sncd
Data Marketing
Industrie

Livre blanc
« Bien sûr, vos données sont parfaites ! »

Version 1.0 - Octobre 2018

Table des matières

1. Introduction

Le monde de la donnée (« ou data ») est un univers inconnu du grand public. Cependant il joue un rôle de plus en plus important. Ne dit-on pas que le futur maître du monde sera celui qui maîtrisera la donnée tant dans sa collecte que dans son exploitation ?

Que cherchons-nous à faire avec elle ? Quels bénéfices peut-on en tirer ? Qui est concerné par la donnée ?
Autant de questions qui doivent nous interpeller. A l'heure d'une réglementation européenne uniforme sur la protection des données à caractère personnel, il est toujours bon de rappeler les fondamentaux sur la donnée **(*)**.

Une donnée n'est exploitable que si elle est exacte et fiable, sinon vos résultats ne seront pas à la hauteur de vos attentes. De plus, une donnée est vivante : en effet derrière cette donnée il y a un être humain qui bouge, achète, consulte, etc. De ces simples constats, découle une grande complexité dans la mise à jour des données se rapportant à une personne.

Le monde change avec sa digitalisation ! En effet cela induit de nouveaux challenges sans pour autant changer ceux qui existaient déjà.

Que cherche à faire un responsable marketing avec sa base de données ?
Un responsable marketing cherche avant tout à proposer ses offres aux bonnes personnes. Donc à être en mesure de les contacter afin de proposer de nouveaux produits ou de nouveaux avantages ou tout simplement l'informer de son actualité. Pour cela il a besoin de pouvoir envoyer une communication directement à son consommateur (via un courrier, un email, un SMS, via les réseaux sociaux...). Autant faut-il que les données d'adressage de son client soient correctes et à jour !

De même un responsable marketing collecte en permanence des données sur ses clients via des formulaires, des jeux, des cartes de fidélité, etc. Or toutes ces remontées peuvent créer une sacrée pagaille dans une base de données. En effet un client peut avoir fourni des informations via plusieurs canaux à différents moments. Nous touchons du doigt la problématique de la mise à jour des données client. Comment s'assurer que la personne dont on met à jour les données est bien la même personne que celle déjà référencée dans la base ? D'autant que les informations collectées peuvent être différentes d'un canal à l'autre !

Notre univers n'est pas « sexy » mais il est indispensable ! Il permet de maintenir une base de données clients à jour et d'éviter de stocker deux fois une même personne sans l'avoir identifiée comme étant la même.

Dans ce document, nous évoquons **des techniques pour entretenir votre base de données**, en détaillant les forces et les faiblesses.
Nous attirons votre attention sur la complexité de ces opérations et **du rôle capital de l'homme pour trancher**.
Enfin nous nous inscrivons totalement dans notre temps et notamment dans **la protection des données à caractère personnel** et dans la **responsabilité** qui nous incombe, en tant que prestataire manipulant de la donnée tous les jours.

(*) *Dès 1978, la loi Informatique et Libertés introduit l'exigence « d'exactitude de la donnée eu égard aux finalités pour laquelle elle a été collectée ».*
Quarante ans plus tard, le RGPD replace cette exigence parmi les grands principes de protection des données, laquelle contribue aux conditions de licéité d'un traitement. Parmi ces grands principes nous retrouvons l'exigence de minimisation des données, jetant de l'ombre sur les pratiques actuelles de rapprochement par croisement avec toutes informations disponibles.
Face à cette double injonction - de minimisation, par la loi et d'enrichissement, à travers le développement de technologies plus performantes - l'entreprise devra développer une stratégie adéquate mariant innovation et conformité. La question centrale se posant alors ici étant celle du niveau impérieux d'enrichissement de la donnée eu égard aux besoins de l'entreprise.

2. La qualité des données

Nous sommes aujourd'hui plongés dans l'ère du Big Data, de l'interconnexion des données et de leur valorisation. La relation client devient de plus en plus proactive, ce qui permet d'anticiper les besoins, d'adapter les stratégies, les opérations, etc.

L'ère numérique marque le passage d'un monde où régnait un seul type de données vers un monde où règnent trois types de données : aux données structurées s'ajoutent les données semi-structurées et non-structurées (issues des réseaux sociaux, mobiles, web).

De plus, il est à noter que les données personnelles sont classées en deux catégories distinctes :

- Les données personnelles identifiantes : ce sont les données rattachées directement à l'identité d'une personne. Ce sont généralement le nom, l'adresse, l'email, la situation familiale, ou encore toutes les données ou fichiers permettant d'identifier indirectement une personne, via par exemple un numéro d'identification, une adresse IP, etc.

- Les données comportementales : ces données sont rattachées à l'ensemble des comportements d'un individu, collectées via le suivi de ses navigations ou mettant en exergue ses comportements d'achats ce qui permet, par exemple, pour un retailer détaillant, d'affiner le profil d'une personne.

Alors pourquoi parler de RNVP (Restructuration Normalisation Validation Postale) qui s'adresse aux adresses postales ? Dans ce contexte devenu très digital l'enjeu reste toujours le même, bien que plus complexe : identifier son client, un consommateur potentiel, quel que soit le mode d'expression et quel que soit le canal.

Si l'on devait classer par ordre d'importance les points de contact d'un individu en fonction de leur pérennité, nous mettrons en tête de la liste l'adresse postale, le téléphone fixe, les e-mails, le téléphone mobile, l'identifiant réseau social, l'adresse IP, l'identifiant appareil, etc.

Pourquoi mettre en avant l'adresse postale dans notre monde de plus en plus digital ? Car une adresse postale reste le lien physique entre un individu et une société.

C'est pourquoi bien maintenir une base postale à jour reste fondamental encore de nos jours. Cependant même si une adresse postale reste importante nous ne devons pas négliger pour autant les autres canaux !

Quels sont aujourd'hui les techniques permettant de maintenir une bonne qualité des points de contact ? Existe-t-il des solutions miracles ? Quels sont les enjeux ? A court terme et à moyen terme ? Comment s'oriente le marché ?

Nous distinguerons les points de contact physiques des autres ou plus particulièrement le marché traditionnel du nouveau.

Nettoyer une base de données traditionnelle (d'adresses postales, de téléphones fixes, de téléphones mobiles) représente un investissement certain. Or cet investissement reste un préalable indispensable pour une exploitation du capital base de données.

2.1 Quel canal ?

Là où le digital ne jure que par la communication email et Web, une base de données se construit encore autour d'un point physique comme l'adresse postale. Certains diront qu'une adresse postale ne sert plus, cependant de nombreux acteurs qui ont fait leur révolution numérique en reviennent. Ce qui laisse présager, non pas un retour en force de ce média, mais un équilibre intéressant entre les différents médias.

- L'email média push est intéressant dans une communication « instantanée » ;
- Le courrier média push est adapté à une communication ROIste ;
- Les réseaux sociaux push et viraux sont à privilégier pour une « communauté » de fans ;
- Le display correspondra à une approche prospective ;
- etc.

Outre le fait qu'entretenir sa base de données multicanale reste primordiale, garder à jour le bon point de contact

2.2 La normalisation

2.2.1 Principe général

Comment identifier un individu si aucun traitement de normalisation de l'information n'est mis en place ? En effet la normalisation de la donnée permet le rapprochement de celle-ci avec une autre.

Par exemple :

Prenons Jean Autrée et JEAN AUTREE. Même si cela vous semble être le même individu du point de vue d'une machine (transformant les lettres en 0 et 1), ce sont 2 individus distincts.

La normalisation va donc consister à mettre au préalable ces « deux individus » sous une forme comparable soit :

Jean Autrée va devenir en Majuscule sans accent JEAN AUTREE qui de ce fait sera comparable au deuxième individu JEAN AUTREE et donc identique.

Comme le montre cet exemple, sans normalisation nous n'aurions pas pu rapprocher ces 2 individus et par conséquent nous aurions créé un doublon dans notre base de données.

Normaliser une information quelle qu'elle soit avant de l'injecter dans une base de données est une nécessité.

IBM n'avait-il pas comme principe « Garbage in Garbage out » ?

2.2.2 Normalisation d'une adresse postale

La normalisation des adresses postales répond à un principe postal défini par La Poste afin de pouvoir automatiser la chaîne de distribution. Aussi lorsque l'on parle en général d'une normalisation d'adresse postale nous faisons référence à la norme **Afnor NF Z10-011** stipulant le format d'une adresse postale.

La norme **Afnor NF Z10-011** stipule :

Un format de l'adresse :

- 6 lignes de 38 caractères maximum (éventuellement une 7ème ligne pour l'international afin d'indiquer le pays).
- Les lignes à blanc – non renseignées – sont supprimées lors de l'édition pour rendre l'adresse plus esthétique.
- L'adresse est alignée à gauche, ce qui facilite la lecture optique mécanisée des adresses par La Poste.

Une typologie des caractères autorisés ou non :

- A partir de la ligne 4 (numéro et libellé de la voie), aucune ponctuation, aucun italique ou souligné n'est autorisé, ce qui perturberait la lecture automatique des adresses.
- Par ailleurs, la norme veut que seule la dernière ligne soit en majuscule, mais La Poste recommande vivement une mise en majuscule des 4 dernières lignes adresses (Complément d'adresse, voie / lieux-dits ou Boîte Postale / Code Postal Commune), afin d'optimiser significativement la reconnaissance des adresses par les lecteurs optiques des centres de traitement de courrier.

Une structure d'adresse :

- Les éléments de l'adresse postale doivent être ordonnés selon une structure bien précise :

Lignes d'une adresse

Ligne 1 : Civilité Nom Prénom
Ligne 2 : Complément de Nom
Ligne 3 : Complément d'adresse
Ligne 4 : N° extension type et libellé de voie
Ligne 5 : Lieu-dit / boîte postale
Ligne 6 : Code postal Ville

Exemple

Ligne 1 : Monsieur Jean AUTREE
Ligne 2 : Etage 3
Ligne 3 : Résidence Patis Micaud
Ligne 4 : 340 RUE DU BOCAGE
Ligne 5 : NOTRE DAME DES LANGUEURS
Ligne 6 : 44440 JOUE SUR ERDRE

2.2.3 Normalisation d'un numéro de téléphone

Il n'y a pas contrairement à l'adresse postale une norme stricte. Nous parlerons plutôt de standards. Les formats les plus répandus sont :

Format national : 01 23 45 67 89
Format international : +33 (0)1 23 45 67 89

Quel que soit le format retenu, le principe reste le même : un seul standard doit être appliqué !

