de la colonne décrit les valeurs manquantes à une propriété donnée ou dans un attribut d'une table.
- La complétude au niveau de la population évalue les valeurs manquantes respectivement à une population de référence. Prenons l'exemple d'une table décrivant les employés d'une entreprise. Si l'entreprise compte soixante employés et que la table représentative n'en décrit qu'une cinquantaine, nous parlons d'incomplétude de la population.

Par ailleurs, [Berti-Equille 07] décrit la complétude non pas comme la simple constatation de présence ou d'absence de valeurs nulles dans la table, mais par rapport au sens que l'on veut déduire de la présence ou de l'absence des données. En effet, en fonction des observations, les valeurs nulles peuvent représenter des objets existants mais inconnus, des objets inexistants ou encore des objets pouvant exister mais dont l'existence n'a pas été prouvée. Ces différentes nuances sont distinguées dans l'« hypothèse du monde fermé » et l'« hypothèse du monde ouvert ».

L'hypothèse du monde fermé pénalise l'absence de données dans le sens où une valeur nulle est interprétée comme étant une valeur manquante et se mesure des deux manières suivantes :

- la complétude horizontale décrit le nombre de valeurs nulles dans un enregistrement ou un ensemble d'enregistrements ;
- la complétude verticale décrit le nombre de valeurs nulles dans un attribut.

En revanche, l'hypothèse du monde ouvert suppose que la connaissance du monde réel est incomplète et que ce qui n'est pas représenté (à savoir les valeurs nulles) est considéré comme inconnu plutôt que faux. Nous distinguons ainsi les notions de couverture et de densité, toutes deux ramenant les données à leurs sources.

- La couverture (*coverage*) de la source *S* mesure le nombre d'enregistrements fournis par *S*, relativement à une relation universelle *U*. [Naumann et al. 04] définissent également la complétude comme la couverture avec laquelle le phénomène observé est représenté dans l'assemblage des données.
- La densité (*density*) mesure le volume moyen de données fournies par une source *S*.

iv- Dimensions de confiance

Plusieurs dimensions peuvent être définies dans ce contexte. Nous en citons les plus courantes.

- Objectivité (*Objectivity*) : Cette dimension est une interprétation des différentes mesures présentées ci-dessus. Ainsi, Agosta considère que l'objectivité est un concept assez

générique regroupant un certain nombre de diverses dimensions telles que l'exactitude (*accuracy*), l'existence (*existence*), la causalité (*causality*), la cohérence (*consistency*), la *timeliness*, la complétude (*completeness*), la non-ambigüité et la précision (*precision*) [Agosta 02]. De plus, Holt insiste sur la généralité de cette dimension en la présentant comme la propriété associée à un ensemble de données exactes, claires, fiables, complètes et non biaisées [Holt 05].

- Crédibilité (*Believability*) : Wang définit la crédibilité comme une dimension subjective déduite par l'analyste sur l'interprétabilité des données [Wang et al. 92]. [Pipino et al. 02] complètent le raisonnement de Wang en présentant deux aspects de la crédibilité :
  - un aspect subjectif reflétant le niveau de confiance accordé aux sources par les experts ;
  - un aspect objectif donné par une moyenne pondérée des métriques évaluant l'objectivité et l'exactitude des données.
- Fiabilité (*Reliability*) : Cette dimension est relative aux sources des données et à leur capacité de représenter avec fiabilité les données réelles [Wand et al. 96].
- Réputation (*Reputation*) : Cette dimension décrit le niveau de confiance des données.

Nous pouvons remarquer, d'après les quelques définitions citées ci-dessus, que les notions sont quelquefois floues (surtout en ce qui concerne les dimensions temporelles et les dimensions de confiance). C'est pour cette raison qu'il est conseillé de bien définir le contexte et le sens attribué à chaque dimension en amont de tout projet d'évaluation de la qualité des données afin d'assurer la bonne interprétation de la valeur de la métrique qui lui est associée.

2.1.3. Classifications des dimensions qualité

Outre la classification sémantique proposée dans le paragraphe précédent, différentes classifications des dimensions qualité ont été proposées dans la littérature. Les travaux initiés par [Wang et al. 96] montrent qu'il n'existe pas de classification universelle des dimensions qualité. En effet, les classifications changent en fonction du contexte de l'étude, des besoins du projet et de l'orientation que l'analyste veut donner à son évaluation. Nous décrivons dans cette section quelques-unes des classifications les plus citées dans la littérature.

Une première classification établie par [Wang et al. 96] distingue les dimensions suivantes.

- Les dimensions intrinsèques concernent la qualité de la donnée elle-même et comprennent les dimensions de crédibilité (*believability*), d'exactitude (*acuracy*), d'objectivité (*objectivity*) et de réputation (*reputation*).
- Les dimensions contextuelles insistent sur les besoins requis en termes de qualité des données respectivement au contexte de l'étude et comprennent les dimensions de valeur ajoutée (*added value*), de pertinence (*relevancy*), de complétude (*completeness*), d'adéquation au volume de données étudié (*appropriate amount of data*) et de *timeliness.*
- Les dimensions de représentativité soulignent l'importance du rôle que joue le système d'information et comprennent les dimensions d'interprétabilité (*interpretability*), de facilité de compréhension (*ease of understanding*), de concision de la représentation (*concise representation*) et de cohérence de représentation (*representational consistency*).
- Les dimensions d'accessibilité intéressent aussi bien les professionnels des systèmes d'information que les consommateurs de données. En effet, l'accessibilité a changé de sens avec la démocratisation des systèmes d'information numériques et sa nécessité en tant que paramètre d'évaluation de la qualité commence à se ressentir comme un vrai besoin. Par ailleurs, [Wang et al. 96] affirment qu'il n'y a pas une grande différence entre traiter l'accessibilité en tant que catégorie et la traiter en tant que dimension, l'important étant qu'elle soit prise en considération. Ainsi, en tant que catégorie, l'accessibilité comprendra les dimensions d'accessibilité et de sécurité de l'accès.

Cette classification est représentée dans la Figure 2.9.

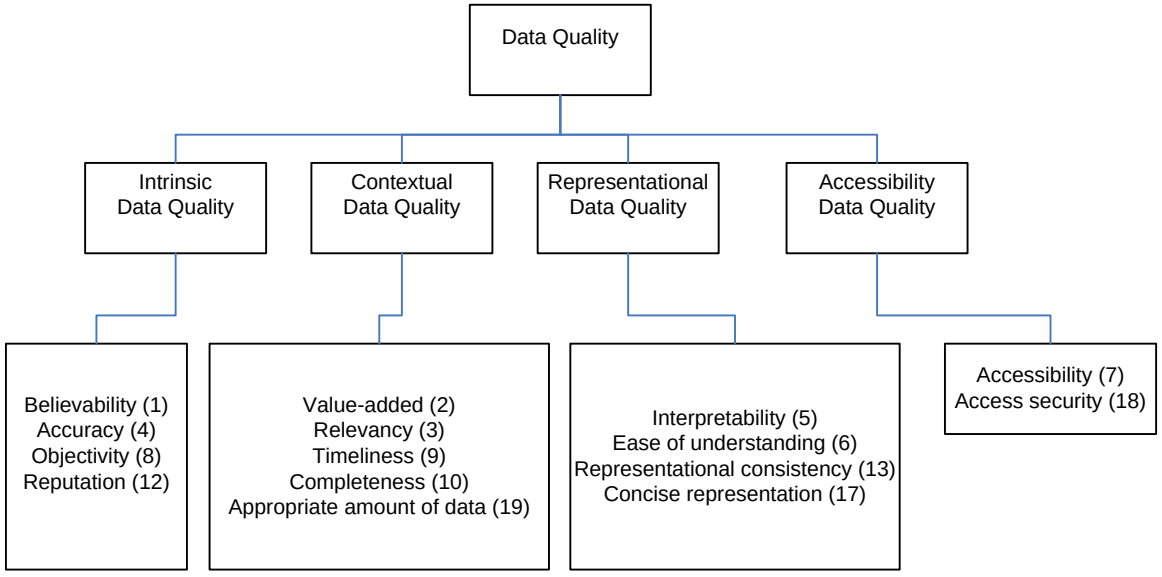


Figure 2.9 Hiérarchie conceptuelle de la qualité des données [Wang et al. 96]

Une autre classification est proposée par Redman qui regroupe les dimensions qualité en trois catégories principales [Redman 01] :

1 le schéma conceptuel (*Conceptual schema*) ;

2 les dimensions qualifiant les valeurs intrinsèques des attributs (*data value view*), indépendamment de la qualité de leur représentation dans la base. Nous retrouvons ainsi les dimensions d'exactitude (*accuracy*), de complétude (*completeness*), de récence (*currency*) et de cohérence (*consistency*) ;

3 les dimensions se rattachant aux formats des données (*data format view*), comme les dimensions d'adéquation (*appropriateness*), d'interprétabilité (*interpretability*), de portabilité (*portability*), de précision du format (*format precision*), de flexibilité du format (*format flexibility*), d'aptitude à représenter les valeurs nulles, d'utilisation efficace de la mémoire et de cohérence de représentation.

Par ailleurs, nous pouvons classifier les dimensions par contexte d'évaluation. Nous distinguons alors :

- les dimensions d'évaluation de la donnée intrinsèque : exactitude (*accuracy*), *timeliness*, fraîcheur (*freshness*), cohérence (*consistency*)…
- les dimensions d'évaluation des sources de données : provenance (provenance), réputation (*reputation*), crédibilité (*reliability*), *timeliness*, …
- les dimensions d'évaluation des processus : temps de réponse (*response time*), délai d'exécution du processus, accessibilité (*accessibility*), coût d'exécution du processus [Peralta et al. 04]

2.2. Métriques qualité

De très nombreuses approches inspirées de différents domaines ont été proposées afin de mesurer la qualité des données dans les bases de données quantifiant, ainsi, les dimensions citées précédemment. Nous distinguons :

- les approches statistiques telles que l'échantillonnage et l'extrapolation, les estimations paramétriques, les techniques d'évaluation de la cohérence allant des règles d'associations et la classification aux questionnaires de validation ;

- les approches relationnelles utilisant des requêtes de la base de données telles que la jointure pour la mesure de l'unicité et les contraintes d'intégrité de Codd pour mesurer la cohérence des données.

Le Tableau 2.3 présente un résumé des métriques les plus utilisées pour l'estimation de la qualité intrinsèque des données et celle des sources. Nous nous sommes intentionnellement limités à ces deux classes de métriques, car ce sont les seules qui nous intéressent dans notre contexte global d'évaluation des données multisources.

Dimension	**Mesure**
Métriques pour la qualité des données	
Exactitude (*Accuracy*)	Fonctions à base de contraintes (déviation par rapport à la normale) [Berti-Equille 99] Fonctions synopsis (résumés statistiques, agrégats, estimations paramétriques) Fonctions exploratrices (techniques de clusterisation, règles d'association, arbres de décision…) Vérification sémantique et syntaxique et évaluation de la précision des données [Akoka et al. 08]
Complétude (*Completeness*)	Taux de valeurs nulles [Batini et al. 06] Complétude horizontale (*horizontal fitness*) : complétude des attributs présents dans la relation par rapport à la demande de l'utilisateur [Naumann et al. 99] ; expimée par [Akoka et al. 08] : taux de déficit : données manquantes par rapport au terrain nominal de référence taux d'excès : données présentes dans le jeu de données mais manquantes ou indéterminées sur le terrain nominal
Fraîcheur (*Freshness*)	Mesurée de deux façons différentes [Akoka et al. 07] : actualité (*currency*) *timeliness* Intervalles entre les dates d'extraction, de mise à jour et d'intégration [Peralta et al. 04]

Dimension	Mesure
Actualité (*Currency*)	= Age des données au moment de la livraison + Date de livraison (ou date d'utilisation) – Date d'insertion des données dans la base [Ballou et al. 98] = Date d'utilisation – date de dernière mise à jour [Ardagna et al. 05]
Volatilité (*Volatility*)	Durée pendant lequel les données demeurent valides [Ballou et al. 98]
Timeliness	$=\left\{\max\left[0,\left(1-\frac{\text{Actualité}}{\text{Volatilité}}\right)\right]\right\}^{s}$ [Lee et al. 06] où s est un facteur de contrôle de la sensibilité du rapport [Ballou et al. 98].
Précision (*Precision*)	Précision=$\frac{nombre\ de\ doublons\ corrects}{nombre\ total\ des\ doublons\ dans\ la\ base}$ dans le contexte d'un processus de fusion des données [Hernandez et al. 95]
Cohérence (*Consistency*)	Contrainte d'intégrité de Codd [Pipino et al. 02] Satisfaction des contraintes [Redman 96]
Crédibilité (*Believability*)	= Min(crédibilité de la source, crédibilité de la donnée comparée à un standard interne (bon sens métier), crédibilité basée sur l'âge des données) [Lee et al. 06]
Adéquation du volume de données (*Appropriate amount of data*)	$=\text{Min}\left(\frac{\text{Nombre d'unités de données fourni}}{\text{Nombre d'unités de données requis}},\ \frac{\text{Nombre d'unités de données requis}}{\text{Nombre d'unités de données fourni}}\right)$ [Lee et al. 06]
Accessibilité (*Accessibility*)	$=\left\{\max\left[0,\left(1-\frac{\text{intervalle de temps mis de la demande à la livraison}}{\text{intervalle de temps mis de la demande à l'obsolescence}}\right)\right]\right\}^{s}$ [Lee et al. 06] où s est un facteur de contrôle de la sensibilité du rapport [Ballou et al. 98].

Dimension	Mesure
Unicité des données (*Unicity*)	Absence de doublons exacts ou approximatifs [Akoka et al. 08]
Métriques pour la qualité des sources	
Fiabilité (*Reliability*)	Méthodes expérimentales (taux d'erreur intrinsèque) [Naumann et al. 99] (choix de la meilleure source dans un contexte multi-source)
Réputation (*Reputation*)	Mesure subjective donnée par l'ordre de préférence donné par l'utilisateur [Naumann et al. 99] (choix de la meilleure source dans un contexte multi-source)
Compréhensivité (ou bien facilité de compréhension) (*Understantability*)	Mesure subjective donnée par l'ordre de préférence donné par l'utilisateur [Naumann et al. 99] (choix de la meilleure source dans un contexte multi-source)
Timeliness	Age moyen des données [Naumann et al. 99] (choix de la meilleure source dans un contexte multi-source)

Tableau 2.3 Métriques les plus courantes associées aux dimensions qualité d'évaluation des données et des sources

2.3. Méthodologies pour l'évaluation de la qualité

Vers le milieu des années 2000, les projets s'intéressant à l'évaluation qualité des systèmes d'information se sont multipliés. Vu les centaines de dimensions et métriques disponibles dans la littérature et la subtilité et le niveau d'expertise requis des responsables qualité pour, dans un premier temps, identifier les bons indicateurs qualité et, ensuite, les déployer au service d'un processus d'assurance qualité, des approches et méthodologies ont été développées pour orienter les experts vers le processus de mesure et d'amélioration qualité le plus adéquat à l'architecture de leur organisation. Ces méthodologies sont en effet des études de cas, exemples ou encore techniques établies par des professionnels du métier ayant fait leur preuve dans l'audit qualité. Nous présentons les plus intéressantes [Batini et al. 06], mais auparavant, nous définissons la notion de méthodologie qualité et les principes sur lesquels elle se base.

2.3.1. Méthodologie qualité : définition et principaux concepts

Une méthodologie qualité de données est définie comme un ensemble de directives et de techniques permettant de caractériser un processus rationnel d'utilisation de l'information pour la mesure et l'amélioration de la qualité des données au sein d'un système d'information. Elle implique une interaction de trois domaines de connaissances ou métiers de l'entreprise [Batini et Scannapieco 06] : la connaissance organisationnelle, la connaissance technique et la connaissance de la qualité des données. Cette interaction est schématisée par la Figure 2.10. Dans la méthodologie décrite par [Batini et al. 06], l'évaluation de la qualité des données est assimilée à un modèle général d'entrées/sorties qui, étant donnés des flux de données, un ensemble de bases de données internes et externes, des processus et macro-processus, un certain nombre de dimensions qualité et un budget, fournit un ensemble de processus reformés et ajustés, des bases de données mesurées et améliorées avec des coûts contrôlés et, éventuellement, des bénéfices. Les objectifs d'une telle approche d'évaluation qualité sont, dans un premier temps, une mesure de la qualité de la base de données et des flux composant le système d'informations sous-jacent, ainsi qu'une réduction des coûts de la non-qualité (dans le cas où la base en question est de faible qualité) et, dans un deuxième temps, une comparaison des mesures qualité objectives obtenues par la méthodologie d'évaluation avec les niveaux qualité de référence subjectifs définis par un expert métier.

Partant de ce postulat, nous remarquons que les méthodologies qualité n'ont pas pour principal objectif de produire ou de générer les dimensions qualité les plus adéquates au processus d'évaluation en question. Ces dimensions qualité servent plutôt d'outils pour définir l'approche qualité la plus adéquate à l'activité, base ou organisation auditée.

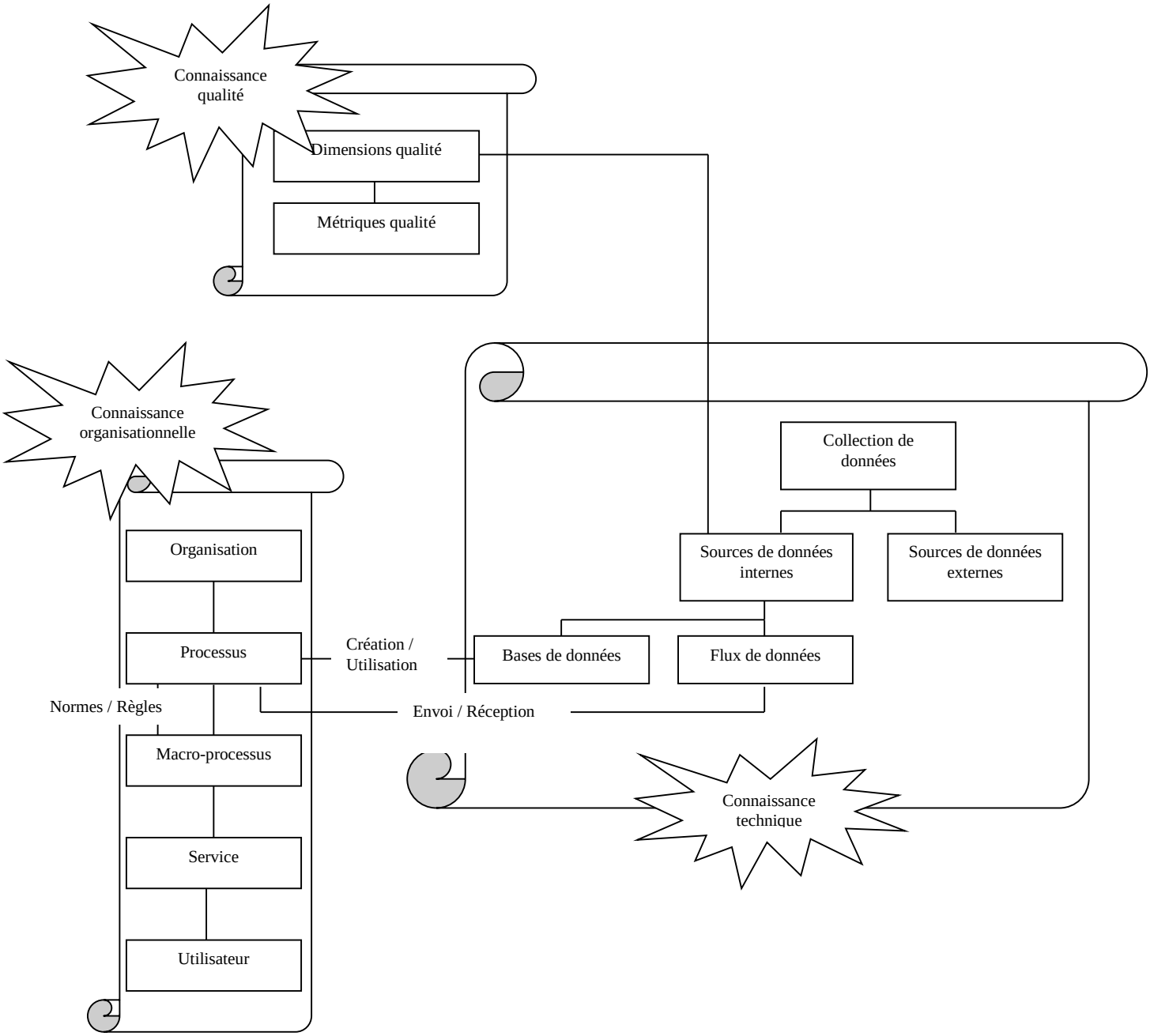


Figure 2.10 Implication des connaissances dans la mesure de la qualité des données et l'amélioration des processus

Les travaux de [Batini et al. 06] distinguent quatre grandes familles de méthodologies qualité des données, définies selon les critères suivants :

1 la stratégie d'amélioration ciblée, où l'on distingue les stratégies orientées données (qui se focalisent sur les sources des données) de celles orientées processus (qui se centrent sur le

processus de production des données pour l'analyser et, éventuellement, le modifier dans le but d'éliminer les problèmes qualité) ;

2 l'objectif attendu, à savoir la mesure (ou l'évaluation) de la qualité des données ou l'amélioration de la qualité. Mesure et amélioration sont généralement liées. L'évaluation de la qualité (aussi appelée *benchmarking*) consiste à comparer les données à des valeurs de référence diagnostiquant ainsi la qualité de la base ;

3 l'objectif envisagé de la méthodologie qualité des données ; envergure générale à large spectre d'activités et dimensions ou bien envergure spécifique ciblant une seule activité de l'organisation, un processus spécifique ou une base de données ;

4 l'envergure voulue par la méthodologie qualité des données, à savoir globale, inter-organisationnelle ou spécifique, intra-organisationnelle, où l'activité d'évaluation et d'amélioration concerne une organisation spécifique ou un secteur d'organisation spécifique ou encore une base ou un processus spécifiques.

Nous nous intéressons, dans le cadre de cet état de l'art, aux approches d'évaluation et d'amélioration qualité. Le but étant de dresser la liste des méthodologies que nous pourrons utiliser par la suite dans notre problématique d'évaluation de la qualité des données multisources.

Ainsi, les approches d'évaluation qualité ont en commun les processus suivants :

1 l'analyse des données, qui consiste à examiner les documentations disponibles sur les données et analyser le schéma logique des données via des entretiens, dans le but de collecter des connaissances sur les données, d'avoir des informations sur les architectures, les flux de données et les règles de gestion ;

2 l'analyse des besoins en qualité des données, qui consiste à collecter les suggestions des utilisateurs des données et des gérants quant aux causes possibles des erreurs, puis à déterminer les objectifs qualité souhaités ;

3 l'identification des données les plus critiques à analyser ;

4 la modélisation du processus d'évaluation cherchant à établir un modèle formel ou semi-formel ;

5 la définition des dimensions qualité adéquates et l'application des métriques correspondantes sur l'ensemble de la base ou bien sur un échantillon (dans le cas où la métrique est inapplicable ou coûteuse) ;

6 l'évaluation du coût de la non-qualité des données défini par le coût de la mise en place du processus d'évaluation qualité puis de son application aux données sous-jacentes ;

7 l'estimation des bénéfices d'évaluation, à savoir les bénéfices d'avoir une base de qualité plus élevée ;
8 l'assignation des tâches de gestion aux responsables de chaque activité faisant partie du processus général de production des données ;
9 l'assignation des responsabilités de contrôle des données aux responsables des méthodes de gouvernance ;
10 l'utilisation des outils et techniques adéquats en respectant les contraintes du budget et les moyens de l'organisation en question.

Parallèlement, les approches d'amélioration qualité requièrent un certain nombre de démarches :

1 la recherche des causes des erreurs : analyse des causes potentielles de déviation des données ;
2 la modélisation des solutions d'amélioration des données ;
3 l'établissement d'un processus de contrôle pour chaque tâche du processus de production des données ;
4 la modélisation des solutions d'amélioration des processus ;
5 la re-modélisation des processus ;
6 la gestion des solutions d'amélioration, à savoir la définition de nouvelles règles améliorant le niveau qualité des données. Ces règles étendent les connaissances du métier de production des données de qualité.
7 le test de l'efficacité de l'approche avec l'établissement de mesures périodiques et d'activités de surveillance qui servent de *feedback* quant à l'approche utilisée.

2.3.2. Exemples de méthodologies qualité

Nous décrivons les approches générales les plus communément utilisées pour l'évaluation et l'amélioration expérimentales (pratique) de la qualité des données.

i- Approche TQdM (Total Quality data Methodology)

Désormais nommée TIQM (*Total Information Quality Methodology*), cette approche était initialement destinée à être utilisée dans les entrepôts de données, notamment dans la phase de consolidation (intégration) des données [English 99]. Elle fournit des directives détaillées afin d'évaluer les coûts de la faible qualité des données, les coûts des processus d'amélioration des données et les bénéfices de l'amélioration de la qualité.

De point de vue de la stratégie managériale, cette approche définit les activités qualité à réaliser, les bases et les flux à considérer ainsi que les techniques à adopter. Concrètement, cela consiste à :

1 évaluer l'aptitude de l'entreprise à appliquer une méthodologie qualité ;
2 étudier la satisfiabilité du client afin de déterminer les problèmes causés par la source ;
3 se concentrer sur un projet expérimental (de test) afin d'ajuster les paramètres de l'approche et éviter ainsi les risques d'échec, conformément au slogan « *think big, start small, scale first* » (penser grand, commencer petit, calibrer progressivement) ;
4 définir la stratégie d'administration de l'information (production et échange des données), en respectant les règles et lois gouvernant les processus métier ;
5 se servir des résultats de l'étude de faisabilité afin d'analyser la résistance de la méthodologie au changement de processus, au contrôle de l'établissement et au partage de l'information et du certificat qualité ;
6 établir une relation spécifique avec les responsables hiérarchiques (*senior managers*) dans le but d'obtenir leur accord et leur participation au processus.

Concrètement, l'approche TQdM est constituée de trois tâches principales centrées autour de l'évaluation, de l'amélioration et de la gestion des solutions d'amélioration. La tâche d'évaluation comprend l'analyse des données, l'analyse des besoins, la mesure de la qualité, l'évaluation de la non qualité des données et l'évaluation des bénéfices (ou gains). La tâche d'amélioration s'articule sur deux axes : l'amélioration des données (l'analyse des données défectueuses, la standardisation et la normalisation des données, la correction et la complétude des données, la mise en correspondance (*matching*), la transformation et la consolidation des données) et l'amélioration des processus qui consiste à vérifier l'efficacité de cette amélioration. Enfin, la tâche de gestion des solutions d'amélioration se rattache à l'évaluation de l'aptitude (*readiness*) de l'organisation à héberger une solution de gestion de la qualité, la création d'une vision pour l'amélioration de la qualité, l'étude de la satisfaction des clients au près des équipes de gestion de l'information, la sélection d'un échantillon de données sur lequel on va appliquer l'approche en question, la définition des problèmes métier à résoudre, la définition du schéma de production de la valeur de l'information, l'établissement d'un niveau de référence pour l'évaluation, l'analyse des réclamations des clients, la définition des coûts qualité dus aux problèmes qualité, la définition de l'intendance de l'information, l'analyse des causes de la non qualité et la recommandation des possibilités d'amélioration et l'établissement

de points réguliers avec les *senior managers* pour la communication et l'éducation qualité au sein de l'organisation dont il en est question.

Les points forts de cette approche, par rapport à d'autres méthodologies définies dans la littérature, résident dans son étude du coût/bénéfice ainsi que dans sa perspective managériale (dans le sens où on y définit une stratégie assurant l'efficacité des choix techniques à prendre par l'organisation).

ii- Approche TDQM (Total Data Quality Management)

Cette approche a été développée dès les années 80 par le MITIQ (MIT Information Quality) pour permettre aux praticiens d'évaluer la qualité de leurs données. Elle n'a, depuis, cessé d'évoluer afin de mieux s'adapter à la réalité des entreprises. Ainsi, la version la plus récente se base sur l'utilisation du modèle IP-MAP (*Information Product MAP*), modèle considérant la donnée comme un produit, et établissant par conséquent le parallèle entre la qualité de la donnée et la qualité du produit ; et du formalisme de modélisation IP-UML, extension du langage UML pour la gestion de la qualité des données grâce à trois concepts : le modèle d'analyse des données, le modèle d'analyse de la qualité et le modèle de la conception de la qualité. L'approche TQdM s'articule en quatre phases primordiales [Shankaranarayan et al. 00] :

1 définition : analyse des besoins en qualité de données (correspondant à la phase « analyse de la qualité » dans le langage IP-UML) ;
2 mesure : évaluation de la qualité des données ;
3 analyse : analyse des données et modélisation des processus ;
4 amélioration : conception et modélisation de la solution d'amélioration des données et des processus (« vérification de la qualité » dans IP-UML) et re-modélisation des processus (« amélioration de la qualité » dans IP-UML).

Ces phases sont itérativement exécutées, constituant ainsi un cycle.

iii- Approche Istat (Istituto nazionale di statistica)

Cette approche vise principalement l'étude des données de type « Adresses/Localisation » dans le cadre d'une structure organisationnelle compliquée composée d'agences locales, centrales et périphériques. Elle s'articule autour de trois activités principales [Falorsi et al. 06] :

1 l'évaluation de la qualité ;

2 l'amélioration interne de la qualité (amélioration intra-organisationnelle), qui concerne principalement les activités locales des agences ;

3 l'amélioration inter-administrative de la qualité (amélioration inter-organisationnelle) qui concerne la globalité du système d'informations coopératif des différentes administrations en termes de flux de données échangés et organisation des bases de données centrales ainsi que les possibilités de coordination.

Ces trois activités interagissent comme le montre la Figure 2.11.

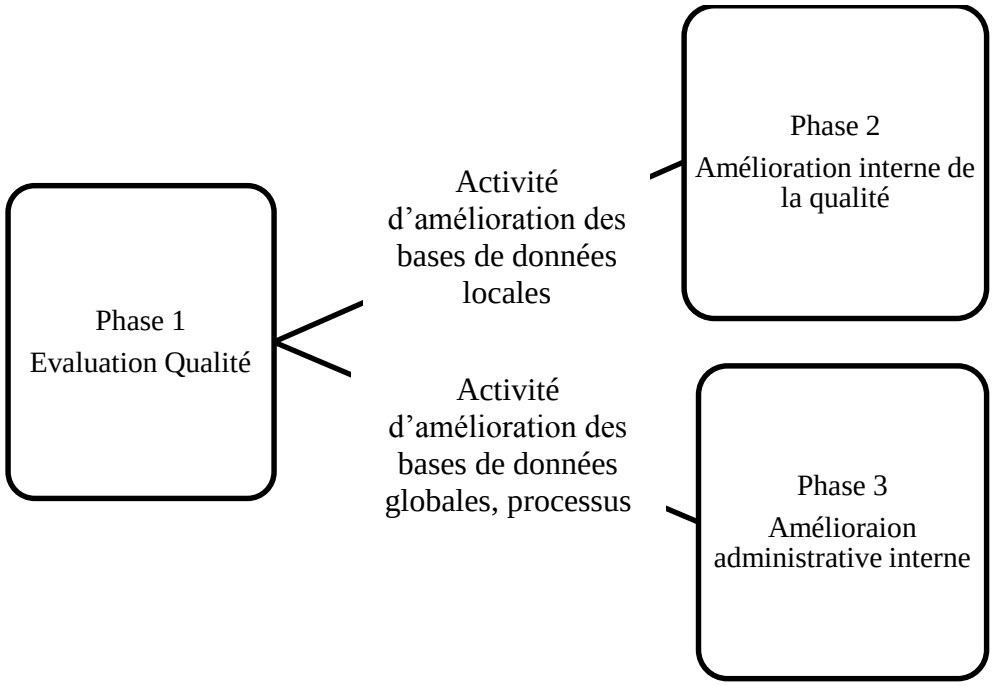


Figure 2.11 Principe général de la méthodologie Istat

Ainsi, selon la méthodologie Istat, l'évaluation globale de la qualité comprend l'analyse des besoins en qualité, l'isolation des dimensions qualité les plus adéquates aux données de type « Adresses/Localisation » (par exemple exactitude et complétude), la détection des données les plus critiques en utilisant des techniques de statistique (choix d'une base nationale, choix d'un échantillon représentatif, détection des données critiques, définition des causes potentielles d'erreur) et la communication des résultats d'évaluation aux différentes agences qui constituent le système d'informations collaboratif.

L'amélioration interne de la qualité des données (au niveau de chaque agence, initiative autonome) passe par la conception et la modélisation des différentes solutions d'amélioration des processus (standardisation du format d'acquisition, standardisation du format d'échange interne en utilisant le langage XML), par l'évaluation locale spécifique de la qualité et la conception et par la modélisation des solutions d'amélioration des processus sur les données

critiques (utilisation des résultats de l'évaluation globale et locales pour décider des interventions spécifiques à effectuer sur les processus internes et étude des résultats de l'évaluation globale et des outils acquis pour décider des interventions à effectuer sur les données, fusion des bases de données internes par exemple).

Enfin, l'amélioration de la qualité des flux inter-organisationnels concerne la standardisation des flux inter-administratifs en utilisant le format XML et la re-modélisation de l'échange des flux en utilisant une architecture publique orientée événements.

iv- Approche CDQ (Complete Data Quality methodology)

Cette approche s'applique à tous les types de connaissances : les organisations, les processus fonctionnels, les services offerts par ces processus, les normes et les règles techniques et fonctionnelles qui régissent le fonctionnement de ces processus, la qualité de ces processus, macro-processus et normes, les données (internes ou externes au système d'informations) ainsi que les dimensions et métriques qualité de données. Elle s'articule en trois phases [Batini et al. 06].

La première phase consiste à recenser l'ensemble des flux organisationnels les plus importants, les données qui y sont échangées, les processus exécutés ainsi que les normes régissant leurs fonctionnements et les services qui en sont générés. Ainsi, elle vise, dans un premier temps, à la reconstitution des flux de données inter-organisationnels les plus pertinents, ainsi qu'à l'établissement des matrices correspondantes (décrivant les organisations et les flux), puis, dans un deuxième temps, à la reconstitution des processus métier des organisations les plus pertinents et la construction des matrices correspondantes, et enfin, à la reconstitution des normes ainsi que les règles qui les contrôlent et les services fournis, et ce, pour chaque processus ou groupe de processus.

La deuxième phase concerne l'évaluation quantitative de la qualité des données et se fait en collaboration avec les utilisateurs internes et finaux des données de l'organisation pour l'identification des problèmes majeurs engendrés par la faible qualité des données. Les processus sont, en effet, scrutés permettant de remonter aux causes et origines des problèmes qualité constatés. Au final, l'analyse génère un ensemble de dimensions et métriques qualité permettant la mesure de la qualité des flux de données organisationnels et des bases de données d'une part, et la définition des objectifs qualité visés par le programme d'amélioration, d'autre part.

La troisième phase définit les processus optimaux pour l'amélioration de la qualité des données. Ainsi, pour chaque base ou flux de données, elle consiste à définir le nouveau niveau de qualité qui améliorerait la qualité du processus et réduirait les coûts en fonction des moyens disponibles, elle modélise les processus de réajustement des activités et choisit les activités qualité qui amélioreraient les objectifs fixés dans l'étape précédente. Par ailleurs, elle choisit des techniques optimales pour les activités qualité, affecte à chaque matrice de données et activités un ou plusieurs processus d'amélioration candidats, et calcule, pour chaque processus d'amélioration, le coût/gain approximatif pour choisir le processus optimal et vérifier que le coût global rejoigne les objectifs fixés.

CDQ se limitant à l'évaluation et à l'amélioration des données structurées, elle a été étendue, cinq ans plus tard, aux données semi-structurées et non structurées. Le modèle HDQM (Heterogeneous Data Quality Methodology) se base sur les mêmes étapes de recensement des connaissances organisationnelles, d'évaluation quantitative de la qualité et d'amélioration quantifiée en terme de ratio qualité/coût des connaissances en question, les dimensions utilisées se reposant de plus en plus sur l'aspect contextuel de l'évaluation (par exemple, pour les données géographiques, la dimension d'exactitude évaluée est celle de l'exactitude spatiale) [Batini et al. 11].

v- Approche DQMDW (Data Quality Meta DataWarehouse)

L'approche DQMDW s'intéresse aux problématiques de surveillance qualité, nettoyage des données et assurance qualité dans la phase d'intégration des données [Maydanchik 07]. Elle distingue quatre catégories des métadonnées qualité :

1 les métadonnées des agrégations, qui fournissent des informations sur la qualité au niveau d'agrégation le plus élevé (définitions, mesure des scores d'agrégation, sommaire et rapports qualité) ;
2 les métadonnées des règles, qui fournissent des informations sur les règles qualité ainsi que des résultats de leur implémentation ;
3 les métadonnées générales, qui fournissent des informations sur la qualité de chaque enregistrement (ou sujet) ;
4 les métadonnées atomiques, qui décrivent le contenu, la signification ainsi que la structure de la donnée elle-même.

L'approche DQMDW s'organise conformément au schéma de la Figure 2.12.

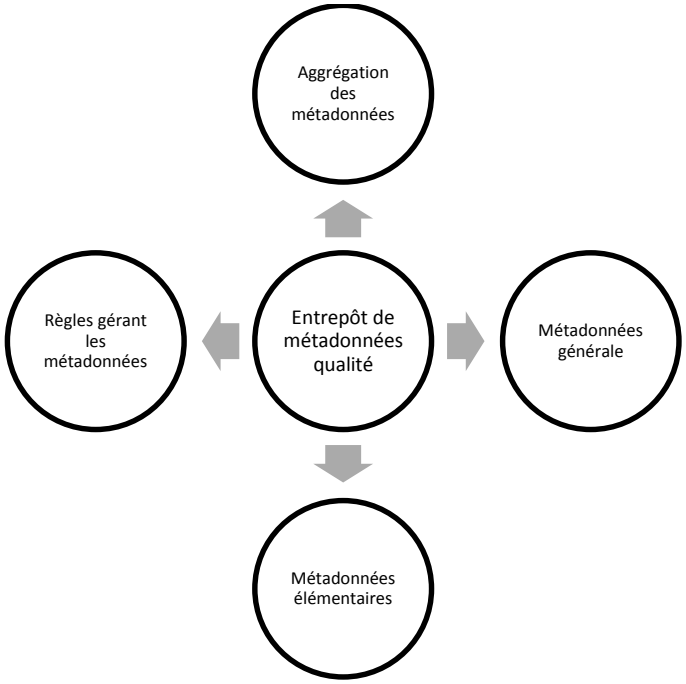


Figure 2.12 Catégories de métadonnées de l'approche DQMDW

2.4. Discussion

Nous décrivions, dans cette section, les outils d'évaluation qualité, à savoir les dimensions et métriques qualité, ainsi que leur déploiement dans le cadre d'un processus de contrôle ou d'assurance qualité d'une organisation ou d'un système d'information donnés.

Les méthodologies décrites dans ce contexte fournissent un ensemble d'instructions logiques et intuitives permettant aux managers d'intégrer les connaissances à évaluer en les guidant, de cette manière, dans la prise de décisions complexes. Le Tableau 2.4 résume les caractéristiques des méthodologies d'évaluation et/ou d'amélioration de la qualité des données auxquelles nous nous sommes intéressés. Elles sont évaluées conformément aux critères de comparaison établies par [Batini et al. 06] (Section 2.3.1).

Méthodologies	**Stratégie d'amélioration**	**Objectif attendu**	**Envergure**	**Envergure**
TQdM	Processus seulement	Evaluation et Amélioration	Activité : intégration des données	Intra-organisationnelle

Méthodologies	Stratégie d'amélioration	Objectif attendu	Envergure	Envergure
TDQM	Données et Processus	Evaluation et Amélioration	Activité : intégration des données	Intra-organisationnelle
Istat	Données	Evaluation et Amélioration	Activité (données de type « Adresse/Localisation »)	Inter-organisationnelle
CDQ	Données et Processus (tous types de connaissance de l'entreprise)	Evaluation et Amélioration	Générale	Inter-organisationnelle
DQMDW	Données seulement	Evaluation	Activité : intégration des données	-

Tableau 2.4. Comparaison des méthodologies qualité

Dans la pratique, les recommandations et les instructions doivent être considérées avec un regard critique. Le mot d'ordre est l'instanciation d'une méthodologie simple qui soit à la fois compréhensible par les différents acteurs du projet d'évaluation et adaptée aux types de connaissances sujets à cette évaluation. C'est pour cette raison que la plupart des exemples d'évaluation des données organisationnelles utilise plutôt la méthodologie TDQM, la plus ancienne des approches étudiées. Cette ancienneté permet en effet de nombreux retours d'expériences et témoignages publiés par les entreprises ayant eu recours à cette méthodologie, améliorant ainsi son utilisabilité et accroissant son intuitivité [Kovac et al. 97, Davidson 04, Wijnhoven et al. 07]. Ceci dit, l'évaluation de la qualité des données n'est généralement pas un objectif en soi. En effet, la faible qualité des données est généralement perçue par la constatation d'un processus défectueux ou d'un produit (dans le sens produit informatique) défaillant. Dans le cas de cette thèse, l'évaluation se place comme prémisse à la sélection multisources des données en réponse à une requête utilisateur, garantissant de cette manière l'adéquation du

résultat, entres autres, aux exigences qualité de ce dernier. Ce cas d'étude est, en effet, un cas particulier de la problématique de la gestion des données multisources.