2.2.4 Normalisation d'une adresse email

Les adresses emails sont constituées des trois éléments suivants, dans cet ordre :

- Une partie locale, identifiant généralement une personne (lucas, Jean.Dupont, joe123) ou un nom de service (info, vente, postmaster) ;
- Le caractère séparateur @ (arobase), signifiant at (« à » ou « chez ») en anglais ;
- L'adresse du serveur, généralement un nom de domaine identifiant l'entreprise hébergeant la boîte électronique (exemple.net, exemple.com, exemple.org).

Le nom de domaine sert à identifier le serveur de messagerie auquel doit être acheminé un message via le protocole Simple Mail Transfer Protocol (SMTP). La transformation du nom de domaine en adresse IP se fait grâce au système de résolution de noms DNS.

Le standard généralement appliqué est l'adresse en minuscules.

2.3 La restructuration

2.3.1 Principe général

Il est nécessaire de positionner les éléments à comparer dans les mêmes cases pour avoir une comparaison efficace.

Reprenons l'exemple de Jean Autrée qui après normalisation est devenu JEAN AUTREE. Si nous comparons JEAN AUTREE avec AUTREE JEAN peut-on dire que c'est la même personne ?

Pour une machine ces deux individus ne sont pas identiques car AUTREE n'est pas égal à JEAN et JEAN n'est pas égal à AUTREE.

Une restructuration consiste donc à remettre dans la bonne case les éléments à comparer soit AUTREE avec AUTREE et JEAN avec JEAN.

2.3.2 Restructuration d'une adresse postale

La restructuration d'une adresse postale répond à des critères bien spécifiques.

L'adresse postale doit être ordonnée selon une structure bien précise stipulée dans la norme Afnor NF Z10-011.

Elle est composée de 6 lignes distinctes dans lesquelles nous trouvons :

Lignes d'une adresse

Ligne 1 : Civilité Nom Prénom
Ligne 2 : Complément de Nom
Ligne 3 : Complément d'adresse
Ligne 4 : Adresse
Ligne 5 : Lieu-dit / boîte postale
Ligne 6 : Code postal Ville

Cet exemple reste simple cependant lorsque vous manipulez une adresse postale vous manipulez 6 lignes différentes qui peuvent être désordonnées :

Adresse 1

Ligne 1 : Civilité Nom Prénom	Monsieur Jean Autrée
Ligne 2 : Complément de Nom	Etage 3
Ligne 3 : Complément d'adresse	Résidence Patis Micaud
Ligne 4 : Adresse	340 rue du Bocage
Ligne 5 : Lieu dit / boîte postale	ND des Langueurs
Ligne 6 : Code postal Ville	44440 Joue sur Erdre

Adresse 2

Ligne 1 : Civilité Nom Prénom	Monsieur Jean Autrée
Ligne 2 : Complément de Nom	Etage 3
Ligne 3 : Complément d'adresse	340 rue du Bocage
Ligne 4 : Adresse	Résidence Patis Micaud
Ligne 5 : Lieu dit / boîte postale	
Ligne 6 : Code postal Ville	44440 Joue sur Erdre

L'opération de restructuration va donc consister à inverser les éléments contenus dans les lignes 3 et 4 (complément d'adresse et adresse) de l'adresse 2 avant de lancer la comparaison avec l'adresse 1. Ce qui permettra d'identifier que les adresses 1 et 2 sont identiques et que Jean Autrée à l'adresse 1 et Jean Autrée à l'adresse 2 ne sont au final qu'une seule et même personne.

2.3.3 Restructuration d'une adresse email

La structuration d'une adresse email tient plus à des conventions qu'à une réelle application d'une norme spécifique.

Le RFC 3696 résume la syntaxe des adresses électroniques et est basé sur les RFC 2821 et RFC 2822. De nombreuses applications ne supportent pas l'ensemble des adresses valides (par exemple, en refusant l'utilisation d'une apostrophe) ou acceptent des adresses non valides. Seuls les lettres sans accent, les chiffres et le point sont très communs. Toutefois le paragraphe 5 de ce RFC 3696 stipule que les systèmes devraient accepter les accents dans les noms de domaines internationalisés IDN (noms de domaines avec accent). A titre d'exemple, il est maintenant possible d'enregistrer des domaines tels que voilà.fr depuis mai 2012 pour l'extension .fr et plusieurs autres.
Exemples d'adresses valides :

- Abc@example.com
- Abc@10.42.0.1
- user+mailbox/department=shipping@example.com
- !#$%&'*+-/=?^_`.{|}~@example.com
- «Fred Bloggs»@example.com
- Loïc.Accentué@voilà.fr8

Exemples d'adresses non valides :

- Abc.example.com (Le caractère @ manque)
- Abc.@example.com (Le caractère . est situé juste avant le caractère @)
- Abc..123@example.com (Le caractère . apparaît deux fois de suite)

La restructuration d'une adresse email reste simple et elle répond au format suivant :
Monemail @ mondomaine . extension

Cela consiste à vérifier que tous les éléments censés être présents sont bien à leur place. Certain procédés permettent de mettre en avant des incohérences dans les types de noms de domaine ou extensions.

2.4 La validation

2.4.1 Principe général

Valider une information (adresse postale, email, téléphone, individu) consiste à confronter l'information initiale avec un référentiel fiable.

Qu'est-ce qu'un référentiel ?

Un référentiel est une source de données mise à jour régulièrement contenant des informations de contact (ici dans notre sujet) d'un individu, d'une adresse, etc. (voir chapitre 6 - Les référentiels)
Comme toute chose évolue (une voie peut changer de nom, un individu change d'adresse, etc.), confronter vos données à un référentiel permet de maintenir votre fichier client à jour. Pour maintenir votre fichier à jour plusieurs confrontations par an sont nécessaires.

Plusieurs référentiels existent en France avec chacun leurs spécificités.

2.4.2 Référentiels postaux

La Poste entretient un référentiel des voies en France. Ce référentiel permet notamment de valider l'existence d'une adresse. Les logiciels de RNVP utilisent ce référentiel afin de pouvoir valider une adresse (**Validation Postale**) :

- **Hexavia** : Référentiel des voies en France ;
- **Hexaclé** : Référentiel des numéros dans une voie.

De même La Poste via le SNA dispose d'un **BNCA référentiel des changements d'adresses postales**. Ce référentiel vous permet de détecter et/ou de retrouver les adresses des personnes qui ont déménagé.

D'autres référentiels peuvent être utilisés dans le traitement d'une adresse postale comme :
- **Les annuaires téléphoniques** issus des opérateurs téléphoniques ;
- **Des référentiels privés** issus de la compilation de plusieurs sources (adresses de livraison) ;
- **Des référentiels géographiques** : issus des outils de navigation (tomtom, google maps, etc.).

Bien que très importants, ces référentiels ne permettent pas de compiler l'ensemble des individus français et de ce fait ne sont pas infaillibles.
Cependant « au règne des aveugles les borgnes sont rois ».

2.4.3 Référentiels emails

Il n'existe pas réellement de référentiels spécifiques aux emails. La plupart du temps ce sont des extensions de référentiels privés.

La complexité réside dans l'exhaustivité et les mises à jour régulières de ces bases.

Les routeurs d'emails détiennent des référentiels d'emails servant essentiellement à repousser de « mauvais » emails avant de les router. Cependant ces référentiels ne sont pas accessibles pour les autres prestataires.

Il existe des solutions permettant d'analyser votre fichier d'emails. Ces solutions sont basées sur l'interrogation des serveurs des FAI.

Grâce à ces solutions vous pouvez « valider » vos emails. Cela se traduit bien souvent par un code adossé à votre email renvoyant la réponse des serveurs.

Type de réponse	Nombre de code client	%
antispam_system	1	77,8%
dead_server	109	10,2%
disposable	43	5,0%
email_disabled	183	2,1%
Error email	7	1,3%
ok	6737	1,1%
ok_for_all	435	0,7%
p_antispam_system	26	0,5%
p_attempt_rejected	3	0,5%
p_email_disabled	64	0,3%
p_relay_error	4	0,3%
p_unknown_email	883	0,1%
smtp_protocol	43	0,1%
spamtrap	24	0,0%
syntax_error	7	0,0%
t_antispam_system	1	0,0%
t_relay_error	1	0,0%
t_unknown_email	1	0,0%
unknown_email	92	0,0%
(vide)		
Total général	8664	

D'où l'interprétation suivante :

Validation	Qté	%
Valid email	7215	83,3%
Invalid email	128	1,5%
Douteux	1297	15,0%
Spam	24	0,3%
	8664	

2.4.4 Référentiel téléphonique

Il fut un temps où France Télécom, avec sa position monopolistique de la téléphonie en France, détenait la liste de tous les abonnés au téléphone.

France Télécom permettait d'accéder à ce référentiel très complet des foyers abonnés au téléphone avec un taux de couverture de 95 % de la population française.

Ce temps est bien révolu avec l'ouverture du marché et la place grandissante des opérateurs alternatifs (Free, SFR, Bouygues…) a bouleversé la donne. Aujourd'hui France Télécom (Orange) ne détient pas tous les abonnés téléphoniques et a perdu de nombreuses parts de marché. Tant et si bien que le référentiel initial de FT ne peut plus être considéré comme tel car n'ayant pas l'exhaustivité des Français.

Pour l'obtenir, il faut passer dorénavant par des acteurs compilant les données annuaires de l'ensemble des opérateurs téléphoniques (opérateurs régionaux).

2.5 Les différentes approches

Le traitement d'entretien des données n'est pas une science exacte mais plutôt une méthode empirique qui reste avant tout un domaine de spécialistes.

Les algorithmes (soundex, phonedex…) bien que très puissants et utilisés régulièrement par l'ensemble de la profession, restent avant tout une histoire d'interprétation humaine.

La complexité tient autant à l'utilisation des algorithmes qu'à la complexité d'une adresse postale.

Les logiciels permettant de traiter des adresses internationales doivent tenir compte des singularités locales afin d'être performants.

L'approche standard

Certaines approches considèrent que l'importance réside dans la confrontation (vérification / correction) vis-à-vis de référentiels.
En s'appuyant sur des algorithmes phonétiques et typographiques, ces approches permettent de traiter 90 % des cas (les 10 % restant pouvant engendrer des coûts non négligeables).

L'approche experte

D'autres affinent via une approche combinée entre référentiels et dictionnaires (regex) permettant un traitement très qualitatif des données, mais encore artisanal malgré les volumes traités.

L'approche spécialiste

D'autres encore créent leur propre référentiel qui par confrontation permet non seulement un traitement standard des données mais également un enrichissement via une source « à jour ».

L'ensemble de ces approches, concernant l'adresse postale, reste malgré tout conditionné par l'évolution et les contraintes postales. Pour être "propre", une base de données doit répondre aux impératifs postaux, qui se traduisent à la fois par des impératifs métiers, économiques et juridiques.

2.6 Les dangers, les craintes et les fausses idées

Les fausses idées autour d'une base de données sont nombreuses et malheureusement peuvent coûter cher dans le temps.

- **Mes adresses sont récentes et ne nécessitent donc pas de traitement !**
  Faux : en effet la data quality permet de normaliser et de vérifier la qualité selon une norme identique des données avant l'intégration dans la base afin de mieux rapprocher donc mieux identifier un individu.

- **Je ne traite mes adresses que pour la recherche des déménagés !**
  Erreur : tous les ans des adresses changent, se créent ou disparaissent, aussi si vous ne traitez pas régulièrement l'intégralité de votre base, vous créez de facto des erreurs dans votre base.
  Par exemple, sur l'année 2017, une vague de fusions de communes a impacté la rédaction des adresses pour 3,5 % de l'ensemble des communes !

- **Traiter une base de données coûte cher :**
  Oui et non : le coût du traitement d'une base de données reste moins cher que de perdre un client ou d'envoyer des communications à une mauvaise adresse.
  Le coût de la non qualité sera :
  → le prix de la communication (incluant la fabrication et la distribution), soit de 0,50€ à 5€ par PND, selon qu'il s'agisse d'un simple mailing, ou d'un catalogue.

- **Traiter ma base de données me préserve des PND :**
  Hélas non : cela permet de limiter la « casse » mais n'élimine pas tous les PND.
  Les référentiels couramment utilisés pour le traitement des déménagés (Estocade, Charade) ne contiennent que 75 % des déménagés. De plus, ils ne contiennent que 5 ans d'historique (durée de conservation maximale autorisée par la CNIL).

Reste une certitude, si vous souhaitez créer une base de données et l'exploiter en marketing, vous devez traiter celle-ci très régulièrement faute de quoi votre « capital » va fondre, comme neige au soleil, très rapidement.

Une base de données est un capital vivant, un patrimoine immatériel de l'entreprise, et donc nécessite un entretien, une utilisation et une mise à jour constants.

2.7 Prospective

Laissons vagabonder notre imagination…

- Supposons une adresse postale uniquement représentée par une coordonnée (X,Y). L"intérêt du traitement de l'adresse postale devient négligeable. Il est par conséquent possible que l'adresse de demain soit significativement différente de ce que l'on peut connaître aujourd'hui !
- Les Postes européennes fusionnent ou bien s'harmonisent, les économies d'échelle semblent phénoménales.
- Plus besoin d'une adresse en clair dans une base de données pour joindre un consommateur mais juste d'un identifiant unique permettant de le joindre quel que soit le moyen.

Si l'on se réfère à l'évolution actuelle, les solutions de demain doivent permettent de :

- Faire des économies d'échelle ;
- Diminuer les erreurs d'interprétation possibles ;
- Identifier avec certitude un consommateur quel que soit le canal d'intégration ;
- Protéger les données personnelles des individus ;
- Respecter le droit d'information des personnes.

Allons-nous tomber dans un gouffre sans fond ou bien existe-t-il des solutions permettant de répondre à l'ensemble de ces besoins ?

La technologie évolue très rapidement et les idées fourmillent !

A suivre.

3. Le rapprochement des données

Le rapprochement des données, à l'instar de la RNVP, s'appuie sur des algorithmes probabilistes (et, par conséquent, inclut une dimension subjective).

La perception de l'efficacité des outils de rapprochement est très dépendante des individus ainsi que des finalités attendues.

La subjectivité de chacun amènera une vision différente de ce qu'est un doublon ou non.

3.1 Le « pourquoi »

Les process de rapprochement des données sont utilisés dans les domaines suivants :

3.1.1 Le nettoyage d'une base de contacts (dédoublonnage de base client, prospects, donateurs…)

Le dédoublonnage a pour but d'obtenir une vision unique du contact.
L'intérêt d'une vision unique est multiple :

→ Éviter le coût d'une communication inutile (envoi de catalogues en double) ;
→ Éviter de « décrédibiliser » une communication (n'expliquez pas à votre client **qu'il est unique… plusieurs fois !**) ;
→ Consolider les données d'achats de dons ou simplement l'historique de la communication client !

Que ce soit pour une simple raison comptable, une raison marketing, ou pour obtenir une base CRM intègre, vous aurez besoin d'une base dédoublonnée.

3.1.2 Préparation d'une base de prospection

Une campagne de prospection (qu'elle soit téléphonique, électronique ou papier) nécessite de :

→ Fusionner vos différentes listes de prospection, puis supprimer (selon une priorité choisie) les doublons ;
→ Appliquer des repoussoirs, c'est-à-dire des fichiers de contacts que vous ne souhaitez surtout pas prospecter ! Ces repoussoirs peuvent se répartir de la façon suivante :

- les repoussoirs dits externes : liste des personnes refusant le démarchage (Bloctel pour le téléphone, Robinson pour le postal) ;
- les repoussoirs dits internes : liste de vos clients, liste de vos prospects les plus récents (afin d'éviter d'appliquer une pression commerciale trop forte, et donc de devoir payer pour un acheminement qui n'aura pas l'efficacité souhaitée), mais également bien sûr les personnes vous ayant demandé de ne plus les prospecter.

3.1.3 Recherche de préexistence dans une base

Il s'agit ici d'une démarche « interactive ».

Lors de la saisie d'une nouvelle entrée en base, il peut être intéressant de vérifier au préalable si celle-ci existe ou non.

Attention, si cette vérification ne pose pas de souci de confidentialité quand la saisie est effectuée par un service interne (comptabilité, call-center…), il est nécessaire de prendre **infiniment plus de précautions** si la saisie est faite par le contact lui-même !
(Il serait fort peu convenable de demander au contact de lever une éventuelle ambiguïté entre lui-même et son voisin de palier !)

Nous verrons ensuite que selon le niveau d'exigence demandé, la recherche de préexistence peut être très simple (identité du nom et du CP par exemple) ou très pointue (en s'appuyant sur des grammaires de déduplication tolérantes aux fautes).

3.1.4 Intégration d'un flux (mise à jour + création de contacts) dans un fichier stock

L'intégration d'un flux est le pendant « batch » de la recherche de préexistence.

L'objet n'est pas de rechercher l'ensemble des doublons dans le fichier stock, mais de faire la part des nouveaux entrants et des contacts déjà présents en base.

3.1.5 L'enrichissement de données

L'enrichissement de données revient à confronter une base à enrichir avec une base de référence (voir le chapitre 4 - Enrichissement de données).

A l'instar du paragraphe précédent, il ne s'agit pas ici de connaître les doublons présents dans le fichier à enrichir ou la base de référence, mais plutôt de connaître les communs entre les deux fichiers, puis ensuite d'enrichir les données de l'un vers l'autre.

Les données à enrichir peuvent être très diverses :

→ Informations postales (retrouver un déménagé, simplement enrichir une adresse depuis un téléphone ou un email) ;
→ Informations téléphoniques, électroniques (le plus souvent en s'appuyant sur les données postales du contact) ;
→ Informations comportementales (issues le plus souvent de mégabases) ;
→ Informations socio-démographiques ;
→ Informations géographiques (zones de chalandises, trajets usuels...).

Les données disponibles sont issues :

→ des fournisseurs historiques (France Télécom, opérateurs téléphoniques alternatifs, La Poste, les hébergeurs de mégabases, l'Insee) ;
→ et de plus en plus, des données gratuites disponibles en open data.

Ces traitements d'enrichissement doivent être réalisés avec beaucoup de précautions afin de préserver l'historique des différents consentements fournis par les personnes quand ceux-ci sont nécessaires (email, sms, etc.). Une attention particulière sera portée à la licéité de la source d'enrichissement et sa fraîcheur.

3.2 Les alternatives

Que vous souhaitiez nettoyer vos doublons, installer un process de recherche de pré-existence, ou simplement lancer une campagne de prospection, vous n'aurez malheureusement pas d'alternatives à l'utilisation d'outils de rapprochement (outils de déduplication).
La question concerne plus l'investissement souhaité, aussi bien dans la mise en place d'une solution, que dans le fonctionnement au quotidien.

3.3 Le « comment »

3.3.1 Les réglages « historiques »

- **BtoB ? BtoC ? Mixte ?**

  Il s'agit ici de la probable première question à se poser sur une problématique de recherche des doublons.
  Les règles de recherche sont en effet différentes.
  En BtoC (fichier de « particuliers »), le poids de la voie est fort, celui du nom de famille également.
  En BtoB (fichier de sociétés ou d'associations), le travail de ressemblance des raisons sociales fait intervenir des gestions d'initiales. Vous pourrez également être amené à associer deux sociétés dont une est connue sous son adresse cedex, et l'autre sous son adresse géographique.