3. Brokering de données multisources

Nous nous intéressons dans cette partie à la sélection des données que nous désignons aussi par son appellation anglophone orientée métier : le *brokering*. Nous définissons d'abord son contexte métier et nous spécifions ensuite l'aspect scientifique de cette problématique, en insistant sur le rôle joué par la qualité des données, d'une part, et la prise en compte de la préférence des décideurs, d'autre part, dans l'assurance qualité de ce processus.

3.1. Brokering pour le marketing

Nous nous plaçons dans le cadre particulier de la construction de plans fichier dans le contexte de réalisation d'une campagne de marketing relationnel B-to-B. En effet, étant donnée une demande de ciblage, le plan fichier se construit suivant deux étapes principales : l'identification ou la sélection des prospects étant donnés le périmètre de ciblage et la qualification des données des prospects (c'est-à-dire l'enrichissement de la sélection) par la consolidation d'informations issues de sources différentes. Cette tâche d'élaboration de plan ficher est critique et s'avère de grande importance dans le sens où un mauvais ciblage, qui se définit généralement par un mauvais choix des critères de filtrage engendrant une sélection inappropriée des prospects, occasionne des pertes financières à l'entreprise car il affecte considérablement la rentabilité de la campagne sous-jacente. C'est pour cette raison que la plupart des entreprises sollicite le service de *brokers d'adresses* pour les accompagner dans la conception et l'établissement de plans fichiers.

Dans le contexte d'intégration de données multisources, le brokering de données se définit comme la sélection appropriée de données à partir d'un ensemble de fichiers multisources étant donné un certain nombre de critères de sélection. Nous décrivons dans ce qui suit les approches de brokering les plus répandues pour la gestion des problèmes de cohérences dans les bases de données multisources. Nous les classons selon leurs critères de sélection. Nous définissons ensuite le rôle de la qualité des données ainsi que la préférence dans le contrôle et l'arbitrage de cette tâche de sélection.

3.2. Approches de brokering

Plusieurs approches de brokering ont été proposées dans la littérature. Nous distinguons les approches de sélection basées sur les métadonnées de celles centrées sur les données elles-mêmes et des approches de sélection orientées schémas de données.

3.2.1. Approches de sélection basées sur les métadonnées

L'approche proposée par [Kashyap et al. 00] a été définie dans le cadre de la gestion des requêtes multimédia et tend à arbitrer la communication entre les brokers et les fournisseurs des données dans le contexte d'un système d'informations hétérogène, en adoptant un raisonnement basé sur les descriptions de métadonnées. Concrètement, la solution se base sur deux composantes principales : une composante de sélection du vocabulaire et une composante de sélection des métadonnées. La sélection du vocabulaire sert à traduire les requêtes des utilisateurs et utilise pour ce faire une ontologie de domaine. La sélection des métadonnées permet d'encapsuler le contenu des données en filtrant tous les détails non pertinents.

Par ailleurs, le brokering ainsi défini est enrichi par un certain nombre de composantes additionnelles. InfoHarness fournit un accès uniforme aux données indépendamment de leur format et de leur localisation grâce à l'utilisation de métadonnées indépendantes du contexte, lequel encapsule les données et les média hétérogènes et représente les informations dans un modèle objet [Shklar et al. 95]. Le système MIDAS (*Multi-sensor Image DAtabase System*) [McKeown et al. 77] est conçu pour sa capacité à gérer des métadonnées spécifiques au domaine, en particulier celles relatives à des données issues de différents types de média relatifs à l'image. Grâce à sa plateforme à base de métadonnées, MIDAS permet de corréler les informations hétérogènes et forme ainsi une approche basique de gestion des métadonnées et de leurs associations respectives.

Les performances de MIDAS sont améliorées grâce à des systèmes comme InfoSleuth [Bayardo et al. 97] et OBSERVER [Mena et al. 96]. En effet, InfoSleuth étends la portée de MIDAS aux données textuelles en utilisant une ontologie appropriée et est utilisé comme base de brokering pour la gestion des données multisources. Ces sources sont alors décrites en utilisant des contextes conceptuels appelés *c-context* et associées aux données qu'elles décrivent moyennant des schémas de mise en correspondance (*mappings*). Par ailleurs, le système OBSERVER (*Ontology Based System Enhanced with (terminological) Relationships for Vocabulary hEterogeneity Resolution*) est, comme son nom l'indique, un système à base d'ontologies dont

les fonctionnalités sont améliorées par une association de terminologies résolvant l'hétérogénéité du vocabulaire. De ce fait, OBSERVER permet à ses utilisateurs de naviguer entre plusieurs ontologies et de percevoir, par conséquent, une vue sémantique conceptuelle de l'ensemble des données. La requête est ainsi d'abord réinterprétée et décomposée afin d'être exécutée et les résultats des sous requêtes ainsi formées reformatés et agrégés.

Dans cette approche de sélection orientée métadonnées, nous assistons à un brokering des données au niveau du contenu de l'information où l'évaluation de la perte des informations sollicite la préférence des utilisateurs finaux. Cette préférence définit un compromis entre précision[11] et rappel[12] et évalue la perte d'information dans les résultats des requêtes ayant subi ces phases de transformation/découpage/assemblage. La perte est définie par la formule suivante :

$$Perte = 1 - \frac{1}{\propto \left(\frac{1}{Précision}\right) + (1-\alpha)\left(\frac{1}{Rappel}\right)}$$

Concrètement, étant donnée la formule de la perte d'informations, la préférence apparaît dans la spécification du coefficient $\propto$ variant dans l'intervalle [0, 1], qui dénote l'importance qu'accorde l'utilisateur à la précision. Cependant, malgré la sophistication de cette plateforme de brokering, le broker n'a aucune visibilité sur les données sélectionnées et, de ce fait, aucun moyen de contrôler la qualité de la sélection. Cet inconvénient est pallié dans les approches de brokering orientées données.

3.2.2. Approches de sélection orientées données

Il existe deux approches de sélection orientées données : la première solution permet de résoudre la problématique d'accès aux données multisources [Ardagna et al. 06], tandis que la deuxième gère les problèmes de cohérence rencontrés lors de l'intégration de sources d'informations hétérogènes [Motro et al. 06].

Dans l'approche proposée par [Ardagna et al. 06], des valeurs qualité sont associées aux valeurs des attributs (dans un contexte de bases de données), permettant par conséquent une intégration orientée qualité. Cette intégration comprend les activités de nettoyage des données telles que la normalisation (c'est-à-dire l'analyse de la forme et du contenu des données), la déduplication

[11] Précision : Dans le contexte de l'évaluation de la classification automatique, la précision est définie par le nombre d'éléments bien classés sur le nombre d'éléments classés dans cette même catégorie.

[12] Rappel : Toujours dans le même contexte de classification automatique, le rappel est défini par le nombre d'éléments bien classés sur le nombre total d'éléments à classer.

des données et la résolution des conflits engendrés par cette intégration. La préférence des utilisateurs se limite dans ce modèle à la spécification des seuils de tolérance/acceptabilité qualité. En effet, le broker choisit les données qui forment la réponse la plus complète et exacte possible, étant donnée une requête de l'utilisateur, et ce moyennant des opérations de jointure. Cependant, l'utilisation de l'unique dimension qualité d'exactitude est insuffisante à elle seule pour assurer l'efficacité des résultats. En effet, une adresse peut être correcte mais obsolète et, par conséquent, inutilisable. De plus, l'évaluation de cette dimension d'exactitude n'est pas d'une simplicité évidente, surtout lorsqu'il s'agit d'évaluer des adresses email qui sont basées sur des pseudonymes et autres informations informelles et non standardisées pour la plupart. Ainsi, la définition d'autres critères qualité s'avère indispensable pour garantir une sélection de qualité acceptable.

La deuxième approche concerne *Fusionplex* [Motro et al. 06], un système d'intégration basé sur la préférence. Des requêtes multisources y sont utilisées pour l'intégration des données et se basent sur des routines de résolution de cohérence qui s'exécutent en deux passes. La première passe consiste à ordonner les enregistrements doublons selon la préférence des utilisateurs experts. Cette préférence est ainsi exprimée par des fonctions d'utilité subjectives, où chaque utilisateur affecte un vecteur de poids préférentiel respectivement au vecteur de métadonnées (dimensions) qualité. Ce vecteur exprime les critères qualité de récence, d'exactitude, de disponibilité et de coût (dimension exprimant à la fois le prix de la donnée et le délai de sa transmission à travers le système d'information hétérogène auquel cette donnée est rattachée). Dans la seconde passe, pour chaque doublon et chaque attribut, les valeurs doubles (sources d'incohérence) sont consolidées via des requêtes SQL. Ces requêtes font appel à des opérateurs simples tels que les opérateurs *max*, *min* et *avg* et traduisent des fonctions d'utilité telle que la moyenne simple pour l'opérateur *avg* par exemple.

Cette approche, bien que palliant la mono-dimensionnalité de l'approche d'[Ardagna et al. 06], a l'inconvénient de se baser sur de simples fonctions d'utilité limitées aux fonctions d'agrégations simples supportées par le langage SQL, qui pourraient ne pas tout à fait correspondre aux véritables préférences des utilisateurs, beaucoup plus complexes.

Notons que la frontière entre les approches de brokering basées sur les métadonnées et celles centrées sur la donnée est très fine vu que les approches orientées données sont principalement basées sur des dimensions de la qualité des données, lesquelles sont finalement des métadonnées.

3.2.3. Approches de sélection orientées schémas

Les approches de brokering orientées schéma sont surtout connues comme des approches logiques où l'intégration se base sur une architecture à trois couches : la couche des sources des données, la couche de médiation et la couche applicative. Comme nous l'avons déjà présenté, trois approches d'intégration se démarquent : les approches LAV, les approches GAV et les approches hybrides BAV. Quelle que soit l'approche adoptée, l'exécution des requêtes nécessite leurs reformulation/transformation/réécriture. Ainsi, dans *Information Maniford*, système d'intégration de type LAV [Levy 98], la reformulation des requêtes nécessite la présentation d'un plan générateur de requêtes afin d'écarter les sources non-pertinentes, de diviser la requête en sous-objectifs, de générer un plan fichier (plan de requêtage) conjonctif et de définir un ordre d'exécution des sous objectifs en question. Dans le système d'intégration de type GAV *TSIMMIS* (*The Stanford-IBM Manager of Multiple Information Sources* [Chawathe et al. 94]), la reformulation des requêtes est basée sur des adaptateurs (*wrappers*) qui jouent le rôle de connecteurs de bases de données et gèrent, de ce fait, la communication entre le schéma médiateur global et les sources de données. Par ailleurs, l'approche d'association de schémas de type BAV définie par [Rizopoulos 10] (Section1.1.2) est aussi basée sur la préférence des utilisateurs. Cette préférence est en effet requise pour gérer les incertitudes perçues lors de la mise en correspondance des schémas et consiste à définir un ordre préférentiel entre les alternatives concurrentes étant données ces *mappings*.

3.3. Rôle de la qualité des données

Les approches de brokering que nous avons décrites illustrent l'importance de la qualité des données dans la résolution des problèmes d'incohérence et de conflits rencontrés lors de l'intégration ou le requêtage d'un système d'information multisources. Ainsi, des dimensions telles que l'exactitude, la fraîcheur, la disponibilité et le coût sont souvent utilisées. Par exemple, Cholvy utilise un processus d'évaluation basé sur le modèle *STANAG 2022* (*STANdardization AGreements 2022* [Cholvy 04]) qui analyse la valeur du couple (fiabilité de la source, crédibilité de la source) déduite des utilisations et 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. 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 évaluent la qualité des données intrinsèques. On y privilégie les données récentes, cohérentes et syntaxiquement exactes [Bleiholder et al. 08].

La qualité des données intervient aussi dans l'évaluation des résultats des requêtes de médiation. En effet, [Kostadinov et al. 05] proposent 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 leurs préférences. Les dimensions qualité choisies sont alors la fraîcheur, le délai et le coût.

Enfin [Batista et al. 07] définissent, toujours dans le cadre de l'amélioration de la qualité des résultats dans les systèmes d'intégration des données, l'algorithme *IQ Manager* d'amélioration de la minimalité du schéma d'intégration, qui se base sur l'élimination de la redondance au niveau des entités, des relations et des attributs. Cet algorithme consiste alors à analyser trois métriques.

- La complétude du schéma est définie par 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 est basée sur la redondance des entités du schéma et la redondance des relations.
- La cohérence des types inclut 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.

En se basant sur l'analyse de ces critères et des scores qualité obtenus, *IQ Manager* procède à des ajustements du schéma d'intégration cible afin d'améliorer sa conception et, par conséquent, l'exécution de la requête.

3.4. Rôle de la préférence

Tout comme la qualité, la préférence des experts métier a également été sollicitée pour la résolution des problèmes d'intégration, au point que certaines approches s'en sont inspirées pour développer un langage de requêtage inspiré de SQL. La difficulté majeure réside dans la gestion de flexibilité apportée par la contrainte de préférence.

En effet, le langage SQLf [Bosc et al. 95] utilise pour ce faire des conditions floues. Ces conditions expriment les préférences des utilisateurs, moyennant des prédicats flous décrits dans la clause WHERE de la requête SQL. Chaque attribut se voit attribuer une valeur dans l'intervalle flou [0, 1]. Par exemple, dans la requête de jointure suivante : *Select R.A, R'.B From R, R' Where c1(R) = c2(R') and R.att1 = R'.att2*, *c1* et *c2* sont deux conditions floues appliquées respectivement aux relations R et R'. Concrètement, étant donnée une relation (voyage, coût, durée), l'utilisateur peut définir un attribut flou (*c1*=rapidité_voyage) et lui affecte un poids entre 0 et 1 traduisant son appréciation de la durée du voyage. L'utilisateur peut aussi créer un attribut flou (*c2*=confort) qui exprimera sa perception du confort en fonction de la durée du voyage. Il pourra ensuite utiliser ces deux attributs (rapidité_voyage,confort) pour sélectionner les voyages courts et confortables.

Ce langage a été étendu pour gérer la bipolarité des préférences. Nous distinguons alors deux types de préférence : une préférence obligatoire et une préférence optionnelle [Tamani et al. 11]. Toujours dans ce contexte, [Liétard et al. 09] expriment la bipolarité en utilisant les expressions lexicographiques où les préférences obligatoires et optionnelles sont séparées et agencées suivant un ordre ascendant ou descendant et où les requêtes s'expriment moyennant une version étendue du langage *SQLf*. Les préférences optionnelles sont ainsi décrites dans une clause *Then*.

Un autre langage similaire à *SQLf, PreferenceSQL* suit le modèle bipolaire. Les préférences optionnelles y sont exprimées dans une clause *Preferring* [Kiebling et al. 02].

Le Tableau 2.5 résume les avantages et inconvénients de ces approches de modélisation de la préférence.

Approches/Critères de performance	**Représentation binaire**	**Représentation bipolaire**	**Représentation multidimensionnelle**
Avantages	Simplicité	Absence de préférence stricte	Grande capacité de représentation des préférences
Inconvénients	Possibilité de contrastes avec les préférences des utilisateurs	Seulement deux niveaux de préférence	Ambigüité en cas de divergences élémentaires dans le n-uplet de préférence ➔ nécessite le calcul

Approches/Critères de performance	Représentation binaire	Représentation bipolaire	Représentation multidimensionnelle
	Faible capacité de représentation des préférences		d'un score de préférence global Complexité de calcul du score global de préférence

Tableau 2.5 Comparaison des approches de modélisation de la préférence

Nous remarquons que plus la modélisation est simple, moins elle représente la préférence des utilisateurs, et que plus elle est fidèle à ces préférences, plus son utilisation à des fins de brokering est complexe et ambigüe. Un compromis doit donc être trouvé entre la fiabilité de représentation des préférences et la simplicité de leur modélisation.

3.5. Discussion

Nous décrivions, au niveau du chapitre introductif, la méthodologie de brokering manuelle telle qu'elle est utilisée dans les départements marketing des annonceurs, où des brokers humains sélectionnent les données de ciblage étant donnée une base de prospection multisources. Nous constations que cette solution de brokering était assez controversée à cause de sa forte subjectivité puisque les données choisies dépendent, d'une part, de la réputation de leurs fournisseurs (réputation acquise par les courtiers grâce à leur expérience) et, d'autre part, des taux de commissions perçus par les courtiers en fonction de la quantité de l'information achetée ou louée.

Alors, afin d'automatiser le brokering et réduire la subjectivité de la sélection manuelle, plusieurs approches sont proposées. Dans ces méthodologies de brokering automatisé, deux critères sont mis en évidence : la prise en compte de la qualité des données et la prise en compte de la préférence de l'utilisateur. Le Tableau 2.6 résume les caractéristiques des approches décrites précédemment (Section 3.2.1).

Approches/Critères de performance	OBSERVER [Kashyap et al. 00]	[Ardagna et al. 06]	[Motro et al. 06]	[Rizopoulos 10]
Type de brokering	Approche orientée métadonnées	Approche orienté donnée (basée sur l'étude l'exactitude des données)	Approche orientée données (basée sur l'étude de la qualité des données)	Approche orientée schéma
Validation	Evaluation du taux de perte des données.	Evaluation de l'exactitude des données.	Evaluation de la récence, de l'exactitude, de la disponibilité et du coût des données.	Validation croisée pour la détection des anomalies dues au *mapping.*
Rôle de la préférence des utilisateurs	Définit l'importance de la précision dans la formule d'évaluation du taux de perte des données.	Les utilisateurs spécifient les seuils de qualité critiques. Les utilisateurs définissent les poids des dimensions qualité	Chaque utilisateur définit sa fonction d'utilité étant donné les attributs qualité.	Les utilisateurs gèrent manuellement et itérativement les problèmes conflictuels au niveau des *mappings*.

Tableau 2.6 Approches de brokering

Nous remarquons que, dans la plupart de ces approches, la préférence est soit définie manuellement par l'expert humain (l'utilisateur) [Ardagna et al. 06, Motro et al. 06, Rizopoulos 10], ce qui empêche toute automatisation du processus, soit basée sur un ensemble de critères génériques qui ne sont pas adaptés au contexte de la sélection de données, ce qui altère la qualité de cette dernière, étant donnée l'absence de visibilité sur les données [Kashyap et al. 00]. Il est donc nécessaire de modéliser rigoureusement la préférence en se basant sur des critères

contextuels, afin de garantir l'adéquation et l'optimalité de la sélection des données dans un environnement multisources.

4 Conclusion

Les grands questionnements de cette thèse se rapportant à la double problématique de l'évaluation et de la sélection des données dans un environnement multisources, notre état de l'art s'est intéressé à la description des principales approches permettant l'accomplissement de ces objectifs. Ainsi, nous avons montré que le succès de telles méthodologies nécessite l'adoption d'une démarche basée sur la qualité des données. Les critères qualité sont, en effet, les principaux garants de l'objectivité et de la contextualité de l'évaluation de la base de données multisources dont il est question dans ce manuscrit, mais aussi de l'optimalité et de l'adéquation de la sélection des données issues d'une telle base.

Chapitre 3 : Evaluation de la qualité d'une base de prospection marketing

« La qualité n'est jamais un accident; c'est toujours le résultat d'un effort intelligent. »
John Ruskin

L'enrichissement du système d'information par des données externes à l'organisation, issues de sources différentes, altère considérablement la qualité de la base multisources ainsi formée. Un processus continu d'évaluation et de surveillance qualité s'impose ainsi comme méthodologie obligatoire à intégrer dans de tels systèmes d'information multisources. Cette intégration nécessite l'extension du modèle de représentation des données par des métadonnées portant sur la qualité, tels que le modèle d'évaluation de la fraîcheur des données de Bouzeghoub et Peralta [Bouzeghoub et al. 04], le projet *QUADRIS* visant une évaluation générique de la qualité des systèmes d'informations [Akoka et al. 08] ou bien le modèle *Polygen* permettant l'analyse de la provenance des données multisources [Wang et al. 90]. D'autres modèles tels que le *DDQ* (*Data and Data Quality*) étendent les modèles d'évaluation aux données semi-structurées en associant des dimensions qualité aux documents XML orientés données. Par ailleurs, les modèles de gestion des systèmes d'informations ont été étendus avec des couches de contrôle qualité telles que le modèle IP-MAP qui décrit le processus de production de l'information dans une organisation donnée et permet de pointer les processus responsables de l'altération de la qualité de l'information en remontant jusqu'aux sources [Shankaranarayanan et al. 00].

Ce chapitre décrit notre méthodologie d'évaluation qualité des bases de prospection multisources, laquelle est fortement inspirée du dénominateur commun aux différentes méthodologies d'évaluation qualité présentées dans le Chapitre 2 (Section 2.3.1). Ainsi, étant donnés des flux de données, un ensemble de bases de données internes et externes, des processus et macro-processus, un certain nombre de dimensions qualité et un budget, ces méthodologies fournissent un ensemble de processus reformés et ajustés, des bases de données mesurées et améliorées avec des coûts contrôlés et, éventuellement, des bénéfices. Ces méthodologies sont très bien structurées et détaillées. Les tâches qui y sont décrites vont de l'analyse des données et des besoins en qualité à la gouvernance de la qualité en passant par la modélisation du processus d'évaluation qualité. Afin de réaliser cette dernière tâche de modélisons du processus d'évaluation qualité, nous nous basons sur la méthode IP-MAP

(*Information Product MAP*) [Shankaranarayanan et al. 00]. En effet, comme nous l'annoncions dans le chapitre introductif, le processus général qui nous intéresse se ramène à la sélection des données multisources qui alimenteront le fichier de ciblage de notre annonceur Maisonphoning. Nous assimilons, alors, cette opération de sélection de cibles à un processus de production de l'information (l'information étant, en l'occurrence, le résultat de la sélection). Le choix de IP-MAP se justifie par la visibilité qu'amène ce modèle au processus de la sélection multisources, facilitant la mise en place des indicateurs de mesure et d'amélioration de la qualité des données.

Le plan du chapitre s'annonce comme suit : nous définissons tout d'abord les principes du modèle IP-MAP sur lequel se basera notre méthodologie d'évaluation qualité ; et nous décrivons, ensuite, les différentes phases de cette méthodologie d'évaluation. En particulier, nous nous concentrons sur la mesure de la qualité des données multisources, laquelle aura lieu en deux temps :

- une évaluation de la qualité intrinsèque des données où nous définissons les dimensions adéquates à l'entité mesurée ainsi que les métriques relatives ;
- puis, une évaluation de la qualité globale d'une donnée, d'un enregistrement, ou d'une sélection (un groupe d'enregistrements) qui permet d'apprécier la valeur réelle de l'entité évaluée.

1. Modèle IP-MAP

Le modèle IP-MAP [Shankaranarayanan et al. 00, Shankaranarayanan et al. 07] est une extension du modèle IMS (*Information Manufacturing System*) [Ballou et al. 98], créé pour rajouter au modèle classique de représentation des flux de données (Diagramme de Flux de Données, par exemple) la dimension qualité des données permettant l'évaluation de ces données sur la base de critères relatifs à leur fraîcheur, leur exactitude ainsi que leur complétude. IP-MAP enrichit, alors, IMS par la représentation systématique des processus de production des données. En effet, IP-MAP considère l'information comme un produit qui s'inscrit dans un cycle de production d'une organisation donnée et dont l'échange occasionne la transmission d'une valeur à son consommateur (utilisateur). Cette analogie de l'information produit nécessite, comme tout processus de production, l'utilisation de procédures de contrôle, d'assurance et d'amélioration de la qualité. Ainsi, IP-MAP se veut comme un modèle décrivant les routines qui permettent de contrôler la qualité de données brutes (provenant directement des sources) ou semi-traitées dans un contexte métier de production de l'information. La production

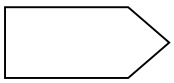
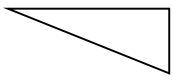
comprend la saisie, la sauvegarde, la création et l'échange des données au sein d'un ou plusieurs systèmes d'informations. De telles informations produit sont par exemple les factures clients, les fiches des étudiants, les relevés bancaires ou les certificats de naissances. L'avantage de l'utilisation d'un modèle tel que l'IP-MAP réside dans la visibilité contextuelle des données qu'il offre aux décideurs quelle que soit la complexité du système d'informations dans lequel se produisent et s'échangent ces données. En effet, cette visibilité permet aux décideurs d'avoir une idée sur les données brutes utilisées, les données nouvellement créées, les sauvegardes intermédiaires utilisées, la manière avec laquelle les éléments de données sont agrégés pour générer les données semi-traitées et les informations produit ainsi que les utilisateurs de ces informations produits ; en un mot, tous les processus, acteurs et données qui leurs permettent de mieux comprendre la qualité des informations produites selon leurs contextes d'utilisation.

Pour ce faire, IP-MAP se base sur huit blocs/concepts principaux.

- Le bloc source décrit les sources des données brutes utilisées dans le processus de création de la donnée et les unités fonctionnelles qui leurs sont associées. Le bloc source décrit ainsi le nom de la source (à savoir le département depuis lequel les données brutes sont issues), la localisation (qui réfère à l'endroit physique dans lequel la source a été stockée), les processus et règles métiers associés aux données ainsi que le format de stockage des données (format papier ou format numérique).
- Le bloc client représente le client de l'information produit. En effet, le client spécifie les caractéristiques du produit, à savoir l'ensemble des données le constituant ainsi que le département (ou unité fonctionnelle) dans lequel il va être déployé et l'entité qui l'utilisera.
- Le bloc qualité des données représente les contrôles qualité essentiels à effectuer pour assurer la production d'informations « zéro défauts ». Ce bloc analyse ainsi les données de production et génère deux sorties possibles : la probabilité *p* que les données soient correctes ou bien la probabilité *1 - p* qu'elles soient de mauvaise qualité.
- Le bloc traitement décrit l'ensemble des opérations et routines impliquant les données brutes ou utilisant des données semi-traitées pour la génération de l'information produit. Le bloc de correction des données en est un exemple. En effet, quand des problèmes qualité sont signalés au niveau des données suite à l'exécution des blocs qualité des données, des actions correctives sont requises. De telles actions sont accomplies par le bloc correcteur de données.

- Le bloc de stockage décrit, entre autres, les processus et règles métiers qui régissent le stockage des données dans les fichiers ou bases de données, telles que les règles de gestion des accès ou publication des données stockées.
- Le bloc décisionnel est requis dans les systèmes de production complexes où une même donnée se voit rattachée à différents blocs afin d'alimenter plusieurs processus. Le bloc décisionnel est alors requis pour identifier les contraintes et règles métiers à valider avant d'associer cette donnée aux processus de production en question.
- Le bloc métier a pour rôle de contrôler l'acheminement de l'information produit quand celle-ci circule d'une unité fonctionnelle à une autre (ou d'une organisation à une autre). En particulier, le bloc métier définit les problèmes qualité pouvant éventuellement surgir lors de la migration de l'information afin d'assigner la responsabilité, en cas d'erreur, à l'unité fonctionnelle (ou organisation) adéquate.
- Le bloc système d'information définit les répercussions de la migration d'une donnée brute d'un système d'information à un autre (telle que la transcription des données du format papier au format digital ou la migration de la donnée d'un SGBD vers un autre). Il décrit alors l'ensemble des règles et des procédures qui régissent cette migration. En effet, dans certains cas, la migration s'effectue par des processus tels que l'email ou le FTP transférant, ainsi, les données d'un système à un autre et d'une unité fonctionnelle à une autre.

Le modèle IP-MAP est un modèle graphique dont les blocs sont modélisés conformément à la Figure 3.1.

Nom du bloc/concept	**Symbole**
Source (Entrée du processus de production de l'information produit)	
Client (Sortie du processus de production de l'information produit)	
Qualité des données (critères d'évaluation qualité)	

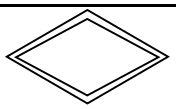
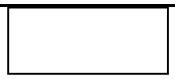
Traitement (Processus)	
Stockage des données et des informations	
Décision	
Métier (unité fonctionnelle ou organisation actuelle / nouvelle unité fonctionnelle ou organisation)	
Système d'information (système actuel / nouveau système)	

Figure 3.1 Blocs IP-MAP [Shankaranarayanan et al. 00]

Dans la suite de chapitre, nous décrivons notre méthodologie d'évaluation de la qualité des données en nous basant sur les principes du modèle IP-MAP.

2. Méthodologie d'évaluation de la qualité intrinsèque des données d'une base de prospection multisources

Pour définir notre méthodologie d'évaluation de la qualité, nous nous basons sur le modèle de diagnostic à dix phases présenté dans le Chapitre 2 (Section 2.3.1). Ce modèle regroupe, en effet, l'ensemble des bonnes pratiques constatées par Batini et Scannapieco après une étude des différentes méthodologies d'évaluation proposées dans la littérature [Batini et al. 06]. Une telle méthodologie nécessite les étapes suivantes :

1 l'analyse des données de la base de prospection ;

2 l'analyse des besoins en qualité des données ;

3 l'identification des données les plus critiques à analyser ;

4 la modélisation du processus d'évaluation cherchant à établir un modèle formel ou semi-formel ;
5 la définition des dimensions qualité adéquates ;
6 l'évaluation du coût de la non-qualité des données ;
7 l'estimation des bénéfices d'évaluation ;
8 l'assignation des tâches de gestion aux responsables de chaque activité faisant partie du processus général de production des données ;
9 l'assignation des responsabilités de contrôle des données aux responsables des données ;
10 l'utilisation des outils et techniques adéquates en respectant les contraintes du budget et les moyens de l'organisation en question.

Nous regroupons les étapes 8,9 et 10 en une seule étape que nous désignerons par : la gouvernance des données[13].

2.1. Analyse des données de la base de prospection

Cette phase consiste, tout d'abord, à examiner les documentations disponibles sur les données. Tout annonceur B-to-B de type Maisonphoning (Chapitre 1) héberge deux types de bases de données :

- la base de données des clients : rassemblant les données d'identification et les données comportementales de consommation et utilisées par le service marketing dans les programmes d'amélioration de la connaissance des clients pour l'amélioration des opérations de ciblage et la proposition d'offres personnalisées. Ces données sont fondamentalement internes à l'entreprise.
- la base de données des prospects : rassemblant principalement des données d'identification qui, contrairement aux données client, sont principalement externes à l'entreprise.

Nous nous intéressons aux bases de données des prospects. En France, ces bases sont généralement, alimentées par trois sortes de fichiers :

[13] La gouvernance des données est, en réalité, un concept beaucoup plus large que les trois tâches 8, 9 et 10 désignées ci-haut et désigne l'ensemble des activités assurant un cadre de contrôle qualité pour la gestion et la protection de ressources d'information clé à travers l'entreprise. Nous voulons en particulier insister sur les fonctions d'assignation de rôles et de responsabilités des acteurs de la donnée au sein l'entreprise.
En savoir plus sur http://www.lesechos.fr/idees-debats/cercle/cercle-74664-la-gouvernance-des-donnees-element-cle-dune-strategie-de-gestion-des-donnees-reussie-1016617.php?eLwRk8gyjOwcfe9X.99

- un fichier socle (la base SIRENE) fourni par l'INSEE[14] recensant toutes les entreprises et les établissements du territoire français décrites par un certain nombre d'informations telles que leur identifiant (SIRET), leur raison sociale, leur adresse, leur numéro de téléphone, leur activité, leur effectif salarié, etc. Ce fichier couvre exhaustivement tous les domaines d'activité, et est considéré, grâce à cette large couverture de l'ensemble des entreprises françaises, une base socle de rapprochement ;
- un (ou plusieurs) fichiers d'enrichissement de l'effectif salarié, des numéros de téléphone, des numéros de fax, de la nominativité ou encore des adresses emails des dirigeants de quelques entreprises françaises définies par leur SIRET ;
- un fichier client fourni par l'entreprise annonceur décrivant l'ensemble de leurs clients. Ce fichier est généralement appelé fichier repoussoir utilisé pour filtrer l'ensemble des clients de la base de prospection afin de ne pas les inclure dans les campagnes de recrutement.

La deuxième étape de cette phase consiste à interviewer les différents acteurs agissant sur la chaîne de production de la base de prospection dans le but de mieux comprendre le schéma logique de chacune de ces données. Pour ce faire, nous avions interviewé un responsable de la base SIRENE. Cette interview nous a permis d'avoir une idée générale sur l'architecture des flux de données et comprendre les principales règles de gestion.

2.1.1. Processus d'alimentation de la base SIRENE

Nous nous intéressons au processus d'alimentation des données de la base SIRENE. En effet, les événements qui affectent la vie des entreprises sont transmis par des liasses à l'INSEE par les CFE[15] (à l'exception des événements du secteur public et les demandes d'immatriculation de certaines associations). L'INSEE n'assure que la gestion du fichier SIRENE. Par contre, il ne se charge pas de la collecte des informations. Deux types de supports de transmission sont distingués :

- le support papier : données saisies et traitées par les directions régionales de l'INSEE ;
- le support numérique (EDI) aussi traité par les directions régionales de l'INSEE.

[14] INSEE : Institut National de la Statistique et des Etudes Economiques

[15] CFE : *Centre de Formalités des Entreprises qui* traite les formalités d'inscription, de modification ou de cessation des entreprises

L'INSEE date ces événements grâce à la date de réception de la liasse, la date de l'événement lui-même et la date de son traitement dans la base SIRENE.

2.1.2 Stockage des données dans SIRENE

Nous distinguons principalement trois bases :

- un répertoire SIRENE « de gestion » stockant les données SIRENE relatives aux entreprises et aux établissements (date de création, activité, numéro SIREN par exemple) ;
- une base de consultation : ce service tout public permet d'obtenir, pour toute entreprise immatriculée au répertoire SIRENE et pour chacun de ses établissements, une « fiche d'identité » comportant les informations à jour au répertoire la veille de la consultation ;
- la base de diffusion SIRENE dont une image trimestrielle est transférée aux clients.

2.1.3. Procédures de normalisation dans SIRENE

Une fois les données stockées, nous nous intéressons à la procédure de normalisation des données dans SIRENE. Ainsi, la normalisation des adresses dans SIRENE consiste à utiliser un automate de formatage des adresses selon la norme postale. Cette procédure est récemment mise en place (depuis 2008) et n'implique pas la vérification des adresses dans le fichier de La Poste (SIRENE n'a pas d'abonnements avec le fichier de La Poste). Par contre, les adresses les plus anciennes ne sont pas normalisées.

2.1.4. Qualité des données dans SIRENE

Les procédures d'assurance qualité de l'INSEE sont assez rudimentaires. En effet, lors de l'intégration d'une nouvelle demande d'immatriculation, l'équipe en charge du répertoire SIRENE effectue un contrôle sur l'émetteur des liasses (CFE) et vérifie la cohérence des données reçues avec les données de la base. Cette mesure de la cohérence se vérifie au moyen :

- d'opérations ponctuelles de vérification de cohérence : rapprochements avec des fichiers d'identification des impôts par exemple en vérifiant que le couple (SIREN, données d'identification) est identique dans les deux fichiers ;
- d'opérations de conformité correspondant à la vérification de certaines règles métiers dans les données de la base (exemple d'opérations Entreprise/Etablissement : une entreprise active doit avoir au moins un établissement actif) ;

- de routines de détection d'anomalies déclenchées suite à des réclamations des utilisateurs du système ou des signalisations d'incohérences faite par les diffuseurs de la base SIRENE ;
- de règles de priorisation hiérarchique. En effet, en cas d'incohérence avec les données au RCS[16], les informations du RCS sont considérées les plus pertinentes.

Par ailleurs, les doublons de la base SIRENE sont détectés au niveau des personnes physiques : la suppression de ces doublons est effectuée, tous les mois, par les gestionnaires de la base. Les doublons sont, en partie, dus à des problèmes d'identification lors de certaines créations d'unités.

Enfin, les indicateurs qualité mis en place par l'INSEE sont principalement d'ordre opérationnels. Nous citons, par exemple :

- les taux de fausses actives mesurées par la date du dernier évènement ;
- le nombre de « doubles » sur les personnes physiques ;
- le traitement des NPAI[17] : l'INSEE peut vérifier des NPAI détectés par l'EAE[18] (une des principales enquêtes auprès des entreprises conduites à l'INSEE) ou en traitant un quota de NPAI signalés par les clients (utilisateurs de la base SIRENE).

2.1.5. Mises à jour INSEE

La fraîcheur des données livrées par l'INSEE aux clients SIRENE peut aller de quelques jours à trois mois. En effet, la livraison d'une base de référence aux clients devient trimestrielle, à l'exception des informations sur l'effectif et les tranches de chiffre d'affaires qui sont recensés moyennant des traitements statistiques (et ne sont disponibles qu'en fin d'année comptable N pour l'effectif au 31 décembre de l'année N-1).

La mise à jour d'un champ n'implique pas forcément la mise à jour de tout l'enregistrement. Les sources de mises à jour sont constituées d'opérations menées à l'initiative de l'équipe en charge de SIRENE ou des retours d'utilisateurs via les EAE.

D'autres opérations sont, par ailleurs, entreprises pour l'amélioration du répertoire :

- demandes de validation des données sur les entreprises/établissements (pour les non répondant des EAE, par exemple) ;

[16] RSC : Registre du Commerce et des Sociétés
[17] NPAI : N'habite Pas à l'Adresse Indiquée
[18] EAE : Enquête Annuelle d'Entreprise

- envoi de questionnaires aux entreprises pour actualiser leur code activité suite au changement du code NAF (300 000 questionnaires envoyés) ;
- investigation sur l'activité d'une entreprise n'ayant pas eu d'événements depuis un certain nombre d'années.

2.2. Analyse des besoins en qualité

Nous proposons de collecter les suggestions des utilisateurs des données et des responsables métier quant aux causes possibles des erreurs, étant donné le contexte d'exploitation et de génération de l'information produit : la sélection des données dans la base de prospection multisources.

En effet, une structure multisources comme celle de la base de prospection implique des problèmes de qualité critiques principalement dus à l'environnement collaboratif imposé par l'approvisionnement de la base socle en fichiers externes multisources. Ainsi, les fournisseurs de données d'enrichissement (vendeurs et annonceurs) manquent de transparence sur leurs propres sources de données, la fraîcheur de leurs données, les fréquences de mises à jour et la manière avec laquelle ces mises à jour se sont déroulées. Par ailleurs, l'environnement collaboratif induit, à cause de la réplication des données, des situations de confusion conflictuelles lors de la mise en correspondance et du déboulonnage des données pour les besoins d'une requête utilisateur quelconque. De plus, les données véhiculées sont, dans la plupart des cas, non normalisées et nécessitent la mise en place de processus de nettoyage et reformatage. Ainsi, les problèmes ciblés par notre évaluation qualité sont rattachés aux domaines suivants.

- La provenance et l'origine de l'information sont nécessaires à l'évaluation de la fiabilité de l'information multisources. La provenance des données permet, en effet, d'étudier l'authenticité des données distribuées étant donné son ascendance et l'ensemble de ses diffuseurs intermédiaires.
- L'intégration et la mise en correspondance des données multisources sont des pratiques obligatoires pour l'efficacité de la prise de décision et l'amélioration du ROI[19] des opérations marketing faisant appel aux données multisources. Cette intégration se base

[19]Le ROI est la mesure de l'efficacité d'un investissement en termes de rentabilité. Il consiste généralement en un simple ratio comparant la valeur du coût de l'investissement avec sa rentabilité.

dans notre cas sur des critères qualité intrinsèques et d'autres critères d'intégration indépendants de la valeur de la donnée, telle que son prix.

- L'efficacité des requêtes interrogeant la base de prospection multisources se mesure par la qualité de la réponse délivrée suite à une requête utilisateur ainsi que par le délai de réponse.

2.3. Identification des données les plus critiques

La base de données à laquelle nous nous intéressons dans le cadre de notre approche d'évaluation est la base de prospection, qui est formée des différents flux présentés dans la Section 2.1. En effet, elle est principalement composée du fichier socle de l'ensemble des entreprises françaises, duquel sont exclues les entreprises clientes. Elle est ensuite enrichie par des noms de contacts et leurs coordonnées respectives. Les données formant cette base multisources sont, cependant, d'une importance relative variable, dépendant du contexte de son utilisation. Par exemple, le numéro de téléphone peut revêtir une grande importance lorsqu'il est sélectionné pour la réalisation d'une campagne marketing téléphonique. Il l'est, cependant, beaucoup moins dans le cadre d'une campagne courrier où il est uniquement sélectionné pour enrichir l'information sur les prospects sélectionnés. Etant donné ce fait, nous regroupons ces données en trois classes de criticité différentes :

1 les attributs d'adressage, qui correspondent concrètement aux données relatives aux canaux des campagnes (les attributs d'adresse pour les campagnes courrier, les attributs de téléphone pour les campagnes téléphoniques et les adresses email pour les campagnes d'emailing) ;
2 les attributs de sélection correspondant aux critères de filtrage de la campagne tels que l'activité de l'entreprise et la masse salariale ;
3 les attributs d'enrichissement (ou de qualification) tel que le chiffre d'affaire de l'entreprise.

Le processus de sélection auquel nous nous intéressons génère, à partir de la base multisources, un fichier de prospects potentiels qui n'est autre qu'une liste de cibles à contacter. Ainsi, les attributs d'adressage sont de criticité la plus importante, étant donnée leur importance pour véhiculer le message marketing à la bonne destination. Ils sont suivis, dans l'échelle de criticité, des attributs de sélection qui assurent la correspondance des prospects sélectionnés par la requête de ciblage à la cible prédéfinie par les responsables stratégiques, puis des attributs d'enrichissement dont l'objectif est d'améliorer la connaissance du prospect.