  La configuration la plus complexe est la présence d'un fichier comportant à la fois des adresses de particuliers et des adresses de sociétés.
  Certains outils sont capable de gérer cela.

- **Mode Individu ? Mode Foyer ?**

  Voulez-vous un contact par individu ?
  « Robert Michu » et « Martine Michu », même s'ils sont à la même adresse, seront considérés comme deux contacts distincts.

  Voulez-vous conserver un contact par foyer ?
  « Robert Michu » et « Martine Michu » seront alors doublons.

  Ce dernier choix (mode Foyer) est plus souvent préféré lors de déduplications préalables à des campagnes de prospection.

- **Mode Contact ? Mode Raison sociale ?**

  Il s'agit du pendant « foyer » ou « individu » pour le BtoB.

  En mode contact, vous choisissez de conserver l'ensemble des contacts d'une même société. Le mode « raison sociale » quant à lui ne conservera qu'une seule entrée par raison sociale (quel que soit le nombre de contacts rencontrés).

3.3.2 Doublon douteux ? ou doublon certain ?

Aussi peaufinés soient les réglages apportés à votre opération, vous obtiendrez des doublons certains, mais également des doublons douteux.

Madame Marie DIGNEL	*Madame Mari-Claire DIGNIEL*
281 rue des 4 Oliviers	*28 rue des 4 Oliviers*
34410 SERIGNAN	*34410 SERIGNAN*

Heureusement, le taux de douteux est généralement bien inférieur à celui des doublons certains.

Mais alors, que voudrez-vous faire de ces douteux ?

- **Levée d'ambiguïté manuelle**

  Si vous souhaitez nettoyer votre base de clients, le plus sage serait de valider ceux-ci manuellement.

  Sur un traitement « one-shot », la mécanique est raisonnable (techniquement). Vous intégrez les doublons certains (et fusionnez leurs historiques), puis temporisez les doublons douteux, le temps de la validation.
  Sur un traitement périodique (ou au fil de l'eau), il faut constituer un puit de « doublons douteux » (attente de réintégration). Ce puit se remplit au fil des traitements de rapprochement, et se vide au fil des validations.

  Deux précautions seront à prendre :

  - dimensionner les moyens humains dédiés à la levée d'ambiguïté pour, au minimum, avoir un puit de données qui se vide au moins aussi vite qu'il se remplit ;
  - historiser les doublons « invalidés », afin d'éviter que les traitements de rapprochement automatique les re-soumettent à validation à chaque fois qu'ils les rencontrent !

  Selon la taille de votre base, il vous faudra prévoir du temps. Peut-être mettre en place une commission qualité dédiée à ce sujet.

- **Choix d'un traitement entièrement automatisé**

  Beaucoup préfèrent en rester à des solutions entièrement automatisées (sans levée d'ambiguïté humaine).

  Vous laissez l'outil de rapprochement décider de ce qui est ou n'est pas doublon.
  Selon la finalité attendue, vous pourrez décider d'un réglage « over-kill » ou « under-kill » (voir glossaire).

  Under-kill
  Si vous ne souhaitez pas prendre le risque de regrouper deux individus différents, alors vous opterez pour un réglage sévère, qui limitera le nombre de doublons douteux.

Over-kill
Ou bien alors vous préférez regrouper le maximum de doublons possible ? Votre réglage devra alors être très ouvert.
Le mode « over-kill » est le mode généralement choisi dans des campagnes de prospection.

Le « just-kill »
Sous le vocable « just-kill » (apparu plus récemment), il faut comprendre un réglage de compromis.
Vous tenterez de limiter les vrais douteux tout en tentant de regrouper le maximum de doublons possible.

Tout ceci n'est évidemment pas un réglage à trois positions, mais plutôt un curseur que l'on positionne selon les objectifs attendus.

3.3.3 Exemples de données exploitées

Historiquement, un dédoublonnage (ou une déduplication) s'appuie sur les informations postales de vos contacts.

Si bon nombre d'utilisateurs se contentent de cette approche, aujourd'hui les outils permettent d'exploiter un grand nombre de données présentes dans vos bases.

- **Données postales**

  Dans tous les cas, une RNVP préalable de vos données est nécessaire. En effet, plus les adresses seront normalisées, plus elles seront comparables.

  La performance des outils doit permettre de comparer vos contacts même si les éléments adresse sont incomplets, ou si des éléments sont inversés.

Madame Marie ROBERT	*Monsieur Robert MARIE*
Residence du Peuple Belge	*12 Avenue du Peuple Belge*
59800 LILLE	*59000 LILLE*

  Selon l'objectif attendu, la tolérance des outils sur les éléments absents ou inversés sera plus ou moins grande.

  Il est probable que si vos attentes sont de nettoyer votre base clients, vous souhaiterez éviter de dédoublonner à tort ! (ce qui reviendrait à supprimer un client de votre base).
  Si, par contre, l'objectif est de dédupliquer des listes de prospection, vous souhaiterez probablement une tolérance plus grande : afin d'éviter d'envoyer des messages en double aux mêmes personnes ; ce qui permettra également d'augmenter l'efficacité des repoussoirs.

- **Données email (ou téléphone, siren, etc.)**

  Selon les bases de données, le taux d'alimentation des emails (ou autre) est variable. De la même manière la fiabilité de ceux-ci peut être variable.
  Plusieurs de vos contacts peuvent-ils utiliser le même email ?

  Selon la configuration, les rapprochements devront croiser avec plus ou moins de tolérance l'email avec le prénom ou le nom (ou un autre élément).

  Attention, les règles d'intégrité associée à votre SI peuvent **imposer une unicité d'e-mail** pour un contact, ou **interdire deux contacts derrière le même e-mail** (assez fréquent sur les environnements e-commerce par exemple).

- **Les informations de filiation**

  La gestion des enfants (ou parents) pour faciliter les rapprochements est très dépendante de la complétude des informations.
  (Le nombre d'enfants est-il fiable ? La base conserve-t-elle les dates de naissance ou les âges ?)

  Si la fiabilité des informations est suffisante, alors, intégrer les informations de filiation peut également vous permettre de retrouver des doublons.

  Attention toutefois à respecter le principe de minimisation des données requis par le RGPD en ne conservant que les données strictement nécessaires à la finalité du traitement.

- **Différentes passes ?**

  Nous comprenons que différents rapprochements peuvent être utilisés pour retrouver les doublons.
  Par exemple, une première passe sur l'adresse postale, une seconde intégrant les emails.

  Par ces mécaniques, il est fréquent de rapprocher deux entrées que nous n'aurions pas trouvées uniquement par l'adresse postale.
  Les déménagements ou les changements de noms maritaux peuvent ainsi être retrouvés.

  Attention néanmoins, la plus grande prudence sera nécessaire s'agissant des couples recomposés.

3.3.4 Quel est le contact maître ? les règles de priorité

La question du choix du premier de groupe n'a d'intérêt que sur des problématiques de nettoyage de base de contacts ou d'opérations de prospection.

Vous devrez faire un choix sur le contact à conserver.

Nettoyage d'une base de contacts :

→ Adresse la mieux remplie ?
→ Contact le plus récent ?
→ Selon le segment ?
→ Selon la présence des téléphones et emails ?
→ Autre ?

Nettoyage d'une base de prospection :

→ Selon le plan fichier ?
→ Selon l'alimentation du prénom ? (afin de personnaliser la communication)
→ Selon le sexe ou l'âge probable du contact ?

3.4 Les différentes méthodes

Si, dans un process de rapprochement de données, les outils sont importants, la capacité de l'opérateur à maîtriser cet outil et son expertise du domaine le sont tout autant.
Sa pédagogie pourra vous fournir un indicateur de sa maîtrise.

La qualité attendue est dépendante de l'objet du rapprochement (« louper » quelques doublons est moins grave pour une prospection que s'il s'agit de nettoyer votre base client).

3.4.1 Méthode « simpliste »

Peut-être avez-vous déjà rencontré un collaborateur ou un prestataire proposant de dédoublonner votre base sous un tableur ?
« mais si Monsieur ! une formule de concaténation sur le CP et les premiers caractères du nom ! et hop ».
(Ah, si la vie pouvait être simple…)

Sauf à ce que vous souhaitiez simplement supprimer des emails en double avant l'envoi d'une communication, une déduplication ne peut se résumer à des comparaisons strictes.
Celle-ci s'appuie sur des algorithmes de prise de décision complexes sur base de probabilité de rapprochements.

3.4.2 Approche phonétique

« Monsieur Lefevre, 12 Rue des Cols Verts » versus « Monsieur Leffebvre, 12 rue des Colverts ».
Cet exemple nous montre que des personnes peuvent être différentes malgré une approche phonétique strictement identique.

A contrario, les écarts phonétiques sont très courants et s'ils peuvent sembler simples à gérer cela représente un réel casse-tête.

Comment prononcez-vous les « …tion » de fin de mots ? (« rue de la Convention » / « Montée du Bastion »),
les « ch » de début de mots (« place du Château » / « route du Chaos »),
ou encore les « t » de fin de mots (« rue d'Alembert » / « rue du 25 Août »).

Dans chacun de ces exemples, les prononciations diffèrent ! Le contexte et la sémantique interviennent dans cette gestion.

3.4.3 Approche typographique

« Madame Martine Michelon, 12 Rue Près Noris » versus « Madame Micjelon, , 12 Rue Près Noirs ».

Les erreurs typographiques (ou fautes de frappe) correspondent à :

→ 1 caractère manquant ;
→ 1 caractère doublé ;
→ Une inversion de deux caractères ;
→ Ou encore le remplacement d'un caractère par un autre (souvent deux touches proches sur le clavier).

Les algorithmes restituent des distances entre les mots (c'est-à-dire un nombre de caractères d'écart).
Les difficultés ne résident pas tant dans le calcul des distances (il existe des méthodes éprouvées pour ce calcul), mais plus dans le nombre de comparaisons à réaliser : en effet, ne sachant pas par avance si l'erreur portera sur le premier caractère d'un mot, ou sur un autre, pour chercher le bon doublon, il sera nécessaire (pour chaque entrée) de comparer toutes les autres entrées de la commune.
Certains peuvent avoir des astuces pour réduire le cercle de comparaison, mais le problème est donc ici un problème de temps de réponse !

3.4.4 Approches mixtes

Chaque méthode ayant son lot d'inconvénients (approximation phonétique erronée « Allée des Ailes » « Allée des Halles », distance (même faible) sur mots très courts « grande rue »/ « grande roue », les meilleurs outils utilisent des méthodes mixtes (recherches phonétiques, puis calcul de distance, puis confrontation des résultats).

3.5 Les dangers, les craintes et les fausses idées

3.5.1 Subjectivité de l'analyse

Sur base d'échantillons de « douteux », nous constatons des différences de jugement sur plus de 30 % des cas présentés entre les différents participants.

Évidemment, un échantillon aléatoire (plutôt qu'un échantillon parmi des douteux) génèrerait une plus grande homogénéité d'avis. Néanmoins, ce petit exercice prouve qu'il n'existe pas de vérité sur la décision à prendre sur les douteux.

La décision de réglage « over » ou « under-kill » pourra reposer sur des raisons qualitatives, ou encore sur des raisons commerciales (ROI)

3.5.2 Les gros volumes génèrent un effet « over-killing »

« Over-kill » (voir également le glossaire)
Un rapprochement (dédoublonnage ou déduplication) n'étant pas une science exacte, il aboutira à un lot de doublons dont quelques uns seront « certains », d'autres « douteux ».
Avoir un réglage « over-kill » revient à réduire les tolérances de comparaison, et donc réduire le nombre de doublons douteux (au risque de ne pas détecter certains vrais doublons).
A contrario, un réglage « under-kill » revient à augmenter les tolérances et donc à augmenter le nombre de doublons douteux (au risque d'obtenir des faux doublons).

Prenons l'exemple d'une « petite » déduplication, où « Madame Juliette Michu » sera comparée à 100 autres contacts (sur la même commune par exemple). Prenons également le cas d'une « grosse » déduplication où cette même dame sera comparée à 5 000 autres contacts.

Puisque, dans le second exemple, « Madame Michu subira 50 fois plus de comparaisons que dans le premier, nous pouvons supposer que la probabilité de trouver un doublon (vrai doublon ou faux doublon) est 50 fois plus importante.

C'est pour cette raison mécanique que la tendance à « l'over-killing » est beaucoup plus importante sur les grosses bases. Attention donc à adapter les seuils souhaités de déduplication.

3.5.3 Les gros volumes génèrent un effet « transitif ».

Joseph Michu 12 rue des Alouettes 59273 FRETIN	**J** Michu 12 rue des Alouettes 59273 FRETIN	Monsieur **Julien** Michu 12 rue des Alouettes 59273 FRETIN

L'exemple ci-dessus met en évidence les dérives dues à la transitivité des rapprochements.

Le danger de la transitivité est évidemment bien plus important sur les déduplications faisant intervenir de très gros volumes. En effet, plus le volume est important, plus le nombre de cas de transitivité augmente, et plus la longueur des chaînes transitives augmente également.

Ce point est loin d'être neutre. Au-delà d'un million d'enregistrements, vous pouvez être certain d'avoir des cas importants de transitivité.

Éviter ces cas est dépendant des outils.
A minima, il sera intéressant d'augmenter le seuil d'acceptation des doublons (avoir une tendance « under-kill ») avec le risque de casser le groupe en exemple ci-dessus.

3.5.4 Fusion des historiques

Outre les problématiques de RNVP et de dédoublonnage, un nettoyage de base de contacts pose également le problème de la réintégration des données.

En effet, la réintégration suppose :

- Soit de supprimer les contacts en double, ainsi que de fusionner l'ensemble des historiques pour rattacher les achats / dons ou autres au contact maître.
- Soit de gérer une nouvelle table des liens doublons et donc de retravailler l'ensemble de l'applicatif afin d'effectuer une consolidation des données dynamiquement (ma table des doublons indique que mon contact n'est pas le maître, donc j'associe l'ensemble des données qui lui sont associées au contact maître).

Dans les deux cas, la réintégration des données devra avoir été anticipée afin d'éviter les mauvaises surprises de faisabilité ou de coûts. De plus, pour limiter les risques de non conformité aux exigences du RGPD (voir paragraphe "La déontologie"), le DPO devra être consulté en amont du projet.

3.6 Prospective

3.6.1 Online

Les recherches de préexistence sont largement répandues, mais…

Elles n'utilisent que très peu la profondeur des algorithmes éprouvés et utilisés en mode batch.
La raison est double :

→ Il est admis d'attendre d'un opérateur, pour que sa recherche aboutisse, qu'il fasse plusieurs essais afin d'arriver au résultat (première recherche par le nom, puis par le CP, etc.).
→ Les technologies éprouvées actuelles, pour beaucoup, n'ont pas des temps de réponses compatibles avec des contraintes de l'online.

Cependant, les technologies évoluent.

Il sera probable que demain, derrière la recherche suivante :
« Mme Dujardin, Lille », une recherche puisse proposer le contact « Jacqueline Michelon, Fretin », simplement parce que la sphère de données comporte des éléments permettant l'association (par exemple, la connaissance d'un conjoint faisant la liaison avec un nom marital, et la connaissance d'un déménagement).

3.6.2 Ouverture de la sphère de comparaison

Les technologies dites « distribuées » permettent déjà, à travers l'utilisation d'un grand nombre de serveurs en parallèles, la multiplication des comparaisons sans augmenter les temps de réponses.

Pour autant, les outils de distribution supposent un travail par index, et nécessitent une simplification des process de recherche.

Cependant, les technologies continuent d'évoluer et il est certain que, dans un avenir proche, dédupliquer plusieurs centaines de millions d'entrées ne prendra qu'une poignée de secondes, tout en ayant la performance des meilleurs outils actuels.

3.7 Rapprochements : les questions à se poser

Voici quelques exemples qui permettront de se faire une idée de la définition souhaitée d'un doublon.

Attention, dans tous les cas, il sera nécessaire de vérifier, sur une liste de rapprochements, si les choix n'apportent pas des effets indésirables (souvent liés à la transitivité).

Infos nominatives	M Robert MICHU 12 rue ...	Mme Juliette Michu 12 rue ...	Un rapprochement au foyer ? à l'individu ?
	M Robert MARIE 12 rue ...	Mme Marie ROBERT 12 rue ...	Les inversions prénom nom sont fréquents, mais... que faire si les éléments postaux font également apparaître un doute ?
	Mme Michu 12 rue ...	Mme Juliette Michu 12 rue ...	Un prénom absent est-il à rapprocher d'un prénom alimenté ? *Attention, risque de transitivité !*
	Monsieur Michu 12 rue ...	Mme Juliette Michu 12 rue ...	Souhaitez-vous tenir compte du sexe supposé ?
Infos postales	Mme Juliette Michu 12 rue des Alouettes 59273 FRETIN 12 rue des Alouettes 59273 FRETIN	Mme Juliette Michu rue des Alouettes 59273 FRETIN 132 rue des Alouettes 59273 FRETIN	Tolérance sur le n° ? La réponse ici peut dépendre de la taille de la commune (ou mieux... du nombre de points de distribution dans la rue !) Un numéro absent ? 12 versus 132 ? (faute de frappe ?)
	12 rue des Alouettes 59273 FRETIN	Porte 12 Res les alouettes 59273 FRETIN	Des éléments autres que la ligne « adresse » doivent-ils être pris en compte ?
	M Robert MARIE 12 rue des Alouettes 59273 FRETIN	M Robert MARIE 59273 FRETIN	Une voie absente est-elle à rapprocher d'une voie présente ? Peut-être la réponse est-elle différente selon que le contact est à Lille ou à Auchy-lez-Orchies (grosse ou petite commune)
	M Robert MARIE 59273 FRETIN	M Robert MARIE 59273 FRETIN	Les deux adresses ici sont parfaitement identiques, mais les considèreriez-vous comme doublons ? (et si plutôt que Fretin le contact était dans une grosse ville ?)
Infos e-mail / tel / autre	Mme Juliette Michu 12 rue des Alouettes 59273 FRETIN j.michu@free.fr	Mme Juliette Michu 17 place de l Eglise 59310 Auchy lez Orchies j.michu@free.fr	Il est probable que vous souhaiterez intégrer les cas de déménagement. La récence (ou un segment) pourra alors être prise en compte pour choisir l'adresse à conserver
	Mme Juliette Michu tribumichu@free.fr	Robert Michu tribumichu@free.fr	Evidemment, les éléments email ou tel devront être croisés avec d'autres éléments. Il faudra déterminer lesquels conserver, et le niveau de tolérance que vous souhaiterez ! (prénom ? nom ? les 2 ?)

Les choix absolus n'existent évidemment pas. Un choix aura dans beaucoup de cas des effets induits indésirables. Néanmoins, ne pas se poser les questions ci-dessus peut s'avérer catastrophique sur la qualité des résultats.

4. L'enrichissement de données

4.1 Le « pourquoi »

Avant de répondre à la question du « pourquoi », il faut tenter de définir ce qu'est l'enrichissement de données.

La définition trouvée sur Wikipedia est la suivante :

> *« Les **enrichissements de données** permettent de procéder à une géo-segmentation de consommateur, offrant une grille de lecture sur de nombreuses appétences. Les enrichissements en centre d'appels peuvent par exemple permettre en temps réel de faire une première analyse du profil de l'interlocuteur et ainsi mieux adapter le message et l'offre. L'enrichissement de données participe à l'amélioration globale de la Relation Client des entreprises. »*

Cette dernière phrase bien que réductrice est la plus juste de cette définition et nous préférons la définition suivante :

L'enrichissement consiste en l'adjonction d'informations manquantes ou complémentaires d'origine interne ou externe.