2.4. Modélisation du processus d'évaluation

Nous nous intéressons à l'évaluation des données de la base de prospection multisources. Pour ce faire, nous recensons l'ensemble des actions entreprises pour la mise en place d'une telle base. Ces actions se rattachent aux processus suivants :

- les processus de transfert des données des fournisseurs à l'hébergeur de la base de prospection
- les processus d'enrichissement des données par des indicateurs de fraîcheur (tels que la date d'intégration des données)
- les processus de normalisation des données (tels que la normalisation des adresses postales et la standardisation des numéros de téléphones, par exemple)

Le modèle IP-MAP de la Figure 3.2 décrit le processus de mise en place d'une telle base de prospection.

2.5. Dimensions qualité

L'analyse des besoins en qualité des données, exprimées au début de cette section (Section 2.2), nous amène à définir les dimensions qualité techniques suivantes. Nous distinguons alors les dimensions d'évaluation des sources de données et les dimensions d'évaluation des données intrinsèques.

1 Les dimensions d'évaluation des sources de données sont aussi appelées dimensions de gestion et permettent d'évaluer objectivement la fiabilité des fournisseurs des données.

2 Les dimensions d'évaluation de la qualité de données intrinsèques qui permettent de résoudre les situations conflictuelles causées par la présence des doublons intra et inter sources et révélées au niveau du moteur de requêtage lors de l'exécution des opérations de sélection en l'occurrence.

Ainsi, nous définissons les dimensions suivantes :

- les dimensions d'évaluation de la qualité au niveau des sources :
  - la réputation ;
  - la valeur ajoutée : une source qui fournit une information exclusive ;

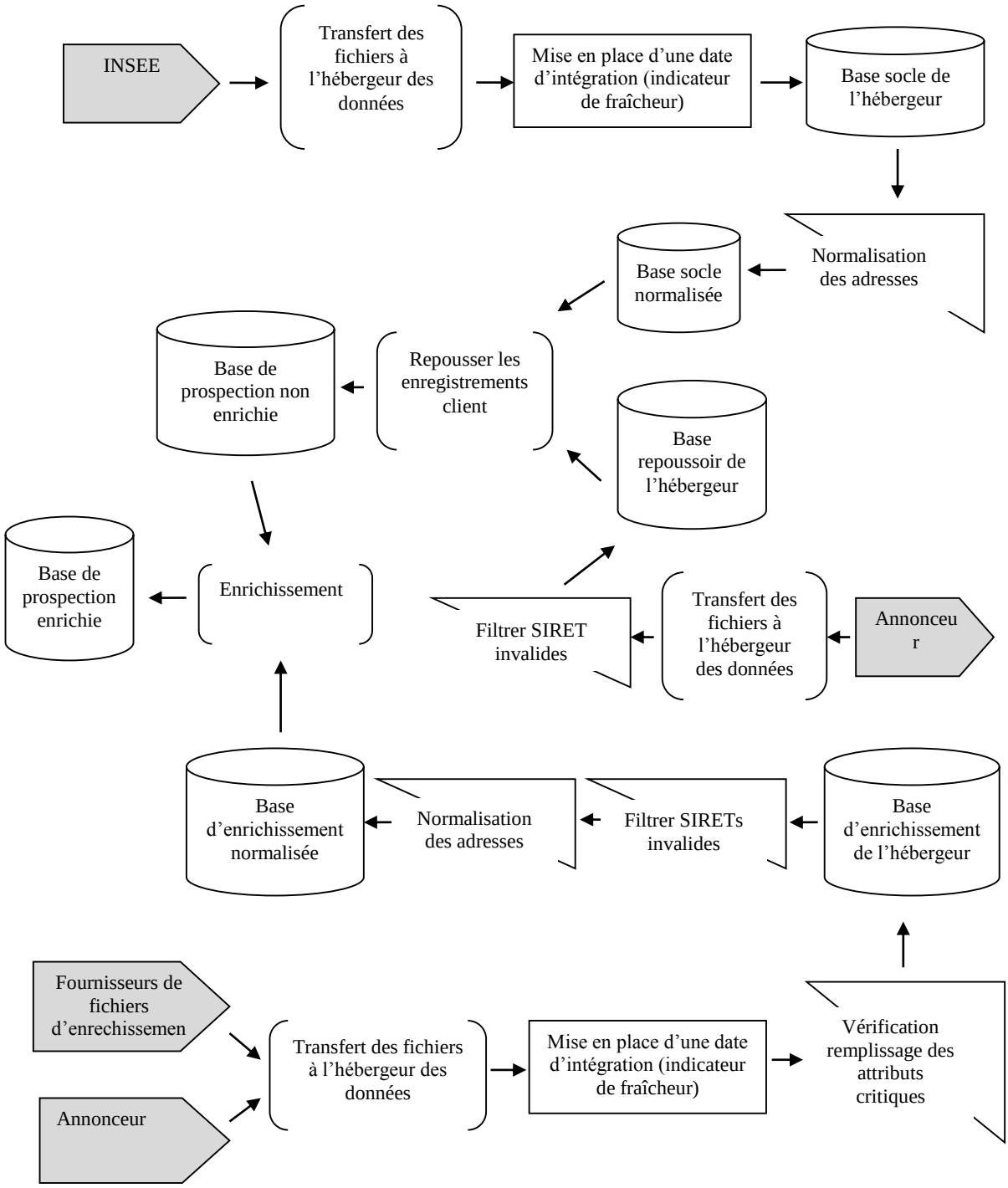


Figure 3.2 Mise en place de la base de prospection

– le prix, qui dépend de nombreux facteurs :

  - le niveau de détail des informations livrées (numéros de téléphone, adresses, emails, fax),
  - la spécialisation (par fonction ou secteur d'activité),

- la richesse des données (données financières très détaillées ou éléments précis sur le comportement d'achat du prospect).

Nous distinguons le prix de la donnée au niveau attribut, comme par exemple la donnée chiffre d'affaires de l'entreprise, et le prix de la donnée au niveau de l'enregistrement ;

- la fraîcheur des sources : c'est la fréquence de mises à jour ainsi que la manière avec laquelle ces mises à jour sont assurées. En effet, certaines données sont plus rapidement obsolètes que d'autres (l'email nominatif en est un exemple) ;
- le volume des données ;
- la fiabilité de la source, qui se définit par une agrégation des différents critères indiqués ci-dessus ;

- les dimensions de résolution de conflit (au niveau de la donnée intrinsèque) :
  - l'exactitude des données définie par :
    - l'exactitude syntaxique, qui vérifie la syntaxe de la donnée en question,
    - l'exactitude sémantique donnée par la probabilité que la valeur en question soit correcte et représente bien l'entité réelle qu'elle représente ;
  - la fraîcheur de la donnée, qui mesure l'âge des données par rapport à la date de dernière mise à jour de l'information en question. Notons ici l'appellation de la dimension où nous privilégions le terme *fraîcheur* au lieu de *timeliness* vu qu'elle est mesurée au niveau de la source et qu'elle servirait, une fois agrégée, à estimer la fraîcheur de la source en question. Notons que la fraîcheur d'un contact peut générer une probabilité plus ou moins importante d'avoir un contact obsolète, donc sémantiquement incorrect ;
  - la cohérence des données, qui se définit par l'absence d'ambiguïté au niveau des relations. Par exemple, si l'effectif d'un établissement est supérieur à l'effectif de l'entreprise dont il est rattaché, on dit qu'il y a un problème de cohérence des données ;
  - l'utilité qui peut être quantifiée soit subjectivement par une note accordée par l'expert selon l'intérêt qu'elle dégage, soit objectivement, moyennant les métriques suivantes :
    - le taux de non utilisation,

- le taux d'erreur (exemple : adresses erronées, adresses non normalisées) ;

– la complétude au niveau du champ et au niveau de l'enregistrement.

Outre ces dimensions techniques intuitives, afin d'assurer la contextualité et la subjectivité de notre évaluation conformément aux règles de *fitness for use* énoncées par Wang et Strong dans [Wang et al. 96], nous ajoutons un ensemble de dimensions empiriques définies par les utilisateurs potentiels de la base de prospection (à savoir, les responsables marketing qui vont définir les critères de ciblage et les brokers qui vont procéder à la sélection multisources). Nous avons, ainsi, établi un questionnaire sur « ce qui fait qu'un fichier de prospection soit de bonne qualité », que nous avons adressé à ces utilisateurs potentiels. Nous avons obtenu les dimensions suivantes :

- la fraîcheur des données,
- le volume de créations de la source par mois. Par exemple, une source avec 15 000 créations par mois est considérée de « bonne qualité » puisqu'elle a un taux considérable de renouvellement des noms et adresses,
- la rentabilité, qui se définit par la gestion de la pression commerciale dans les campagnes marketing. Une base moins sollicitée est, de ce fait, « plus rentable »,
- la complétude des informations clef, qui comprend :
  - la complétude de l'information nom et adresse de la société,
  - l'« utilisabilité » de l'information clef. En effet, en ce qui concerne les campagnes marketing B-to-B, l'utilisabilité se définit par le taux d'entreprises en cours d'activité dans le fichier SIRENE,
- la réputation de l'annonceur : il est plus facile de louer, auprès des courtiers les fichiers fournis par une entreprise de vente de mobiliers et fournitures de bureau que ceux fournis par une entreprise de vente d'accessoires pour sanitaires, étant donné que le domaine d'activité de la deuxième entreprise est moins prestigieux que celui de la première,
- la qualité du système de requêtage :
  - la complétude du résultat,
  - la fiabilité du système,
  - l'accessibilité / sécurité d'accès,
  - l'unicité des données en sortie,

- la disponibilité des données,
- la convivialité de l'interface.

Ainsi, le regroupement des dimensions empiriques et intuitives nous conduit à la définition de l'ensemble des indicateurs qualité qui permettent à la fois l'objectivité et la contextualité de l'évaluation. Le Tableau 3.1. regroupe l'ensemble de ces dimensions en trois classes différentes : les dimensions d'évaluation de la qualité des sources, les dimensions d'évaluation de la qualité de la donnée intrinsèque et les dimensions d'évaluation de la qualité du système de requêtage. Le Tableau 3.1. définit par ailleurs l'ensemble des métriques affectées à ces dimensions qualité.

2.6. Evaluation du coût de la non-qualité et des bénéfices de l'évaluation

Le coût de la non-qualité des données est impérativement perçu au niveau du ROI du processus évalué (en l'occurrence, la sélection de prospects potentiels à partir de la base multisources) puisqu'un message non véhiculé ou arrivant à une mauvaise destination n'est pas rentable. Par ailleurs, la sur-sollicitation d'une même cible entraîne sa lassitude qui s'exprime dans un premier temps par le non-intérêt au message de ciblage et, de ce fait, la perte du prospect. Dans un deuxième temps, la réputation de l'annonceur se dégrade. Une base de prospection polluée implique également des déficiences dans le processus de prise de décision et, de ce fait, des erreurs stratégiques.

Domaine	Dimension	Métrique	
Qualité des sources	Réputation	Information données par les experts en B2B. Poids initial attribué aux sources.	
	Valeur ajoutée	Précision des données : extra information (le nombre de champs en plus)	
	Prix	Prix du fichier source	
		Prix d'un enregistrement par fichier source	
	Fraîcheur (dans le sens *Timeliness* telle que définir par Bouzeghoub [Bouzeghoub et al. 04]	Indique la fréquence de création ou de mise à jour des données dans la source	Deux métriques sont définies : Fréquence de mise à jour : $Date\ Système - Date\ de\ dernière\ MAJ\ de\ la\ source$ (1) Volatilité des données : la durée pendant laquelle les données demeurent valides
	Complétude (Nous utilisons la notion de complétude telle que définie par Naumann [Naumann et al. 04]	Définie par 2 notions : la couverture et la densité.	$C(S) = c(S) * d(S)$ (2)
		La couverture est relative au nombre d'objets présents dans la source relativement à la réalité et décrit ainsi la complétude du schéma (schema completeness).	$c(S) = \frac{\lvert S \rvert}{\lvert W \rvert}$ (3)

Domaine	Dimension	Métrique	
		La densité décrit le nombre des valeurs et attributs manquants.	$d(S) = Avg\big(d(Ai)\big)$ (4)
	Volume de données (nous utilisons la définition de Pipino [Pipino et al. 02]	Le degré d'adéquation du volume de données par rapport à la tâche en question	$\min(\frac{volume\ de\ données\ fourni}{volume\ de\ données\ dont\ on\ a\ besoin}; \frac{volume\ de\ données\ dont\ on\ a\ besoin}{volume\ de\ données\ fourni})$ (5)
	Fiabilité[20]	La fiabilité de la source à un instant *t* est définie par une fonction de pondération des différentes mesures citées ci-haut.[21]	$F_i^t = \sum_{i,j} v_{i,j}(t) \times h_{i,j}(t)$ (6) où : *i* est relatif à la source *j* est un indice relatif à la dimension $v_{i,j}(t)$ est la valeur de la dimension *j* relative à la source *i* à un instant *t*

[20] Nous distinguons la fiabilité de la source et la fiabilité des dimensions (c'est-à-dire la fiabilité des métriques utilisées pour évaluer une métrique donnée. Par exemple, nous citons la métrique de **Cronbach's alpha** que nous détaillons ci-après)

[21] Cette valeur peut être comparée à une valeur estimée calculée par des méthodes de lissage exponentiel. La valeur prédite est déduite à partir de la valeur de la fiabilité à l'instant *t-1*.

Domaine	Dimension	Métrique	
			$h_{i,j}(t)$ est le poids attribué à la dimension *j* de la source *i* à un instant *t*
Qualité des données	Exactitude (Nous utilisons la notion d'exactitude évoquée par Batini et Scannapieco [Batini et al. 06]	C'est la proximité d'une valeur v à une valeur v' considérée comme la représentation correcte du phénomène du monde réel que la valeur v vise à représenter. Batini et Scannapieco distinguent 2 notions : l'exactitude et l'identification.	Métrique pour les erreurs ne gênant pas l'identification $\sum_{i=1}^{N} \frac{\beta((q_i > 0) \times (s_i = 0))}{N}$ (7)
			Métrique définissant les erreurs affectant l'identification $\sum_{i=1}^{N} \frac{\beta((q_i > 0) \times (s_i = 1))}{N}$ (8)
			Métrique calculant le taux des enregistrements exacts $\sum_{i=1}^{N} \frac{\beta((q_i = 0) \times (s_i = 0))}{N}$ (9)
	Récense et cohérence temporelle (timeliness) (Nous utilisons la définition de la	C'est la probabilité que la valeur d'un attribut donné soit à jour.	$Q^{w}_{temps}(t, w_1, \dots, w_n) = P^{w}(\mathrm{T} \geq \mathrm{t} \mid W_1 = w_1, \dots, W_n = w_n)$ $= 1 - P^{w}(T < t \mid W_1 = w_1, \dots, W_n = w_n)$ $= 1 - F^{w}(t \mid w_1, \dots, w_n)$ $= 1 - \int_0^t f^{w}(\theta \mid w_1, \dots, w_n) d\theta$ (10) Où W : la valeur de l'attribut en question

Domaine	Dimension	Métrique	
	récence utilisée par Heinrich [Heinrich et al. 09]		t : l'âge de la donnée = $t_1 - t_0$ w_i: données supplémentaires desquelles dépend l'attribut en question T : la durée de vie de la variable Par ailleurs, Q^w_{temps} est aussi définie par l'équation suivante : $$Q^w_{temps}(t) = e^{(-déclin(A).t)}$$ Où : déclin(A) : le déclin du taux relatif à l'obsolescence de l'attribut A pour une période de temps t donnée.
	Cohérence (Nous adoptons la définition de Batini [Batini et al. 06])	Cette dimension détecte les violations des règles sémantiques (dont les contraintes d'intégrité)	Techniques de détection d'anomalies (Exemple : règles de décision, règles d'associations)
	Utilité (Nous adaptons la définition de	L'utilité de l'information est déterminée en fonction des objectifs et attentes des utilisateurs et définie par le degré de	Mesure traduisant la perte d'utilité d'un attribut étant donnée sa qualité. L'utilité exprimée dans cette équation est dépendante du contexte :

Domaine	Dimension	Métrique	
	Shankaranarayanan et Even sur l'utilité des ressources [Even et al. 08])	pertinence des données vis-à-vis de la tâche en question. Pour Shankaranarayanan et Even, elle est directement rattachée à la qualité de l'entité évaluée. En effet, une mauvaise qualité de l'entité traduit la perte de son utilité [Even et al. 08].	$$Q_m^D = \frac{\sum_{n=1}^{N} u_n q_{n,m}}{\sum_{n=1}^{N} u_n} = \frac{1}{u^D} \sum_{n=1}^{N} u_n q_{n,m} \quad (11)$$ Mesure traduisant la perte d'utilité d'un attribut étant donnée sa qualité. L'utilité exprimée dans cette équation est indépendante du contexte (c'est-à-dire $u_n = \frac{u^D}{N}$) : $$Q_D = \frac{\sum_{n=1}^{N} u_n Q_n}{\sum_{n=1}^{N} u_n} = \frac{1}{u^D} \sum_{n=1}^{N} u_n Q_n \quad (12)$$ où : N : nombre d'enregistrements u_n : mesure non négative traduisant à l'utilité de l'entité mesurée ($u_n \geq 0$) $q_{n,m}$: la mesure qualité d'un attribut (qui décrit le degré de déficience d'un enregistrement donné vis-à-vis de la dimension qualité en question) Q_n : la mesure qualité d'un enregistrement donné Nous soulignons que la définition de l'utilité telle que décrite dans (12) est très semblable à la fiabilité (Equation (6)). Il suffirait de remplacer la mesure d'utilité par le poids

Domaine	Dimension	Métrique	
			d'importance de la mesure qualité pour retrouver la même équation.
Qualité du système de requêtage	Complétude du résultat	Nous reprenons la notion de complétude dans le sens densité définie ci-haut où la complétude est définie comme le taux de champs remplis	$$Q_{completeness} = \frac{\sum_{i=1..N} P(i)}{N} \quad (13)$$ où N est le nombre de champs contenus dans le résultat de la requête $$P(i) = \begin{cases} 1; si\ le\ champ\ i\ est\ renseigné \\ 0; si\ le\ champ\ i\ est\ vide \end{cases}$$
	Fiabilité du système	C'est l'aptitude du système à calculer la requête et générer les résultats en une durée de temps bien définie (ne dépassant pas les 20 secondes dans notre cas). C'est la probabilité de n'avoir aucune défaillance à un instant *t* donné de l'exécution de la requête.[22]	$$F(t) = e^{-\frac{t}{MTTF}} = e^{-\lambda t} \quad (14)$$ où : λ : taux de défaillance du système $MTTF\ (ou\ MTBF)$: Mean Time To failure (ou Mean Time Between Failure) telle que: $$MTTF = \frac{\sum(temps\ de\ fonctionnement - temps\ de\ pannes)}{nombre\ de\ pannes + 1} \quad (15)$$
	Accessibilité VS Sécurité d'accès	Unicité des données en sortie	

[22] http://fr.wikipedia.org/wiki/Fiabilit%C3%A9

Domaine	Dimension	Métrique
		Disponibilité des données
		Convivialité de l'interface

Tableau 3.1 Les métriques d'évaluation qualité

Prenons l'exemple de l'incohérence des effectifs entreprise, attribut de sélection crucial faisant partie des principaux critères de ciblage, donnés par deux sources s_1 et s_2. Un tel problème de cohérence définit deux cas de figures :

- soit les effectifs appartiennent à la même tranche[23] (attribut utilisé lors de la sélection de la phase 2), dans ce cas le problème de cohérence est anodin et n'a pas de graves conséquences sur la qualité de la sélection ;
- soit les effectifs n'appartiennent pas à la même tranche, dans ce cas, le problème de cohérence est plus sérieux puisque le choix de la mauvaise valeur induira des pertes financières conséquentes quant à la rentabilité de la campagne marketing en question.

2.7. Gouvernance

Cette dernière phase est relative à l'assignation des tâches de gestion aux responsables de chaque activité faisant partie du processus général de production des données, l'assignation des responsabilités de contrôle des données aux responsables des données et l'utilisation des outils et techniques adéquats en respectant les contraintes du budget et les moyens de l'organisation. Dans notre contexte, deux personnes sont affectées à ce projet : un ingénieur informatique faisant aussi office de consultant junior en qualité des données et un directeur qualité.

2.8. Discussion

Nous avons établi dans cette section les bases de notre méthodologie d'évaluation de la qualité des données où nous avons analysé l'ensemble des ressources (données, processus et flux) impliquées dans la création de la base de prospection multisources, et recensé l'ensemble des dimensions et métriques qui nous permettent de quantifier la qualité intrinsèque des entités mesurées. Dans la suite de ce chapitre, nous nous focalisons sur ce dernier point de mesure de la qualité des données de la base de prospection ainsi formée, où nous proposons d'apprécier sa qualité globale à travers une agrégation contextuelle des métriques intrinsèques. La contextualité est assurée par l'intégration de la dimension de préférence des utilisateurs dans l'agrégation. Nous montrons, en effet, la complexité de l'évaluation de la valeur réelle de l'entité mesurée par la simple analyse individuelle de ses métriques (surtout lorsque ces

[23] Tranche telle que définie par le champ TEFEN de l'INSEE. Par exemple, les effectifs 1, 2 appartiennent à la tranche d'effectifs '01', les effectifs de 3 à 5 appartiennent à la tranche '02' et les effectifs de 6 à 9 appartiennent à la tranche '03'.

métriques sont de valeurs contradictoires) et nous mettons en évidence le rôle de cette évaluation globale dans la résolution de notre problématique de sélection des données multisources.

3. Evaluation de la qualité globale des données : vers une appréciation de la valeur de la donnée

Nous décrivons dans cette section la deuxième contribution de nos travaux de thèse, à savoir l'agrégation des dimensions qualité en scores d'évaluation globaux. En particulier, nous nous intéressons à la mise en place d'un processus d'agrégation de critères qualité intrinsèques et nous mettons en évidence le rôle de la préférence dans la gestion des confusions impliquées par l'agrégation de critères bipolaires ou contradictoires.

Pour ce faire, nous définissons, d'abord, les besoins d'agrégation pour l'évaluation qualité d'une base de données multisources et nous détaillons les principes de notre approche d'agrégation préférentielle. Cette dernière est basée sur l'utilisation des intégrales floues, notamment l'intégrale de Choquet. Une formalisation permet ensuite d'expliciter les différents concepts de notre approche.

3.1. Besoins d'agrégation

L'évaluation de la qualité des données pour des besoins d'intégration logique d'un ensemble de données multisources nécessite, outre l'identification des principaux critères qualité et le choix de l'outil de mesure approprié, la quantification de l'ensemble de ces mesures en un indicateur de performance qualité unique. L'intérêt consiste principalement à identifier, lors de l'exécution d'une requête utilisateur, et étant donné un ensemble de doublons répondant à cette requête, l'alternative de meilleure qualité. La procédure d'évaluation adoptée dans la plupart des méthodologies qualité consiste à mettre en place, pour chaque indicateur qualité, un ensemble de valeurs seuils définissant leur adéquation quant aux attentes des utilisateurs métier. Le problème avec cette stratégie est qu'elle n'est pas globale (elle évalue les indicateurs qualité d'une valeur donnée un à un). Elle ne traite pas non plus le cas bipolaire où une valeur satisfait un indicateur qualité mais en viole un autre.

Prenons l'exemple d'un ensemble de valeurs de numéros de téléphones (Tableau 3.2) et plaçons-nous dans le cadre d'une intégration LAV où ces numéros sont des doublons issus de deux sources différentes.

Numéros de téléphone	Fraîcheur (nombre de mois)	Exactitude
0639233923	500	0.9
112342345	5	0.1

Tableau 3.2 Exemple de valeurs qualité bipolaires

Le premier numéro de téléphone (0639233923) est vraisemblablement correct mais obsolète, tandis que le deuxième (112342345) est récent mais incorrect (car il ne contient que 9 chiffres sur 10, si nous nous référons à la normalisation française). Dans ce cas, trois possibilités de résolution de conflits sont envisageables :

1 soit l'expert métier choisit manuellement l'alternative la plus adéquate parmi les doublons à chaque fois qu'une situation conflictuelle se présente, ce qui est une tâche fastidieuse, quand il s'agit de gérer l'incohérence de milliers de doublons ;
2 soit l'expert métier choisit le critère qualité le plus important étant donné un ensemble d'indicateurs qualité, ce qui constitue une tâche décisionnelle exclusive et fortement subjective (qui dépend considérablement de l'expert) ;
3 soit un score global d'agrégation est automatiquement calculé pour l'ensemble des indicateurs d'évaluation qualité.

Dans ce dernier cas, l'évaluation de la qualité d'une donnée décrite par un ensemble de métriques qualité est vue comme une problématique d'agrégation multicritère et son déploiement pour la gestion des incohérences dues à la présence de doublons lors de l'intégration de fichiers (ou bases de données) multisources n'est autre qu'une opération de décision multicritère.

3.2. Méthodologie d'agrégation

3.2.1. Définition du contexte d'agrégation

La définition du contexte de l'agrégation nécessite l'étude des propriétés des métriques qualité, qui définissent le noyau de notre agrégation.

Ainsi, une première propriété des métriques qualité se rattache à la dépendance qui pourrait exister entre quelques unes d'entre elles et où la valeur d'une métrique est intrinsèquement reliée à la valeur d'une deuxième métrique. Par exemple, de par même leurs définitions, la complétude nette (qui comptabilise les données renseignées et correctes) dépend fortement de l'exactitude. Dans ce contexte, [Helfert et al. 09] synthétisent l'ensemble des dépendances entre les dimensions qualité et distinguent deux types de dépendances : des dépendances à corrélation positive où l'amélioration de la qualité d'une dimension améliore celle d'une autre dimension, et des dépendances à corrélation négative où l'amélioration de la qualité d'une dimension qualité dégrade celle d'une autre dimension. Ils établissent, ainsi, l'analyse de corrélation résumée dans le Tableau 3.3.

Dimension1	**Dimension2**	**Type de corrélation**
Convenance temporelle (*Timeliness*)	Exactitude (*Accuracy*)	Négative
Convenance temporelle (*Timeliness*)	Crédibilité (*Believability*)	Positive
Convenance temporelle (*Timeliness*)	Cohérence de représentation (*Consistent representation*)	Négative
Convenance temporelle (*Timeliness*)	Complétude (*Completeness*)	Négative
Complétude (*Completeness*)	Cohérence de représentation (*Consistent representation*)	Négative

Dimension1	Dimension2	Type de corrélation
Complétude (*Completeness*)	Concision (*Conciseness*)	Négative
Accessibilité (*Accessibility*)	Sécurité (*Security*)	Positive
Accessibilité (*Accessibility*)	Exactitude (*Accuracy*)	Négative

Tableau 3.3 Corrélation entre dimensions qualité [Helfert et al. 09]

Une autre caractéristique des dimensions qualité est que leurs métriques ne sont pas commensurables (additives). Ainsi, l'unité de mesure de la métrique de fraîcheur se valorise en nombre de jours ou de mois, alors que la métrique d'exactitude est une probabilité dont la valeur varie dans l'intervalle [0, 1].

Finalement, comme l'évaluation de la qualité des données est avant tout une tâche subjective [Pipino et al. 02] et contextuelle, l'agrégation des dimensions qualité doit également être subjective et contextuelle dans le sens où le score d'agrégation final doit conforter l'expert métier et correspondre au jugement qu'il a porté sur la qualité de la donnée évaluée. Dans le contexte marketing, en fonction du type de la campagne de prospection, une métrique qualité peut voir son importance changer, impliquant un changement du score qualité. Ainsi, dans l'exemple de la donnée téléphone, la métrique de fraîcheur est considérée comme la plus importante dans le contexte d'une campagne téléphonique, elle est, cependant, beaucoup moins importante que la métrique d'exactitude, dans le cas d'une campagne courrier.

En résumé, les contraintes d'agrégation sont :

1 la dépendance des métriques ;
2 la non-commensurabilité des métriques ;
3 la contextualité et la subjectivité de l'agrégation affectant l'importance relative de certaines dimensions/métriques.

3.2.2. Approches d'agrégation

Nous élaborons à présent un aperçu des approches d'agrégations proposées dans la littérature. Nous distinguons deux approches principales d'agrégation [Lemaître 07] :

- agréger puis comparer : cette approche privilégie l'utilité multi-attribut, largement utilisée pour la résolution des problèmes de décision multicritère ;
- comparer puis agréger : cette stratégie est utilisée dans les approches de surclassement où l'aide à la décision multicritère est considérée comme un processus d'élaboration d'une structure de préférences. En effet, contrairement à la première approche, cette démarche ne suppose pas l'existence d'une préférence sur les décisions. Les alternatives sont comparées par paires, la préférence n'est ainsi ni totale, ni transitive, et l'agrégation est réalisée sur les résultats de comparaison des différentes paires. La méthode de Condorcet[24] et la méthode ELECTRE I [Roy 68] sont deux exemples d'approches de type « comparer puis agréger ».

L'approche « comparer puis agréger » peut s'avérer coûteuse en termes de temps d'exécution dans le cas où il s'agit d'un large volume de données multisources. Nous nous intéressons donc à la première approche d'agrégation « agréger puis comparer » et nous en présentons quelques exemples.

i- Première approche : fusion des classifieurs

Notre contexte d'agrégation nécessite le calcul d'un score global d'agrégation (numérique ou discret) des métriques d'évaluation qualité. Pour ce faire, plusieurs pistes de résolution sont explorées. La première consiste à assimiler l'agrégation des critères qualité à la fusion des classifieurs où les métriques, prises séparément, sont considérées comme des classifieurs partiels d'évaluation de la donnée. Les métriques n'étant pas forcément homogènes (puisqu'elles peuvent être soit des valeurs binaires, soit des scores de valeurs continues), nous optons pour l'agrégation de classifieurs hétérogènes. Ainsi, des approches comme le vote majoritaire [Black 58], la méthode bayésienne [Michie et al. 94], la méthode de Dempster-Shafer [Dempster 67, Shafer 76], la méthode BKS (behavior-knowledge space) [Huang et al. 95] et la régression logistique [Jacquet-Lagrèze et al. 87] sont considérées. Notons que le choix de la méthode dépend fortement du résultat souhaité par l'agrégation :

- soit le résultat est binaire ou à valeurs discrètes, auquel cas des valeurs de segmentation seuils sont à définir par les experts métiers autant au niveau des métriques qualité élémentaires qu'au niveau du résultat ;
- soit le résultat est un score numérique.

[24] http://www.bibmath.net/dico/index.php?action=affiche&quoi=./c/condorcet.html

ii- Deuxième approche : théorie de l'incertain

La théorie de l'incertain a été explorée dans la littérature afin d'estimer la fiabilité des sources et de prédire l'exactitude de l'information [Cholvy 04, Delavallade et al. 10]. Cette discipline se décline en trois théories :

- la théorie des probabilités, basée sur l'étude de l'incertitude pour modéliser les situations dont la connaissance est imparfaite. Malgré sa notoriété, cette pratique ne permet pas de modéliser l'ignorance et nécessite des probabilités *a priori* « subjectives, délicates à obtenir » [Delavallade et al. 10] ;
- la théorie des possibilités, basée sur la prise en compte combinée de l'incertitude et de l'imprécision dans les connaissances. Cette théorie, compatible avec la théorie des sous-ensembles flous (modélisation des imprécisions), offre ainsi un large ensemble d'opérateurs d'agrégation, donc, plus de souplesse pour gérer la fusion des informations multisources.
- la théorie de l'évidence, introduite par Dempster puis reprise sous un formalisme mathématiquement plus abouti par Shafer [Dempster 67, Dempster 68, Shafer 76], généralise la théorie des probabilités et la théorie des possibilités. Elle permet de manipuler des évènements non nécessairement exclusifs ce qui lui confère l'avantage de pouvoir représenter explicitement l'incertitude sur un événement.

iii- Troisième approche : décision multicritère

Cette approche a été évoquée pour la gestion de l'intégration de données multisources [Naumann 98]. Dans un tel contexte, la décision concerne le choix des alternatives multisources concurrentes. Elle dépend ainsi de deux types de critères :

- des critères qualité tels que la compréhensibilité, l'étendue (nombre d'attributs fournis par une source donnée) et la disponibilité des données ;
- des critères de coût s'exprimant au niveau du temps de calcul des requêtes et du prix associé.

Quatre méthodes sont alors étudiées :

1 la méthode SAW de poids additifs (*Simple Additive Weighting*) où les scores sont d'abord normalisés avant de se voir affecter des poids. Le score final est alors la somme des valeurs multipliées par leurs poids respectifs ;

2 la méthode TOPSIS (*Technique for Order Preference by Similarity to Ideal Solution*), qui définit la meilleure source en calculant deux distances Euclidiennes : entre les scores normalisés et celui de la meilleure source fictive d'une part, et entre les scores normalisés et la pire des sources fictive d'autre part. La source retenue est la plus proche de la meilleure source fictive et la plus éloignée de la pire source fictive ;

3 le processus de hiérarchie analytique AHP (*Analytical Hierarchy Process*), qui consiste à hiérarchiser l'objectif absolu représentant la satisfaction de l'utilisateur en contraintes de coûts et de qualité, sous-objectifs représentés par les dimensions qualité (Section 2.5). Des matrices de comparaison sont ensuite établies entre les sources pour chacune des dimensions qualité. Ces scores de comparaison sont finalement agrégés moyennant un vecteur poids pour chacun des objectifs et sous-objectifs fixés ;

4 la méthode DEA (*Data Envelopment Analytics*), encore appelée méthode du point extrême, détermine l'efficacité de chaque source moyennant des programmes linéaires de maximisation de performance. Ainsi, contrairement aux autres méthodes évoquées dans cette étude, la méthode DEA ne permet pas d'établir un classement des alternatives concurrentes, en l'occurrence les sources.

La comparaison de ces quatre méthodes repose sur cinq critères : le niveau d'interaction avec l'utilisateur (expert métier), la mise en place des poids des critères, la détection de la dominance d'un élément (source de données dans cet exemple), la fonction de normalisation utilisée et le type de résultat obtenu. Le Tableau 3.4 résume le résultat de cette comparaison.

	SAW	**TOPSIS**	**AHP**	**DEA**
Interaction avec l'utilisateur	Définition du vecteur de poids	Définition du vecteur de poids	Prise de décision	_
Mise en place des poids	Vecteur de poids	Vecteur de poids	Comparaison normée	_
Détection de la dominance	Oui	Oui	_	Oui
Fonction de normalisation	Distance normalisée	Distance Euclidienne	Vecteur normalisé	_

	SAW	TOPSIS	AHP	DEA
Type du résultat	Classement	Classement	Classement	Classification

Tableau 3.4 Comparaison des méthodes de décision multicritère [Naumann 98]

L'applicabilité de la méthode DEA suppose la prise en compte d'un certain nombre d'hypothèses, telles que l'existence d'une fonction d'utilité linéaire et l'indépendance des critères agrégés, hypothèses irréalisables dans le contexte d'agrégation de dimensions qualité. De plus, comme elle n'utilise pas de vecteur poids, la méthode DEA ne permet pas aux utilisateurs d'exprimer leurs préférences quant aux critères d'agrégation. Enfin, DEA ne permet pas de calculer un score d'agrégation global. Elle est en effet efficace pour identifier le meilleur élément étant donné un certain nombre d'alternatives et permet juste d'établir une classification générale. Les autres méthodes reposent sur l'expression d'une préférence des utilisateurs induisant une subjectivité importante lors de l'agrégation. De plus, elles nécessitent une grande interaction avec les utilisateurs, soit pour mettre en place les vecteurs de poids, soit pour la comparaison élémentaire de tous les critères. Cette interaction induit, malgré l'implication des utilisateurs dans le processus d'agrégation, une perte de temps considérable, d'où l'importance de l'automatisation de tels processus. Ainsi, [Bleiholder et al. 08] définissent un ensemble de stratégies de prise de décisions pour l'automatisation de l'intégration des données multisources. Ces stratégies sont principalement axées sur la problématique de la gestion des conflits dans le contexte de l'intégration de données multisources, problématique nécessitant généralement la prise en compte de l'avis d'un expert métier. Ils discernent ainsi :

- 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, et 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. Par ailleurs, le

principe « ne pas déformer l'information » (*no gossiping*) considère uniquement les valeurs cohérentes et plausibles ;

- l'approche à base de méta-données utilise le principe « fait confiance à tes amis » (*trust your friends*). Ici, les données d'une source sont privilégiées étant donnés des critères tels que le prix, la fiabilité, le volume des données fournies et d'autres critères qualité ;

- les stratégies de résolution des conflits et d'intégration des informations, comme 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 stratégies, la donnée imputée au système intégré résultant peut être différente des données sujettes au 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 est 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 ;
  - la stratégie « à jour » (*keep up-to-date*) qui préconise la valeur la plus récente.

iv- Discussion

Etudions la faisabilité de chacune des trois approches dans notre contexte d'agrégation de métriques qualité et commençons par l'approche d'agrégation des classifieurs. La méthode du vote majoritaire est vite écartée de notre champ d'intérêt car fortement subjective et très simpliste. Par ailleurs, les métriques qualité étant des variables dépendantes et non commensurables, nous écartons la méthode bayésienne et celle de Dempster-Shafer qui supposent l'indépendance des classifieurs. Il en est de même pour la régression logistique, qui se retrouve en dehors de notre périmètre d'intérêt étant donné ses résultats mitigés [Grabisch et al. 00]. Reste alors à analyser les performances de la méthode BKS aussi appelée « agrégation multi-nominale » (multinominal combination) et principalement utilisée pour la reconnaissance de motifs. Cette méthode qui a le grand avantage de ne pas travailler sous l'hypothèse d'indépendance des classifieurs, réalise de bonnes performances d'apprentissage, mais

nécessite cependant la présence d'un échantillon de données qui soit large et représentatif et provoque, le cas échéant, un risque de sur-apprentissage [Huang et al. 95].

Pour les mêmes raisons, nous écartons également les approches de décision multicritères. En effet, SAW, TOPSIS et AHP se basent sur des vecteurs poids lesquels sont spécifiés par les décideurs, impliquant alors une grande subjectivité. Par ailleurs, l'approche DEA permet plutôt de définir une classification, plutôt que de calculer un score d'agrégation ce qui ne résout pas notre problème. De plus, DEA suppose l'indépendance des critères agrégés [Sant'Anna 02], ce qui n'est pas le cas des dimensions qualité (Section 3.2.1).

Ainsi, sur la base de cette brève étude des approches d'agrégation multicritère, nous privilégions les approches basées sur la théorie de l'incertain beaucoup plus souples et beaucoup plus rigoureuses car beaucoup moins dépendantes des utilisateurs. En effet, les préférences des décideurs y sont généralement interprétées et non directement définies par les décideurs eux-mêmes. En particulier, nous nous intéressons aux fonctions d'agrégation non-additives pour le calcul d'un score qualité qui soit le plus proche possible des préférences de l'expert métier. Le but ultime étant d'utiliser ce score pour la comparaison des alternatives concurrentes dans le contexte d'une intégration des données multisources.

3.2.3. Choix de l'approche d'agrégation multicritère

i- Intérêt des fonctions non-additives

Nous nous définissons dans ce paragraphe notre opérateur ou fonction d'agrégation multicritère. Dans ce contexte, plusieurs techniques ont été développées depuis les premières recherches élaborées dans le domaine de l'agrégation multicritère au début des années quatre-vingt dix. Nous distinguons :

- les opérateurs conjonctifs où l'agrégation se fait en utilisant le "et" logique. Ces opérateurs sont les t-normes (les normes triangulaires) telles que le minimum ;
- les opérateurs disjonctifs où l'agrégation se fait en utilisant le "ou" logique. Ces opérateurs sont les t-conormes (les conormes triangulaires) telles que le maximum ;
- les opérateurs de compromis qui se situent entre les opérateurs conjonctifs et disjonctifs. Ces opérateurs sont les sommes pondérées, les opérateurs de moyenne, le minimum et le maximum pondérés, la somme pondérée ordonnée (OWA) et les intégrales floues ;
- les opérateurs hybrides qui n'appartiennent à aucune des classes ci-dessus.

Les opérateurs conjonctifs et disjonctifs telles que le minimum, le maximum, le minimum lexicographique et le maximum lexicographique sont écartées de notre champ d'intérêt car ils ne respectent pas le principe de séparabilité des fonctions d'agrégation. En effet, ils ne distinguent pas les profils ayant les mêmes minimaux (ou maximaux). Nous nous focalisons donc sur les opérateurs de compromis, qui visent, contrairement aux opérateurs conjonctifs et disjonctifs, à résumer un ensemble de critères numériques en un score global et donc à prendre en compte l'ensemble des critères et non une valeur extrême. De plus, ces fonctions de compromis permettent de prendre en compte l'importance des attributs ou critères considérés, en la modélisant à l'aide de poids. Ainsi, nous distinguons, dans cette catégorie :

- les techniques additives qui résument des données commensurables via une fonction monotone croissante, comme par exemple, la moyenne ou la somme pondérée ;
- les techniques non-additives qui calculent des fonctions de consensus à partir d'un ensemble de valeurs numériques non commensurables, comme par exemple la somme pondérée ordonnée (OWA), les minimum et maximum pondérés, les intégrales floues (intégrale de Choquet et intégrale de Sugeno), etc.

Jusqu'à récemment, les fonctions les plus répandues étaient les fonctions additives (de types moyennes pondérées) en raison de leur intuitivité et de leur facilité d'utilisation. Cependant, ces fonctions présentent quelques faiblesses. Ainsi, l'agrégation additive, qui se base sur l'étude de la somme pondérée, suppose implicitement que tous les critères peuvent s'exprimer indirectement dans la même unité. Elle véhicule, de ce fait, l'idée de la compensation possible entre critères (les poids représentant des taux de substitution entre les critères). De plus, une faible variation des poids peut entraîner de grandes conséquences sur la préférence globale. La fonction d'agrégation additive élimine également les alternatives Pareto-optimales (c'est-à-dire les alternatives qui ne sont dominées par aucune autre) puisque la seule alternative existante est le maximum. Enfin, l'inconvénient majeur de ces fonctions réside surtout dans le fait qu'elles ne sont pas capables de modéliser une quelconque interaction entre attributs ou critères [Marichal 02].

Par ailleurs, l'objectif final de l'agrégation étant la prise de décision pour sélectionner la meilleure alternative parmi un ensemble de doublons, notre fonction d'agrégation doit respecter les principes de base des règles de décision [Fargier et al. 09] :

- principe d'universalité : pour toute paire d'alternatives (a, b), on doit pouvoir décider si $a \geq b$;

- principe d'unanimité : si a est au moins aussi bon que b sur tous les critères, alors a ≥ b ;
- principe d'efficacité : si a est au moins aussi bon que b sur tous les critères et meilleur sur certains (c'est-à-dire a *domine* b), alors a > b.
- principe d'indépendance mutuelle (ou séparabilité ou Pareto-optimalité) : la préférence entre les alternatives a et b ne dépend pas des critères sur lesquels a et b reçoivent la même évaluation. Ainsi, elle exprime, dans un certain sens, l'indépendance des critères.

Les fonctions de minimum et de maximum pondérés, d'une part, et les sommes pondérées ordonnées (OWA), d'autre part, nécessitent la définition d'un vecteur poids fixant au préalable l'importance des performances des critères et influençant grandement le résultat de l'agrégation. Une solution a été d'explorer les fonctions non-additives issues de la théorie de l'utilité multi-attribut (*Multi Attribute Utility Theory* - MAUT) car ces fonctions conduisent à l'indépendance préférentielle mutuelle parmi les critères [Keeney *et al.* 76], notamment les intégrales floues. Ainsi, dans le but d'obtenir une représentation flexible des phénomènes complexes d'interaction parmi les critères, nous substituons au vecteur poids (utilisé principalement dans le cas additif) une fonction d'ensemble non additive, appelée mesure floue, un concept introduit en aide à la décision par [Sugeno, 74].

Notre choix s'est fixé, en particulier, sur l'intégrale de Choquet discrète comme méthode d'agrégation des métriques qualité au niveau de la donnée en raison de sa généricité, de sa non-additivité et de sa prise en compte des synergies éventuelles entre ces différents critères qualité. En effet, les techniques additives sont inapplicables sous les contraintes de dépendance et de synergies entre les critères, la plupart travaillant sous l'hypothèse d'indépendance entre les dimensions.

ii- Définition de l'intégrale de Choquet

La notion d'intégrale floue a été introduite avec la définition du concept de mesure floue [Sugeno 74], qui permet de définir une importance relative non seulement pour chaque critère, mais aussi pour chaque sous-ensemble de critères. On appelle mesure floue (ou mesure non-additive) sur un ensemble de critères N, une application d'ensemble $\mu : P(N) \rightarrow [0, 1]$ satisfaisant les axiomes :

- $\mu(\phi) = 0$; $\mu(N) = 1$;
- $\forall S,T \subset N / S \subset T$; $\mu(S) \leq \mu(T)$ (monotonie).

Dans un contexte d'analyse multicritère, le coefficient μ (*K*), pour $K \subseteq N$, est interprété comme le poids ou l'importance de la coalition des critères de *K*. Ainsi, en plus de la distribution de poids usuels sur les critères pris individuellement, des « poids » sur toutes combinaisons de critères sont également définis. L'axiome de monotonie (moins fort que le classique axiome d'additivité) signifie alors simplement que le fait d'ajouter un critère à une combinaison ne peut faire décroître l'importance de celle-ci.

L'intérêt de définir ce genre de mesures consiste principalement à s'épargner de la difficulté liée à la définition des niveaux d'importance des critères intrinsèques, tout en respectant les préférences du décideur expert vis-à-vis du score final des différentes alternatives et en écartant l'hypothèse idéaliste de l'indépendance des critères sous jacents.

Parmi cette famille des intégrales floues, outre les propriétés usuelles des opérateurs d'agrégation et la modélisation de l'importance relative des critères, la famille de l'intégrale de Choquet a la particularité de permettre la représentation des phénomènes d'interaction mutuelle qui peuvent exister entre certains critères. Les interactions s'étendent de la synergie négative (interaction négative) à la synergie positive (interaction positive). Le concept de l'intégrale de Choquet a été proposé dans [Schmeidler 86] puis par Murofushi et Sugeno dans [Murofushi et al. 91] utilisant le concept de la capacité introduit par Choquet [Marichal 98].

Ainsi, l'intégrale de Choquet se base sur deux concepts fondamentaux :

- l'utilité est une fonction de gain qui a pour but de modéliser numériquement les préférences du décideur. Les fonctions d'utilité peuvent être vues comme permettant de traduire les valeurs de l'attribut en degré de satisfaction [Kojadinovic 06]. Elles sont commensurables [Labreuche et al. 08], monotones et ascendantes dans le sens où si une alternative *a* est préférée à *b* alors $u(a) \geq u(b)$ [Labreuche 09];
- la capacité modélise la mesure floue sur laquelle se base l'intégrale et résume l'importance des critères (le vecteur poids traditionnellement utilisé dans les méthodes additives) en agrégeant les fonctions d'utilité.

Dans un contexte d'agrégation, l'intégrale de Choquet peut être considérée dans le contexte discret comme un opérateur d'agrégation n-aire où la partie intégrée est un ensemble de N valeurs $x_1, \dots, x_n$ de $\mathbb{R}$ [Marichal 00]. L'intégrale de Choquet discrète de $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ respectivement à une capacité $\mu \in F_N$ est alors définie par la formule suivante :

$$C_\mu(x) = \sum_{i=1}^{n} x_{(i)}[\mu(A_{(i)}) - \mu(A_{(i+1)})]$$

où (.) est une permutation sur N telle que $x_{(1)} \leq \cdots \leq x_{(n)}$ et $A_{(i)} = \{(i), \ldots, (n)\}$ et $A_{(n+1)} = \emptyset$.

Etant donnée cette définition, si nous disposons de trois critères x_1, x_2et x_3 dont l'importance est telle que $x_3 \leq x_1 \leq x_2$, l'intégrale de Choquet C respectivement à la capacité μ est donnée par :

$$C_\mu(x_1, x_2, x_3) = x_3[\mu(3,1,2) - \mu(1,2)] + x_1[\mu(1,2) - \mu(2)] + x_2\mu(2).$$

3.3. Agrégation et apprentissage préférentiels

L'intégrale de Choquet, telle que définie dans la Section 3.2.3 permet de garantir une agrégation des métriques qualité modélisant la préférence des décideurs (grâce au concept de l'utilité) et respectant les caractéristiques théoriques intrinsèques des dimensions qualité (grâce au concept de la mesure floue représenté par la capacité). Cette section détaille ces mécanismes et met en évidence l'utilisation de l'intégrale de Choquet en tant que classifieur capable d'apprendre une fonction d'agrégation préférentielle sur un échantillon d'apprentissage, puis de la prédire sur des exemples non classifiés (c'est-à-dire où la préférence du décideur n'a pas été exprimée). Ce dernier point sera, en effet, un argument pour l'automatisation d'un tel processus d'agrégation pour son application dans l'agrégation des métriques qualité de notre base de prospection multisources.

3.3.1. Formalisation

Nous considérons la terminologie suivante.

- Un ensemble de sources S = {s_i}, i ∈ $[1, |S|]$. est ensemble de sources de données. En pratique, une source est un fichier décrivant des données d'entreprises. La description du fichier en attributs diffère d'une source à l'autre. Ainsi, chaque source s_i est décrite :
  - en intention (schéma) par un ensemble d'attributs $A_i = \{a_{ij}\}$, j ∈ $[1, m_i]$. L'ensemble de tous les attributs est $A = \cup A_i, i \in [1, |S|]$; sachant que $\cap A_i \neq \emptyset$ étant donnépuisque que deux sources peuvent fournir des informations semblables pour une entreprise donnée ;.
  - en extension (données) par un ensemble d'enregistrements $D_i = \{R_{ik}\}, k \in [1, n_i]$, où chaque enregistrement est un ensemble de données : $R_{ik} = \{d_{ijk}\}$, avec $d_{ijk} \in dom(a_{ij})$ le domaine (ensemble de valeurs possibles) de a_{ij}. Chacun de ces

enregistrements est identifié par un identifiant unique des entreprises (le numéro SIRET) noté id_{ik} ; où $i \in [1, |S|]$et $k \in [1, n_i]$ (Tableau3.5).

- Un ensemble de dimensions qualité Δ_{ij} est un ensemble de dimensions qualité associées à un attribut a_{ij} : $\Delta_{ij} = \{\delta_{ijl}\},\ l \epsilon [1,\ \lambda_{ij}]$.
- Un ensemble de valeurs de métriques Ω_{ijk} est un ensemble de valeurs de métriques qualifiant la qualité d'une donnée d_{ijk} tel que $\Omega_{ijk} = \{\omega_{ijkl}\}$.

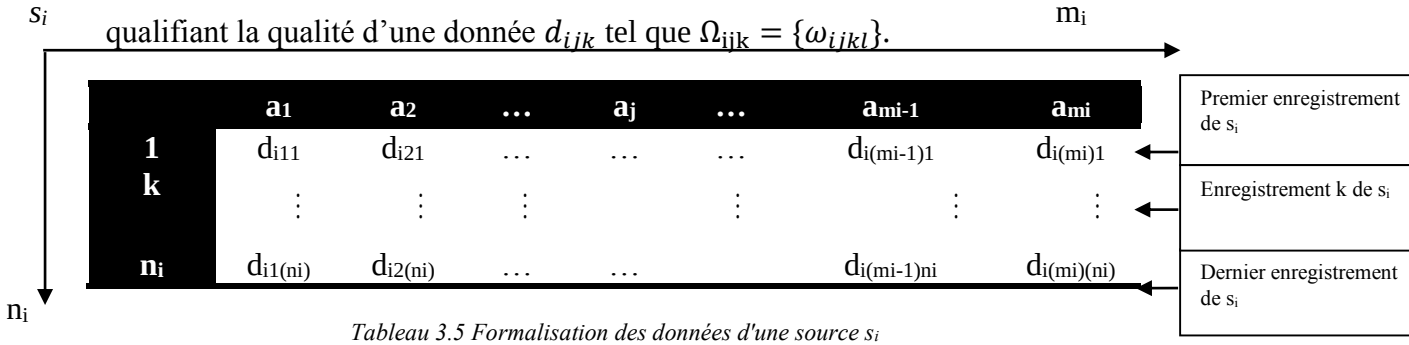


Tableau 3.5 Formalisation des données d'une source s_i

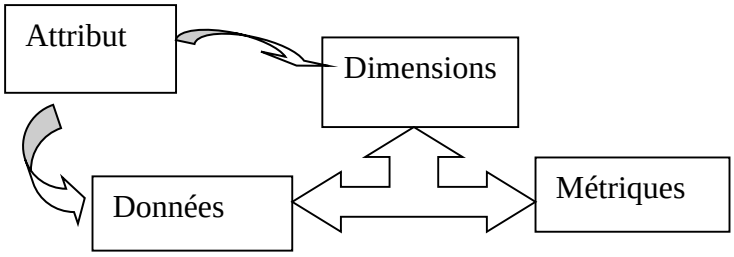


Figure 3.3 Relations entre les concepts

Ainsi, comme le montre la Figure 3.3, un attribut a_{ij} donné est qualifié par λ_{ij} dimensions et chaque donnée d_{ijk}se voit affecter un tableau de valeurs de métriques qualité $\Omega_{ijk} = \{\omega_{ijkl}\}$, tableau à une dimension et de longueur taille λ_{ij}.

Etant donnés ces paramètres, la fonction d'agrégation des métriques qualité relatives à une donnée en particulier se définit par :

- $\sigma_{ijk} = agregD(\omega_{ijkl}), l \in [1, \lambda_{ij}],$ où agregD(ω_{ijkl})*est une* fonction d'agrégation des métriques qualité calculées au niveau de la donnée d_{ijk} .

3.3.2. Agrégation préférentielle des métriques qualité à l'aide de l'intégrale de Choquet

Nous utilisons l'intégrale de Choquet discrète et nous nous basons uniquement sur les fonctions d'utilité et de capacité pour modéliser les préférences de l'utilisateur/décideur ainsi que les synergies existant entre ces critères. En effet, l'intégrale de Choquet consiste à agréger des mesures évaluant un ensemble de critères (les critères sont, dans notre exemple, les métriques qualité assignées à un attribut a_{ij}). Cette agrégation a la particularité de prendre en considération les synergies qui peuvent exister entre les différents critères ainsi que l'importance relative de certains de ces critères par rapport à d'autres.

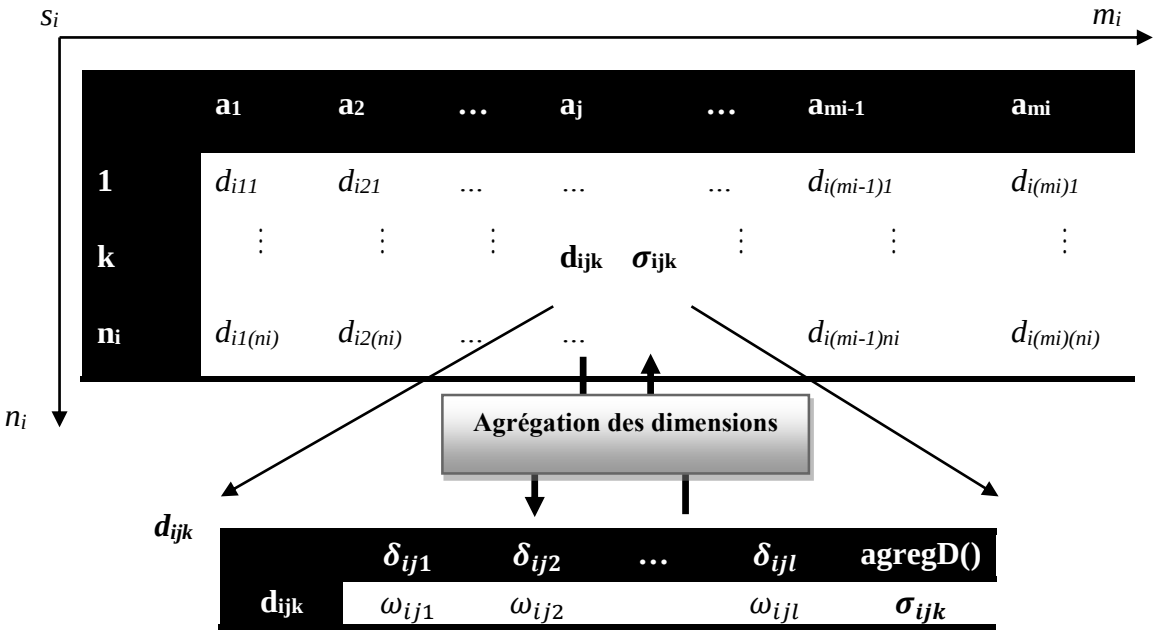


Tableau 3.6 Calcul des dimensions qualité d'une donnée d_{ijk}

i- Données du problème d'agrégation

Soit un ensemble de données d_{ijk} décrivant un attribut a_{ij} donné (décrivant la colonne a_{ij} de la source s_i) qualifié par un ensemble de valeurs de métriques $\Omega_{ijk} = \{\omega_{ijkl}\}$. L'objectif est de résumer cet ensemble de valeurs en un score qualité σ_{ijk}permettant d'évaluer la qualité de la donnée d_{ijk}. Pour ce faire, nous notons ρ_{ijkl} l'utilité élémentaire et $P_{ijk} = \{\rho_{ijkl}\}$ l'ensemble des utilités relatives à une métrique ω_{ijkl}.

Par souci de simplification, nous nous proposons d'illustrer le principe de la fonction *AgregD()* en allégeant la terminologie proposée dans la formalisation (Section 3.3.1). Ainsi nous proposons :

- *s* pour décrire la source s_i
- *a* pour décrire l'attribut a_{ij}
- d_k pour décrire la donnée d_{ijk}
- δ_l pour décrire la dimension δ_{ijl}, Δ pour décrire l'ensemble des dimensions Δ_{ij}et λ pour décrire le nombre de dimensions λ_{ij}
- ω_{kl} pour décrire les valeurs des métriques ω_{ijkl} et Ω_k pour décrire l'ensemble des valeurs des métriques Ω_{ijk}
- ρ_{kl} pour décrire l'utilité ρ_{ijkl} et P_{kl} pour décrire P_{ijk}
- μ_{kl} pour décrire la capacité μ_{ijkl}
- σ_k au lieu de σ_{ijk}

Supposons que l'attribut *a* se voie assigner 5 dimensions qualité (l =5). Afin de paramétrer la fonction d'apprentissage (fonction d'agrégation), nous construisons un échantillon prototype de taille n_i=*4*. L'échantillon prototype doit représenter toutes les combinaisons possibles des valeurs des dimensions δ_l (Tableau 3.7).

d_k	δ_1	δ_2	δ_3	δ_4	δ_5
d1	ω_{11}	ω_{12}	ω_{13}	ω_{14}	ω_{15}
d2	ω_{21}	ω_{22}	ω_{23}	ω_{24}	ω_{25}
d3	ω_{31}	ω_{32}	ω_{33}	ω_{34}	ω_{35}
d4	ω_{41}	ω_{42}	ω_{43}	ω_{44}	ω_{45}

Tableau 3.7 Prototype d'apprentissage de la fonction d'agrégation : tableau des données

L'intégrale de Choquet consiste à additionner des capacités élémentaires μ_{kl} et ensemblistes $(\mu_{kl'} \ldots \mu_{kl''})$ $(l' \in [1,\lambda] et\ l'' \in [1,\lambda]; l' \neq l'')$ des différentes utilités respectives ρ définissant les critères qualité ω_{kl} d'une donnée d_k. Capacité et utilité sont ainsi les principales composantes de l'intégrale de Choquet.

ii- Définition de la fonction d'utilité et calcul de ρ

La fonction d'utilité est utilisée pour transformer les données (à agréger) en degrés de satisfaction modélisant ainsi les préférences de l'utilisateur. L'utilité étant commensurable et ascendante, elle peut être modélisée comme une fonction de transformation T des métriques qualité ω_{kl} définie comme suit.

$$T: \Omega_k \rightarrow \mathrm{P}_k$$

$$\omega_{kl} \longmapsto \rho_{kl}$$

Cette transformation T est donc une fonction de normalisation permettant d'obtenir des métriques ρ_{ijkl} commensurables et, de ce fait, prêtes à l'agrégation. Nous définissons deux fonctions de normalisation $T(\omega_{kl}) = \rho_{kl}$ selon la monotonie de la métrique ω_{kl} :

- si la métrique ω_{kl} est une fonction ascendante :

$$\rho_{kl} = \frac{|\omega_{kl} - Max(w_l) + \Delta'(\omega_l)|}{\Delta(\omega_l)} ;$$

- si la métrique ω_{kl} est une fonction descendante :

$$\rho_{kl} = 1 - \frac{|\omega_{kl} - Max(w_l) + \Delta'(\omega_l)|}{\Delta(\omega_l)} ;$$

où

- w_l désigne l'ensemble des métriques qualité quantifiant les dimensions qualité δ_l qui évaluent l'attribut a fourni par la source $s.$;
- $\Delta(\omega_l) = Max(w_l) - Min(w_l)$;
- $\Delta'(\omega_{kl}) = Max'(w_l) - Min'(w_l)$;
- $Max'(w_l) = Max(w_l) - \Delta(\omega_l) * 0.01$;
- $Min'(w_l) = Min(w_l) + \Delta(\omega_l) * 0.01$.

De la même manière que l'utilité, nous représentons par μ_{kl} la capacité et par $\mathrm{M}_k = \{\mu_{kl}\}$ l'ensemble des capacités relatives à une donnée d_k.

Par ailleurs, la fonction d'utilité que nous employons pour l'apprentissage est une fonction de discrétisation de la valeur de la métrique qualité ω_{kl} étant donnée une (ou des) valeurs seuil θ; les seuils étant définis au préalable par l'expert métier.

Supposons que l'expert ait défini deux valeurs seuils θ_1 et θ_2 pour discrétiser les valeurs ω_{kl} de la dimension δ_l et que $\theta_1 > \theta_2$. La fonction est ainsi définie de la sorte :

- si v> θ_1 ; alors ρ=1
- si θ_2<v< θ_1 ; alors ρ=0.5
- si v< θ_2 ; alors ρ=0

La table d'apprentissage est donc représentée par le Tableau 3.8.

$\boldsymbol{d_k}$

	$\boldsymbol{\delta_1}$	$\boldsymbol{\delta_2}$	$\boldsymbol{\delta_3}$	$\boldsymbol{\delta_4}$	$\boldsymbol{\delta_5}$
d1	ρ_{11}	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{15}
d2	ρ_{21}	ρ_{22}	ρ_{23}	ρ_{24}	ρ_{25}
d3	ρ_{31}	ρ_{32}	ρ_{33}	ρ_{34}	ρ_{35}
d4	ρ_{41}	ρ_{42}	ρ_{43}	ρ_{44}	ρ_{45}

Tableau 3.8 Prototype d'apprentissage de la fonction d'agrégation - Tableau des utilités

iii- Calcul de la capacité

L'intégrale de Choquet consiste à sommer l'ensemble des capacités élémentaires et collectives (notées respectivement μ_{kl} et $(\mu_{kl'} \dots \mu_{kl''})$ telles que $l' \in [1, \lambda]$ et $l'' \in [1, \lambda]; l' \neq l''$) relatives aux métriques normalisées (ou utilités) ρ_{ijkl}. Capacités et utilités représentent en effet les piliers de notre fonction d'agrégation, qui permet d'affecter des poids à des critères (les dimensions qualité d'évaluation) ou à des ensembles de critères δ_l étant données les préférences des utilisateurs (ou décideurs). En théorie, ces préférences peuvent s'exprimer par un ordonnancement (total ou partiel) des critères de décision ou des alternatives, par des scores d'importance quantitative des critères telles que les poids ou encore par la quantification de la synergie pouvant exister entre les paires de critères. Afin de limiter la subjectivité de la fonction d'agrégation, nous considérons uniquement l'expression des préférences par ordonnancement des alternatives.

Supposons, alors, que $\rho_3(d_k) \geq \rho_4(d_k) \geq \rho_1(d_k) \geq \rho_2(d_k) \geq \rho_5(d_k)$ pour un d_k donné. L'intégrale de Choquet calculée par rapport à la capacité μ s'exprime ainsi de la manière suivante :

$$C_\mu(d_i) = \rho_3[\mu(\delta_1, \delta_2, \delta_3, \delta_4, \delta_5) - \mu(\delta_1, \delta_2, \delta_4, \delta_5)] + \rho_4[\mu(\delta_1, \delta_2, \delta_4, \delta_5) - \mu(\delta_1, \delta_2, \delta_5)] + \rho_1[\mu(\delta_1, \delta_2, \delta_5) - \mu(\delta_1, \delta_5)] + \rho_1[\mu(\delta_1, \delta_5) - \mu(\delta_5)] + \rho_5\mu(\delta_5)$$

Ainsi, étant donné un ensemble d'objets ordonnés, l'intégrale de Choquet relative à la donnée d_k est plus généralement définie par l'expression suivante :

$$\sigma_k = C_\mu(\rho_{kl}) = \mathrm{C}_\mu\big(\rho_{k(l)}\big) = \sum_{l=1}^{\lambda} \rho_{k(l)}[\mu\big(\mathrm{d}_{\mathrm{k}(l)}\big) - \mu(\mathrm{d}_{\mathrm{k}(l+1)})]$$

où

- (.) définit une permutation λ sur un ensemble d'alternatives d_{ijk} telle que $\rho_{(1)} \leq \cdots \leq \rho_{(\lambda)}$, $d_{\mathrm{k}(l)} = \{(l), \ldots, (\lambda))\}$ et $d_{\mathrm{k}(\lambda+1)} = \emptyset$;
- ρ est l'utilité de la métrique qualité normalisée.

iv- Conclusion

L'agrégation que nous proposons se base sur une approche subjective où les capacités sont déduites des préférences de l'utilisateur. Ainsi, la fonction d'agrégation calculant σ_k diffère selon l'attribut d_k. De la même façon, la fonction d'agrégation calculant σ_k diffère selon le type de canal de la campagne (étant donné que l'importance des attributs diffère). La qualité globale du fichier de sélection σ n'est autre qu'une moyenne arithmétique des scores σ_k individuels.

3.3.3. Apprentissage des préférences d'agrégation

Nous nous intéressons dans ce paragraphe à l'apprentissage des préférences d'agrégation des décideurs exprimées lors du calcul de l'intégrale de Choquet. Notre objectif est de reproduire les décisions préférentielles des utilisateurs de manière automatique, et ce afin d'éviter les sollicitations ponctuelles des décideurs à chaque fois qu'une confusion est rencontrée lors de l'intégration des données multisources. Ainsi, nous nous basons sur la capacité d'apprentissage de l'intégrale de Choquet qui a été démontrée, entre autres, dans [Grabisch 95]. Des fonctions traitant les problématiques de fouilles de données telles que les moindres données, la programmation linéaire et les moindres carrées ont d'ailleurs été utilisées dans ce contexte [Grabisch et al. 03, Grabisch et al. 00, Bebcakova, 11].

i- L'agrégation préférentielle : une problématique d'apprentissage

L'apprentissage des préférences est en effet une discipline récente dans le domaine de l'intelligence artificielle et l'apprentissage automatique qui consiste à observer et apprendre les préférences d'un individu, notamment lors de l'ordonnancement d'un ensemble d'alternatives, pour prédire l'ordonnancement automatique d'un nouvel ensemble d'alternatives [Fürnkranz et al. 10].

Afin d'accomplir sa fonction d'apprentissage, l'intégrale de Choquet se repose sur un ensemble de concepts qui permettent de gérer aussi bien la prise en considération des préférences (ou décisions) des utilisateurs (préférences entre les critères et préférences entre les alternatives ou objets) que l'interaction et la synergie entre les différents critères au sein de l'agrégation des données, les plus importants sont l'indice de Shapley et l'indice d'interaction.

- L'indice de Shapley (aussi appelé indice d'importance) d'un critère i par rapport à une capacité μ est définie par :

$$\phi_\mu(i) = \sum_{T \subseteq N \setminus i} \gamma|t|(n)[\mu(T \cup i) - \mu(T)] \ ; \ \forall i \epsilon N,$$

où

$$\gamma_t(n) = \frac{(n-t-1)!\,t!}{n!} \ (t = 0,1, \dots, n-1).$$

Il s'interprète dans ce contexte comme le poids moyen du critère i en question.

- L'indice d'interaction, introduit dans [Murofushi et al. 91] afin de représenter le degré d'interaction positif ou négatif entre deux critères de décision donnés. [Wendling et al. 08], définit l'interaction entre deux critères i et j par :

$$I(\mu, ij) = \sum_{T \subseteq X \setminus ij} \frac{(n-t-2)!\,t!}{(n-1)!} \Delta_{ij}\mu(T)$$

où

$$\Delta_{ij}\mu(T) = \mu(T \cup ij) + \mu(T) - \mu(T \cup i) - \mu(T \cup j)\,.$$

De cette manière, étant donné un ordonnancement préférentiel sur un échantillon fictif d'apprentissage, l'intégrale de Choquet discrète est capable de quantifier, puis d'apprendre, les poids relatifs des différentes métriques qualité.

ii- Apprentissage des préférences : vers une automatisation de l'évaluation qualité

L'automatisation du processus d'agrégation préférentielle des métriques qualité est principalement requise pour la résolution des incohérences lors de l'intégration (logique ou physique) d'un grand volume de données multisources. En effet, un score qualité estimant les préférences du décideur sera automatiquement assigné à chaque entité concurrente (doublon) éliminant, de cette manière, toute confusion quant au choix de l'alternative à intégrer. L'automatisation consiste, alors, dans la prédiction des valeurs d'agrégation des métriques

qualité telles qu'elles étaient définies par le décideur sur un échantillon d'apprentissage représentatif.

Notre méthodologie d'apprentissage est illustrée dans la Figure 3.4.

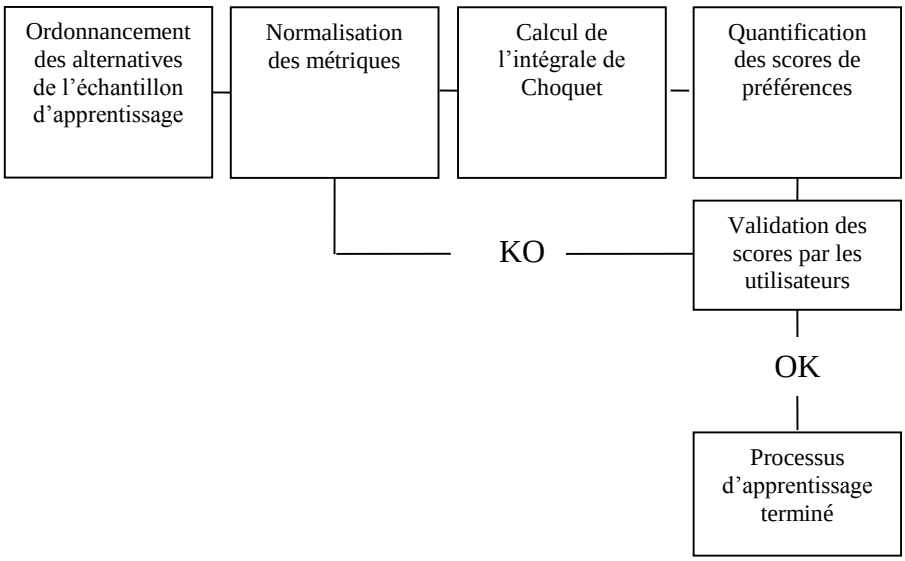


Figure 3.4 Apprentissage des préférences d'agrégation

Comme dans toute procédure d'apprentissage, le décideur dispose d'un échantillon de données expérimentales qu'il doit ordonner selon sa préférence. Idéalement, pour que la fonction d'apprentissage soit efficace, l'échantillon doit représenter tous les cas de figures possibles qui correspondent, dans notre cas, à toutes les combinaisons possibles des valeurs des métriques qualité. Prenons l'exemple de deux métriques ω_1 et ω_2 et supposons que :

- ω_1 admet uniquement deux nuances d'évaluation : bonne qualité (1) et mauvaise qualité (0,1),
- ω_2 admet trois nuances d'évaluation : bonne qualité (1), qualité acceptable (0,5) et mauvaise qualité (0,1).

La table d'apprentissage des deux critères qualité quantifiés par ω_1 et ω_2 est alors donnée dans le Tableau 3.9.

Alternatives	ω_1	ω_2
a	1	1
b	1	0,5
c	1	0,1

Alternatives	ω_1	ω_2
d	0,1	1
e	0,1	0,5
f	0,1	0,1

Tableau 3.9 Table d'apprentissage

Les nuances d'évaluation sont des données subjectives et dépendent principalement du décideur. Par exemple, ce dernier peut assigner deux classes d'appréciation à la métrique d'exactitude : correcte et incorrecte ; et trois classes d'appréciation à la métrique de fraîcheur : fraîche, acceptable et obsolète. De plus, les classes d'évaluation dépendent de la donnée rattachée à la métrique. Ainsi, les classes d'appréciation de la fraîcheur d'un numéro de téléphone ne sont pas les mêmes que celles d'évaluation d'un nom d'entreprise.

Une fois l'apprentissage terminé, les scores d'agrégation prédits sont analysés par le décideur. Si les scores prédits ne quantifient pas ses préférences avec fiabilité, la fonction d'apprentissage doit être réajustée.

4. Conclusion

Ce chapitre décrit notre approche d'évaluation globale de la qualité des données multisources. L'évaluation qualité est en effet utilisée pour la gestion des incohérences imposées par le contexte de sélection de données de prospection marketing. Le but étant de définir un critère qui soit à la fois pragmatique et rigoureux pour la gestion des doublons dans le processus de sélection des données multisources. Deux contributions ont alors été décrites :

1 la description d'une méthodologie vigoureuse et complète pour la mise en place du processus d'évaluation intrinsèque et contextuelle de la qualité des données ;
2 la quantification de la qualité globale à différents niveaux (donnée, enregistrement et sélection) à travers l'apprentissage de la préférence d'agrégation du décideur métier.

Nous nous sommes principalement intéressés à l'utilisation de l'intégrale de Choquet discrète pour agréger les métriques qualité des alternatives multisources concurrentes, de façon à prendre en considération les contraintes préférentielles des décideurs et à déduire un score qualité global qui en désignera la meilleure candidate.

Chapitre 4 : Optimisation préférentielle de la sélection des données d'une base de prospection marketing

« Ce qui sauve, c'est de faire un pas, encore un pas. »
Antoine de Saint-Exupéry

Nous abordons dans ce chapitre le problème de la sélection des données dans les bases multisources. Ce problème implique, au même titre que l'intégration logique des données multisources, la résolution des problèmes de cohérence causés par la présence de doublons dans la base interrogée, amenant ainsi le décideur à choisir entre un ensemble d'alternatives concurrentes.

Dans notre contexte particulier d'aide à la prospection, la sélection forme le fichier de ciblage que l'on utilisera pour effectuer les campagnes marketing. Alors, afin d'optimiser la fiabilité du fichier de ciblage, cette sélection sera contrainte par deux facteurs, la qualité et le coût des données multisources. Nous assimilons cette problématique à un problème de décision multicritère, et plus particulièrement, à un problème d'optimisation multicritère (dite aussi multiobjectifs). Les critères de coût et de qualité représentent la fonction utilité (encore appelée gain ou compromis) à optimiser.

Dans cette thèse, nous n'abordons pas la problématique d'estimation des coûts globaux de l'opération marketing (dans le sens optimisation du ROI de la campagne marketing, par exemple) ni celle des coûts de la non qualité des données. En effet, nous nous limitons à la considération du prix de vente (ou de location) des données multisources comme unique contributeur au calcul de notre variable coût. La qualité est donnée par le score global obtenu à l'issue de notre phase d'évaluation décrite dans le Chapitre 3.

Notre démarche d'optimisation s'inscrit dans le cadre d'une méthodologie globale qui s'articule en trois phases principales :

1 la construction du modèle ;
2 l'optimisation ;
3 la décision (l'articulation de la préférence de sélection).

En effet, l'optimisation aura pour but de définir un ensemble de solutions Pareto-optimales (c'est à dire un ensemble de solutions admissibles, non dominées, efficaces qui répondent aux

contraintes d'optimisation exprimées dans la définition de la problématique. Ensuite, le décideur choisira la meilleure alternative parmi cet ensemble de solutions Pareto-optimales [Branke et al. 08] : c'est la phase décisionnelle de notre méthodologie. Auparavant, le modèle d'optimisation sous-jacent sera rigoureusement défini pour formaliser, d'une part, la fonction objectif et les contraintes d'optimisation et spécifiant, d'autre part, l'algorithme de résolution adéquat.

Le plan du chapitre s'énonce alors comme suit. Nous décrivons, dans la première section, une formalisation de notre problème d'optimisation. Nous détaillons, ensuite, les principes de notre démarche d'optimisation de la sélection des données multisources en définissant, d'une part, la fonction d'optimisation, et d'autre part, la procédure de prise de décision utilisée pour le choix de la solution optimale. Une section décrivant un bref état de l'art des techniques d'optimisation multiobjectifs sera préalablement proposée.

1. Formalisation du problème d'optimisation

Nous reprenons les bases de la formalisation décrite dans le Chapitre 3 (Section 3.3.1) et nous définissons les fonctions d'agrégation suivantes.

- $\sigma_{ijk} = agregD(\omega_{ijkl}), l \in [1, \lambda_{ij}]$, où $agregD()\ est\ une$ fonction d'agrégation des métriques qualité calculée au niveau de $la\ donnée\ d_{ijk}$.
- $\sigma_{ik} = agregR(\sigma_{ijk}), j \in [1, m_i]$, où $agregR()$ est une fonction d'agrégation de métriques qualité au niveau d'un enregistrement.
- $\sigma = agregF(\sigma_{ij}), j \in [1, m_i]\ et\ i \in [1, |S|]$, où $agregF()$ est une fonction d'agrégation de métriques qualité au niveau d'un ensemble de données composant le fichier de prospection.
- Une utilité élémentaire ρ_{ijkl} (telle que $P_{ijk} = \{\rho_{ijkl}\}$) représente l'ensemble décrivant les utilités attribuées à ω_{ijkl}) et une capacité élémentaire μ_{ijkl} (telle que $M_{ijk} = \{\mu_{ijkl}\}$ regroupe l'ensemble des capacités attribuées à ω_{ijkl}. ρ_{ijkl} et μ_{ijkl} sont calculées pour chaque métrique ω_{ijkl} en prémisse d'une agrégation par l'intégrale de Choquet.

Par ailleurs, nous définissons les fonctions :

- π_{ijk}, une fonction prix exprimant le prix unitaire d'une donnée. De la même façon, nous définissons les fonctions $\pi_k$et π qui représentent respectivement le prix d'un

enregistrement k (relatif à un identifiant id_k) et celui de l'ensemble de la sélection formant le fichier de ciblage ;

- ψ, une fonction de gain en qualité/prix telle que $\psi = f(\sigma, \pi)$ que l'on se propose d'optimiser.

Finalement, nous définissons les données de la campagne à savoir le budget, les attributs du fichier de prospection et le nombre d'enregistrements à fournir.

- P est le potentiel de prospection pour la campagne c telle que $P = \{d_{ijk}\}$. Le potentiel représente le nombre de données de la base multisources satisfaisant les contraintes *C1* de ciblage.
- a_c fait référence à l'ensemble des attributs constituant le fichier de ciblage c tel que $a_c \subset A$.
- η_c est la taille du fichier c tel que $\eta_c = \{\eta_{ij}\}$ et η_{ij} est le nombre de données par attribut et par source. $i \in [1, |S|]$ et $j \in [1, |a_c|]$.
- Une donnée est toujours référencée par ses coordonnées tridimensionnelle d_{ijk} où $i \in [1, |S|], j \in [1, |a_c|]$ et $k \in [1, \eta_c]$
- σ_c est le niveau minimum de qualité imposé par l'utilisateur (le client dans le cas de notre application) pour la campagne c.
- $coûtFixe_i$ est le coût fixe imposé par la source s_i. ; $i \in [1, |S|]$.
- $prixMin_i$ est le prix minimum à payer pour une source s_i ($i \in [1, |S|]$) si le nombre d'enregistrements commandé (loué) est inférieur à nb_i.
- $prixUnitaire_{ij}$ représente le prix unitaire d'achat ou de location d'une donnée d_{ijk} ou d'un enregistrement de données d'une source i.
- nb_i est le nombre minimum de données (en l'occurrence, données emails) commandées (louées) auprès d'une source S_i ($i \in [1, |S|]$).
- $Budget_c$ est le budget défini par le client pour la campagne c.

Dans notre problématique d'optimisation de la sélection multisources, maximiser le gain revient à minimiser le prix total (coût du fichier de ciblage) en maximisant sa qualité globale. Nous pouvons alors écrire le problème sous la forme suivante :

$$Maximiser \sum_{j} \sum_{i} \psi(\pi_{ij}, \sigma_{ij}) \ (F)$$

où

$$\begin{cases} \sigma_{ij}(\{d_{jk}\}) = {\sum_{k=1}^{\eta_{ij}} \sigma_{ijk}}/{\eta_{ij}} \ (F1) \\ \pi_{ij}(\{d_{jk}\}) = Co\hat{u}tFixe_{ij} + prixMin_i + (\eta_c - nb_i) * prixUnitaire_{ij} \ (F2). \end{cases}$$

Une première étape de résolution consiste à identifier la fonction de profit ψ à double objectif.

2. Définition de l'algorithme d'optimisation

L'objectif de cette étude est double : modéliser, dans un premier temps, la fonction objectif de profit ψ traduisant le compromis entre le prix π et la qualité σ ; et définir, ensuite, en fonction de ψ, l'approche d'optimisation la plus adéquate.

Pour ce faire, nous nous inspirons du problème de sélection de portefeuille multiobjectif où il est aussi question de minimiser le risque d'investissement tout en maximisant le profit. La revue de la littérature montre que les fonctions de profit sont généralement réduites à de simples fonctions linéaires [Arnone et al. 93, Mansini et al. 03, Di Gaspero et al. 07] et la résolution proposée varie des démarches déterministes qui cherchent à définir la solution optimale exacte aux démarches heuristiques qui cherchent à l'approcher moyennant des techniques de recherches locales basées, entres autres, sur les algorithmes évolutionnaires et les algorithmes de fourmis [Di Gaspero et al. 07].