Chacun trouvera la réponse à SON « pourquoi » selon ses objectifs, sa stratégie et ses enjeux.

L'enrichissement de données permet :

- D'augmenter et d'affiner les clés de déduplication (par exemple, une date de naissance ou un numéro de téléphone mobile qui permettront de rapprocher ou séparer des doublons) ;
- De disposer d'un potentiel d'analyse interne plus pertinent et puissant (comparer les paniers moyens avec le niveau de revenu) ;
- De comparer une population cible à une population complémentaire ou plus vaste pour identifier des forces et des faiblesses ;
- De s'inscrire dans une relation multicanale avec ses clients ;
- etc.

Vous trouverez quelques exemples de bénéfices tirés de l'enrichissement :

- Connaître le client pour un marketing plus efficient ;
- Cartographier ses clients pour quadriller un territoire ;
- Optimiser son ROI en utilisant le meilleur canal de contact ;
- Personnaliser les offres selon la cible et les médias utilisés.

En listant vos besoins et vos projets d'actions vous pourrez déterminer les données nécessaires. Ces données seront soit indispensables à la réalisation, soit permettront d'optimiser vos actions. Il vous faudra garder en permanence à l'esprit le bénéfice attendu et obtenu.

4.2 Les alternatives

La seule alternative est de ne pas pratiquer d'enrichissement.

Plutôt qu'une alternative, la question devient : pourquoi enrichir ? ou pourquoi pas ?

- Impossibilité technique ;
- Pas de budget ou ROI incertain ;
- Les données nécessaires n'existent pas ;
- Pas le temps ni les compétences ;
- C'est un coup d'épée dans l'eau ;
- etc.

Alors comment faire autrement, dès lors qu'un besoin a été identifié ?
Contourner l'apport d'informations en masse par une collecte à l'unité, comme par exemple faire participer le client à la complétude des informations de son compte en ligne.
Récolter des informations via le SAV ou le service client.

Cela reste de l'enrichissement mais au fil de l'eau…

4.3 Le « comment »

L'enrichissement peut se réaliser pour un fichier BtoC comme pour un fichier BtoB.
Les logiques restent les mêmes et la nature des données et des points de contact diffère. Ici, nous ne traiterons que de données BtoC.
Dans cette partie, vous pourrez suivre le déroulé logique de la mise en place d'une action d'enrichissement, qu'elle soit massive et ponctuelle ou réalisée petit à petit.

Pour commencer vous pouvez mettre en œuvre ce projet seul ou vous faire accompagner aux différentes étapes comme durant tout le process.

Après la phase d'étude et de cadrage du besoin ou des objectifs, les principales étapes sont :

- La faisabilité ;
- L'identification des données nécessaires ;
- Le sourcing ;
- La qualité des différentes sources ;
- L'usage des données ;
- Le ou les bénéfices constatés et avérés des traitements mis en place.

4.3.1 La faisabilité

L'enrichissement de données revient à confronter une base à enrichir avec une base de référence (voir le chapitre 6 « Les référentiels »).

Le socle technique de vos données est-il ouvert, adaptable et prêt à recevoir le fruit de cet enrichissement ? Devra-t-il être interprété, confronté à d'autres données ? Pourrez-vous exploiter ces données brutes ou faudra-t-il un retraitement ?
Vos données sont-elles exportables simplement ? ou un traitement de construction des données avant export est-il nécessaire en interne ?

Une fois cette étape validée, le second point crucial réside dans la qualité des clés de rapprochement pour obtenir le meilleur taux de communs avec la base de référence.
Les différentes clés de rapprochement, comme vu plus haut, doivent être de bonne qualité pour un bon taux de rapprochement : nom prénom, adresse postale, date de naissance, numéros de téléphone, adresse mail, Id, code Iris, id carreau, coordonnées XY Lambert, id cookies, id IOT, etc.

Même si chaque clé permet un travail plus ou moins précis sur la majeure partie de la volumétrie, c'est bien l'utilisation combinée de plusieurs clés qui permet de gagner en précision et de lever les doutes sur les faux doublons. Prenons l'exemple de Monsieur Jean Dupont à Paris 14. Son adresse postale permet déjà de le distinguer parmi les X Jean Dupont de l'arrondissement. Il peut encore y avoir plusieurs Jean Dupont dans un même immeuble. Sa date de naissance ou son numéro de téléphone permettront de savoir précisément de quel Jean Dupont nous parlons.

Essayez de partager des expériences avec des confrères, cela vous permettra de vous situer sur la faisabilité et d'anticiper les éventuelles difficultés.

4.3.2 L'identification des données nécessaires

Historiquement, l'enrichissement client consiste à intégrer des données complémentaires dans une base de données, comme un CRM, une base de données marketing (BDM) ou plus récemment une Data Management Plateform (DMP).

Avec l'essor du numérique, des multiples connexions et du collaboratif, de nouveaux modes d'enrichissement arrivent :

- L'utilisation de données par un prestataire qui ne vous restitue que le fruit de l'analyse basée sur ces données ;
- Le traitement d'un fichier de cibles par une segmentation plus ou moins volatile comme dans le programmatique ;
- L'utilisation d'Internet pour capter des données.

C'est souvent l'usage et la pérennité qui changent mais la finalité reste toujours la même.

Les avancées technologiques permettent des usages de plus en plus variés et nombreux qui ont le mérite de satisfaire les différents besoins et budgets.

Ces mêmes avancées technologiques augmentent le nombre, la diversité et la circulation de données disponibles tout en les rendant plus rapidement périmées, par les possibilités de mise à jour quasi-instantanées.
Il vous faudra commencer par la hiérarchisation des types de données et de leur apport dans votre stratégie.

Les premières données auxquelles vous penserez seront les données complémentaires de contact pour mieux communiquer avec vos cibles :

- L'e-mail appending qui enrichit votre fichier d'adresses postales des emails que vous n'avez pas ;
- Les numéros de téléphone (fixe, portable) pour joindre directement ou transmettre des messages vocaux ;
- Les adresses postales si vous n'avez que l'email (reverse appending).

Ensuite, pour commencer à analyser vos populations, les données de profil, sociodémographiques comme :

- La date de naissance, la composition du foyer, les niveaux de revenus, le type d'habitat.
- Ou simplement géographiques (densité d'habitat, zones de chalandises, trajets usuels...).

Une fois la population cernée dans son cadre de vie, vous pouvez avoir besoin de mieux comprendre et anticiper les comportements.
Vous allez chercher les centres d'intérêt et les habitudes de consommation :

- Les données comportementales ou transactionnelles on et offline (achats en magasin, VAD/e-commerce, collecte de fonds, centres d'intérêts, canaux de commande, moyens de paiement, etc.) ;
- Les données de navigation pour compléter les centres d'intérêt ou les intentions ;
- Les données contextuelles (nuages de tag).

Et vous pouvez aussi partir à l'aventure de l'Open Data et du Big data.
Le champ est vaste, alors il faut raisonner en « entonnoir » et adapter la partie large de cet entonnoir à sa volumétrie interne et à son budget tout en gardant une analyse fine du ROI.

4.3.3 Le sourcing

Où et comment vous procurer ces données ?
Obtenir des données peut se pratiquer de deux manières :

- en interne ;
- et/ou en externe.

Il faudra donc identifier les différentes sources et comparer l'accès, les coûts (indirects aussi...), la qualité et les délais.
Il peut être plus simple d'aller chercher des adresses mails complémentaires chez un prestataire extérieur que dans les services voisins au sein de votre entreprise.

La principale justification est le time-to-market, le coût extérieur étant compensé par l'agilité du partenaire.
Un autre frein est l'incapacité à assurer, en interne, le travail technique de rapprochement des deux bases.

4.3.3.1 Les données internes

En interne vous pourrez trouver des données dans d'autres bases (un ERP, une base Web, ou tout autre SI qui capte, gère ou analyse des données).
Vous pourrez aussi capter des données pendant un traitement ou pendant une action.
Ces données pourront venir au fil de l'eau (informations récoltées dans le compte client, données issues d'enquêtes, saisie d'informations collectées dans les tickets de caisse, par le service après-vente, etc.).

Attention toujours à bien analyser le contexte de réponse pour s'assurer que ces données sont bien le reflet de la réalité et non des informations erronées car transmises en réaction à une situation (fausse déclaration intentionnelle dans un contexte de litige par exemple).

4.3.3.2 Les données externes

Il existe de nombreuses sources de données externes.
Adoptons la nomenclature des éditeurs de DMP ou CDP (Customer Data Platform).

- Vos données internes constituent la 1st part.
- Les données externes sont distinguées en 2nd et 3rd part.

Dans la « Second part », vous trouverez des données chez des partenaires comme un réseau de revendeurs, un fournisseur, un client ou un confrère avec lequel vous décidez d'échanger. En général, dans ce cas, il n'y a pas de notion mercantile autour de la donnée. Il faut seulement garder à l'esprit que tout traitement a un coût (visible ou masqué) qui est bien réel et peut s'envoler si le sujet n'est pas sous contrôle.
Il faut aussi se poser la question de la pérennité de cette source ainsi que du maintien de sa qualité, si la nature de la donnée est plus ou moins périssable.

Le nouveau règlement européen vient renforcer le droit d'information des personnes. Les partenaires seront vigilants et devront informer les personnes de la possibilité d'une communication à des tiers lors de la collecte, et collecter un opt-in s'il s'agit de canaux soumis au consentement (email et SMS en BtoC) (voir chapitre 5 - La déontologie).

La « Third part » est composée des fournisseurs de données dont c'est l'activité principale.
Les données disponibles sont issues de fournisseurs historiques comme :

- les opérateurs téléphoniques ;
- La Poste ;
- les éditeurs de mégabases ;
- l'Insee ; et de plus en plus,
- les données gratuites fournies en open-data.