Dans cette section, nous décrivons notre approche de définition de la fonction de profit et nous dressons un état de l'art des métaheuristiques proposées pour résoudre de tels problèmes d'optimisation multiobjectifs.

2.1. Historique de l'optimisation multiobjectifs

Les fondements mathématiques de l'optimisation multiobjectifs ont commencé vers la fin du XIX^{e} siècle (1895) où Cantor, puis Hausdorff avaient établi les bases et principes des espaces de dimensions infinies. C'est en introduisant le concept de la maximisation vectorielle (*Vector Maximum Problem*), en 1951, que Kuhn et Tucker ont fait basculer l'optimisation multiobjectifs dans le monde mathématique [Khun & Tucker 51, Coello Coello et al. 07].

Les premières applications ont alors commencé à affluer, dès les années 60. Les ingénieurs autant que les décideurs ont particulièrement apprécié la puissance du concept "compromis" pour la résolution des problèmes d'ingénierie dits difficiles tels que la gestion des réseaux hydrauliques, la gestion de la production ainsi que les problèmes économiques. Plus tard, une pléthore d'approches d'optimisation multiobjectifs a été définie et plusieurs classifications ont été proposées pour distinguer les approches interactives, locales, de voisinage ou encore globales. Nous retenons en particulier la classification de [Lampinen 00] qui répartit les algorithmes d'optimisation multiobjectifs en trois groupes : les approches énumératives, les approches déterministes et les approches stochastiques. Le premier groupe des approches énumératives consiste à lister l'ensemble des solutions possibles du problème d'optimisation en question, puis de choisir celle qui l'optimise le plus. Malheureusement, ces approches ne sont pas applicables aux problèmes à variables continues qui admettent une infinité de solutions. Le second groupe des algorithmes déterministes établit sa stratégie d'optimisation sur des hypothèses sur la fonction objectif telles que la linéarité, la convexité et la dérivabilité. Nous y trouvons les algorithmes gloutons, *branch and bound*, profondeur d'abord (*depth-first*) ou encore meilleur d'abord (*best-first*). Enfin, la classe des algorithmes stochastiques, contrairement au second groupe des approches déterministes, n'établit pas d'hypothèses *a priori* sur la fonction objectif, mais se base uniquement sur l'analyse des valeurs prises par cette fonction au fur et à mesure de l'exécution. Les approches stochastiques se caractérisent des autres classes d'approches par leur capacité à apprendre automatiquement une stratégie de recherche, ce qui permet de converger efficacement vers un optimum étant donné un large espace de solutions. De telles fonctions sont le recuit simulé, l'algorithme de Monte Carlo, la recherche tabou et les algorithmes évolutionnaires.

A l'issue de cette section, notre objectif consiste à déterminer la fonction objectif la plus adéquate à notre problématique d'optimisation, ainsi que l'algorithme de résolution correspondant.

2.2. Définition de la fonction objectif pour la sélection d'un fichier de ciblage

2.2.1. Réduction de la dimensionnalité de la fonction objectif

Une première approche consiste à simplifier la fonction de gain en réduisant sa dimensionnalité, celle-ci étant bidimensionnelle, s'exprimant en fonction de la qualité et du prix. Pour ce faire,

nous tentons d'exprimer l'une des dimensions en fonction de l'autre, en d'autres termes, d'exprimer la qualité en fonction du prix ou inversement. Nous menons, pour cela, une étude expérimentale sur un échantillon de données faisant intervenir deux attributs de l'ensemble a_c des attributs du fichier de prospection : le numéro de téléphone (a_1) et l'effectif (a_2). Afin de bien simuler l'aspect multisources de notre base de prospection, les données de a_1 et a_2 proviennent de deux sources de qualité globale et de prix différents : s_1 et s_2.

i- Définition de l'expérience

L'idée est de générer plusieurs sous-échantillons aléatoires de s_1 et s_2 en faisant varier à chaque fois les proportions des données de ces deux sources pour les deux attributs a_1 et a_2 ; et de mesurer pour chaque échantillon la qualité et le prix sous-jacents. Pour ce faire, nous effectuons deux séries de tests.

Dans la première série de tests, nous fixons les données de a_1 et nous faisons varier les données de a_2 comme suit.

- Echantillon 1 : a_1 fixe, 80 % de s_1;20 % de $s_2$pour a_2.
- Echantillon 2 : a_1 fixe, 70 % de s_1;30 % de $s_2$pour a_2.
- Echantillon 3 : a_1 fixe, 50 % de s_1;50 % de $s_2$pour a_2.
- Echantillon 4 : a_1 fixe, 30 % de s_1;70 % de $s_2$pour a_2.
- Echantillon 5 : a_1 fixe, 20 % de s_1;80 % de $s_2$pour a_2.
- Echantillon 6 : a_1 fixe, 100 % de $s_1$pour a_2.

Inversement, la deuxième expérience consiste à fixer les effectifs (a_2) en faisant varier l'appartenance des téléphones (a_1) aux sources s_1 et s_2 :

- Echantillon 1 : a_2 fixe, 80 % de s_1;20 % de $s_2$pour a_1.
- Echantillon 2 : a_2 fixe, 70 % de s_1;30 % de $s_2$pour a_1.
- Echantillon 3 : a_2 fixe, 50 % de s_1;50 % de $s_2$pour a_1.
- Echantillon 4 : a_2 fixe, 30 % de s_1;70 % de $s_2$pour a_1.
- Echantillon 5 : a_2 fixe, 20 % de s_1;80 % de $s_2$pour a_1.

ii- Résultats de l'expérience

L'étude montre que les fonctions qualité et prix sont proportionnelles et corrélées (Figure 4.1 et Figure 4.2). Ainsi, plus la qualité est élevée, plus le prix est cher et inversement.

Ainsi, les fonctions linéaires de compromis de type $prix = \alpha * qualité^{\beta}$ ne sont visiblement pas idéales pour résoudre la problématique multiobjectifs d'optimisation du gain qualité/prix (maximisation de la qualité et de diminution de prix).

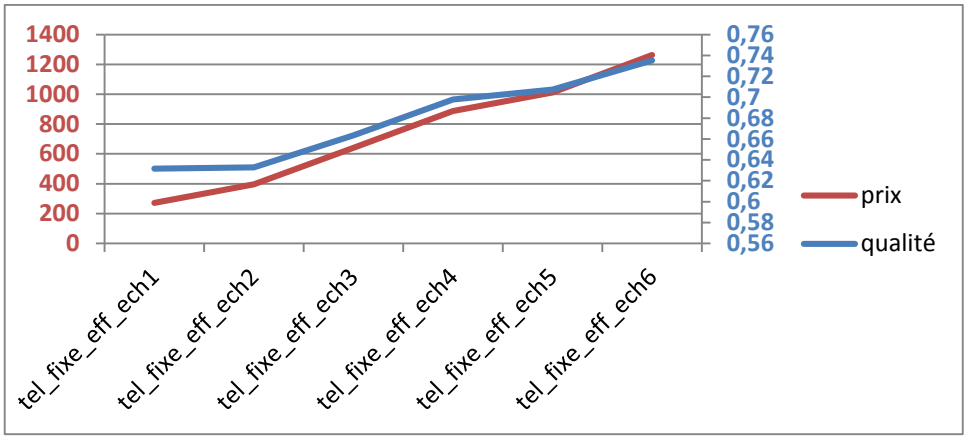


Figure 4.1 Evolution du prix et de la qualité avec a_1 fixe

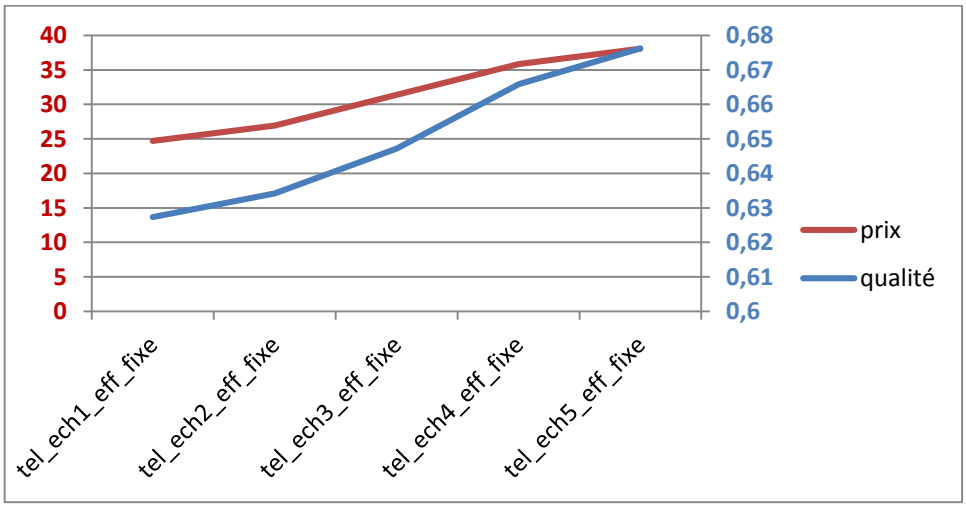


Figure 4.2 Evolution du prix et de la qualité avec a_2 fixe

2.2.2. Etude de la fonction commerciale d'élasticité

Afin de montrer la non-linéarité des fonctions de qualité et de prix, nous nous inspirons des fonctions utilisées en gestion de projets pour la détermination des coûts et des prix. Nous étudions, en particulier, la fonction d'élasticité. Cette fonction mesure la sensibilité de la demande aux variations du prix en calculant la variation de la quantité demandée d'un produit à la suite de la variation de son prix. Elle est quantifiée par la formule :

$$e = \frac{\frac{dQ}{Q}}{\frac{dP}{P}} = -\frac{dQ}{dP} * \frac{P}{Q}$$

La simulation expérimentale de la variation de la qualité des données en fonction de leurs prix montre que l'élasticité mesurée n'est pas stable. Elle dépend des variations de prix et de qualité (Figure 4.3 et Figure 4.4).

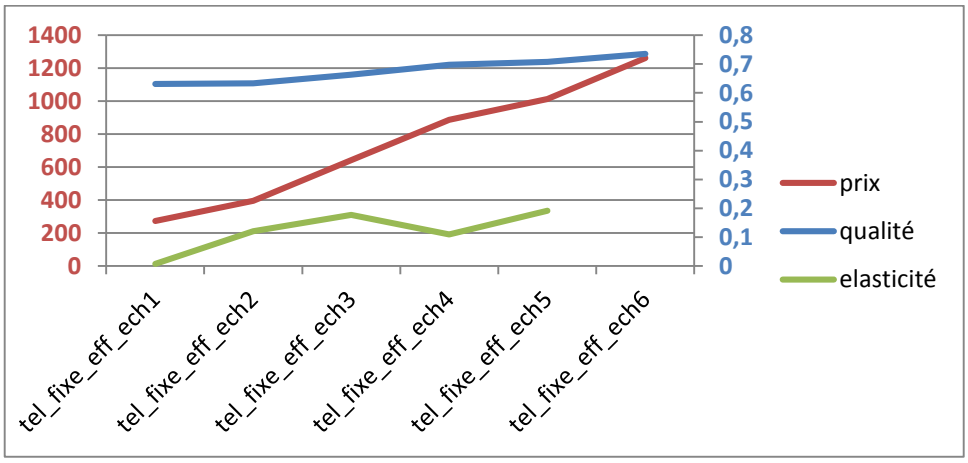


Figure 4.3 Elasticité en fonction de la qualité et du prix avec a_1 fixe

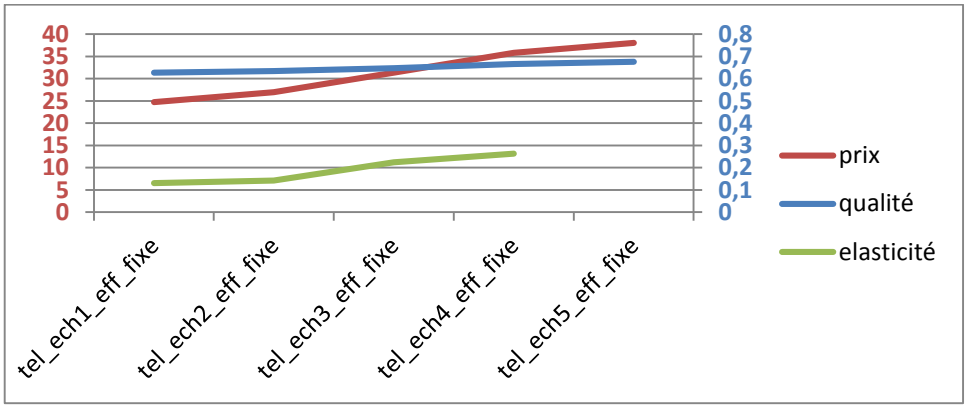


Figure 4.4 Elasticité en fonction de la qualité et des prix avec a_2 fixe

Etant donné ce constat de non-linéarité du prix et de la qualité et donc de non-linéarité de la fonction objectif, la résolution linéaire du problème d'optimisation par une méthodologie de type Simplexe linéaire n'est pas envisageable. De plus, la simplification du problème d'optimisation (F) en un seul problème de minimisation de prix sous la contrainte de qualité ne résout en rien la linéarité du problème puisque la fonction prix (*F1*) est non linéaire. Le recours à des fonctions heuristiques est donc indispensable.

2.2.3. Simulation de la fonction objectif

En l'absence de fonction classique de modélisation du gain qualité/prix, nous proposons l'utilisation d'une fonction d'optimisation approchée. Cette fonction simule la synergie existant entre les deux paramètres d'optimisation (à savoir le prix et la qualité) afin d'approcher la

fonction de gain telle qu'elle est désirée par le décideur. Une des possibilités consiste à utiliser l'intégrale de Choquet discrète précédemment utilisée dans la modélisation de l'agrégation préférentielle des dimensions qualité. L'utilisation de l'intégrale de Choquet discrète est, en effet, un moyen simple pour définir la fonction de profit tout en garantissant son adéquation aux besoins du (ou des) décideur(s). Le décideur définit alors les poids flous à attribuer aux deux critères d'optimisation en exprimant ses préférences de pré-ordre sur un échantillon d'apprentissage. L'intégrale calcule, moyennant l'agrégation floue des capacités, le profit estimé.

2.3. Détermination de l'approche d'optimisation

Dans la suite de ce chapitre, nous nous intéressons à la définition de l'approche d'optimisation dans le contexte de sélection du fichier de ciblage. La fonction objectif étant une fonction non-additive floue (d'après la définition même de l'intégrale de Choquet), nous optons pour des algorithmes d'optimisation heuristiques qui permettent de gérer la complexité de la fonction objectif.

2.3.1. Introduction aux fonctions heuristiques

Généralement, le recours à l'heuristique a lieu dans deux cas de figure :

1 quand la résolution des problèmes d'optimisation difficiles, non linéaires, combinatoires ou encore NP-Complets se ramène à l'examen d'un nombre infini de combinaisons et s'avère coûteuse en termes de temps ou de ressources ;

2 quand on ignore l'existence d'un algorithme de résolution exact rapide qui puisse retrouver l'optimum global en un temps polynomial de l'ordre de N^n où N désigne le nombre de paramètres inconnus du problème et n un entier.

L'optimisation par heuristiques a été expérimentée dans d'innombrables domaines aussi variés que la gestion, l'ingénierie, la conception, les télécommunications, les transports et l'informatique pour traiter des problèmes réels relatifs à l'optimisation des portefeuilles, l'identification des experts sur les réseaux sociaux et, par exemple, le calcul du plus court chemin multicritères [Skolpadungket et al. 07 ; Ahmad et al. 08 ; Fouchal et al. 10]. Les heuristiques proposées sont alors des méthodes approximatives, fournissant une solution réalisable au problème d'optimisation NP-difficile en question, en un temps polynomial. La solution n'est pas nécessairement optimale. Ces techniques de résolution sont, pour la plupart,

inspirées des problèmes génériques de maximisation de profit sous contraintes de ressources limitées. Ces problèmes génériques ont, en effet, été posés par l'informatique théorique et servent souvent de *benchmarks* pour comparer les performances de différents algorithmes de résolutions. De tels problèmes d'optimisation sont, par exemple, le problème du voyageur de commerce (*Travelling Salesman Problem* – TSP) et le problème du sac à dos (*Knapsack Problem* - KP).

S'inspirant du raisonnement heuristique utilisé dans ces deux problèmes classiques d'optimisation, un grand nombre d'approches de résolution ont vu le jour. Ces approches sont appelées métaheuristiques. Ce sont des stratégies générales applicables à un grand nombre de problèmes, à partir desquelles on peut dériver un algorithme heuristique pour un problème particulier. Nous distinguons les méthodes de programmation dynamiques à base d'heuristiques, les approches locales (telles que les algorithmes gloutons et les méthodes stochastiques), les algorithmes de recherche locale (telles que la méthode de descente du gradient), les métaheuristiques de voisinage telles que la recherche taboue et les algorithmes évolutifs, dits aussi évolutionnaires), les métaheuristiques agrégatives telles que l'approche MOGA (*Multiple Objective Genetic Algorithm*) et, les algorithmes à heuristiques non agrégatives tels que la méthode VEGA *Vector Evaluated Genetic Algorithm* [Collette et al. 02 ; Mavrotas 08].

D'après [Collette et al. 02], ces métaheuristiques ont en commun un certain nombre de caractéristiques :

- elles sont, pour la plupart, stochastiques et permettent de faire face à l'explosion combinatoire des possibilités ;
- elles sont d'origine combinatoire : elles ont l'avantage, décisif dans le cas continu, d'être directes, c'est-à-dire qu'elles ne recourent pas au calcul, souvent problématique, des gradients de la fonction objectif ;
- elles sont inspirées par des analogies : avec la physique (recuit simulé, diffusion simulée, etc.), avec la biologie (algorithmes génétiques, recherche tabou, etc.) ou avec l'éthologie (colonies de fourmis, essaims de particules, etc.) ;
- elles sont capables de guider, dans une tâche particulière, une autre méthode de recherche spécialisée (par exemple, une autre heuristique, ou une méthode d'exploration locale).

Par ailleurs, elles présentent des inconvénients relatifs aux difficultés de réglage des paramètres mêmes de la méthode ainsi que l'importance du temps de calcul.

Nous nous focalisons, en particulier, sur les algorithmes évolutionnaires, dont le principe est de faire évoluer progressivement une solution initiale jusqu'à ce qu'un optimum soit atteint. En effet, ces algorithmes ont fait leurs preuves dans les problématiques de maximisation de profit, en particulier, le problème de sélection de portefeuille multiobjectif. Notre intérêt est justifié par les similarités que représente le problème de sélection de portefeuille avec notre problématique de sélection des données multisources [Di Gaspero et al. 07].

2.3.2. Algorithmes évolutionnaires

i- Principe et aperçu des algorithmes évolutionnaires

Les algorithmes évolutionnaires sont des procédures de recherche locales qui s'appliquent à un large spectre de problématiques d'optimisation, notamment celles dont la fonction objectif est irrégulière, à savoir, non convexe, ni linéaire ni quadratique. Ils appartiennent à la famille des approches par population, puisqu'ils manipulent un ensemble de solutions, par opposition aux approches par point qui cherchent à optimiser une solution unique par itération. Cette caractéristique confère à la famille des algorithmes évolutionnaires l'avantage de la rapidité d'exécution des approches de résolution parallèles, et par conséquent, la possibilité de trouver multiples solutions optimales [Kalyanmoy 08].

Les algorithmes évolutionnaires ont la particularité de travailler sur une population de solutions et cherchent à optimiser plusieurs solutions Pareto en une même itération. Il s'agit d'une technique d'optimisation itérative et stochastique, car utilisant itérativement des processus aléatoires. Chaque individu est caractérisé par un génotype (codage), lequel détermine un phénotype (l'interprétation de ce codage). Nous distinguons les algorithmes génétiques, les stratégies évolutionnaires et la programmation évolutionnaire. Tous sont basés sur le principe de l'évolution naturelle et de la théorie darwinienne et guidés par « l'instinct de survie » et le principe de la « sélection naturelle ».

Un algorithme évolutionnaire consiste à manipuler un ensemble de solutions (appelées individus) d'une population donnée, moyennant un ensemble d'opérateurs (dits opérateurs de variation) ; puis de le faire évoluer vers un optimum, moyennant des fonctions d'adaptation (appelées encore fonction d'évaluation ou fonction fitness). Nous distinguons deux opérateurs de variation principaux :

- l'opérateur de croisement qui permet de créer de nouvelles solutions en combinant des parties de deux (ou plusieurs) solutions ;
- l'opérateur de mutation est un opérateur unaire, qui permet de générer de nouvelles solutions en modifiant l'une des solutions de la population.

Grâce aux changements qu'ils opèrent sur les solutions courantes, ces opérateurs jouent un rôle prépondérant dans l'assurance de la diversité de la solution générée. Ils lui permettent, de cette manière, de s'éloigner des optima de voisinage locaux et d'explorer d'autres espaces.

Le schéma traditionnel d'un algorithme évolutionnaire s'articule comme le montre la Figure 4.5.

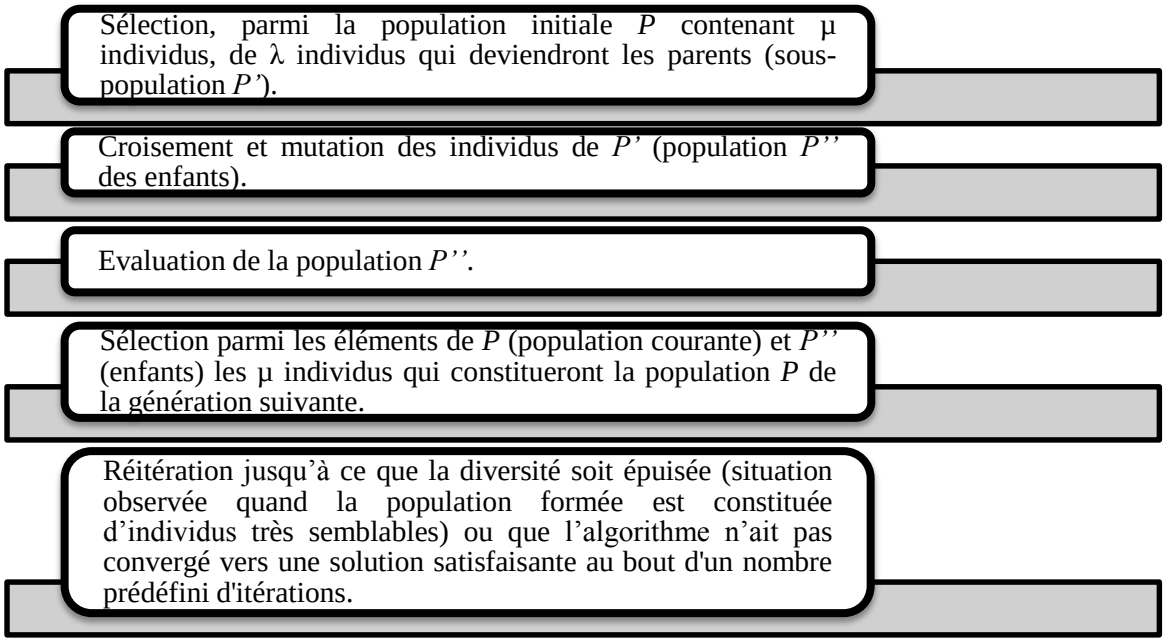


Figure 4.5 Schéma générique d'un algorithme évolutionnaire

La famille des algorithmes évolutionnaires comprend :

- l'algorithme génétique générationnel ;
- l'algorithme génétique stationnaire (le *steady-state algorithm*) ;
- l'algorithme EP (programmation évolutionnaire) ;
- les stratégies d'évolution.

Ces algorithmes sont décrits dans le Tableau 4.1.

Programmation évolutionnaire (Evolutionary Programming) **[Fogel 66, Fogel 95]**	**Algorithmes génétiques** **(Genetic algorithms)** **[Holland 75, Goldberg 89]**	**Les stratégies d'évolution (Evolution strategy)** **[Rechenberg 73, Schwefel 81]**	**La programmation génétique (Genetic programming)** **[Koza 92]**	**L'algorithme génétique stationnaire « Steady-State genetic algorithm »** **[Whitley et al. 88]**
Imaginée par L.J. Fogel dans les années 60 et reprise par son fils D.B. Fogel trente ans plus tard, la programmation évolutionnaire a initialement été conçue pour la découverte d'algorithmes à états finis pour l'approximation de séries temporelles. La PE s'intéresse plus à l'évolution qu'à la programmation et se différencie des algorithmes génétiques par la manière avec laquelle elle encode les paramètres du problème. Ainsi, au lieu d'allouer une représentation spécifique aux individus en codifiant tous leurs paramètres en une chaîne de bits, l'algorithme PE encode uniquement l'état initial, la table de transition et l'alphabet, mettant l'emphase sur la relation entre les parents et leurs descendants plutôt que la simulation des opérateurs	Les algorithmes génétiques sont des algorithmes stochastiques itératifs qui opèrent sur des individus codés, à partir d'une population initiale. L'évolution de la population d'une génération *k* à une génération *k+1* nécessite les trois opérateurs génétiques de croisement, mutation et sélection, la mutation pouvant donner un individu aussi bien voisin qu'éloigné et la sélection modélisant l'adaptation. La démarche la plus courante consiste à croiser un individu, exercer une mutation sur les deux enfants et remplacer les parents par les enfants.	Les stratégies d'évolution sont des algorithmes itératifs. Elles ont été initialement conçues pour résoudre des problèmes d'optimisation continus. Dans un algorithme SE, les individus sont des points (vecteurs de réels). Comme la programmation évolutionnaire, les SE n'utilisent que la mutation et la sélection. L'algorithme le plus simple, noté (1+1)-ES, manipule un seul individu. A chaque génération (itération), l'algorithme génère par mutation un individu enfant à partir de l'individu parent et sélectionne le meilleur des deux pour le conserver dans la population. La mutation dans un tel algorithme est aléatoirement appliquée à tous les composants de l'individu pour produire un enfant. Cette mutation fonctionne selon les principes suivants : un enfant	La programmation génétique est une spécialisation d'algorithmes génétiques où chaque individu est un programme informatique caractérisé par une séquence d'opérations (fonctions) appliquées à des valeurs (arguments) et dont le but est de permettre à cet individu (programme informatique) d'apprendre, grâce à l'algorithme évolutionniste, à optimiser peu à peu une population d'autres programmes pour augmenter leur degré d'adaptation (fitness) à réaliser une tâche demandée par un utilisateur. Les opérateurs de mutation et de croisement correspondent à des échanges de codes entre deux programmes.	L'algorithme SSGA se caractérise par le fait qu'à chaque génération λ parents (généralement choisis par sélection stochastique) produisent μ enfants qui remplacent, dans la génération suivante, les μ pires parents.

Programmation évolutionnaire (Evolutionary Programming) [Fogel 66, Fogel 95]	Algorithmes génétiques (Genetic algorithms) [Holland 75, Goldberg 89]	Les stratégies d'évolution (Evolution strategy) [Rechenberg 73, Schwefel 81]	La programmation génétique (Genetic programming) [Koza 92]	L'algorithme génétique stationnaire « Steady-State genetic algorithm » [Whitley et al. 88]
génétiques d'inspiration naturelle. De cette manière, étant donnée une population initiale, la descendance est générée par mutation appropriée (au problème à résoudre). Une méthode de sélection permet, ensuite, d'établir des tournois pour ne retenir que le meilleur élément.		ressemble à ses parents, et plus (moins) un changement est important, moins (plus) la fréquence de changement est élevée. Cet algorithme (λ+μ)-ES se généralise en un algorithme (m+l)-ES dans lesquels λ parents génèrent μ enfants et où une sélection ramène les λ+μ individus à λ individus (les meilleurs survivent). [Hao et al. 99] Les stratégies d'évolution s'écartent de l'évolution naturelle en modifiant tous les individus à chaque génération. Par ailleurs, elles se distinguent des algorithmes génétiques par leur désir de conserver un codage lisible, et, contrairement aux algorithmes évolutionnaires, elles n'exigent pas d'hypothèses par rapport à l'espace d'état.		

Tableau 4.1 Les grandes familles des algorithmes évolutionnaires

ii- Algorithme de fourmis artificielles (Ant Colony Optimization-ACO)

L'algorithme ACO est une procédure de recherche stochastique constructive d'inspiration éthologique qui imite le comportement des fourmis réelles dans leurs activités de fourragement, notamment pour trouver le plus court chemin reliant le nid à un point de nourriture et où, par l'action autonome de chacune des fourmis, un comportement global émergeant apparaît (phénomène appelé communication coopérative ou encore stigmergie). Cette stigmergie observée chez les insectes sociaux en général et les fourmis en particulier est mise en œuvre grâce à une communication chimique moyennant le dépôt de phéromones. Cette substance permet d'échanger des informations qualitatives et assure, grâce à sa stabilité relative, une certaine permanence. Deux phénomènes sont alors observés : l'évaporation et le renforcement des phéromones. L'évaporation permet aux fourmis d'explorer de nouveaux trajets et le renforcement permet d'assurer la convergence vers le trajet optimum.

Les fourmis artificielles se comportent de la même façon, mimant la communication chimique basée sur l'évaporation et le renforcement des phéromones. Ainsi, l'évaporation permet d'effacer l'historique des solutions précédemment découvertes, permettant l'exploration de nouveaux espaces. Le renforcement est utilisé, quant à lui, pour assurer la fonction d'apprentissage permettant d'orienter la recherche vers l'espace des solutions Pareto-optimales. A la différence des insectes réels, ces fourmis artificielles (approche aussi appelée intelligence collective artificielle ou intelligence en essaim *Swarm Intelligence* [Abraham et al. 06]) possèdent une mémoire, ne sont pas totalement aveugles et évoluent dans un espace temporel discret qui permet l'appréciation de la performance de la solution aux instant t et $t+1$.

Outre son efficacité à résoudre les problèmes combinatoires les plus complexes en un temps raisonnable [Dorigo et al. 04, Dorigo et al. 96, Dorigo et al. 05], l'algorithme de fourmis (ACO) a, entre autres, été utilisé dans les problèmes d'optimisation de la sélection, notamment, pour la sélection des variables d'apprentissage optimales dans les processus décisionnels (problématique connue sous le terme *feature selection* qui consiste à réduire la dimensionnalité de l'échantillon d'apprentissage en choisissant les attributs d'apprentissage les plus prédictifs) [Jensen et al. 03, Jensen 06, Al Ani 05, Abraham et al. 06] ; d'où notre intérêt.

Le principe de l'algorithme générique de colonie de fourmis (algorithme ACO) a été défini dans le cadre de résolution de l'algorithme TSP et se présente comme le montre la Figure 4.6.

Début

- Initialisation des paramètres, initialisation des taux de phéromones.

Début de l'activité

- Construction des solutions pour chaque fourmi
- Recherche locale (définition de la meilleure solution de l'itération courante, étape optionnelle)
- Mise à jour des phéromones

Fin de l'activité

Fin

Figure 4.6 Principe des colonies de fourmis

Ainsi, chaque fourmi se voit attribuer aléatoirement le premier élément de la solution (la ville dans l'exemple TSP). A chaque itération, chaque fourmi *k* utilise une règle probabiliste (règle de transition dans le cas TSP) lui permettant de choisir le nouvel élément *j* à inclure à la solution actuelle (la nouvelle ville dans l'exemple TSP). Une telle règle est définie par la formule suivante :

$$p_{S_k(j)} = \frac{{\zeta_{S_k(j)}}^{\alpha} {\eta_{S_k(j)}}^{\beta}}{\sum_{l \in Candidates^k} {\zeta_{S_k(l)}}^{\alpha} {\eta_{S_k(l)}}^{\beta}} \quad (1)$$

où

- S_k est la solution actuelle construite par la fourmi *k* ;
- $\zeta_{S_k(j)}$ représente le nouveau taux de phéromones résultant de l'ajout de l'élément *j* à la solution courante (la probabilité de transition vers une ville *j*, dans l'exemple du TSP) ;
- $\eta_{S_k(j)}$ représente l'information heuristique assurant le caractère aléatoire de la recherche qui encourage l'exploration de nouveaux espaces de solutions ;
- α et β sont deux paramètres permettant de jauger l'influence relative des taux de phéromones et de l'information heuristique ;
- $Candidates^k$ représente le voisinage possible de la solution actuelle S_k.[25]

Les deux phénomènes biologiques d'évaporation et de renforcement des phéromones sont modélisés par une mise à jour du taux de phéromone.

- L'évaporation est modélisée par la formule $\zeta_{S_k(j)} = (1-\rho)\zeta_{S_k(j)}, \forall j \in S_k$; ρ représentant le taux d'évaporation.

[25] Dans l'exemple du problème du voyageur de commerce, S_k définit la liste des villes non encore explorées par *k*, ne faisant pas partie de S_k.

- Le renforcement est modélisé par $\zeta_{S_k(j)} = \zeta_{ij} + \sum_{k=1}^{m} \Delta\zeta_{S_k(j)}, \forall j \in S_k$; où $\Delta\zeta_{S_k(j)}$ définit la quantité de phéromones déposée par la fourmi *k* sur les éléments de la solution courante S_k.

La règle probabiliste (1) montre que le choix d'inclusion d'un élément *j* à la liste actuelle des solutions dépend de deux éléments principaux : le taux de phéromones et l'information heuristique lesquels sont conditionnés par les deux paramètres α et β. Ainsi, si α=0, les éléments sélectionnés sont ceux qui améliorent la performance de la fonction heuristique $\eta_{S_k(j)}$ et les performances de l'algorithme ACO sont comparables à celles d'un algorithme glouton. Par contre, si β=0, l'algorithme chercherait uniquement à maximiser les taux de phéromones représentés par la fonction $\zeta_{S_k(j)}$, écartant tout biais heuristique et toute recherche exploratoire. Par ailleurs, toute valeur de α>1 engendrerait un renforcement de phéromones, et donc une convergence rapide vers un état stationnaire où toute les fourmis choisissent la même solution. Cette caractéristique modélise ainsi l'apprentissage artificiel [Dorigo et al. 05]. En conclusion, l'utilisation des traces de phéromone permet d'exploiter l'expérience de recherche acquise par les fourmis et de renforcer ainsi l'apprentissage pour la construction de solutions dans les itérations futures de l'algorithme (apprentissage par renforcement) alors que l'information heuristique peut guider les fourmis vers les zones prometteuses de l'espace de recherche. Enfin, toutes les fourmis travaillent pour construire *k* solutions en une même itération. Ces solutions peuvent se construire soit parallèlement, soit itérativement (au cas où une synchronisation est nécessaire).

iii- Discussion

Malgré les avantages offerts par les algorithmes évolutionnaires, qui facilitent leur convergence vers une solution satisfaisante, ces algorithmes offrent à l'utilisateur (décideur) le « leurre » d'une solution satisfaisante à tous les problèmes d'optimisation difficiles qu'ils soient non linéaires, à contraintes et objectifs non différentiables, à variables de décisions incertaines ou encore à multiples optima. Une approche de résolution plus judicieuse consiste à envisager une approche hybride profitant à la fois des avantages respectifs d'un algorithme classique rigoureux et d'un algorithme évolutionnaire heuristique. Notons que certaines études [Coello Coello 02] considèrent l'algorithme ACO comme une forme d'hybridation puisqu'il s'agit d'un système multi-agents (SMA) où les interactions entre les agents s'inspirent du comportement biologique de la colonie de fourmis.

2.3.3. Approche hybrides d'optimisation

Plusieurs hybridations ont été proposées pour la résolution des problèmes d'optimisation multiobjectifs dits difficiles, en particulier, ceux exprimant des contraintes complexes non linéaires ou des contraintes d'inégalités. L'hybridation se justifie par le fait que l'algorithme évolutionnaire seul ne peut pas résoudre efficacement les problèmes d'optimisation sous contraintes [Coello Coello 02]. Les hybridations les plus répandues sont celles combinant les algorithmes génétiques (ou plus généralement les algorithmes évolutionnaires) aux algorithmes de recherche locale connus sous le nom de GRASP *Greedy Randomized Adaptative Search Procedure* où un ensemble de solutions est, d'abord, construit selon un principe glouton aléatoire, puis amélioré par une méthode de recherche locale. Dans ce contexte, des approches ont été proposées pour améliorer les performances des algorithmes génétiques dans la résolution de l'algorithme et TSP en les combinant avec la métaheuristique des ACO [Abbatista et al. 96].

i- Quelques exemples d'hybridation

En 2007, [Di Gaspero et al. 07] proposent d'améliorer les performances de l'algorithme quadratique, utilisé dans le cadre de l'optimisation de la sélection de portefeuilles d'actifs, en le combinant à un algorithme glouton de recherche locale, et ce au cas où aucune solution faisable n'est trouvée par l'algorithme quadratique initial. L'expérience a été réalisée avec des algorithmes de recherche locale, tels que la descente du gradient et la recherche tabou, et montre des résultats meilleurs que les approches classiques, traditionnellement utilisées jusqu'alors pour l'optimisation de la sélection de portefeuilles d'actifs, et ce, en terme de temps d'exécution et de minimisation du risque du portefeuille généré.

Un autre exemple d'hybridation [Meyer 08] couple ACO avec un algorithme de propagation de contraintes : l'algorithme CP[26]. ACO est en effet choisi grâce à sa capacité à apprendre rapidement une stratégie de recherche adéquate d'une solution optimum parmi un large espace de solutions. Par ailleurs, l'utilisation de l'algorithme CP permet de réduire efficacement l'espace de recherche en éliminant les sous-espaces superflus ne satisfaisant pas la problématique d'optimisation.

[26] L'algorithme CP est un algorithme de résolution des problèmes par contraintes qui utilise les contraintes du problème pour réduire la taille de l'espace des solutions à explorer.

ii- Approches interactives : un cas particulier d'hybridation

Nous nous penchons sur un cas particulier des algorithmes hybrides : les algorithmes interactifs. Généralement, ces approches s'intéressent à l'articulation des préférences dans un contexte d'optimisation décisionnelle et le concept d'optimalité est remplacé par celui de l'efficacité, laquelle est exprimée par la Pareto-optimalité.

Dans ce contexte, nous citons l'algorithme ROR *Robust Ordinal Regression* [Branke et al. 10] qui combine un algorithme d'optimisation évolutionnaire (NSGA-II) avec une procédure interactive d'aide à la décision multicritères dans le but d'améliorer la convergence du premier. Les préférences du décideur y sont modélisées par des fonctions additives (encore appelées fonctions d'utilité), le but étant d'approcher une solution Pareto-optimale. Ces préférences sont exprimées soit par un classement de pré-ordre entre les alternatives ou via des comparaisons de l'intensité de préférence par couple d'alternatives.

[Klamroth et al. 08] définissent aussi une approche d'optimisation hybride où un algorithme interactif de type a posteriori est combiné à un algorithme d'optimisation approchée pour affiner, au fur et à mesure de la consultation du décideur, un ensemble de solutions approximatives préalablement définies. Ce dernier détermine les points de référence (ou intervalles de valeurs) à explorer pour améliorer l'efficacité de la solution d'optimisation. Les points de référence en question sont détectés, puis mis à jour, moyennant des outils de visualisation qui permettent d'exclure et filtrer les solutions inintéressantes. Par ailleurs, le processus étant itératif, il permet au décideur de mieux apprendre le problème sous-jacent et d'approfondir, par conséquent, les solutions qui lui semblent intéressantes, sans avoir à comparer plusieurs solutions à la fois. De plus, comme l'utilisateur a la possibilité d'ajuster ses préférences, la résolution du problème est peu coûteuse en temps de calcul.

Un autre exemple d'hybridation interactive concerne la combinaison d'un algorithme évolutionnaire et d'un algorithme de programmation linéaire pour l'optimisation de la gestion de portefeuilles et la maximisation des retours sur investissements [Subbu et al. 05]. Les algorithmes génétiques de type PSEA *Pareto Sorting Evolutionary Algorithm* et TOGA *Target Objective Genetic Algorithm* sont d'abord utilisés pour la génération d'un ensemble de solutions initiales, qui sont stockées dans un entrepôt. Ces solutions sont ensuite épurées moyennant un filtre de dominance. Un ensemble de solutions Pareto-optimales non-dominées est alors généré. Le décideur exprime alors ses préférences afin de ne garder que la (ou les) meilleure(s). Pour ce faire, des prédilections ordinales, beaucoup plus intuitives que les cardinales, sont

progressivement introduites par critère d'importance (dans le contexte d'une prise de décision multicritères).

2.3.5. Conclusion

Dans cette section, nous avons présenté une revue de l'ensemble des approches d'optimisation évolutionnaires. Cette étude montre l'intérêt des approches hybrides et interactives pour assurer la rapidité de convergence des algorithmes d'optimisation. Comme le montre le paragraphe ii, l'hybridation se rapporte à un algorithme glouton évolutionnaire dont la fonction objectif se mesure soit par une approche déterministe, soit par une fonction additive [Branke et al. 10, Subbu et al. 05]. L'interaction apparaît dans les préférences du décideur qui sont modélisée par différentes approches : un classement ordinal [Subbu et al. 05], une définition des points de référence [Klamroth et al. 08] ou encore un pré-ordre entre les alternatives [Branke et al. 10].

3. Broker ACO, une approche d'optimisation guidée par les fourmis

Dans cette section, nous décrivons notre proposition d'optimisation : BrokerACO, une approche interactive et hybride basée sur la métaheuristique de l'algorithme de fourmis dont l'objectif est de maximiser le gain qualité/prix d'une sélection multisources que représente un fichier de ciblage marketing. A la différence des approches précédemment décrites dans la revue de littérature (Paragraphe ii, Section 2.3.2), où les préférences des décideurs sont définies à postériori de la génération des solutions, nous modélisons nos préférences pour qu'ils représentent le compromis qualité/prix et nous les injectons dans l'étape de calcul de la fonction de survie (*fitness*) afin de contrôler, automatiquement, la production des générations.

3.1. Modélisation des préférences

Contrairement aux approches définies dans les Sections 2.3 et 2.4, les préférences du décideur sont quantifiées par une fonction *f* représentant le compromis qualité/prix, contraintes de notre problématique d'optimisation.

Selon les principes de base des règles de décision [Fargier et al. 09] (Chapitre 3, Section 3.2.3), une telle fonction *f* se caractérise par deux critères principaux :

- le principe d'unanimité (appelé aussi principe de dominance ou procédure de Pareto), qui exige que *f* soit monotone croissante ;
- le principe d'efficacité, qui exige que *f* soit monotone, strictement croissante.

L'intégrale de Choquet, étant monotone, ascendante d'une part, et permettant la modélisation des préférences d'autre part, apparaît comme la fonction adaptée pour modéliser et automatiser cette fonction *fitness*. Le compromis qualité/prix est alors assimilé à une agrégation bi-critères (les critères étant la qualité et le prix), et les préférences sont apprises grâce à un échantillon fictif qui représente l'ensemble des valeurs possibles que peuvent prendre ces deux critères.

3.2. Principe de BrokerACO

Nous rappelons la fonction objectif de notre problématique d'optimisation, définie dans la Section1.

$$Maximiser \sum_{j} \sum_{i} \psi(\pi_{ij}, \sigma_{ij}) \ (F)$$

Comme dans tout algorithme de fourmis, BrokerACO se base sur les deux paramètres η et ζ pour améliorer ses performances de convergence vers la solution optimale. η modélise l'attractivité qui incite les fourmis artificielles à la recherche locale. ζ définit la visibilité qui encourage à l'exploration. Etant données ces définitions, nous assimilons η à la variable prix (incitant l'algorithme à choisir des enregistrements de la même source) et ζ à la variable qualité (privilégiant l'exploration de nouvelles sources en quête d'une meilleure qualité).

Par ailleurs, nous choisissons une stratégie élitiste afin d'optimiser les chances de convergence rapide de l'algorithme vers une solution optimale. Cette stratégie consiste à mettre à jour le taux de phéromones (ζ) de la meilleure solution intermédiaire, celle réalisant le meilleur score de la fonction objectif sur un cycle donné. En effet, les cycles sont exécutés itérativement, permettant aux fourmis d'une itération *t* d'apprendre à mieux résoudre le problème d'optimisation sous-jacent à partir des performances des générations précédentes (*t-1* en l'occurrence). Ainsi, en augmentant le taux de phéromones de la meilleure solution de la génération *t-1*, nous augmentons leur attractivité par les fourmis de la génération *t*. Inversement, les solutions non performantes voient leurs taux de phéromones s'évaporer progressivement afin de diminuer leurs chances d'être choisies par les fourmis de la génération *t*.

La Figure 4.7. définit l'algorithme BrokerACO.

Début

Etape 0

- **Initialisation de η et ζ**

Etape 1

Répéter

Pour chaque fourmi de la colonie, faire

- Construire une solution itérativement et sauvegarder les performances de la solution (en termes de qualité et de prix).
- Mettre à jour les phéromones afin d'augmenter l'attractivité de la meilleure solution locale (indicateur η) : mise à jour locale.

Fin Pour

- Calculer l'intégrale de Choquet, identifier la solution ayant la meilleure performance et sauvegarder ses performances (prix, qualité).
- Mettre à jour le taux de phéromones pour les composants de la meilleure solution : mise à jour globale

Jusqu'à ce qu'un nombre maximal d'itérations soit atteint

Etape 2

- Identifier la meilleure solution sur l'ensemble des itérations

Fin

Figure 4.7 L'algorithme BrokerACO

Les Figure 4.8 et Figure 4.9 détaillent la première étape de BrokerACO.

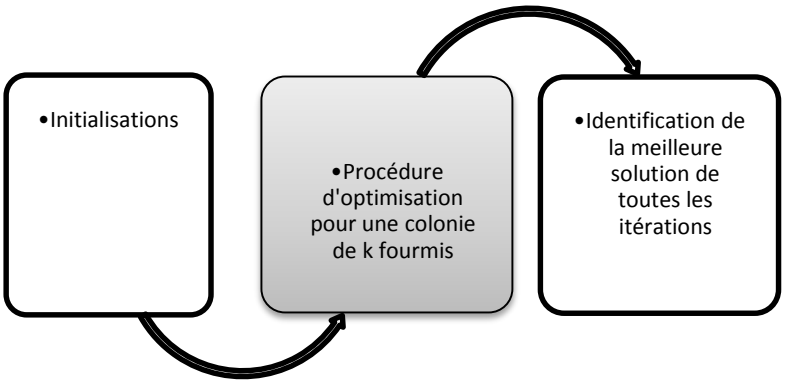


Figure 4.8 Etape1 de BrokerACO

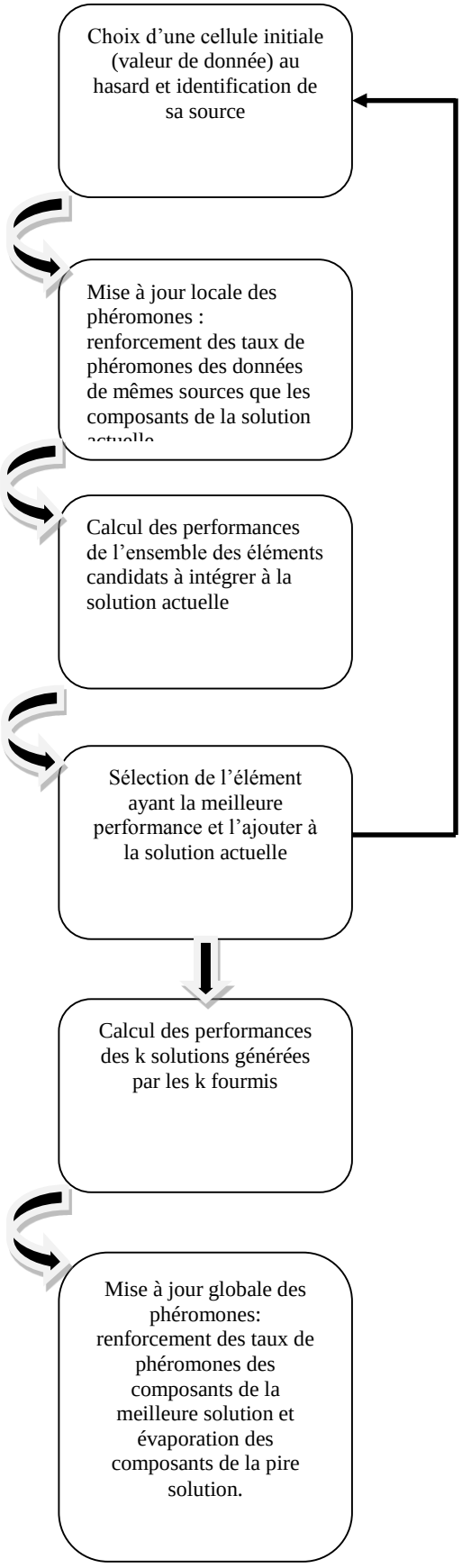


Figure 4.9 Détails de la procédure d'optimisation

4. Conclusion

Ce chapitre détaille les principes de BokerACO, notre approche d'optimisation de la sélection de données multisources en fonction des critères de qualité et de prix. L'approche s'articule sur deux phases :

- la modélisation de la fonction objectif à partir de la simulation des préférences des décideurs. La modélisation des préférences se base, pour ce faire, sur l'intégrale de Choquet discrète. L'intégrale de Choquet est en effet capable de reproduire et quantifier l'agrégation des scores de qualité et de prix telle qu'elle est souhaitée par les décideurs ;
- l'optimisation de cette fonction objectif moyennant l'algorithme de fourmis qui privilégie la recherche locale et collaborative, augmentant ainsi les chances de l'approche de converger rapidement vers une solution optimale.

A notre connaissance, cette hybridation de l'algorithme de fourmis et de l'intégrale de Choquet discrète pour la sélection (ou intégration) sous contraintes de données multisources est une première dans le domaine de la qualité des données. En effet, l'utilisation de l'intégrale de Choquet pour l'expression et la modélisation de la fonction objectif garantit le respect des contraintes et la satisfaction des attentes des décideurs métier. En l'absence de fonctions modélisant rigoureusement ce profit (qualité, prix) (cf. expériences de la Section 2.2.1), la simulation par l'intégrale de Choquet semble être la meilleure alternative. De plus, l'utilisation de l'algorithme de fourmis permet de garantir la convergence rapide de la solution vers un optimum de l'espace Pareto.

Chapitre 5 : Application au processus de brokering en prospection marketing monocanal

« La théorie c'est quand on sait tout et que rien ne fonctionne. La pratique c'est quand tout fonctionne et que personne ne sait pourquoi. »
Albert Einstein

Nous nous intéressons dans ce chapitre à la validation de notre approche d'intégration de données multisources guidée par la qualité des données. Pour ce faire, nous appliquons l'agrégation préférentielle de dimensions qualité au *brokering* de fichiers d'adresses de MaisonPhoning, l'opérateur de télécommunication Français dont nous parlions en introduction. En effet, MaisonPhoning souhaite réaliser une compagne marketing de prospection B-to-B et nous l'accompagnons, en particulier, dans le ciblage de ces prospects. Il s'agit de sélectionner les prospects (les entreprises dans le contexte B-to-B) à partir d'un ensemble de fichiers d'adresses multisources et multi-fournisseurs. Ces fichiers sont qualifiés par des prix et des niveaux de qualité différents. Des contraintes budgétaires ainsi que d'autres relatives au niveau minimum de qualité requis conditionnent cette sélection et un compromis prix/qualité doit alors être exprimé, modélisé et quantifié.

Le plan du chapitre est le suivant. Nous décrivons tout d'abord le contexte de notre application, à savoir la campagne de prospection marketing au sein de l'entreprise MaisonPhoning. Nous appliquons ensuite notre méthodologie d'évaluation de la qualité des données dans les bases de prospection multisources et nous déterminons les dimensions qualité les plus pertinentes à utiliser, une fois rigoureusement agrégées, comme critères de la sélection multisources des données de ciblage. Nous décrivons, pour ce faire, le fonctionnement de BrokerACO, notre outil d'aide à la prospection, dont l'objectif est d'optimiser la sélection multisources en fonction de contraintes de coûts et de qualité. Nous évaluons finalement les performances de BrokerACO sur des campagnes marketing B-to-B réelles en les comparant aux performances d'un broker humain.

1. Analyse contextuelle de l'environnement de prospection

Pour le recrutement de nouveaux clients et l'enrichissement du fichier récurrent, l'entreprise MaisonPhoning effectue une série de campagnes, dites de prospection. Elle adresse ses offres à ses prospects moyennant divers canaux de communication tels que l'adresse postale, l'email, le téléphone fixe ou mobile et le fax. Dans le cas particulier du marketing B-to-B, la base de prospection est principalement constituée des données du répertoire officiel SIRENE fourni par l'INSEE. Le répertoire SIRENE est considéré, grâce à son exhaustivité, comme le socle de toute base de prospection B-to-B. Ce fichier, initialement dédié aux entreprises immatriculées au répertoire des entreprises et des établissements, ainsi qu'aux créateurs d'entreprises pour les aider dans leurs démarches administratives, est aussi enrichi par un ensemble de fichiers ou autres « mégabases de données » permettant :

- en cas d'enrichissement horizontal, d'avoir des informations complémentaires telles que les liens capitalistiques des entreprises ou les adresses email de certains employés stratégiques des entreprises ;
- en cas d'enrichissement vertical, d'ajouter une liste d'entreprises récemment créées qui n'ont pas encore été intégrées à la base SIRENE.

Les fichiers d'enrichissement sont généralement des fichiers de compilation et de comportements avérés consistant en une liste d'identifiants, de noms et de coordonnées de personnes physiques ou morales, classées selon différents critères (géographique, profession, domaine d'activité, etc.). Ces fichiers sont soit loués, soit achetés auprès de fournisseurs officiels ou opportunistes.

Dans la suite de cette section, nous analysons la base de prospection de MaisonPhoning et décrivons ses données les plus critiques relativement au contexte de la réalisation de campagnes marketing.

1.1. Analyse des données

L'analyse de la base de prospection de MaisonPhoning montre que celle-ci est formée d'un ensemble de 27 sources réparties comme suit :

- le fichier SIRENE de l'INSEE (appelé source s_1) comprenant l'ensemble des entreprises françaises qualifiées par un certain nombre d'informations (m_1 attributs : $a_{11},\ldots, a_{1m1}$)

telles que leur identifiant (SIRET), leur raison sociale, leur adresse, leur numéro de téléphone, leur activité, leur effectif salarié, etc. Cette base socle de rapprochement compte aujourd'hui environ 7 millions de lignes (n_1 enregistrements) ;

- un fichier d'enrichissement (appelé source s_2). s_2 désigne le fichier de qualification des entreprises en contacts stratégiques, effectif, numéros de téléphone et fax et compte environ 3 millions d'enregistrements. Ce fichier est fourni par une agence de conseil en marketing compile des données provenant de :
  - l'INSEE (les évènements mensuels de déménagement et les cessations d'activité),
  - le fichier « Kairos » des journaux d'annonces légales (JAL) dénombrant les évènements qualifiant les entreprises qui vont de leur création à leur disparition, en passant par tous les événements qui jalonnent leur existence (modifications, cessions, fusions, locations gérances...) et qui se différencie du fichier de l'INSEE par sa fraîcheur,
  - l'INPI (Institut National de la Propriété Industrielle) décrivant les brevets et marques déposés par les entreprises,
  - d'autres fichiers provenant du greffe des tribunaux et bilan des entreprises ;
- 25 sources de qualification des entreprises en emails définissant les sources s_3, s_4, …, s_{27}. Ces sources sont décrites par m_i attributs $i \in [3,27]$ décrivant des contacts de certaines entreprises françaises définies par leur SIRET, avec des informations de qualification telles que l'email du contact et sa nominativité, tels que $m_3=m_4=\ldots=m_{27}$. L'ensemble de cette mégabase contient environ 3 millions d'enregistrements au total.

1.2. Analyse du processus général de prospection

Pour la réalisation de leurs campagnes de prospection, les responsables marketing de MaisonPhoning doivent tout d'abord déterminer le canal marketing. Dans le marketing relationnel, nous distinguons les canaux suivants :

- l'adresse postale pour la réalisation de campagnes courrier (*mailing ou publipostage*) : il s'agit de l'envoi postal d'une enveloppe, d'une lettre, d'un coupon-réponse ou encore d'un catalogue. L'entreprise MaisonPhoning aurait ainsi principalement besoin des informations de la source s_1. Notons que cette campagne pourrait solliciter la source s_2 au cas où l'entreprise ciblerait les dirigeants des entreprises ;

- le téléphone pour la réalisation de campagnes téléphoniques (*phoning*) : il s'agit de contacter le prospect téléphoniquement afin de lui présenter l'offre ou le produit. Dans ce cas de figure, MaisonPhoning solliciterait les données des sources s_1 et s_2 ;
- le fax pour la réalisation de campagnes *faxing* : il s'agit d'envoyer le message par télécopie. Dans ce cas, MaisonPhoning solliciterait les informations des sources s_1 et s_2 ;
- l'email pour la réalisation de campagnes courriel (*emailing* ou *email*) où MaisonPhoning solliciterait les sources s_1, s_3, s_4, …, s_{27}.

Une fois le canal marketing sélectionné, les responsables de la campagne sélectionnent les prospects à approcher. Il s'agit d'identifier, dans la base de prospection, un certain nombre d'entreprises prospect définies par leur code SIRET unique. Notons id_{ik} cet identifiant (tel que $i \in [1, |S|]$ et $k \in [1, n_i]$). Ces identifiants SIRET sont ensuite enrichis par les informations d'adressage et autres attributs de qualification selon les besoins des gestionnaires de la campagne marketing. L'ensemble de ces informations forme notre fichier de ciblage, dont le processus de création est décrit dans la Figure 5.1.

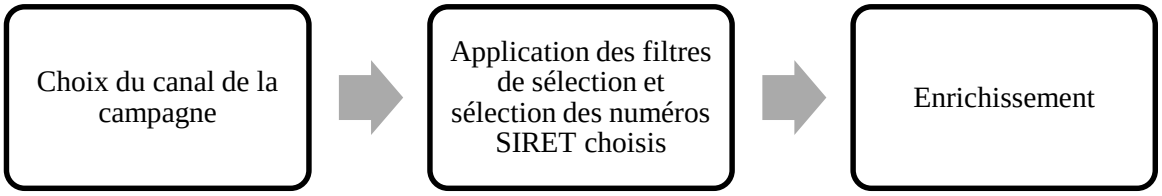


Figure 5.1 Processus de ciblage de l'entreprise MaisonPhoning

Le choix du (ou des) canal (canaux) marketing constitue un premier filtre de sélection dans la base de prospection. En effet, les informations des données d'adressage telles que les adresses postales, les adresses emails ou encore les numéros de téléphones ne sont pas forcément renseignées par tous les fichiers de prospection. Ainsi, après la sélection du canal de prospection, nous obtenons une population P_1 d'identifiants SIRET. Dans l'exemple d'une campagne emailing, la population P_1 recense l'ensemble des entreprises dont l'email du contact principal est renseigné. Les entreprises de P_1 sont alors celles données par s_1 et s_3, s_4, …, s_{27} (fournisseurs des adresses emails).

La deuxième phase du processus consiste en l'application des filtres de ciblage à la population P_1. Ces filtres concernent, par exemple, le secteur d'activité des entreprises prospectées et leur

effectif salarié. Les identifiants SIRET des entreprises satisfaisant ces critères sont alors sélectionnés et forment la sous-population P_2.

Les SIRET de P_2 sont par la suite enrichis, d'une part, par les informations de ciblage (informations d'adressage et détails des critères de sélection telles que la raison sociale de l'entreprise, les nom et prénom du contact principal) et, d'autre part, par des informations de qualification telles que l'enseigne d'une entreprise ou encore son chiffre d'affaire.

1.3. Analyse des besoins en qualité des données

1.3.1. Exigences qualité imposées par la stratégie de ciblage

La stratégie de ciblage telle que décrite dans la Section 1.2 impose des contraintes qualité minimales sur la base de prospection interrogée. Pour expliciter notre propos, reprenons notre exemple directeur de réalisation d'une campagne emailing de prospection.

1 La première phase consiste à sélectionner les numéros SIRET des entreprises ayant une adresse email renseignée. A ce stade, nous ne cherchons pas à savoir si les adresses email affectées aux contacts des entreprises sont exactes ou cohérentes. Seule leur présence nous importe.

2 La deuxième phase consiste à prendre, parmi les numéros SIRET sélectionnés, ceux qui répondent aux critères de filtrage ; à savoir l'effectif salarié et le secteur d'activité de l'entreprise. Or, l'information sur l'effectif salarié peut provenir de deux sources différentes : s_1 et s_2 et peut ainsi présenter une incohérence. Par exemple, dans notre requête de ciblage (que nous appelons désormais *Q*), nous nous intéressons aux entreprises employant moins de 50 salariés. Supposons, pour un SIRET donné, que l'effectif donné par s_1 soit de 40 (cœur de cible) et celui donné par s_2 soit de 160 (hors cible). Si nous faisons confiance à l'information donnée par s_1 et que cette information s'avère fausse, nous risquons de contacter des entreprises hors cible, ce qui affectera directement la rentabilité de la campagne. Si, par contre, nous faisons confiance à l'information donnée par s_2 (et donc nous ne sélectionnons pas le numéro SIRET en question) et que cette information s'avère fausse, nous risquons de passer à côté d'un client potentiel.

3 La troisième phase est encore plus critique, principalement parce qu'elle se rapporte à l'enrichissement de la population par les données d'adressage. Ainsi, toujours en se référant à l'exemple de la campagne emailing, privilégier une alternative qui s'avèrera fausse constitue une perte conséquente d'argent étant donné que certains attributs comme

l'adresse email sont loués et dans certains cas achetés et donc impliquent un certain coût par utilisation. De plus, cela représente une perte de client potentiel.

Outre les problèmes de cohérence potentiels induits par le rapprochement des données multisources, d'autres problèmes qualité rattachés à la valeur intrinsèque de la donnée peuvent apparaître. Ces problèmes de qualité sont par exemple l'inexactitude, l'obsolescence, l'incomplétude, etc.

1.3.2. Analyse des données

Nous étudions les 27 sources mises à disposition pour les besoins de prospection et nous remarquons les problèmes de qualité suivants :

- au niveau de s_1 (Tableau 5.1) :
  - problèmes d'exactitude au niveau de l'attribut tel décrit dans les 2e et 4e enregistrements, où le premier numéro de téléphone est une séquence automatique de 0 et de 1 et où le deuxième ne respecte pas le modèle de standardisation français,
  - problèmes de cohérence (exemple : la tranche d'effectif salarié de l'entreprise décrite indique une information différente de celle donnée par la variable effectif salarié[27] comme c'est le cas pour l'entreprise d'identifiant 00032345600032 définie dans le 3e enregistrement),
  - problème de complétude (exemple : le champ tel manque dans la description de l'entreprise d'identifiant 00034444600032 (5e enregistrement), impliquant la non complétude de l'attribut tel) ;
- au niveau de s_2 (Tableau 5.2) :
  - problème de complétude (la variable Fonction n'est pas souvent renseignée dans la source s_2, comme c'est le cas pour l'entreprise d'identifiant 00123422900021, 2e enregistrement),
  - problème de doublons entraînant un problème de cohérence où le même téléphone appartient à deux entreprises de SIREN[28] et adresses différents ;
- au niveau de s_3, s_4, …, s_{27} (Tableau 5.3) :

[27] La tranche d'effectif salarié correspond à la variable TEFEN/TEFET (selon qu'il s'agisse d'une entreprise ou d'un établissement) du répertoire SIRENE de l'INSEE. Ici, il s'agit de la variable TEFEN (tranche d'effectif entreprise) où les effectifs 1, 2 appartiennent à la tranche d'effectifs '01', les effectifs de 3 à 5 appartiennent à la tranche '02' et les effectifs de 6 à 9 appartiennent à la tranche '03'.

[28] Le SIREN est composé des 9 premiers caractères du SIRET

- problème de complétude (les attributs Email, Nom_codé et Prenom_codé ne sont pas souvent renseignés dans les fichiers s_3, s_4, …, s_{27}). En effet le code "DA390023" n'est autre que le cryptage de la valeur "NULL" dénotant une absence d'information ;
- problème de doublons intra et inter-sources ;

- problème de cohérence au niveau de l'effectif des sources s1 et s2 :
  - soit les effectifs appartiennent à la même tranche (attribut utilisé lors de la sélection de la phase 2 (Section 1.3.1)), dans ce cas le problème de cohérence est anodin et n'a pas de graves conséquences sur la qualité de la sélection ;
  - soit les effectifs n'appartiennent pas à la même tranche. Dans ce cas, le problème de cohérence est plus sérieux, puisque le choix de la mauvaise valeur induira des pertes financières conséquentes dues à la prospection de cibles inappropriées.

SIRET	Nom_entreprise	Effec-tif	Tranche d'effectif	Activité	Adresse	Tel	Date_MAJ
01223211900012	EXAMPLE1	0	00	Activités financières	78000 Versailles	0139233900	2010
00345342200014	EXAMPLE2	2	01	Santé et action sociale	75016 Paris	0101010101	2000
00032345600032	EXAMPLE3	20	02	Activités financières	69005 Lyon	0134337341	1995
00123422900021	EXAMPLE4	4	02	Hôtels et restaurants	75003 Paris	8990994344	2003
00034444600032	EXAMPLE5	1	01	Activités financières	55003 Paris		1990

Tableau 5.1 Exemple de données de s_1

Siret	Nom_dirigeant	Prenom_dirigeant	Fonction_dirigeant	Effectif	Tel
01223211900012	Dupont	Jean	Directeur technique	1	0134329909
00123422900021	Alban	Julien		14	0322712281
00032345600032			Gérant	3	0434337341
00034444600032	Durand	Marion	Gérant	1	0434337341

Tableau 5.2 Exemple de données de s_2

Source	Siret	Nom_codé[29]	Prenom_codé	Email
s_3	01223211900012	XE31SEDE	ED22XDER	1
s_4	00345342200014	DA390023		1

[29] Les sources s_3, s_4, …, s_{27} sont alimentées par des nominativités codées. Un nom codé est un nom crypté par un programme de codage irréversible (dans notre cas il s'agit d'un programme hashcode). L'alimentation de la base en noms/prénoms irréversiblement codés permet au fournisseur des fichiers de qualification de communiquer des informations confidentielles en toute sécurité.

Source	Siret	Nom_codé[29]	Prenom_codé	Email
s_7	01223211900012	DERTB221	EDRTUH2S	1
s_{22}	01223211900012	DA390023	TREF2FRE1	1
s_{23}	00032345600032	YYTHTH2	TRFEG33U	1
s_{24}	01223211900012	4RRTRTU	DA390023	1
s_{25}	00032345600032	TREFDRE3	TREFERTU	1
s_{27}	00032345600032	34FTEENN	TRETYUOP	1

Tableau 5.3 Exemple de données de s3

1.3.3. Questionnaire auprès des fournisseurs : connaissance des fournisseurs

Afin d'avoir une idée plus complète sur la qualité des données d'enrichissement, nous avons interrogé notre fournisseur principal de données. Il s'agit, en effet, d'une agence de conseil en marketing direct dont le métier est la commercialisation de la donnée B-to-B.

La stratégie d'évaluation de la qualité des données adoptée par cette agence est principalement conditionnée et guidée par la rentabilité des campagnes exprimée en termes de retours mailings pour les campagnes courrier et de nombre de clicks pour les emailings. Ainsi, outre le nettoyage de leurs données, comme le traitement RNVP (Restructuration, Normalisation et Validation Postale) de l'adresse postale et la déduplication des données, la mesure de la qualité se fait principalement par des tests sur échantillons, lesquels sont réalisés en permanence afin de valider l'exactitude des informations d'adressage.

Nous remarquons par contre l'absence d'indicateurs de qualité intrinsèques à la donnée. Les dimensions de fraîcheur sont, malgré leur importance, peu présentes et se rapportent principalement à la date d'intégration des données dans le système d'information. Par ailleurs, la provenance des données n'est pas reportée au niveau de la donnée, toutes les informations étant agrégées afin de former une sorte de mégabase[30] de prospection.

En cas d'incohérence relevée par des doublons multisources, le fournisseur se base sur la réputation des sources. Cette information est principalement appréciée par « les échos » des clients et usagers et non mesurée d'une manière rationnelle et objective qui synthétiserait sous forme d'indicateurs techniques le niveau de confiance à accorder à la source.

[30] Une mégabase est une base de données comportant de très nombreux enregistrements (parfois plusieurs dizaines de millions). Elles sont gérées directement par les annonceurs – essentiellement des entreprises ou des groupes intervenant sur des marchés de biens de grande consommation ou par des organismes spécialisés qui revendent leurs résultats pour un produit particulier à un annonceur concerné

Par ailleurs, le fournisseur attribue une grande importance au taux de renouvellement des données de la source (mesuré par la fréquence de mise à jour des fichiers[31]) puisque, compte tenu du contexte de prospection marketing, les nouvelles données permettraient d'élargir le domaine de ciblage, étant dépourvues de contraintes de pression commerciale[32].

En conclusion, exactitude, réputation et fréquence de mises à jour demeurent les principaux critères d'évaluation des données dans les agences marketing. L'entité mesurée est la source et les moyens mis en œuvre pour la quantification de ces critères sont très pragmatiques et se basent pour la plupart sur des indicateurs mesurés à postériori de leur utilisation (à l'instar des dimensions de réputation et d'exactitude).

1.4. Identification des données les plus critiques

Outre l'identifiant SIRET (ou SIREN) et les attributs d'adressage (l'adresse pour le mailing, le téléphone pour le phoning ou encore l'email pour l'emailing) qui sont indispensables à l'envoi des campagnes marketing, d'autres attributs sont critiques pour la construction du plan fichier :

- l'effectif salarié. L'entreprise MaisonPhoning vise une catégorie bien précise d'entreprises : les très petites entreprises (TPE). Cette caractéristique se traduit par un effectif salarié inférieur à 10;
- l'activité de l'entreprise : critère déterminant pour la pertinence de la cible quant au message marketing de la campagne ;
- la nominativité des contacts de l'entreprise ciblée (nom_codé et prenom_codé de s_3, s_4, …, s_{27} ; et nom_dirigeant et prenom_dirigeant de s_2) ;
- la fonction des contacts (fonction_dirigeant de s_3, s_4, …, s_{27}) ;
- avec un degré moindre, le chiffre d'affaire (TCA de s_1), car il permet d'étudier le profil des entreprises prospectées.

[31] La fréquence de mise à jour des fichiers ne veut pas forcément dire la fréquence de mise à jour des données ou des valeurs de ces données.

[32] La pression commerciale est définie par la pression ressentie par le client (ou prospect) résultant des multiples sollicitations marketing (publicité, email, démarchage physique et téléphonique, etc..). On parle plus souvent dans ce cadre de pression publicitaire ou pression marketing.

2. Evaluation de la qualité de la base de prospection

L'évaluation de la qualité de notre base de prospection s'inscrit dans une démarche en deux phases :

- une évaluation de la qualité intrinsèque de la donnée,
- une appréciation de la qualité globale d'un attribut ou d'un ensemble d'attributs, laquelle est exprimée à travers l'agrégation préférentielle calculée par l'intégrale de Choquet discrète.

Dans cette section, nous démontrons l'efficacité de l'intégrale de Choquet discrète, d'abord, comme outil d'agrégation préférentielle des dimensions qualité, ensuite comme outil d'apprentissage et de modélisation des préférences d'agrégation.

2.1. Définition des dimensions intrinsèques et des métriques correspondantes

Nous nous concentrons dans un premier temps sur l'évaluation de la qualité des attributs critiques. Ces attributs sont en effet responsables du bon acheminement de la campagne vers les cibles marketing. Le Tableau 5.4 détaille l'ensemble de ces attributs, les dimensions qualité ayant été choisies pour les évaluer ainsi que les métriques sous-jacentes. Les dimensions retenues sont :

- la provenance de l'information,
- l'exactitude syntaxique[33],
- la cohérence (aussi définie comme l'exactitude syntaxique),
- la fraîcheur,
- l'unicité,
- la complétude.

[33] Exactitude syntaxique : distance entre la valeur v et la valeur v' considérée comme la représentation exacte de la réalité dont v est le représentant

Attributs	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarques/Difficultés
Activité	s1 : – APEN700 – CJ	*Fraîcheur	*La fraîcheur est calculée à partir de la date de dernière mise à jour du champ. Cette date peut être calculée de 2 manières: – à partir de la notification des évènements entreprise livrée par l'INSEE (champ **DAPEN** de l'INSEE qui indique la date de validité de l'activité principale) – à partir du questionnaire annuel de mise à jour de l'INSEE.	*Calculée en nombre de mois à partir de la date système.	**Remarques** *Si l'information sur la date de mise à jour provient des questionnaires annuels de l'INSEE, nous considérons l'information fraîche les 12 mois suivant l'enquête (puisqu'il s'agit d'un questionnaire annuel). *Les résultats du questionnaire annuels sont diffusés sous forme de fichiers de mise à jour de la base. **Difficultés** *L'information DAPEN n'est pas tout le temps renseignée, ainsi en l'absence d'information sur la date de mise à jour du champ, nous utilisations comme date de référence la date de mise à jour de l'enregistrement. *En cas d'absence d'informations sur la date de MAJ de l'enregistrement, nous utilisons la date d'intégration de l'enregistrement dans la base(ou le système d'informations); dans ce cas, nous pénalisons la dimension fraîcheur puisque l'information risque d'être beaucoup plus obsolète qu'elle ne l'est réellement.
		*Provenance de l'information	*Questionnaire aux fournisseurs et jugement attribué à la méthode d'intégration, modification et suppression de l'information.	*Nous distinguons 2 types d'origines : – une origine directe de l'information (le fournisseur des données par rapport à l'entreprise	**Remarque** *Cette dimension, combinée à d'autres dimensions telles que la fraicheur, nous permet de calculer rationnellement la fiabilité des différentes sources.

Attributs	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarques/Difficultés
				MaisonPhoning) telle que les sources s_1, s_2 et s_3. – l'origine initiale de l'information (premier propriétaire de l'information) telle que le BODACC, les JALs, les questionnaires, …etc.	**Difficultés** *Dans le cas où toutes les sources sont officielles, comment évaluer les différentes origines ? Nous avons demandé l'avis des fournisseurs (par exemple pour l'INSSEE, le RCS est le plus fiable).
		Exactitude syntaxique	*Vérification de la contrainte d'intégrité: validation du format du code activité (APEN700 appelée ecnore code Naf) qui est sur 5 caractères (l'ancien code est sur 4 caractères))	*Le nouveau code Naf est constitué de 5 caractères: 4 chiffres + 1 lettre	
		Cohérence	*Confrontation des différentes valeurs en cas de rapprochement possible *Extraction de règles sémantique à support important (Règles d'association - RA)	Exemple de règles : "Si Codpos=75 alors activité≠agriculture". Ainsi, tous les artisans dont le code postal=75 seront soumis au questionnaire	
		Exactitude sémantique	*Questionnaire interne de validation d'activité *Vérification de la présence par rapport à un référentiel		

Attributs	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarques/Difficultés
		Complétude	*Taux de remplissage de la colonne		
Fonction	*S_2 : – dir1_code_fonction, – dir1_lib_fonction	Provenance	*Questionnaire aux fournisseurs		
		Fraîcheur	*Date de dernière MAJ du champ, si celle-ci est renseignée, date insertion (création du champ ou de l'enregistrement dans e système d'informations) sinon.	*Nombre de mois	
		Exactitude syntaxique	*Vérification de la valeur du champ dans un fichier référentiel (décrivant la fonction et le code fonction)		
		Exactitude sémantique	*Questionnaire externe		
		Cohérence	*Règles sémantiques (telles que les règles d'association) *Règles métier *Rapprochement avec la base alliance	*Exemple de RA: un PDG rattaché à une activité=agriculture => incohérent *Exemple de règles métier: si une même source fournit 2 PDG => incohérent	*Dans le cas de l'utilisation de règles d'associations, il faut choisir les champs (attributs) à analyser
		Complétude	Taux de remplissage de la colonne		
Contact (Nom/Prénom)	*S_2 : – dir1_nom	Exactitude (syntaxique et sémantique)	*Questionnaire sur la fonction puis rapprochement avec nom/prenom hashcodés		

Attributs	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarques/Difficultés
	– dir1_prenom *s_3 : – nom_code – prenom_code		*Vérification manuelle sur un échantillon test, et établissement d'un modèle d'exactitude.		
		Fraîcheur	*Date de mise à jour si elle est fournie par la source.	*Nombre de mois	
		Unicité du nom/prénom			**Remarque** Si le couple nom/prénom est présent plus de deux fois dans la base, nous pouvons nous poser des questions quant à sa qualité. Il pourrait, en effet, s'agir de noms/prénoms génériques.
Effectif	*effectif : – s_1 : efencent / efefetcent – s_2 : effectif *tranche d'effectif : – s_1 : tefen / tefet – s_2 :variabe calculée à partir de l'effectif	Exactitude syntaxique	*Règles syntaxiques	Exemple : Les effectifs sont des entiers positifs	
		Fraîcheur	*Date de dernière de mise à jour du champ **DEFEN**, si la provenance de la source est l'INSEE). *Questionnaire interne, si cette date n'est pas renseignée.	*Exprimée en nombre de mois	**Remarque** *Sachant que la date de dernière mise à jour de l'INSEE correspond à "Décembre N-1" (avec N-1: année précédente) et que le résultat est diffusé l'année N+1, **le décalage minimal est donc de 13 mois, généralement nous estimons ce décalage à 2 ans**.

Attributs	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarques/Difficultés
					Difficulté *L'effectif de l'entreprise varie en 2 ans. Dans quelle mesure peut-on affirmer la fiabilité de l'information provenant de l'INSEE.
		Cohérence	*Si source=s_1; alors utilisation de règles de logique (comparaison de l'effectif établissement (efetcent) par rapport à l'effectif entreprise (efencent), l'efencent étant toujours supérieur ou égal à l'efetcent) *sinon, utilisation de règles sémantiques du type RA en analysant un ensemble d'attributs tels que l'activité.	*Exemple de règles logiques : efencent ≥efetcent Exemple de règles sémantique : une entreprise dans l'agriculture où le nombre d'employés est supérieur à 10➔ peu cohérente	**Difficultés** *L'utilisation des RA n'est pas souvent facile à mettre en œuvre impliquant la préparation d'un échantillon d'apprentissage adéquat, le choix des paramètres de contrôle (filtrage des règles) et des attributs significatifs.
		Provenance (en particulier, les méthodes de collecte/mise à jour des effectifs utilisées par les sources)	*Questionnaire		**Remarques** *Les questionnaires montrent que l'information de s_2, provient des bilans déposés par les entreprises. *La méthode de mise à jour des effectifs de s_1 sont celles de l'INSEE qui se font par une analyse des bilans correspondant généralement à l'année N-2 de l'année courante. Ces effectifs sont parfois mis à jour par des traitements ponctuels, cependant, l'information n'est pas bien référencée dans le fichier de livraison.
					Difficultés *La difficulté d'obtention de telles informations dépend des sources. Il est plus facile d'obtenir des informations de fournisseurs secondaires (tels que Kairos ou s_2) que de sources officielles telles que l'INSEE.

Attributs	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarques/Difficultés
		Complétude	*Taux de remplissage de la colonne		
Chiffre d'affaires	*s_1 : – TCA	Exactitude syntaxique	*Vérification de la syntaxe du champ.		
		Fraîcheur	*Date de dernière mise à jour du champ, si l'information existe *Questionnaire, sinon.		**Remarque** *La date de mise à jour de la variable TCA provient de la date du dernier bilan entreprise qui correspond généralement à N-2 de l'année courante.
		Complétude	*Taux de remplissage de la colonne		
SIREN/SIRET	*s_1, s_2 et s_3.	Complétude	*Taux de remplissage de la colonne		**Remarque** *La complétude du SIRET est un facteur de la qualité de la source.
		Exactitude sémantique	*Entier de 9 chiffres pour le SIREN. *Entier de 9 chiffres pour le SIRET.		
		Unicité (variable calculée pour le SIRET)	*Taux de doubles : utilisation d'un algorithme interne de détection des doublons.		
Adresse	*s_1 : – Etablissement : l3_cadr,l4	Exactitude syntaxique	*Algorithme de validation postale (utilisé en interne chez AID).	*L'adresse est considérée valide si le code postal et la ville sont corrects et que l'adresse répond à la norme postale.	

Attributs	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarques/Difficultés
	_voie,l5_disp,l6_post,l7_etrg,rpet,depet,aronet,ctonet,comet,libcom,codpos,code_voie,numvoie,typvoie,libvoie – entreprise: rpen,depcomen	Exactitude sémantique	*Analyse du fichier NPAI de La Poste pour détecter les adresses erronées. *Analyse du fichier des déménagés fournis par l'INSEE.		
		Fraîcheur	*Etude du fichier d'évènements de l'INSEE pour le repérage des déménagements des entreprises. *Utilisation de la date de mise de l'enregistrement, sinon.	*Questionnaire externe si l'information provient initialement de l'INSEE. *Questionnaire interne sinon.	
Téléphone	s_1 et s_2	Fraîcheur	*Date de derrière MAJ, si source ≠ insee ; sinon, questionnaire interne		
		Provenance	*Questionnaires		
		Exactitude syntaxique	*utilisation de règles syntaxiques *utilisation d'un algorithme de normalisation disponible chez AID	*Exemple de règles syntaxiques : Si le numéro de téléphone est sur 10 caractères et que le premier numéro = '0', alors le numéro est considéré correct, autrement le numéro est faux.	
		Exactitude sémantique	*Questionnaire		

Attributs	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarques/Difficultés
		Utilité	*Indice d'appréciation ou de dépréciation subjective, définie par le décideur.	*Exemple : les numéros de portables sont plus importants que les numéros fixes puisque, dans le premier cas, le contact est directement établi avec la cible.	
		Complétude	*Taux de remplissage de la colonne		**Remarque** *La complétude est un indicateur de fiabilité de la source.
Email	s_3	Provenance	*Questionnaire		
		Exactitude	*Taux de *bounces* (correspondant au taux d'emails erronés)		**Remarque** *Ce taux est facilement calculable à partir des fichiers de retours des campagnes emailing fourni par le routeur des campagnes d'emailing. *L'exactitude syntaxique de l'email ne peut être vérifiée. S_3 donne uniquement l'information sur la présence ou absence de l'email.
		Fraîcheur	*Calculée par rapport à la date de réception des données.	*Nombre de mois	**Remarque** *L'information sur la date d'intégration des fichiers n'est pas livrée par le fournisseur. Nous supposons alors que la date de mise à jour correspond à la date de livraison des données (ce qui biaise l'appréciation qualité de la donnée email).
		Complétude	*Taux de remplissage de la colonne		
Fax	s_2	Exactitude sémantique	*Analyse des fax erronés retournés *Questionnaire		
		Exactitude syntaxique	*Format correct (utiliser la normalisation des téléphones)		

Attributs	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarques/Difficultés
		Fraîcheur	*Calculé à partir de la date de la dernière mise à jour des données de s_2.	*Nombre de mois	

Tableau 5.4 Dimensions et métriques intrinsèques des données

Remarques

La difficulté principale ayant partiellement compromis l'accomplissement de cette évaluation est l'absence de quelques informations clef. Certaines sont rattachées à la valeur d'attributs tels que les noms des dirigeants, les numéros de téléphone et les emails, d'autres sont essentielles au calcul de quelques métriques telles que la fraîcheur. En effet, dans notre exemple, la date technique de mise à jour des données (date à laquelle la donnée est intégrée dans le système d'information), n'a simplement pas été renseignée dans les fichiers d'enrichissement. Ainsi, malgré leur conscience de l'importance de la qualité dans l'appréciation des données de la base, certains réflexes de journalisation des données, pourtant primordiaux pour le calcul de certaines métriques qualité, ne sont pas encore intuitifs chez les administrateurs des bases de données ; l'exactitude et la cohérence demeurant les dimensions les plus populaires, donc les plus analysées.