Ces deux dernières catégories nécessitent d'être capable d'intégrer et d'interpréter ces données pour les rendre « comestibles ».
Certains acteurs savent aussi massifier et combiner des sources de type 2nd, 3rd, Insee et Open data. Ils pourront vous livrer des solutions personnalisées à vos enjeux.

4.3.4 La qualité des différentes sources

La première étape consiste, en partant des clés dont la qualité aura été validée, à procéder à une étude statistique. Quel volume et quel pourcentage de correspondance ?

- Est-ce la première fois ou est-ce pour un complément ? Dans ce dernier cas, il faudra s'attendre à trouver moins d'informations que lors de la première action.
- La complétude absolue est rarement possible. La raison en est simple, il reste toujours des individus qui souhaitent communiquer moins d'informations que d'autres.

Il est conseillé de réaliser un POC (proof of concept) pour tester les sources sur un échantillon avant de se lancer dans des traitements volumineux.
Cela permet aussi de valider la bonne intégration et exploitation dans votre BDD.

La qualité n'est pas à négliger pour les raisons suivantes :

- Vous allez exploiter ces données pour traiter ce que vous avez de plus précieux.
- Revenir en arrière sur une intégration de mauvaises données peut s'avérer plus coûteux que le traitement initial.

Parmi les différentes sources d'enrichissements il faut aussi appréhender la structuration des données : des données non ou mal structurées ne peuvent produire qu'un enrichissement médiocre.

Pour valider la qualité, ne pas hésiter à demander des références, des cas d'usage, de la transparence sur la provenance des données et sur la politique en matière de protection des données.

Vous pouvez aussi mettre des réserves juridiques et financières.

4.3.5 L'usage des données

Les données récoltées doivent être exploitables et vous devez en obtenir l'autorisation d'usage via le fournisseur, que ce soit un intermédiaire, un « propriétaire » ou même le consommateur lui-même.
Gardez à l'esprit cette image sur vos données : sont-elles chaudes, tièdes ou froides ?
Cela vous aidera à anticiper les besoins de mise à jour ou la péremption éventuelle de chaque donnée.
De même, le traitement initial d'enrichissement doit conduire à se poser deux questions :

- Les données sont-elles périssables ou évolutives ?
- Est-il possible de disposer de mise à jour de ces données, sous quelle fréquence et à quelles conditions ? Ai-je un besoin ponctuel ou permanent en regard de mon retour sur investissement ?

Dans votre exploitation de ces données, vous allez les stocker, peut-être les louer ou les échanger. Anticipez aussi cet usage dans votre démarche.
Vous devez être en mesure de suivre et de tracer vos données :

- D'où viennent-elles ?
- Quelle utilisation vais-je en faire ?
- Combien de temps puis-je les conserver et les exploiter ?
- Est-il possible de répondre à une demande de renseignement ou d'opposition à leur utilisation ?

C'est un périmètre complet à intégrer dans votre organisation.

Voir le chapitre 5 - La déontologie.

4.3.6 Le ou les bénéfices constatés et avérés des traitements mis en place

Économiquement et rationnellement, toute action qui représente un coût doit être justifiée par le bénéfice qu'elle produit.
Dans le cadre de la collecte de données sur des individus, le volet règlementaire s'invite aussi dans la perspective de la protection définie par les nouvelles dispositions européennes.

Les textes du RGPD (ou GDPR en anglais), tout comme la loi Informatique & Libertés précédemment, reconnaissent l'intérêt légitime du responsable de traitement ou d'un tiers.
Concrètement, collecter des informations sur la consommation d'eau d'un ménage est légitime si cela vous permet d'adapter une offre basée sur cette analyse. Il faudra toutefois que ledit ménage soit informé de votre action et bénéficie de la possibilité de s'y opposer.

Même si la contrainte légale existe, elle ne bride pas la créativité dès lors que les intentions sont bonnes, transparentes, respectueuses et justifiables.

Il vous reste à vous concentrer sur la maîtrise de vos enjeux et l'analyse de votre performance.

- Disposer de toutes les coordonnées pour contacter un client doit se ressentir en termes de fidélité, de valeur et de qualité d'échange.
- Disposer de ses revenus ou de données financières pour mieux cerner son profil de consommateur doit vous permettre d'adapter vos montants moyens de propositions ainsi qu'éventuellement les conditions ou facilités de paiement (3 fois sans frais, par exemple).

Si vous n'adaptez pas vos sollicitations et n'en tirez pas de bénéfices, vous pourrez penser légitimement que l'enrichissement ne sert à rien.
L'enrichissement n'est donc pas une fin en soi, mais participe à l'évolution du mix marketing.

Vous entendrez que c'est la donnée qui pilote le marketing : "data drive marketing". Notre positionnement est plus nuancé : nous croyons au "data *driven* marketing", c'est-à-dire au marketing enrichi et amélioré par le traitement efficace de la masse de données disponibles et exploitables.

4.4 Les différentes approches

Comme dit plus haut il est possible de récolter des données au fil de l'eau durant les différentes interactions avec un client : par sa saisie personnelle dans son espace client en ligne, par la saisie d'un téléopérateur lors d'un appel SAV ou d'une sollicitation commerciale, par l'interrogation lors d'un passage en caisse, etc.

Nous allons traiter ici les enrichissements « massifs », donc volumineux sur une grande quantité de contacts en même temps.
Deux approches s'ouvrent à vous :

- les données embarquées ;
- les données utilisées temporairement.

Ces deux approches ne s'opposent pas particulièrement, elles répondent à des logiques, organisations et enjeux différents et/ou complémentaires.
On peut aussi parler de données chaudes ou froides, comme déjà exposé.

4.4.1 Les données embarquées

Ce sont les données intégrées dans une BDM, un CRM, une DMP ou toute base qui regroupe des informations sur des clients.
Il s'agit de données plus ou moins périssables :

- L'âge, par la date de naissance ou par le score sur les prénoms sera à traiter de manière différente.
- Le niveau de revenu, le nombre d'enfants, le type d'habitat évoluent plus rapidement, surtout dans les tranches d'âges plus jeunes.
- Les centres d'intérêts, les paniers moyens et enfin la navigation sont beaucoup plus volatiles et nécessitent, outre une mesure de récence, un véritable process de gestion dans le temps.

Plus vous souhaiterez intégrer de données variées, plus leur gestion deviendra complexe.

Les différentes bases peuvent faire évoluer leur « modèle de données », plus ou moins facilement et rapidement. Il n'est donc pas de frein à intégrer, par exemple, des dates de naissance, des tranches de revenus et des analyses de montants moyens basées sur les X dernières transactions.

En revanche, il devient compliqué de changer régulièrement la nature et la cohabitation des données complémentaires car une perte d'historique peut devenir préjudiciable.
Devrez-vous tout intégrer dans vos bases ?
Une partie de la solution peut se trouver dans le paragraphe suivant.

4.4.2 Les données utilisées temporairement

Vous pourrez être amené à fournir des données complémentaires à votre prestataire ou à vos services de datamining. Ces données ne serviront qu'à des calculs et ne vous seront pas utiles dans votre base de données. Il est d'usage chez certains fournisseurs d'adapter la tarification à cet usage à fin d'étude.
Dans le même esprit, vous ne pourrez enrichir votre base qu'avec le fruit de l'analyse de ces données (une échelle de valeur de potentiel d'achat ou de fidélisation, par exemple).

Vous pourrez aussi utiliser des données « à la volée » sans les conserver, comme, par exemple, pour l'adaptation d'un message selon l'âge du destinataire, mais sans embarquer la date de naissance dans une base de prospection ou de fidélisation.
Ces données servent à qualifier une personne ou un segment d'audience pour l'activation d'une campagne digitale, sur un genre (femme ou homme), une tranche d'âge, un pouvoir d'achat, etc.

4.5 Les dangers, les craintes et les fausses idées

4.5.1 La profondeur des données

Le premier danger est de croire à l'exhaustivité d'un référentiel.

Par exemple, les fichiers Estocade ou Charade ne regroupent pas l'intégralité des foyers qui déménagent mais seulement ceux qui souscrivent un contrat de réexpédition avec La Poste.
D'une manière générale, il ne faut pas vous attendre à trouver 100 % de votre besoin en matière d'enrichissement. La multiplicité de sources peut vous aider à vous approcher de l'exhaustivité.

Gardez à l'esprit que les données les plus rares sont les plus chères et aussi les plus risquées. La multiplicité des sources reste une garantie de qualité de la donnée.

4.5.2 L'usage des données

Sortir ses données pour les confronter à des référentiels ou des mégabases représente un risque puisqu'elles vont :

- Voyager ;
- Être temporairement hébergées par un tiers ;
- Être manipulées, etc.

Vous devez vous assurer de toutes les bonnes pratiques en matière de sécurité et de confidentialité avant tout traitement (voir paragraphe "La déontologie").
La pratique de l'adresse piège à intégrer dans votre fichier dès son extraction est aussi une bonne manière de contrôler l'usage éventuel ou "accidentel"…

Concernant l'usage des données, il ne faut pas oublier la notion de pérennité des données, que ce soit dans un esprit de fraîcheur ou de limite dans le temps de leur droit d'utilisation (durée de conservation).

4.6 Prospectives

En parallèle d'un cadre réglementaire qui protège mieux le consommateur, les technologies vont permettre une intensification de l'usage des données.
Prenons pour structure les 4 V du big data inventés par IBM : volume, vélocité, variété et véracité.

De plus en plus de données disponibles seront utilisées, traitées et échangées. La circulation des données tendra vers le temps réel et vers l'usage "à la volée". Il ne sera plus nécessaire de tout embarquer dans une base. Il faudra gérer les liens pour accéder à la donnée qui sera de plus en plus fiable et "fraîche".
Cette accélération permettra aussi de qualifier de mieux en mieux et les pratiques insuffisantes seront amenées à disparaître.

Un risque sera la perte du standard et la multiplicité des données non structurées. Ce manque de structure pourra influencer l'interprétation et la perception de l'analyse des données.

Nous assisterons très certainement à une forme d'Uberisation du marché des données. Il est possible d'imaginer une certaine maîtrise du consommateur sur l'usage qu'il accepte de ses données personnelles, selon les marchés, les canaux et les offres.