2.2. Estimation de la qualité globale d'une donnée à partir de métriques qualité intrinsèques

Une donnée étant qualifiée par un ensemble de dimensions et métriques qualité (Tableau 5.4), nous nous proposons dans cette section de construire, grâce à l'intégrale de Choquet, un modèle d'agrégation des métriques qualité. Pour ce faire, nous nous basons sur notre exemple de réalisation d'une campagne emailing pour l'entreprise MaisonPhoning. Ainsi, une fois les numéros SIRET correspondants à la population ciblée définis (Section 1.3.1), nous sélectionnons les emails des contacts à prospecter. Cette tâche est anodine quand un et un seul email est attribué à un contact relatif à un numéro SIRET donné. Elle l'est cependant beaucoup moins dès lors que plus d'un email est affecté à un même identifiant. Dans ce cas particulier, la technique la plus communément adoptée par les brokers consiste à se référer aux sources des différentes alternatives et de choisir celle appartenant à la source de meilleure réputation. Cependant, si les deux alternatives appartiennent au même fournisseur, les brokers en choisissent une au hasard. Cette stratégie de choix est donc exclusivement subjective et ne permet pas de justifier rationnellement le choix effectué, d'où l'intérêt d'une approche objective telle que l'intégrale de Choquet, qui utilise les dimensions qualité intrinsèques pour la génération d'un indicateur robuste de décision.

Nous détaillons ces propos avec un exemple. Tout d'abord, nous définissons notre exemple d'apprentissage. Nous décrivons ensuite le paramétrage de l'intégrale de Choquet pour une

agrégation préférentielle des dimensions/métriques qualité. Nous montrons enfin l'efficacité et les performances de l'intégrale de Choquet en les comparant aux préférences initiales du décideur/broker.

Remarque

Nous nous focalisons dans cet exemple sur l'agrégation des dimensions qualifiant un seul attribut, en l'occurrence l'email. Notons que dans la pratique, cette opération s'applique à tous les attributs impliqués dans le processus de sélection multisources.

2.2.1. Exemple

Supposons qu'un contact relatif à une entreprise ciblée se voit affecter trois adresses email différentes provenant de deux sources distinctes. Une illustration est donnée par le Tableau 5.5.

ID	Source	Email
0299	S_3	Oui
0299	S_3	Oui
0299	S_{11}	Oui

Tableau 5.5 Exemple de doublons d'adresses emails

Rappelons que, dans le cas de l'entreprise MaisonPhoning, les emails sont loués. Ainsi, au lieu d'avoir l'information stockée en clair dans la base de données, une valeur booléenne (Oui/Non) indique la présence ou l'absence d'un email pour un contact donné. Il s'agit d'une mesure de sécurité décidée par le fournisseur, qui restitue la valeur de l'email lors de l'exécution de la campagne.

D'après le Tableau 5.4, la qualité de la donnée email est généralement appréciée par :

- l'exactitude calculée par rapport à l'analyse d'un échantillon test aléatoirement sélectionné et définie par :
  - l'analyse du code retour SMTP[34] de l'email envoyé, qui permet de détecter les *hard bounces*[35], des *soft bounces*[36] et des emails livrés ;

[34] SMTP : Simple Mail Transfer Protocol, protocole de communication utilisé pour transférer le courrier électronique vers les serveurs de messagerie électronique.

[35] Les *hard bounces* sont les emails qui ne sont pas arrivés à destination pour motif de refus permanent (adresse email incorrectement photographiée ou inexistante).

[36] Les *soft bounces* sont les emails qui ne sont pas arrivés à destination pour motif de refus temporaire (boîte de réception du destinataire pleine ou serveur de messagerie équipé d'un système de filtrage empêchant la réception des messages d'un expéditeur donné).

- l'analyse de l'email (quand il est disponible), qui permet d'estimer l'exactitude sémantique de l'email. Ainsi, s'il s'agit d'un email générique de type info@domaine.com, l'email est considéré un soft bounce, s'il est de valeur aaaa@aaaa.com, il est considéré un hard bounce;

- la réputation ou plus généralement la fiabilité de la source qui résume sa qualité globale (sa fraîcheur, le taux de chute des emails, c'est à dire le taux d'emails valides qu'elle livre, son taux de complétude, etc.) ;
- la fraîcheur de la donnée ;
- la complétude, qui qualifie la source des données plutôt que la valeur intrinsèque de l'email.

Les dimensions qualité étant principalement contextuelles, nous modifions cette liste de sorte que les dimensions choisies soient plus adaptées aux emails utilisées par MaisonPhoning dans le cadre d'une campagne emailing. Ainsi, le but final étant la sélection d'individus (dans le sens statistique du terme) qui optimiseraient à la fois le prix et la qualité du fichier de ciblage, nous rajoutons le prix comme indicateur ajustant l'appréciation de la qualité globale des emails. Ensuite, la complétude étant un indicateur qualifiant un attribut plutôt qu'un indicateur intrinsèque à une valeur donnée, nous la supprimons de notre liste de dimensions d'évaluation qualité. Par ailleurs, la fraîcheur des adresses emails de l'entreprise MaisonPhoning n'étant pas fiable car indiquant la date de création de la donnée dans la base au lieu de la date de création de la donnée elle-même, nous la supprimons de notre liste de dimensions. Rappelons que l'exactitude de l'email est analysée à partir des codes retour SMTP de campagnes tests à partir d'échantillons extraits aléatoirement des fichiers s_3,…, s_{27}. La proportion testée par source est proportionnelle au volume d'emails proposée par cette source.

En conclusion, les dimensions qualité finalement choisies pour décrire les emails sont :

- l'exactitude ;
- la fiabilité de la source ;
- et le prix.

Aussi, notons que dans notre exemple, la fiabilité de la source est calculée à partir des indicateurs suivants :

- le taux de *hard bounces* ;
- le taux de chute ;
- le taux de doublons ;

- l’exactitude du contact, appréciée par la qualité du nom/prénom du contact. Ces données étant cryptées, nous réduisons cette appréciation à la détection des valeurs nulles et des initiales ;
- le prix.

2.2.2. Construction de la table d’apprentissage des préférences d’agrégation

L’agrégation que nous proposons (Chapitre 3), se base sur l’apprentissage des préférences du décideur (ou expert métier), lesquelles sont exprimées par des relations de pré-ordre entre les alternatives d’un échantillon de données. Dans notre exemple, cet échantillon d’apprentissage est défini de manière à représenter toutes les combinaisons possibles que peuvent prendre les métriques servant à l’évaluation d’un email donné. Ceci nous permettra de bien apprécier toutes nuances des préférences des décideurs.

Pour ce faire, nous définissons des niveaux d’appréciation pour chacune des dimensions comme suit :

- deux niveaux d’appréciation pour le prix de l’email : bon marché (1) et cher (0) ;
- deux niveaux d’appréciation pour la fiabilité de la source : non fiable (0) et fiable (1) ;
- trois niveaux d’appréciation pour qualifier l’exactitude de l’email : bonne qualité (1), qualité inconnue (0,5) et mauvaise qualité (0).

Remarquons que le codage des niveaux d’appréciation est défini de telle sorte que plus une métrique est appréciée, plus sa valeur est élevée (proche de 1).

Le Tableau 5.6 donne un exemple de table d’apprentissage.

Les appréciations du décideur sont exprimées sous formes d’astérisques. Plus le nombre d’astérisques est élevé, plus l’alternative est appréciée. Par ailleurs, nous remarquons que les alternatives dont les adresses emails sont de mauvaise qualité (0) ne sont pas analysées. Le décideur considère en effet que les mauvaises adresses emails sont directement renvoyées par le routeur d’emailing impliquant un échec d’envoi. La valeur « 0 » de l’attribut « qualité de l’email » est alors déterministe quant à l’échec ou la réussite de la campagne emailing.

Exemples	Exactitude de l'email C3	Fiabilité de la source C2	Prix de l'email C1	Préférences du décideur
l	1	1	1	****
h	0,5	1	1	***
d	0	1	1	ils sont directement filtrés; pas dans l'optimisation
j	1	0	1	****
f	0,5	0	1	**
b	0	0	1	
k	1	1	0	****
g	0,5	1	0	**
c	0	1	0	
i	1	0	0	****
a	0	0	0	

Tableau 5.6 Table d'apprentissage des préférences

2.2.3. Quantification des préférences et apprentissage de la fonction d'agrégation

i- Choix de l'utilité

La fonction d'utilité modélise le degré de satisfaction des différents critères. Comme les métriques qualité sont une représentation mathématique des critères qualité, elles peuvent être considérées comme des fonctions d'utilité à condition d'être **croissantes**. Cependant, la fonction prix (C1) est une fonction décroissante (plus le prix est élevé, moins il est apprécié), contrairement aux critères d'exactitude (C3) et de fiabilité (C2) qui, eux, sont bien croissants (plus la valeur de la métrique est élevée, plus elle est appréciée).

D'autres part, l'utilisation de l'intégrale de Choquet en tant que fonction de capacité générant une utilité globale nécessite que les fonctions d'utilité soient commensurables [Kojadinovic 06]. Nous remarquons dans ce contexte que, contrairement aux critères *C2* et *C3*, les valeurs du critère *C1* ne sont pas comprises entre *0* et *1*. Ces dernières sont alors normalisées selon la fonction (F2) définie dans le Chapitre 3. Notons que les critères *C2* et *C3* peuvent aussi être normalisées afin d'avoir une meilleure distribution de leurs valeurs respectives entre 0 et 1. Pour ce faire, nous utilisons la fonction de normalisation (F1).

$$\rho_{ijkl} = \frac{\left|w_{ijkl} - Max\left(w_{ijl}\right) + \Delta'\left(w_{ijl}\right)\right|}{\Delta\left(w_{ijl}\right)} \ (F1)$$

$$\rho_{ijkl} = 1 - \frac{\left|\omega_{ijkl} - Max(w_{ijl}) + \Delta'\left(\omega_{ijl}\right)\right|}{\Delta\left(\omega_{ijl}\right)} \ (F2)$$

ii- Choix de la capacité

Plusieurs fonctions de capacité ont été définies dans la littérature [Grabisch et al. 12]. Nous distinguons :

- les approches fondées sur les moindres carré. Ce type de fonctions cherche une capacité qui minimise l'erreur quadratique moyenne entre les scores théoriques et celui calculé par l'intégrale de Choquet ;
- les approches de programmation linéaire. Ce type de fonctions tend à trouver une capacité qui sépare le mieux les différentes alternatives ;
- la méthode du minimum de variance. Ce type de fonctions tend à calculer une capacité qui maximise l'entropie et minimiserait la variance ;
- la méthode du minimum de distance. Il s'agit de déterminer une capacité qui minimise la distance entre les scores théoriques définis par l'utilisateur et les scores calculés par l'intégrale de Choquet.

Notre objectif étant discriminatoire (puisque nous cherchons à bien distinguer les alternatives de bonne qualité des alternatives de qualité moindre), nous optons pour une capacité basée sur la programmation linéaire.

iii- Apprentissage des préférences

Nous nous basons sur la fonction de capacité déterminée à la Section ii et la table d'apprentissage (Tableau 5.6) pour calculer l'intégrale de Choquet correspondant aux préférences du décideur. Pour ce faire, nous avons développé un programme R (interprétable par le logiciel R-Gui-Project) (Figure 5.2). Nous avons utilisé en particulier le package *Kappalab*, où les préférences du décideur sont modélisées par la fonction *Choquet.preorder* et dont la valeur est déduite à partir des relations de préordre définies par l'expert humain lors de l'appréciation des différentes alternatives (ou objets) (Tableau 5.6). Nous obtenons alors les scores de référence du Tableau 5.7.

```
##chargement des librairies nécessaires au package Kappalab
library(lpSolve)
library(quadprog)
library(kernlab)
library(kappalab)

##Apprentissage de la fonction d'agrégation de la qualité de l'email
##qualite_email#fiabilite_source#prix_unitaire
a <- c(0,0,0)
b <- c(0,0,1)
c <- c(0,1,0)
d <- c(0,1,1)
e <- c(0.5,0,0)
f <- c(0.5,0,1)
g <- c(0.5,1,0)
h <- c(0.5,1,1)
i <- c(1,0,0)
j <- c(1,0,1)
k <- c(1,1,0)
l <- c(1,1,1)

delta.C <- 0.4
delta.B <- 0.001
delta.D <- 0.0001
Acp <- rbind(c(l,k,delta.D),
             c(k,j,0),
             c(j,i,0),
             c(l,h,delta.B),
             c(j,h,delta.B),
             c(k,h,delta.B),
             c(i,h,delta.B),
             c(h,g,delta.B),
             c(h,f,delta.B),
             c(g,f,0),
             c(g,e,delta.B),
             c(f,e,delta.B),
             c(e,d,delta.C),
             c(e,c,delta.C),
             c(e,b,delta.C),
             c(e,a,delta.C))

s <- mini.var.capa.ident(3,3,A.Choquet.preorder=Acp)
mu <- zeta(s$solution)
```

Figure 5.2 Utilisation du package Kappalab pour l'apprentissage de la fonction d'agrégation

Exemples	**Exactitude de l'email** C3	**Fiabilité de la source** C2	**Prix de l'email** **C1**	**Préférences du décideur**	**Scores de qualité globale**
l	1	1	1	****	1
h	0,5	1	1	***	0,5
d	0	1	1	ils sont directement filtrés; pas dans l'optimisation	$2*10^{-6}$
j	1	0	1	****	0,82

Exemples	Exactitude de l'email C3	Fiabilité de la source C2	Prix de l'email **C1**	Préférences du décideur	Scores de qualité globale
f	0,5	0	1	**	0,401
b	0	0	1		10^{-6}
k	1	1	0	****	0,82
g	0,5	1	0	**	0,401
c	0	1	0		10^{-6}
i	1	0	0	****	0,8
e	0,5	0	0	*	0,4
a	0	0	0		0

Tableau 5.7 Scores qualité globale calculés par Kappalab pour la table d'apprentissage

2.3.4. Remarques

Remarque 1 : Ajustement de la fonction d'agrégation

Dans notre exemple, les scores calculés reproduisent bien le préordre établi par le décideur. Cela dit, dans certains cas, les résultats de l'intégrale de Choquet peuvent ne pas quantifier fidèlement le préordre du décideur. Dans ce cas, des ajustements doivent être effectués au niveau des écarts entre les alternatives exprimées par les variables *delta.x*. L'apprentissage est alors répété jusqu'à ce que les résultats satisfassent les contraintes de préférence.

Remarque 2 : Gestion de la qualité des données codées

L'évaluation de la qualité des bases de données est un processus itératif et continu et notre exemple d'évaluation de la base de prospection nous le montre bien, notamment en ce qui concerne l'évaluation de l'adresse email. En effet, comme l'email n'est pas fourni en clair (seul un indicateur de sa présence est délivré), aucune information sur sa qualité n'est connue au préalable. C'est uniquement après l'exécution de la campagne emailing, et en analysant les fichiers logs des routeurs, que l'on pourra découvrir la valeur de l'email et se rendre compte de sa qualité. Ainsi, pour évaluer la qualité de l'email, nous utilisons le processus itératif suivant.

- La valeur de la métrique « qualité de l'email » est initialement neutre (fixée à 0.5).
- Quand un email est utilisé dans une campagne emailing, la valeur de la métrique en question est mise à jour comme suit : 1 si l'email est livré et 0 si c'est un *hard bounce*.

Au fur et à mesure des campagnes, l'exactitude de certains emails est mise à jour, permettant le calcul d'un indicateur de qualité moyenne de l'email par source.

Remarque 3 : Calcul de la qualité d'un enregistrement de données

La qualité d'un enregistrement de données est généralement évaluée par l'agrégation de la qualité intrinsèque des différents attributs qui le composent. Par exemple, si l'enregistrement est composé des attributs SIRET, Nom_entreprise, Nom_codé, Prenom_codé, Email, sa qualité est calculée à partir de la qualité des attributs Nom_entreprise, Nom_codé, Prenom_codé, Email[37]. A cette qualité globale agrégée s'ajoute un ensemble de dimensions relatives à l'enregistrement, telles que la complétude, le prix et autres dimensions décrites dans le Tableau 5.8.

2.3. Difficultés rencontrées

Deux types de difficultés ont été rencontrés lors de l'évaluation de la qualité de la base de prospection :

1 des difficultés liées à la mise en place des métriques qualité ;
2 des difficultés dues à l'utilisation du package Kappalab lors du calcul de l'intégrale de Choquet.

2.3.1. Difficultés dans le calcul de certaines métriques qualité

Ces problèmes s'expliquent par l'ambigüité dans la définition de quelques métriques qualité. Deux dimensions peuvent décrire le même problème qualité, par exemple, l'exactitude et la complétude nette, où en effet la complétude nette décrit l'ensemble des données exactes d'un attribut donné (Chapitre 3).

Par ailleurs, certaines métriques sont difficilement mesurables dans la pratique. Prenons l'exemple de la métrique quantifiant la dimension *timeliness*, généralement définie par le délai entre la *date d'utilisation* et la *date de la dernière mise à jour* de la donnée. Le problème provient de la manière dont l'attribut *date de dernière mise à jour* est défini. Trois cas de figures sont rencontrés.

[37] Dans le contexte des campagnes marketing B-to-B, le Siret de l'entreprise doit être correct étant la clef du ciblage, sa qualité n'est donc pas discutable.

1 Aucune date de mise à jour n'est renseignée pour l'attribut mesuré. Le fichier est en effet livré en mode « annule et remplace » et la date de dernière mise à jour est déduite par la comparaison de la donnée du fichier livré à la date *t* avec celui livré à la date *t-1*. Ce mode de déduction est forcément biaisé. Ainsi, une donnée dont la valeur n'a pas été modifiée entre les versions *t* et *t-1* se verra assigner *t-1* comme date de dernière mise à jour, même si elle a fait l'objet d'une validation (par questionnaire, par exemple) lors de la mise à jour *t*. Par ailleurs, une mise à jour technique (relative à un changement de codification ou à une mise en forme des données) ne doit pas être considérée comme une vraie mise à jour, auquel cas elle fausserait le calcul de l'indicateur de fraîcheur.
2 Une date de mise à jour est fournie pour l'ensemble de l'enregistrement et non pour chacun des attributs le composant. Cette date est souvent biaisée. En effet, dans la plupart des systèmes marketing, il suffit qu'un attribut de l'enregistrement fasse l'objet d'une enquête d'enrichissement pour que la date de dernière mise à jour de tous les attributs formant l'enregistrement soit modifiée.
3 Une date de mise à jour est fournie pour chacun des attributs de la base. Il s'agit de la meilleure des configurations, très rare en pratique.

Toujours concernant l'ambiguïté de la mise en place des indicateurs de *timeliness* et de fraîcheur de l'information de manière générale, nous devons faire la différence entre la date de création de la donnée et la date à laquelle l'information est livrée ou intégrée dans le système d'informations. Prenons l'exemple de l'information « effectif entreprise » de l'INSEE. Cette information est calculée à l'année N-2 pour être diffusée l'année N, créant de cette manière un décalage non anodin de deux ans entre la date de création de l'information et la date de sa diffusion aux partenaires de l'INSEE.

2.3.2. Difficultés rencontrées quant à l'utilisation de Kappalab

Le package Kappalab a été développé par Michel Grabisch, Ivan Kojadinovic et Patrick Meyer pour le calcul de l'intégrale de Choquet discrète dans l'environnement de développement R-Project [Grabisch et al. 12].

L'utilisation de ce package, bien que simplifiant considérablement le calcul de l'intégrale de Choquet, son utilisation présente quelques difficultés.

1 La première difficulté est relative à la complexité de paramétrage du package. En effet, Kappalab offre plus d'une cinquantaine de fonctions et de méthodes pour la modélisation des préférences des utilisateurs telles que le pré-ordre des alternatives, la synergie des

capacités, l'importance relative des critères ainsi que les interactions relatives entre ces critères. Ces possibilités nécessitent une bonne compréhension des subtilités existant entre chacune des fonctions.

2 La capacité reste, malgré sa complexité, une fonction d'attribution de poids. Ainsi, elle ne permet pas de modéliser toutes les préférences des utilisateurs, notamment dans le cas où ces contraintes comportent des contradictions.

3 La fonction d'agrégation doit être définie pour chacun des attributs de la base de données multisources et pour chaque type de campagne (puisque les préférences des décideurs diffèrent selon le contexte et la criticité de l'attribut dans ce contexte). Ainsi, le paramétrage nécessite pour chaque critère le recueil des préférences des utilisateurs/décideurs, ce qui représente une sollicitation importante. Ceci peut difficilement être mis en place dans la pratique.

Attributs concernés	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarque	Difficulté/Jugement
Enregistrement	Enregistrement lié à un SIRET (quelle que soit la source)	Fraîcheur	*Entreprise vivante : • entreprise achète (repérage fichier clients) • active (repérage fichier des évènements)			Facile
		Récence	*Date de MAJ ou création d'un enregistrement donné			Utilisée pour repérer les entreprises récemment créées. En effet, ces entreprises peuvent faire l'objet d'une campagne marketing spécifiques intéressant en particulier les entreprises de téléphonie de type MaisonPhoning.
		Complétude	Complétude horizontale			

Attributs concernés	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarque	Difficulté/Jugement
	Enregistrement lié à 1 seule source	Utilité	*Utilisation d'une mesure traduisant la perte d'utilité d'un record étant donné sa qualité/fiabilité *OU: les enregistrements naturellement repoussés car clients ou stop_contacts			*La difficulté est conditionnée par la facilité/difficulté de modélisation de la formule
		Prix				Facile
		Coût MAJ	Pondération des poids des différents attributs qualité relatifs aux variables (champs) du record	Dans le cas de demande de traçabilité d'une info telle que la date de MAJ d'un champ, par exemple.	Information obtenue niveau fournisseurs	Facile
		Fiabilité				Difficile, il faut trouver la bon modèle de pondération
		Durée de validité	*Poids initial donné attribué par l'expert *Utilisation des connaissances sur les fournisseurs (d'où est-ce qu'ils obtiennent leurs	Pour un record dont la fraîcheur < 5 ans (par exemple) et qui n'est pas utilisé, est ce qu'on peut		*La première difficulté concerne la détermination du seuil à partir duquel

Attributs concernés	Variables liées	Dimension Qualité	Métrique	Formules/Exemples	Remarque	Difficulté/Jugement
			données) pour modifier le score *Utilisation des résultats des rapprochements avec d'autres sources (contradiction/similarité) pour modifier le score *Utilisation des retours des campagnes pour modifier les scores	considérer qu'il n'est plus valide?		un record n'est plus considéré comme valide
		Réputation				Difficile, il faut trouver le bon modèle de pondération

Tableau 5.8 Dimensions et métriques utilisées pour l'évaluation de la qualité des attributs de la base de données multisources

3. Optimisation de la sélection de ciblage avec BrokerACO

Les fonctions de calcul de scores qualité étant définies pour les différentes entités composant le fichier de ciblage (attributs, enregistrements et l'ensemble de la sélection), nous décrivons dans cette section le déroulement de la phase d'optimisation du processus de sélection des données multisources pour les besoins de ciblage marketing.

3.1. Objectif d'optimisation

Il s'agit d'optimiser une fonction de gain $\psi = f(\sigma, \pi)$ dénotant le compromis qualité/prix observé au niveau du fichier de ciblage. σ exprime la qualité moyenne de la sélection et π son prix. Dans le cadre de ce mémoire, nous étudions en premier lieu le cas d'une campagne monocanal, en l'occurrence une campagne emailing. En effet, dans la pratique, les annonceurs peuvent programmer plusieurs campagnes pour le même fichier de ciblage : une campagne emailing et une campagne téléphonique, par exemple. Nous parlons dans ce cas de campagnes multicanal. De plus, pour simplifier l'explication de la méthodologie de sélection, nous optons pour un ciblage mono-attribut où il s'agit de sélectionner les adresses emails qui optimiseraient le profit qualité/prix de la campagne emailing. L'illustration de notre approche se réfère aux concepts définis dans le Chapitre 3, Section 3.3.1.

i- Quantification de la fonction qualité $\boldsymbol{\sigma}$

Le fichier de ciblage est composé d'un ensemble d'enregistrements de prospection, où chacun décrit une cible marketing. Ces individus sont indépendants les uns des autres. Ainsi, la qualité globale σ est la moyenne arithmétique de la qualité agrégée σ_{ijk} calculée au niveau de chaque enregistrement R_{ik} du fichier de ciblage. La qualité de l'enregistrement est définie par l'agrégation par l'intégrale de Choquet des scores qualité calculés pour chacun des attributs d_{ijk} composant l'enregistrement en question (Section 2).

ii- Quantification de la fonction prix $\boldsymbol{\pi}$

Tout comme l'agrégation, la fonction d'optimisation dépend fortement de la contextualité du problème, à savoir du type de la campagne et des attributs qui y sont sélectionnés. De plus, la

fonction de calcul des prix diffère d'une source à une autre. Certaines sources imposent, par exemple, l'utilisation de formules de prix dégressives où le prix décroit en fonction de la quantité commandée. Dans notre cas, les sources s_3, s_4, …, s_{27} adoptent une stratégie différente[38] :

- le prix est fixe ($co\hat{u}tFixe_i + prixMin$) quel que soit le nombre d'adresses commandées si celui-ci est inférieur à un certain seuil, et ce pour chacune des sources ;
- au-delà de ce seuil, la notion de prix unitaire est appliquée et le prix total est proportionnel à la quantité de données commandées.

La fonction de prix par source π_i est alors définie comme suit :

$$\begin{cases} \pi_i = co\hat{u}tFixe_i + prixMin\ ; si\ nbre\ d'emails\ commandés \leq\ nb_i \\ \pi_i =\ co\hat{u}tFixe_i +\ prixUnitaire_i * nb\ \ \forall nb\ \geq\ nb_i \end{cases}$$

où

- nb_i est le nombre minimum d'emails à commander pour une source s_i donnée;
- $co\hat{u}tFixe_i$ est le coût fixe imposé par la source s_i, $i \in \{3, 4, \dots, 27\}$;
- $prixMin_i$ est le prix minimum à payer pour une source s_i ($i \in \{3, 4, \dots, 27\}$) si le nombre d'enregistrements commandé (loué) est inférieur à nb_i ;
- nb_i est le nombre minimum de données (en l'occurrence, données emails) commandées (louées) auprès d'une source s_i ($i \in \{3, 4, \dots, 27\}$).

Le prix total π est la somme des prix élémentaires et est défini par la formule suivante :

$$\pi = \pi_1 + \pi_2 + \sum_{i=3}^{27} \pi_i\ .$$

$\pi_1$et $\pi_2$étant des prix fixes (indépendants de la quantité des données utilisées par la campagne), nous nous utilisons la fonction variable, à savoir les prix des données de location ($\sum_{i=3}^{27} \pi_i$), dans le calcul de la fonction de compromis ψ.

[38] Notons que ces contraintes ne se posent pas dans le cas de sélection de données achetées comme c'est le cas des informations provenant des sources s_1 et s_2.

3.2. Définition des contraintes d'optimisation

3.2.1. Contraintes réelles

Nous distinguons les contraintes contextuelles suivantes.

- Nombre minimum de données commandées : les fournisseurs dont il est question dans cette étude imposent des tarifs partiellement dégressifs, dans le sens où le prix d'achat est fixe en deçà d'un certain nombre d'emails commandés par source. Concrètement, chaque fournisseur de données définit trois paramètres : le coût fixe, le nombre minimal de données à commander et le coût unitaire à utiliser si le nombre de données dépasse le seuil minimal. Par exemple, supposons que pour s_3 le coût fixe soit de 50 €, le nombre minimal de 500 adresses et le coût unitaire de 0,1 € l'adresse. Le coût de location de 200 adresses est le même que celui de 500 adresses, à savoir 50€. Par contre, le coût de location de 600 adresses est de 50+(600-500)*0,1=60 €. Ceci induit une sorte de dépendance de sélection quant à la source des données sélectionnées *(C1)*.
- Contraintes budgétaires : une campagne de prospection est généralement limitée par un budget de prospection, lequel conditionne le choix des sources des données cibles *(C2)*.
- Contraintes de qualité minimale : la rentabilité des campagnes marketing dépend fortement de la pertinence et de la justesse du ciblage (caractéristique exprimée dans le jargon marketing par l'*actionnabilité*/l'utilisabilité des données de ciblage). Les données ciblées doivent ainsi s'adapter au contexte de prospection, mais surtout, être de bonne qualité *(C3)*.

Se rajoutent à ces contraintes contextuelles les contraintes techniques suivantes, imposées par le contexte multisources.

- Une première contrainte est imposée par la nature de certaines données, telles que celles fournies par s_2 ou encore s_3, s_4, …, s_{27}, où la sélection d'un attribut impose la sélection de tous les autres attributs appartenant au même bloc de dépendance, entrainant ainsi une dépendance des attributs a_{ij}, et $a_{ij'}$, fournis par la même source s_i. Un bloc de dépendance est, en effet, un ensemble d'attributs contextuellement liés, donc indissociables, comme le nom et le prénom, le numéro de la rue, le nom de la rue et le code postal… Dans notre exemple, la sélection de Nom_dirigeant de la source s_2 impliquerait la sélection des attributs Prénom_dirigeant et Fonction_dirigeant. De

même, pour les données des sources s_3, s_4, …, s_{27}, la sélection de l'email impliquerait la sélection des attributs Nom_Codé et Prénom_Codé $(C4)$.

- Une deuxième contrainte est relative à la dépendance de sélection, où la sélection de l'attribut a_j conditionne la sélection des autres attributs de la sélection $a_{j'}$ tel que $j \neq j'$ pour l'enregistrement k. Par exemple, dans le contexte des campagnes emailing, en sélectionnant des adresses emails, nous sélectionnons (implicitement) leurs identifiants SIRET (idik). Ces idik vont ainsi restreindre l'univers de sélection des attributs Effectif et Tel, d'où la dépendance $(C5)$.

3.2.2. Relaxation des contraintes

Les contraintes décrites dans le paragraphe 3.2.1 ne sont pas facilement modélisables dans le contexte de résolution de notre problème d'optimisation, notamment à cause des relations d'interdépendances décrites dans $C4$ et $C5$. Des relaxations aux contraintes du problème sont alors opérées. Ainsi, nous nous désencombrons de l'aspect multi-attributs de notre sélection et nous nous intéresserons, dans un premier abord, à l'optimisation de la sélection mono-attribut des données prospects éliminant, ainsi, la contrainte (C4). Du coup, en relaxant la contrainte (C4), qui opte pour la sélection mono-attribut, la contrainte (C5) se trouve libérée de toute dépendance d'identifiants.

Notre problématique d'optimisation devient alors :

$$P_{BrokerACO} = Maximiser \sum_i \psi(\pi_{ij}, \sigma_{ij})$$

$$s.c. \begin{cases} \eta_{ij} \geq minimum_i \ (C1) \\ \sum_i \pi_{ij} \leq Budget_c \ (C2) \\ Moyenne\left(\sum_i \sum_{k=1}^{\eta_{ij}} \sigma_{ijk}\right) \geq \sigma_c \ (C3) \end{cases}.$$

Le prix des données de l'attribut a_{ij} issues de la source *i* se calcule donc de la manière suivante :

$$\pi_{ij}(\{d_{jk}\}) = Co\hat{u}tFixe_{ij} + \left(\frac{minimum_{ij}}{1000}\right) * tarifMille_{ij} + \left(\frac{supplément_{ij}}{1000}\right) * tarifMille_{ij}$$

où

- $Co\hat{u}tFixe_{ij}$ est le coût fixe d'achat (ou de location) de l'attribut a_{ij} imposé par le fournisseur de la source i ;
- $minimum_{ij}$ est le nombre de données minimum imposé par le fournisseur de la source i pour l'attribut a_{ij} ;
- $supplément_{ij} = \begin{cases} \eta_{ij} - minimum_{ij}; si\ \eta_{ij} \geq minimum_{ij} \\ 0; sinon \end{cases}$;
- $tarifMille_{ij}$ est le tarif d'achat (ou de location) au mille de données de la source i.

3.3. Résolution du problème d'optimisation par l'algorithme BrokerACO : expérimentation et validation

3.3.1. Objectifs de l'expérience

Dans cette section, nous implémentons et validons l'algorithme BrokerACO avec la fonction objectif $\psi(\pi_{ij}, \sigma_{ij})$ décrite dans la Section 3.2.2, en simulant des campagnes test. L'objectif de cette série de test est, en réalité, double.

1 En premier lieu, nous cherchons à montrer l'efficacité de BrokerACO et à mettre en évidence l'importance d'une approche contextuelle vis-à-vis d'une approche générique.

2 En second lieu, nous cherchons à trouver, à partir de cette série de campagnes tests, les valeurs adéquates des paramètres utilisés dans l'algorithme de fourmis, à savoir le nombre de fourmis, le nombre d'itérations pour chaque fourmi et les valeurs des paramètres α et β.

Nous simulons deux séries de tests.

1 Dans la première série, nous montrons l'efficacité de BrokerACO en le comparant à d'autres algorithmes d'optimisation dont la fonction objectif est plus simple, plus générique et moins contextuelle, en l'occurrence, la fonction minimisant le prix et la fonction maximisant la qualité.

2 Dans la deuxième série, nous optimisons les paramètres utilisés dans BrokerACO afin que le fichier de ciblage généré représente le mieux le compromis qualité/prix exprimé par le décideur.

Pour chacun de ces tests, les paramètres de validation sont les suivants :

- le nombre d'emails corrects ;

- le nombre d'emails erronés (*hard bounces*) ;
- la qualité moyenne du fichier de ciblage généré ;
- le prix du fichier de ciblage généré.

3.3.2. Validation de BrokerACO

Nous simulons trois campagnes emailing. Dans la première campagne, les emails sont sélectionnés de façon à minimiser le prix total de la sélection (P_{Prix}). Dans la deuxième campagne, le fichier de ciblage est composé de façon à maximiser la qualité globale des emails ciblés ($P_{Qualité}$). Enfin, dans la troisième campagne, le fichier de ciblage vise à représenter un compromis qualité/prix exprimé par le décideur (P_{Broker}). Cette série de tests a pour objectif de montrer la capacité de BrokerACO à modéliser tous les types de préférences possibles, s'adaptant de cette manière à tous les types de compromis qualité/prix. Cet avantage est garanti par l'intégrale de Choquet.

Dans la suite de cette section, nous décrivons les préférences du décideur que nous estimerons avec l'intégrale de Choquet. Nous comparons ensuite les performances de chacun des trois fichiers de ciblage et leur adéquation aux contraintes exprimées par les décideurs.

i- Modélisation de la fonction de compromis

A l'instar des préférences d'agrégation, la fonction de compromis modélise les préférences du décideur exprimées sur un échantillon d'apprentissage représentatif de l'exhaustivité des valeurs prises par le couple (qualité, prix). En effet, afin de détecter toutes les nuances d'appréciation, nous définissons cinq niveaux de discrétisation de la qualité : excellent (1), bon (0,7), acceptable (0,5), faible (0,3) et médiocre (0,1). De la même manière, nous identifions trois niveaux de discrétisation du prix : cher (0,1), moyen (0,5) et bon marché (1). Les préférences du décideur sont exprimées par le Tableau 5.9.

Nous remarquons que les préférences du décideur exprimées dans cet exemple sont plutôt modérées : le décideur préfère avoir un fichier de qualité et de prix moyen plutôt qu'un fichier de ciblage moins coûteux et de mauvaise qualité ou encore un fichier de ciblage de bonne qualité et très cher, l'idéal étant d'avoir un fichier de ciblage de très bonne qualité et bon marché.

Nous apprenons ensuite ces préférences en utilisant l'intégrale de Choquet du package Kappalab. Les scores estimés sont exprimés dans le Tableau 5.9.

Observations	Qualité	Prix	Préférences du décideur	Scores calculés par Kappalab
a	1	1	****	1
b	0,7	1	****	0.7875
c	1	0,5	****	0.7625
d	0,7	0,5	****	0.605
e	0,1	1	**	0.3625
f	0,5	1	***	0.6458333
g	0,1	0,5	**	0.2166667
h	0,5	0,5	***	0.5
i	0,3	1	**	0.5041667
j	0,3	0,5	**	0.3583333
k	1	0,1	***	0.5725
l	0,3	0,1	*	0.205
m	0,1	0,1	*	0.1
n	0,5	0,1	*	0.31
o	0,7	0,1	**	0.415

Tableau 5.9 Tableau d'apprentissage de la fonction de compromis qualité/prix

ii- Comparaison de l'intégrale de Choquet à des fonctions d'optimisation non-contextuelles

Le but de ces campagnes est de mettre en évidence l'importance de la contextualité de l'intégrale de Choquet, dans le contexte d'un problème d'optimisation, en la comparant à d'autres algorithmes moins contextuels tels que les algorithmes de sélection privilégiant les prix faibles ou encore la qualité la plus élevée. Pour ce faire, trois campagnes tests sont effectuées. Dans la première, nous optons pour l'optimisation avec BrokerACO dans la deuxième, nous utilisons l'optimisation qui privilégie les prix les moins chers et dans la troisième, l'optimisation guidée par la qualité la plus élevée.

Le fichier d'adressage ciblé par les campagnes tests se compose de 100 individus (emails de contacts) sélectionné à partir d'une base de prospection composé de 2465 individus et 25 sources. Les résultats sont illustrés dans le Tableau 5.10.

En effet, les performances de BrokerACO permettent de bien représenter le compromis qualité/prix souhaité par le décideur. Les critères de performance *Nombre d'emails corrects* et *Nombre de hard bounces* sont calculés à partir des retours de campagnes précédentes. Ils ne sont donc pas renseignés pour les emails n'ayant pas été sujets à des campagnes de prospection. Ainsi, pour le fichier de ciblage utilisé dans la campagne orientée prix, par exemple, seul un des emails sélectionnés a déjà fait partie de campagnes précédentes. Cet email était correct.

	P_{Broker}	$P_{Qualié}$	P_{Prix}
Nombre d'emails corrects*	99	99	1
Nombre de *hard bounces**	0	1	0
Qualité moyenne normalisée	0,91	0,92	0,01
Prix du fichier (€)	85,47	87,17	85,31

Tableau 5.10 Comparaison de BrokerACO avec des algorithmes d'optimisation non contextuels

iii- Optimisation des paramètres de BrokerACO

Dans cette série de tests, nous validons les valeurs des paramètres utilisés dans BrokerACO. Nous nous intéressons en particulier aux paramètres suivants :

- α (solution guidée par le taux de phéromones, donc dépendante de l'historique de la solution) et β (solution guidée par l'information heuristique, donc indépendante de l'historique de la solution) ;
- le nombre d'itérations ;
- le nombre de fourmis.

Nous faisons varier à chaque fois l'ensemble des paramètres et nous retenons les critères les plus performants.

Validation des valeurs de α et β

Dans BrokerACO, les éléments composant la solution sont choisis un à un jusqu'à atteindre la taille η_c ciblée. Le rôle des paramètres α et β y est très important. D'une part, la solution générée doit garder en mémoire les sources des éléments déjà sélectionnés (α) afin de garantir le respect de la contrainte $C1$. D'autre part, le rôle de β est primordial pour garantir la diversité de la solution et éviter d'avoir un fichier de ciblage où tous les éléments appartiennent à la même source. Un compromis entre ces deux paramètres doit alors être trouvé. Pour ce faire, nous effectuons une série de tests où nous faisons varier à chaque fois les valeurs de α et β. Les performances sont évaluées par la moyenne du prix et la moyenne de la qualité (cf. Tableau 5.11).

α et β	α=0,7 β=0,3	α=1 β=1	α=0,5 β=0,5	α=0,3 β=0,7
Moyenne qualité	1,65	1,19	1,06	1,24
Moyenne prix normalisé	0,01906	0,01907	0,01906	0,01905

Tableau 5.11 Validation des paramètres α et β

Nous remarquons que la solution donnant le meilleur compromis qualité/prix est celle où $\alpha = 0{,}7$ et $\beta = 0{,}3$.

Validation du nombre de fourmis

Le nombre de fourmis est un paramètre important, puisqu'il participe à la rétroaction positive principale du système. Selon [Dréo et al. 03], les algorithmes de fourmis sont assez peu sensibles à un réglage fin du nombre de fourmis. En même temps, ni l'utilisation d'une seule fourmi ni le parallélisme naturel de l'algorithme ne semblent être la bonne approche. En effet, la première fait perdre l'effet d'amplification des solutions et la seconde peut s'avérer néfaste car noyant la recherche dans un ensemble innombrable de possibilités. Dans cette expérience, nous réalisons des campagnes tests en faisant varier le nombre de fourmis utilisé dans la solution afin d'atteindre ce compromis expérimental. Les valeurs testées sont 1, 3 ,5 et 10.

Nombre de fourmis	**1**	**3**	**5**	**10**
Moyenne qualité	1,19	1,65	1,22	1,47
Moyenne prix normalisé	19,06	19,06	19,07	18,7

Tableau 5.12 Validation du nombre de fourmis

La meilleure solution est celle obtenue avec un nombre de fourmis égal à 3 (Tableau 5.12), ce qui confirme la théorie énoncée dans le paragraphe précédent.

Validation du nombre d'itérations

Une itération de l'algorithme BrokerACO génère une solution en partant d'une configuration initiale aléatoire. En générer plusieurs augmenterait nos chances d'obtenir un fichier de ciblage

mieux optimisé, car explorant des solutions à partir de plusieurs configurations initiales aléatoires. Le Tableau 5.13 décrit les performances de BrokerACO avec un nombre d'itérations variant entre 1, 3, 5 et 10. Le maximum d'itérations testé est de 10 afin de permettre la réalisabilité de l'algorithme dans le contexte d'une vraie campagne marketing.

Nombre d'itérations	1	3	5	10
Moyenne qualité	1,65	1,22	1,22	1
Moyenne prix normalisé	19,06	28,67	28,05	18,7

Tableau 5.13 Validation du nombre d'itérations

L'expérience montre l'inutilité de ce paramètre dans l'efficacité de la solution. Ce résultat est plausible, puisque les itérations qui génèrent les solutions intermédiaires sont complètement indépendantes les unes des autres. La différence réside dans le premier enregistrement sélectionné dans la solution, le seul élément complètement aléatoire de l'algorithme.

4. Conclusion

Dans ce chapitre, nous avons montré l'intérêt de l'intégrale de Choquet et de l'algorithme BrokerACO dans la sélection contextuelle des données de ciblage. Nous avons montré sa validité en le comparant à des algorithmes d'optimisation non contextuels grâce à une première série de campagnes tests et nous avons validé les paramètres qui y sont utilisés en effectuant une autre série de campagnes tests. Les paramètres à retenir sont :

- α=0,7et β=0,3 ;
- Nombre de fourmis=3 ;
- Nombre d'itérations=1 ;

Ainsi, grâce à BrokerACO, nous sommes parvenus à définir une approche complètement automatique et automatisée de gestion de prise de décision dans la sélection de données multisources et ce quelle que soit la préférence du décideur métier. Cette approche est utile, notamment chez les brokers d'adresses, pour la sélection rigoureuse des fournisseurs de données en fonction de la qualité globale de leurs fichiers et des prix de location qu'ils

proposent. De plus, l'intégrale de Choquet étant une fonction d'agrégation multicritères, nous pouvons modifier ces critères ou en ajouter d'autres sans changer l'algorithme. Il suffit alors d'apprendre les nouvelles préférences.

Cependant, l'intégrale de Choquet étant une addition de capacités, elle ne permet pas de résoudre toute sorte d'agrégation préférentielle, notamment les préférences contradictoires. Aussi, l'apprentissage de ces préférences requiert la définition d'un tableau d'apprentissage exhaustif comportant tous les niveaux d'appréciation des différents critères, afin de couvrir toutes les possibilités d'agrégation quelle que soit la préférence.

Chapitre 6 : Conclusions et perspectives

« Dès l'arrivée, le départ se profile. »
Ylipe

1. Bilan

Nous avons abordé dans cette thèse des problématiques liées à la qualité des données dans le contexte de campagnes marketing, où une donnée de mauvaise qualité peut fausser l'objectif de la campagne et engendrer des pertes financières importantes. Les trois contributions principales que nous proposons sont les suivantes.

1 La première contribution est une méthodologie d'évaluation contextuelle et intrinsèque de la qualité des données multisources [Ben Hassine et al. 09], qui a permis d'établir une liste de dimensions qualité appropriées à la problématique d'intégration (sélection) de données multisources.

2 La deuxième contribution est une approche d'agrégation et de quantification préférentielle de la qualité des données [Ben Hassine-Guetari et al. 10], qui a permis de quantifier le degré d'appréciation de la donnée.

3 La troisième contribution est une technique d'optimisation préférentielle de sélection multicritères de données multisources qui a donné lieu à BrokerACO, un outil de sélection de données de prospection préférentielles.

BrokerACO, un outil d'aide à la prospection marketing qui tend à imiter les préférences du décideur métier pour sélectionner des données de brokering multisources. Cette approche originale, flexible et complètement automatique permet de modéliser et d'optimiser un compromis multicritère (les critères étant, dans notre contexte applicatif, la qualité et le prix), et ce grâce à une hybridation entre l'algorithme de fourmis et l'intégrale de Choquet. Le choix de l'intégrale de Choquet, d'abord comme outil d'agrégation des critères qualité, puis comme outil de modélisation du compromis qualité/prix, s'est avéré pertinent. Ainsi, les résultats expérimentaux montrent que les préférences d'agrégation y sont reproduites, notamment celles de l'agrégation des dimensions qualité où les critères sont parfois dépendants les uns des autres.

2. Perspectives

Les perspectives ouvertes par nos travaux visent à évaluer les performances de BrokerACO dans le cadre plus complexe d'une campagne multicanal impliquant plusieurs attributs cruciaux. L'objectif sera de mettre l'accent sur l'aspect contextuel de notre solution et exploiter les capacités de l'intégrale de Choquet à modéliser et quantifier les préférences.

Par ailleurs, afin d'atteindre des temps de réponse acceptables, nous prévoyons d'implémenter une version parallélisée de BrokerACO, où chaque plan fichier s'établira indépendamment l'un de l'autre. Une phase de conciliation permettra par la suite de choisir les prospects communs à adresser.

De plus, le prototype de BrokerACO ayant intéressé certains acteurs du marché B-to-B, nous nous proposons de développer un produit commercialisable offrant les différentes possibilités de prospection : mono-canal mono-attribut, mono-canal multi-attributs et multi-canal multi-attributs. Ce produit comporterait éventuellement une première interface d'apprentissage des préférences de l'utilisateur. Ainsi, en fonction du nombre d'attributs et du nombre de seuils pour chaque attribut, la table d'apprentissage est automatiquement générée, permettant d'abord à l'utilisateur de définir et saisir son pré-ordre, et à l'intégrale de Choquet ensuite d'apprendre et quantifier ses préférences. Une deuxième interface permettrait à l'utilisateur de connecter sa base de prospection et de sélectionner le fichier de ciblage adéquat en fonction des préférences précédemment définies.

L'usage de BrokerACO ne se limite pas uniquement à la prospection marketing. En effet, nous pouvons l'utiliser comme outil d'ETL au moment de l'intégration de données multisources dans une base de données centralisée. Cette intégration se ferait, par exemple, sur la définition d'un ensemble de critères qualité lesquels seraient agrégés et scorés selon les préférences de l'administrateur de la base de données.

Références bibliographiques

[Abraham et al. 06] Abraham, A.; Grosan, C.; Ramos, V. Swarm Intelligence and Data Mining. In *Studies in Computational Intelligence*, Springer Verlag, Germany, 2006.

[Agosta 02] Agosta, L. The meaning of data quality: the data strategy advisor. Information management magazine. 2000. Disponible sur :http://www.information-management.com/issues/20021001/5817-1.html.

[Ahmad et al. 08] Ahmad, M.A.; Srivastava, J. An Ant Colony Optimization Approach to Expert Identification in Social Networks, In *First International Workshop on Social Computing, Behavioral Modeling, and Prediction*, Arizona, 2008, p. 120-128.

[Akoka et al. 07] Akoka, J.; Berti-Équille, L.; Boucelma, O.; Bouzeghoub, M.; Comyn-Wattiau, I.; Cosquer, M.; Goasdoué-Thion, V.; Kedad, Z.; Nugier, S.; Peralta, V.; Sisaid-Cherfi, S.; A framework for data quality evaluation in data integration systems. In *ICEIS'07 10th Int. Conf. on Enterprise Information Systems*, Madeira, Portugal, 2007, p. 170-175.

[Akoka et al. 08] Akoka, J. ; Berti-Equille, L. ; Boucelma, O. ; Bouzeghoub, M. ; Comyn-Wattiau, I. ; Cosquer, M. ; Gasdoué, V. ; Kedad, Z. ; Nugier, S. ; Peralta, V. ; Quafafou, M. ; Sisaïd-Cherfi, S. Evaluation de la qualité des systèms multisources: une approche par les patterns. *In QDC'08, 4ème Atelier Qualité des données et des connaissances*, Nice, France, 2008.

[Al Ani 05] Al Ani, A. Feature Subset Selection Using Ant Colony Optimization, In *International Journal of Computational Intelligence*, 2005, Vol. 2, No. 1, p. 53-58.

[Ardagna et al. 05] Ardagna, D.; Cappiello, C.; Comuzzi, M.; Francalanci, C.; Pernici, B. A broker for selecting and provisionning high quality syndicated data. In *the 10th ICIQ International Conference on Information Quality*, MIT, Cambridge, MA, USA, 2005.

[Ardagna et al. 06] Ardagna, D.; Cappiello, C.; Francalanci, C.; Groppi, A. Brokering Multisource Data with Quality' in Constraints On the Move to Meaningful Internet Systems. In *CoopIS, DOA, GADA, and ODBASE, Lecture Notes in Computer Science*, 2006, Vol. 4275/2006, p. 807-817.

[Arnone et al. 93] Arnone, S.; Loraschi, A.; Tettamanzi, A. A genetic approach to portfolio selection. In *Neural Network World 6*, 1993, p. 597-604.

[Ballou et al. 98] Ballou, D. P.; Wang, R. Y.; Pazer, H.; Tayi, G. K. Modeling Information Manufacturing Systems to Determine Information Product Quality. In *Management Science*, No. 4, Vol. 44. 1998, p. 462-484.

[Batini et al. 06] Batini. C.; Scannapieco. M. Data quality: concepts, methodologies and techniques. *Data-Centric Systems and Applications*, Springer 2006. ISBN 978-3-540-33172-8.

[Batini et al. 11] Batini, C.; Barone, D.; Cabitza F.; Grega, S. A data quality methodology for heterogeneous data. In *International Journal of Database Management Systems (IJDMS)*, Vol.3, No.1, 2011, p.60.

[Batista et al. 07] Batista, M. C. M.; Salgado, A. C. Information quality measurement in data integration schemas. In *proccedings of the 33th conference of VLDB*, Vienna, Austria, 2007, p.61-72.

[Bayardo et al. 97] Bayardo, R.; Bohrer,W.; Brice, R.; Cichocki, A.; Fowler, G.; Helal, A.; Kashyap, V.; Ksiezyk, T.; Martin, G.; Nodine, M.; Rashid, M.; Rusinkiewicz, M.; Shea, R.; Unnikrishnan, C.; Unruh, A.; Woelk, D. Infosleuth: Semantic Integration of Information in Open and Dynamic Environments. In *Proceedings of the 1997 ACMInternational Conference on the Management of Data (SIGMOD), Tucson, Arizona*, 1997, p.195-206.

[Ben Hassine et al. 09] Ben Hassine, S. *D*ata Quality Evaluation in an E-Business Environment: A Survey. In *ICIQ'09: the 14th International Conference on Information Quality*, Potsdam, Germany, 2009, p.189-p201.

[Ben Hassine-Guetari et al. 10] BenHassine-Guetari, S.; Darmont, J.; Chauchat, J.H. Aggregation of data quality metrics using the Choquet integral, *In QDB'10: the 8th International Workshop on Quality in Databases (VLDB)*, Singapore, 2010.

[Bent et al. 08] Bent, G.; Dantressangle, P.; Vyvyan, D.; Mowshowitz, A.; Mitsou, V. A dynamic distributed federated database. In the 2nd annual conference of ITA, Imperial College, London, UK, 2008. Disponible sur: https://www.usukitacs.com/papers/3864/A%20Dynamic%20Distributibuted%20Federated%20Database.pdf.

[Berry et al. 83] Berry, L.T.; Shostack, G.L.; Upah, G.D. Emerging Perspectives on Services Marketing. In *Proceedings of Services Marketing Conference: American Marketing Association*, Chicago, IL, 1983.

[Berti-Equille 99] Berti-Equille, L. *La qualité des données et leur recommandation : modèle conceptuel, formalisation et application à la veille technologique*. Thèse de doctorat en informatique. Toulon, 1999, 241p.

[Berti-Equille 07] Berti Equille, L. *Quality awareness for managing and mining data*. Habilitation à Diriger des Recherches, Univ. of Rennes 1, 2007. Disponible sur :

http://pageperso.lif.univ-mrs.fr/~laure.berti/pub_files/Habilitation-Laure-Berti-Equille.pdf 2007.

[Black 58] D. Black. *The Theory of Committees and Elections*. Cambridge University Press, 1958, Kluwer Academic Publishers: New edition (1986). ISBN-13: 978-0898381894.

[Bleiholder et al. 08] Bleiholder, J.; Naumann, F. Data fusion, In *ACM computing surveys* (CSUR), Vol. 41, N. 1, 2008, pp1-41.

[Boisdevésy 96] Boisdevésy, J.C. *Le marking relationnel : à la découverte du conso-acteur*, Les Editions d'Organisation, 1996.

[Bosc et al. 95] Bosc, P.; Pivert, O. SQLf: A relational database language for fuzzy querying. In *IEEE transactions on Fuzzy systems*, Vol. 3, no. 1, 1995, p. 1–17.

[Bouzeghoub et al. 04] Bouzeghoub, M.; Peralta, V. A Framework for Analysis of Data Freshness, In *Proceedings of the 2004 international workshop on Information quality in information systems: IQIS 2004*, ACM New York, NY, USA ©2004, ISBN:1-58113-902-0, 2004, p. 59-67.

[Branke et al. 08] Branke, J.; Deb, K.; Miettinen, K.; Słowinski, R. *Multiobjective optimization*, Springer-Verlag Berlin Heidelberg, Germany, 2008.

[Branke et al. 10] Branke, J.; Greco, S.; Słowiński, R.; Zielniewicz, P. Interactive evolutionary multiobjective optimization driven by robust ordinal regression. *Bulletin of the Polish Academy of Sciences*, Technical Sciences, Vol. 58, No.3, 2010, p. 347–358.

[Busse et al. 99] Busse, S.; Kutsche, R.D.; Leser, U.; Weber, H. *Federated information systems: concepts, terminology and architectures,* Technical report, TU Berlin, No. 99-9, 1999.

[Coello Coello 02] Coello Coello, C. A. Theoretical and Numerical Constraint-Handling Techniques used with Evolutionary Algorithms: A Survey of the State of the Art, In *Computer Methods in Applied Mechanics and Engineering*, Vol. 191, No. 11-12, 2002, p. 1245-1287.

[Calvanese et al. 05] Calvanese, D.; Giacomo, G. D.; Lembo, D.; Lenzerini, M.; Rosati, R. Inconsistency tolerance in P2P data integration: an epistemic logic approach. In *International conference on database programming languages (DBPL)*, Trondheim, Norway, 2005.

[Chawathe et al. 94] Chawathe, S.; Garcia-Molina, H.; Hammer, J.; Ireland, K.; Papakonstantinou, Y.; Ullman, J.; Widom, J. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In *the 16th Meeting of the Information Processing Society of Japan (IPSJ)*, Tokyo, Japan, 1994, p.7-18.

[Cholvy 04] Cholvy, L. Information evaluation in fusion: a case study. Information processing and management of uncertainty. In *Knowledge-Based Systems (IPMU 2004)*, Perugia, 2004, p. 993–1000.

[Coello Coello et al. 07] Coello Coello, C. A.; Lamont, G. B.; Van Veldhuizen, D. A. *Evolutionary Algorithms for Solving Multi-Objective Problems*. 2nd Edition, Springer, 2007, 800p. ISBN 978-0-387-36797-2.

[Collette et al. 02] Collette, Y.; Siarry, P. *Optimisation multiobjectif*, Editions Eyrolles, 2002, 322p. EAN13 : 9782212111682.

[Davidson 04] Davidson, B. Developing data production maps: meeting patient discharge data submission requirements. In *International Journal of Healthcare Technology and Management*, Vol.6, No.2, 2004, p. 223-240.

[Delavallade et al. 10] Delavallade, T. ; Akdag, H. ; Bärecke, 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 (Cotation des informations), Journées Francophones d'Ingénierie des Connaissances, IC'10*, 2010, p. 87-98.

[Dempster 67] Dempster, A. P. Upper and Lower Probabilities Induced by a Multivalued Mapping. In *Annals of Mathematical Statistics*, Vol 38, 1967, p. 325-339.

[Dempster 68] Dempster, A. P. A Generalisation of Bayesian Inference (with discussion). In *Journal of Royal Statistical Society*, ser. B 30, 1968, p. 205-247.

[Dey 04] Dey, A. *Data Integration System based on both GAV and LAV query processing approaches*. Submitted in partial fulfillment of the requirements for the MSc Degree in Advanced Computing of the University of London and for the Diploma of Imperial College of Science, Technology and Medicine, 2004, 115p.

[Di Gaspero et al. 07] Di Gaspero, L.; Di Tollo, G.; Roli, A.; Schaerf, A. Hybrid Local Search for Constrained Financial Portfolio Selection Problems. In *Proceedings of the 4th international conference on Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems (CPAIOR '07)*, Brussels, Belgium, Springer-Verlag, Berlin, Heidelberg, 2007, p. 44-58.

[Dong et al. 09a] Dong, X. L.; Berti-Equille, L.; Srivastava, D. Integrating conflicting data: the role of source dependence. In *Proceedings of the PVLDB endowment*, Vol 2, No. 1, 2009, p. 550-561.

[Dong et al. 09b] Dong, X. L.; Naumann, F. Data Fusion: resolving data conflicts for integration. In *Proceedings of the PVLDB endowment*, Vol. 2, No. 2, 2009, p. 1654-1655.

[Dorigo et al. 96] Dorigo, M.; Maniezzo, V.; Colorni, A. Ant system: optimization by a colony of cooperating agents. In *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, Vol. 26, No. 1, 1996, p. 29-41.

[Dorigo et al. 04] Dorigo, M.; Stützle, T. Ant Colony Optimization, *MIT Press*, Cambridge, 2004, 319p. ISBN: 9780262042192.

[Dorigo et al. 05] Dorigo, M.; Blum C., Ant Colony Optimization Theory: A Survey. In *Theoretical Computer Science*, Vol. 344, No. 2-3, 2005, p. 243-278.

[Dréo et al. 03] Dréo, J. ; Petrowski, A. ; Taillard, E. ; Siarry P. *Métaheuristiques pour l'optimisation difficile*, Édition Eyrolles, 2003, ISBN : 2-212-11368-4.

[English 99] English, L. P. *Improving Data Warehouse and Business Information Quality : Methods for reducing costs and increasing profits*. Wiley & Sons, 1999, 544p. ISBN: 978-0-471-25383-9.

[Even et al. 08] Even, A.; Shankaranarayanan, G. Comparative analysis on data quality and utility inequality assessments. In *ECIS'08*, 2008, p. 1835-1846.

[Fargier et al. 09] Fargier, H. ; Lemaître, M., Décision multicritères, ISAE, Cours, 2009, 93p. Disponible sur : *ftp://ftp.irit.fr/IRIT/ADRIA/PapersFargier/MCDMssp.pdf.*

[Falorsi et al. 06] Falorsi, P. D. ; Scannapieco, M. Principi Guida per la Qualita dei Dati Toponomastici nella Pubblica Amministrazione (in Italian). In *ISTAT*, serie Contributi, Vol. 12. Disponible sur : http://www.istat.it/dati/pubbsci/contributi/Contr anno2005.htm, 2006.

[Fogel 66] Fogel, L. J.; Owens, A. J.; Walsh, M. J. *Artificial Intelligence through Simulated Evolution*. New York: John Wiley, 1966, 170p.

[Fogel 95] Fogel, D. B. *Evolutionary Computation. Toward a New Philosophy of Machine Intelligence*. IEEE Press Piscataway, NJ, USA ©1995, 1995, 296p, ISBN:0-7803-1038-1.

[Fouchal et al. 10] Fouchal, H.; Gandibleux, X. ; Lehuede, F. Algorithme de Martins et intégrale de Choquet pour le calcul de plus courts chemins multi-critères préférés, In *11ème congrès de la Société Française de Recherche Opérationnelle et d'Aide à la Décision (ROADEF 2010)*, Toulouse, France, 2010.

[Friedman et al. 99] Friedman, M.; Levy, A.; Millstein, T. Navigational plans for data integration. In *Proceedings of the 16th National Conf. on AI*, AAAI Press, 1999, p. 67–73.

[Goldberg 89] Goldberg, D.E. *Genetic algorithms in search, optimization and machine learning* (1st ed.), Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA ©1989, 1989. ISBN:0201157675.

[Grabisch 95] Grabisch, M. A new algorithm for identifying fuzzy measures and its application to pattern recognition. In Proceedings of the *Int. Joint Conf. of the 4th IEEE Int. Conf. on Fuzzy Systems and the 2nd Int. Fuzzy Engineering Symposium*, Yokohama, Japan, 1995, p. 145-150.

[Grabisch et al. 00] Grabisch, M.; Roubens, M. Application of the Choquet integral in multicriteria decision making. In M. Grabisch, T. Murofushi, and M. Sugeno, editors,*Fuzzy Measures and Integrals - Theory and Applications*, Physica Verlag, 2000, p. 348-374.

[Grabisch et al. 03] Grabisch, M.; Labreuche C. Capacities on lattices and k-ary capacities. In *Proceedings of the 3rd International Conference on Fuzzy Logic and Technology (EUSFLAT 2003)*, 2003, p. 473 – 490.

[Grabisch et al. 12] Grabisch, M.; Kojadinovic, I.; Meyer, P. *Package 'kappalab'*,version 04-5, 2012. Disponible sur: http://cran.r-project.org/web/packages/kappalab/kappalab.pdf.

[Hacid et al. 04] Hacid, M. S. ; Reynaud, C. L'intégration de sources de données. In *Revue information, interaction, intelligence*, Numéro Hors série Web sémantique, 2004.

[Hao et al. 99] Hao, J. K. ; Galinier, P. ; Habib, M. Méthaheuristiques pour l'optimisation combinatoire et l'affectation sous contraintes. In *Revue d'Intelligence Artificielle*, Vol. 2, No. 13, 1999, p. 263-324.

[Heinrich et al. 09] Heinrich, B.; Klier, M. A novel data quality metric for timeliness considering supplemental data. In *Proceedings of the 17th European conference on information systems (ECIS 2009)*, Verona, Italy, 2009, p. 2651-2662.

[Helfert et al. 09] Helfert, M.; Foley, O.; Ge, M. ; Cappiello, C. Analysing the effect of security on information quality dimensions. In *Proceedings of the 17th European conference on information systems (ECIS 2009)*, Verona, Italy, 2009, p. 2785-2797.

[Hernandez et al. 95] Hernandez, M. A.; Stolfo, S. J. The merge/purge problem for large databases. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, Vol. 24, No. 2, 1995, p. 127-138. ISBN:0-89791-731-6.

[Holland 75] Holland, J. *Adaptation in natural and artificial systems*, University of Michigan Press, Ann Harbor, 1975, 183p.

[Holt 05] Holt, K.G . The Data Quality Act. *In Proceedings of the 1st Annual National Foodborne Epidemiologists Meeting*, 2005, 15p. Disponible sur:

http://www.aphl.org/conferences/proceedings/Documents/2005_1st_National_Foodborne_Epid_Meeting/15_Holt.pdf.

[Huang et al. 95] Huang, Y. S.; Suen, C. Y. A method of combining multiple experts for the recognition of unconstrained handwritten numerals, In *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , Vol.17, No.1, 1995, p.90-94.

[Jacquet-Lagrèze et al. 87] Jacquet-Lagrèze, E.; Meziani, R.; Słowiński, R. MOLP with an interactive assessment of a piecewise linear utility function. In *European Journal of Operational Research*, Vol. 31, 1987, p. 350–357.

[Jensen et al. 03] Jensen, R.; Shen, Q. Finding Rough Set Reducts with Ant Colony Optimization. In *Proceedings of the 2003 UK Workshop on Computational Intelligence*, 2003, p 15-22.

[Jensen 06] Jensen, R. Performing Feature Selection with ACO. In *Studies in Computational Intelligence (SCI)*, Springer Verlag, Germany, Vol. 34, 2006, p. 45–73, 2006.

[Kalyanmoy 08] Kalyanmoy, D. Introduction to Evolutionary Multiobjective Optimization, In *Multiobjective Optimization*, Lecture Notes in Computer Science, Vol. 5252, 2008, p. 59-96.

[Kashyap et al. 00] Kashyap, V.; Sheth, A. Information Brokering across Heteroge-neous Digital Data: A Metadata-based Approach, In *Advances in Database Systems*, Kluwer Academic Publishers, Boston, London, 2000, 224p. ISBN 978-0-306-47028-8.

[Khun & Tucker 51] Kuhn, H. W.; Tucker, A. W. Nonlinear Programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, Calif., 1951, p. 481-492. Disponible sur: http://projecteuclid.org/euclid.bsmsp/1200500249.

[Kiebling et al. 02] Kiebling, W.; Köstler, G. Preference SQL − Design, Implementation, In *Proceedings of the 28th International Conference on Very Large Databases (VLDB 2002)*, Hong Kong, China, 2002, p. 990-1001.

[Klamroth et al. 08] Klamroth, K.; Miettinen, K. Integrating Approximation and Interactive Decision Making in Multicriteria Optimization, In *Operations Research*, Vol 56, No. 1, 2008, p. 222-234.

[Kojadinovic 06] Kojadinovic, I. *Contributions to the interpretation of non-additive measures and to the identification of decision-making models based on the Choquet integral*, Habilitation à diriger les recherches, Université de Nantes, 2006, 91p.

[Kolovos et al. 06] Kolovos, D. S.; Paige, R.F.; Polack, F.A.C. Merging models with the epsilon merging language (EML), In *Model driven engineering languages and systems*, Lecture Notes in Computer Science, Vol. 4199, 2006, p. 215-229.

[Kostadinov et al. 05] Kostadinov, D. ; Peralta, V. ; Soukane, A. ; Xue, X. Intégration de données hétérogènes basée sur la qualité. In *Actes des 23ème congrès INFORSID'05*, Grenoble, France, 2005, p. 471-486.

[Kovac et al. 97] Kovac, R.; Lee, Y.W.; Pipino, L.L. Total Data Quality Management: The Case of IRI. In *the Proceedings of the 1997 Conference on Information Quality*, 1997, p. 63-79.

[Koza 92] Koza, J.R. *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press Cambridge, MA, USA ©1992, 1992, 835p. ISBN 0-262-11170-5.

[Labreuche et al. 08] Labreuche, C.; Grabisch, M. The Choquet integral for the aggregation of interval scales in multicriteria decision making, In *CoRR*, Vol abs/0804.1762, 2008. Disponible sur: http://arxiv.org/pdf/0804.1762.pdf.

[Labreuche 09] Labreuche C. (2009), On the completion mechanism produced by the Choquet integral on some decision strategies, Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, Lisbon, Portugal, pp 567-572.

[Lampinen 00] Lampinen, J. *Multiobjective Nonlinear Pareto Optimization, A Pre-Investigation Report*. Lappeenranta University of Technology, Laboratory of Information Processing, Lapperanta, Finland, 2000, 31p.

[Lee et al. 06] Lee, Y. W.; Pipino, L.; Funk, J. D.; Wang, R. Y. *Journey to Data Quality*. The MIT Press ©2006, 2006, 226p. ISBN:0262122871.

[Levy 98] Levy, A. Y. The Information Manifold Approach to Data Integration. In *IEEE Intelligent Systems*, Vol 13, 1998, p. 12-16.

[Liétard et al. 09] Liétard, L. ; Rocache, D. Requêtes à Préférences et Bipolarité. In *15ème Colloque National de la Recherche en IUT CNRIUT 2009*, Lille, 2009, 8p. Disponible sur : http://cnriut09.univ-lille1.fr/articles/Articles/Fulltext/253a.pdf.

[Mansini et al. 03] Mansini, R.; Ogryczak, W.; Speranza, M. G. Lp solvable models for portfolio optimization a classification and computational comparison. In *IMA Journal of Management Mathematics*, 2003, p.187-220.

[Marichal 98] Marichal, J. L. *Aggregation Operators for Multicriteria Decision Aid*. PhD thesis, Institute of Mathematics, University of Liège, Liège, Belgium, 1998.

[Marichal 00] Marichal J. L. An axiomatic approach of the discrete Choquet integral as a tool to aggregate interacting criteria. In *IEEE Transactions on Fuzzy Systems*, Vol 8, No 6, 2000, p. 800-807.

[Marichal 02] Marichal, J. L. Aggregation of interacting criteria by means of the discrete Choquet integral. In *Aggregation operators: new trends and applications*, Physica-Verlag GmbH, 2002, p. 224-244.

[Maydanchik 07] Maydanchik, A. *Data quality assessment*, Technics Publications, LLC , USA ©2007, 2007, 336p. ISBN:0977140024 9780977140022.

[Mavrotas 08] Mavrotas G., Generation of efficient solutions in Multiobjective Mathematical Programming problems using GAMS. Effective implementation of the ε-constraint method. In *Applied Mathematics and Computation 01/2009*, Vol. 213, 2008, p. 455-465.

[Mc. Brien et al. 03] Mc. Brien, P.; Poulovassilis, A. Data integration by bi-directional schema transformation rules, In *19th International conference on data engineering*, 2003, p. 227-238.

[McKeown et al. 77] McKeown, D. M. Jr.; Reddy, D. R. The MIDAS sensor database and its use in performance evaluation, In *Proceedings of Image Understanding Workshop*, Palo Alto, CA, 1977.

[Mecella et al. 02] Mecella, M. ; Scannapieco, M. ; Virgillito, A. ; Baldoni, R. ; Catarci, T. ; Batini, C. Managing data quality in cooperative information systems. In *Proceedings of the 10th International conference on cooperative information systems*, LNCS 2519, 2002, p. 486-502.

[Mena et al. 96] Mena, E.; Kashyap, V.; Sheth, A.; Illarramendi, A. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. In *Proceedings of the First IFCIS International Conference on Cooperative Information Systems (CoopIS'96)*, IEEE Computer Society Press, 1996, p. 14-25.

[Micheaux 07] Micheaux, A. *Perception et comportement du consommateur face à la pression marketing direct : Recherche empirique appliquée dans un contexte d'envoi d'emailings publicitaires*, Mémoire de thèse, Université Paris1, 2007, 254p.

[Michie et al. 94] Michie, D. ; Spiegelhlter, D. J. ; Taylor, C. *Machine learning, Neural and Statistical classification*. Ellis Horwood series in artificial intelligence, Ellis Horwoord, 1994, 298p. ISBN:0-13-106360-X.

[Murofushi et al. 91] Murofushi, T.; Sugeno, M. A theory of fuzzy measures. Representation, the Choquet integral and null sets, In *Journal of Mathematical Analysis and Applications*, Vol. 159, No. 2, 1991, p. 532–549.

[Motro et al. 06] Motro, A.; Anokhin, P. Fusionplex: resolution of data inconsis-tencies in the integration of heterogeneous information sources. In *Information Fusion journal*, Vol. 7, No. 2, 2006, p. 176-196.

[Naumann 98] Naumann, F. Data fusion and data quality. In *The new techniques and technologies for statistics seminar (NTTS)*, Sorrento, Italie, 1998, p. 147–154.

[Naumann et al. 99] Naumann, F.; Leser, U.; Freytag, J.C. Quality-driven integration of heterogeneous information systems. In *Proceedings of the 25th VLDB Conference*, Edinburgh, Scotland. 1999, p. 447-458.

[Naumann et al. 04] Naumann, F; Freytag, J. C.; Leser, U. Completeness of Integrated Information Sources. In *Information Systems*, Vol. 29, No. 7, 2004, p. 583–615.

[Olson 03] Olson, J. E. *Data quality: the accuracy dimension*, Morgan Kaufmann publishers, 2003, 314p. ISBN: 978-1-55860-891-7.

[Peralta et al. 04] Peralta, V.; Ruggia, R.; Kedad, Z.; Bouzeghoub, M. A Framework for Data Quality Evaluation in a Data Integration System. In *Proceedings of the 19th Brazilian symposium on databases (SBBD 2004)*, Brasilia, Brazil, 2004, p. 134-147.

[Peralta 06] Peralta, V. *Data Quality Evaluation in Data Integration Systems*. Mémoire de thèse, Université de Versailles Saint-Quentin-en-Yvelines, France, 2006, 176p.

[Pipino et al. 02] Pipino, L. L.; Lee ,Y. W.; Wang, R. Y., Data quality assessment, In *Communications of the ACM - Supporting community and building*, Vol. 45, No. 4, 2002, p. 211-218.

[Rechenberg 73] Rechenberg, I. *Evolutionasstrategie : optimierung technischer systeme nach prinzipien der biologischen evolution*. Department of Process Engineering, Technical University of Berlin, Stuttgart, 1973, 170p. ISBN-10: 3772803733.

[Redman 96] Redman, T. *Data Quality for the Information Age*. Artech House: Boston, MA, 1996, 303p.

[Redman 01] Redman, T.C. *Data Quality: The field guide.* Digital Press, Boston, 2001, 241p.

[Rizopoulos 10] Rizopoulos, N. *Schema matching and schema merging based on uncertain semantic mappings*. Thèse de doctorat, Imperial College London, 2010, 267p.

[Roy 68] Roy, B. Classement et choix en présence de points de vue multiples (la méthode ELECTRE). In *RIRO* 8, 1968, p. 57-75.

[Sant'Anna 02] Sant'Anna, A. P. *Data envelopment analysis of randomized ranks*. Pesqui. Oper, Vol. 22, No. 2, 2002, p. 203-215. Disponible sur: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0101-74382002000200007&lng=en&nrm=iso. ISSN 0101-7438.

[Schmeidler 86] Schmeidler, D. Integral representation without additivity. In Proceedings of the Amer. Math. Soc., Vol. 97, No. 2, 1986, p. 255-261.

[Schwefel 81] Schwefel, H. P. *Numerical optimization of computer models*, John Wiley & Sons Ltd (June 17, 1981), 1981, 398p. ISBN-13: 978-0471099888.

[Shafer 76] Shafer, G. A *Mathematical Theory of Evidence*. Princeton University Press, Princeton N.J, 1976, 314p. ISBN-13: 978-0691100425.

[Shklar et al. 95] Shklar, L.; Sheth, A.; Kashyap, V.; and Shah, K. Infoharness: Use of Automatically Generated Metadata for Search and Retrieval of Heterogeneous Information. In *Proceedings of CAiSE '95*. Lecture Notes in Computer Science, No 932, 1995, p. 217-230.

[Skolpadungket et al. 07] Skolpadungket, P.; Dahal, K.; Harnpornchai, N. Portfolio optimization using multi-objective genetic algorithms, In *IEEE Congress on Evolutionary Computation (CEC 2007)*, 2007, p. 516-523. E-ISBN: 978-1-4244-1340-9.

[Shankaranarayanan et al. 00] Shankaranarayanan, G.; Wang, R.; Ziad, M. IPMAP: Representing the Manufacture of an Information Product, In *the Proceedings of MIT Data quality Conference (IQ 2000)*, Boston, MA, 2000, p. 1-16.

[Shankaranarayanan et al. 07] Shankaranarayanan, G.; Wang, R. Y. IPMAP: Current State and Perspectives. In *Proceedings of the 12th International Conference on Information Quality (ICIQ)*, Boston, Massachusetts, U.S.A, 2007.

[Sheth et al. 90] Sheth, A. P.; Larson, J. A. Federated database systems for managing distributed, heterogeneous and autonomous databases. In *ACM Computing Surveys (CSUR) - Special issue on heterogeneous databases*, Vol. 22, Vol. 3, 1990, p. 183-236.

[Strong et al. 97] Strong, D. M.; Lee, Y. W.; Wang, R. Y. Data quality in context. In *Commun. Of the ACM*, Vol. 40, No. 5, 1997, p. 103-110.

[Subbu et al. 05] Subbu, R.; Bonissone, P. P.; Eklund, N. H. W.; Bollapragada, S.; Chalermkraivuth, K. C. Multiobjective financial portfolio design: a hybrid evolutionary

approach. In *Proceedings of IEEE Congress on Evolutionary Computation*, Vol. 2, 2005, p. 1722-1729.

[Sugeno 74] Sugeno, M. *Theory of fuzzy integrals and its applications.* Ph.D. Thesis, Tokyo Institute of Technology. 1974.

[Talburt 11] Talburt, J. R. *Entity resolution and information quality*. Morgan Kaufmann Publishers, 2011, 256p. ISBN-13: 978-0123819727.

[Tamani et al. 11] Tamani, N.; Liétard, L.; Rocacher, D. Bipolarity in flexible querying of information systems dedicated to multimodal trans-port networks. In *Proceedings of the 10th international symposium on programming and systems (IPSP'11)*, Algeria, 2011, p. 108-115.

[Tari et al. 98] Tari, Z.; Zalavsky, A.; Savnik, I. Supporting cooperative databases with distributed objects. In *Parallel and distributed systems: theory and applications*, J.L. Aguilar Castro publisher, 1998.

[Wand et al. 96] Wand, Y.; Wang, R. Y. Anchoring data quality dimensions in ontological foundations. In *Communications of the ACM.*, Vol. 39, No. 11, 1996, p. 86-95.

[Wang et al. 90] Wang, R. Y.; Madnick, S. E. A polygen model for heterogeneous database systems: the source tagging perspective. In *VLDB 90*, Brisbane, Australia, 1990, p. 519-538.

[Wang et al. 92] Wang, R. Y.; Reddy, M. P.; Kon, H. B. *Data quality requirements analysis and modeling*. Cambridge, Mass.: Alfred P. Sloan School of Management, Massachusetts Institute of Technology, 1992, 15p. Disponible sur: http://mitiq.mit.edu/documents/publications/TDQMpub/IEEEDEApr93.pdf.

[Wang et al. 96] Wang, R. Y.; Strong, D. M. Beyond Accuracy : What data quality means to data consumers. In *Journal of Management Information Systems*, Vol. 12, No. 4, 1996, p. 5-33.

[Wang et al. 01] Wang, R. Y.; Ziad, M.; Lee, Y. W. *Data quality*. Kluwer academic publishers, 2001, 167p. ISBN-13: 978-0792372158.

[Wiederhold 93] Wiederhold, G. Intelligent integration of information. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, 1993, p. 434-437.

[Wendling et al. 08] Wendling, L.; Rendek, J.; Matsakis, P. Reconnaissance de symboles graphiques par le biais de l'intégrale de Choquet. In *Colloque International Francophone sur l'Ecrit et le Document - CIFED 08*, 2008, p. 175-180.

[Whitley et al. 88] Whitley, D. Genitor: A different genetic algorithm. In *Proceedings of Rocky. Mountain Conference on Artificial Intelligence*, Denver, 1988, 18p.

[Wijnhoven et al. 07] Wijnhoven, F.; Boelens, R.; Middel, R.; Louissen, K. Total data quality management: a study of bridging rigor and relevance. In *Proceedings of the 15th European Conference on Information Systems (ECIS 2007),* St. Gallen, Switserland, 2007, p. 925-937.

[Wu et al. 11] Wu, M.; Marian, A. A framework for corroborating answers from multiple web sources. In *Information systems journal*, Vol. 36, No. 2, 2011, p. 431-449.

[Yong 04] Young, W. A. *Evaluation of Peer-to-Peer Database Solutions*, 2004, 34p. Disponible sur: http://www.tonyyoung.ca/cs654paper.pdf.

Références web

[1] http://www.sncd.org/nos-publications/termes-du-marketing-direct/, consultée le 16/09/2014

[2] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31984L0450:fr:HTML, consultée le 16/09/2014

[3] http://www.adproxima.fr/glossaire-13-b-to-b.html, consultée le 16/09/2014

[4] http://www.insee.fr/fr/, consultée le 16/09/2014

[5] http://www.e-marketing.fr/Marketing-Direct/Article/LE-COUPLAGE-AU-SERVICE-DU-B-TO-B-23429-1.htm, consultée le 16/09/2014

[6] http://www.commercial-database.fr/fichiers-france-btob, consultée le 26/09/2014

[7] http://www.qas.com/, consultée le 16/09/2014

[8] http://www.journaldunet.com/management/marketing-commercial/enquete/07/070926-qualite-bases-donnees.shtml, consultée le 16/09/2014

[9] http://www.paperblog.fr/224431/b2b-la-qualite-des-bases-de-donnees-fait-defaut/, consultée le 16/09/2014

[10] http://www.experian.fr/ressources/actualites/cp-livre-blanc-data-quality-en-2012.html, consultée le 16/09/2014

[11] https://www.iso.org/obp/ui/fr/#iso:std:iso:19157:ed-1:v1:fr, consultée le 16/09/2014