Les objets connectés et les données qu'ils produiront ou collecteront vont participer à la finesse d'analyse des comportements de consommation.
Il faut s'attendre à une croissance exponentielle.

Enfin, l'intelligence artificielle et le machine learning trouvent leur carburant dans les données, pour produire aussi des données.

Allons-nous assister à une "multiplication des pains" ?

5. La déontologie

5.1 Préambule

Le nouveau règlement européen sur la protection des données (adopté en avril 2016, entré en application en mai 2018) remet, dans beaucoup d'entreprises, le dossier déontologique sur le dessus de la pile. Ce règlement s'applique à toute entreprise stockant ou utilisant des données à caractère personnel (que ce soit des données BtoC ou BtoB, que les traitements soient automatisés ou manuels).
S'imprégner de ce règlement semble un minimum indispensable.

Définition

La déontologie est « l'ensemble des règles et des devoirs qui régissent une profession, la conduite de ceux qui l'exercent, les rapports entre ceux-ci et leurs clients et le public » (cf. dictionnaire Larousse).

Il faut comprendre qu'il y a une double profondeur derrière les règles de déontologie :

1. une profondeur législative (les obligations légales sur la sécurité, la protection des données et le respect des individus salariés, clients et fournisseurs) ;
2. mais également une profondeur plus philosophique concernant sa propre éthique (ou le point de vue d'une entreprise, ou syndicat) sur les bonnes pratiques.

La Charte du Sncd se range dans cette seconde catégorie.

Réserves

Ce paragraphe constitue une synthèse des usages en vigueur. Le lecteur pourra se référer aux publications du Sncd pour une information plus complète, ou encore se référer aux derniers travaux de la CNIL.
Évidemment, détenir un numéro de déclaration CNIL n'est pas suffisant pour être en conformité avec la législation, ni avec une démarche déontologique.
Pour cette raison, le RGPD introduit la notion « d'Accountabilty » (responsabilisation de l'entreprise) et a supprimé la plupart des formalités auprès de la CNIL au profit du registre des traitements du responsable de traitement et du sous-traitant.

5.2 Pourquoi adhérer à une déontologie ?

Afficher l'adhésion de son entreprise à une éthique déontologique est intéressant à plusieurs titres :

- La morale d'une entreprise est de plus en plus intégrée par nos clients, lors de la recherche de leurs fournisseurs. Cette tendance vertueuse permet naturellement de limiter certaines dérives.
- Elle permet de limiter la dégradation de l'image du marketing client perçue par le grand public.
- Une démarche d'éthique n'est plus le moteur unique, en effet, la législation devient contraignante, et les pénalités sévères !

Attention, une déontologie n'est plus aujourd'hui suffisante, la législation (et en particulier le nouveau règlement européen) encadre les us et coutumes.
L'objectif de ce nouveau règlement est de « redonner aux citoyens le contrôle de leurs données personnelles, tout en harmonisant l'environnement réglementaire des entreprises ».

5.3 Règlement européen sur la protection des données personnelles

Le règlement a été définitivement adopté par le Parlement européen le 14 avril 2016, et entre en application le 25 mai 2018 dans les 28 Etats membres.

Vis-à-vis des principes prévus par la directive de 1995, le règlement prévoit un élargissement du périmètre d'application en matière de protection des données et une échelle géographique plus englobante.

5.3.1 Responsabilité partagée

La réglementation tend à partager la responsabilité entre :

- le responsable du traitement ;
- le sous-traitant ;

(Que nous appellerons, dans ce paragraphe : **les ACTEURS**).

5.3.2 Le périmètre d'application du règlement

Le RGPD s'applique dès lors que :

→ Un des ACTEURS (responsable de traitement ou sous-traitant) est établi sur le territoire européen ;
→ Un des ACTEURS fournit un bien ou un service à un résident européen ou suit son comportement sur internet.

5.3.3 Le DPO (délégué à la protection des données)

Le responsable de traitement et le sous-traitant peuvent (ou doivent, voir ci-après) nommer un Délégué à la Protection des données.

Son rôle est de :

- Informer et conseiller les ACTEURS ;
- Contrôler le respect de la réglementation ;
- Coopérer avec les autorités de contrôle ;
- Être le point de contact entre l'ACTEUR et les personnes concernées.

Le DPO est dans certaines conditions facultatif, mais recommandé. Il est obligatoire pour :

- les administrations ;
- le traitement de données sensibles ;
- le suivi systématique et régulier à grand échelle des personnes.

5.3.4 Le renforcement des droits des personnes

Le droit à l'information

Le RGPD renforce l'obligation d'information des ACTEURS à l'égard des individus auprès desquels ils collectent des données, dans le cadre d'une transparence accrue des traitements réalisés.
L'individu doit être informé au moment de la collecte des données ou lors de la première communication des données. Cette information doit être facilement accessible et aisément compréhensible.
Le nombre d'informations à fournir aux individus lors de la collecte augmente. Les politiques de confidentialité doivent être tenues à jour et faciles à trouver.

Le droit d'accès

La personne concernée a le droit d'obtenir du responsable du traitement la confirmation que des données à caractère personnel la concernant sont ou ne sont pas traitées et, lorsqu'elles le sont, l'accès auxdites données.

Le droit à l'effacement (« droit à l'oubli »)

Le règlement confère le droit à l'individu de demander l'effacement de ses données dans certaines circonstances.
En marketing, le droit à l'effacement doit s'articuler avec le droit d'opposition à la prospection (voir ci dessous "Le droit d'opposition à la prospection (opt-out)"). En effet, si une personne demande à une entreprise d'effacer les données qui la concernent parce qu'elle ne souhaite plus recevoir de prospection commerciale de sa part, l'entreprise devrait intégrer les coordonnées de cette personne à sa liste d'opposition à la prospection et non les effacer. En agissant de la sorte, l'entreprise ne pourra plus utiliser les données de la personne à des fins de prospection. Si au contraire elle les effaçait, il y aurait alors un risque de nouvelle prospection de la personne dans le futur, ce que justement celle-ci ne souhaite pas.

Le droit à la limitation du traitement

La personne concernée a le droit d'obtenir du responsable du traitement la limitation du traitement lorsque l'un des éléments suivants s'applique :

- l'exactitude des données à caractère personnel est contestée par la personne concernée, pendant une durée permettant au responsable du traitement de vérifier l'exactitude des données ;
- le traitement est illicite et la personne concernée s'oppose à leur effacement et exige à la place la limitation de leur utilisation ;
- le responsable du traitement n'a plus besoin des données à caractère personnel aux fins du traitement mais celles-ci sont encore nécessaires à la personne concernée pour la constatation, l'exercice ou la défense de droits en justice ;
- la personne concernée s'est opposée au traitement, pendant la vérification portant sur le point de savoir si les motifs légitimes poursuivis par le responsable du traitement prévalent sur ceux de la personne concernée.

Le droit de rectification

La personne concernée a le droit d'obtenir du responsable du traitement, dans les meilleurs délais, la rectification des données à caractère personnel la concernant qui sont inexactes. Compte tenu des finalités du traitement, la personne concernée a le droit d'obtenir que les données à caractère personnel incomplètes soient complétées, y compris en fournissant une déclaration complémentaire.

La portabilité des données

Ce droit redonne aux personnes la maîtrise de leurs données obtenues dans le cadre d'un contrat ou sur la base d'un consentement. Toute personne pourra ainsi récupérer les données qui la concernent (et éventuellement, demander leur transfert à un tiers).

Le droit d'opposition à la prospection (opt-out)

Les individus ont le droit de s'opposer à tout moment au traitement de leurs données à des fins de prospection. Lorsqu'un individu exerce ce droit, le responsable de traitement ne peut plus traiter ses données à ces fins. Il s'assurera donc d'inclure les coordonnées de cet individu dans sa liste d'opposition.

Les entreprises ont l'obligation d'informer les individus de leur droit d'opposition au traitement de leurs données. Cette information doit être explicite et distincte des autres informations fournies.

Le droit d'opposition au profilage à des fins de prospection

Le règlement intègre des dispositions relatives au profilage, c'est-à-dire l'utilisation des données personnelles afin d'évaluer l'intérêt des personnes pour certains produits ou services, leur comportement ou leur position géographique.

Le règlement instaure un droit d'opposition au profilage à des fins de prospection.

Remarque : le rapprochement de données, pour des finalités de mise à jour de bases de données, peut être apparenté dans certains cas à du profilage.

5.3.5 Devoirs et responsabilités des acteurs

Minimisation des données

Les ACTEURS devront veiller à minimiser la quantité de données traitées, dès la conception d'un projet. Ainsi, seules les données nécessaires au regard de la finalité du traitement devront être collectées.

Sécurité des données

Les ACTEURS mettent en œuvre des mesures techniques et organisationnelles pour assurer la sécurité des données. La pseudonymisation et le chiffrement font partie des mesures qui réduisent les risques liés à la protection des données.

Pseudonymisation des données

On parle de pseudonymisation de données dès lors que celles-ci ne peuvent plus être rapportées à l'individu qu'elles concernent, sauf recours à des informations additionnelles.

Limitation de la conservation des données

Les données personnelles (ou a minima, les liens permettant d'identifier les personnes) ne doivent être conservées que le temps nécessaire à la poursuite de l'objectif pour lequel elles ont été collectées. Elles devront ensuite être supprimées (sans restauration possible) ou rendues anonymes, sauf pour archivage à des fins de preuve.

Les données (ou traitements) sensibles

Une donnée est dite sensible, dès lors qu'elle concerne :

- les origines raciales ou ethniques ;
- les opinions politiques, religieuses ou philosophiques ;
- une appartenance syndicale ;
- les données de santé ;
- l'orientation sexuelle ;

Et (plus récemment) avec le RGPD :

- les données génétiques ou biométriques.

Dès lors qu'un traitement fait apparaître des données sensibles à grand échelle, il est considéré comme susceptible d'engendrer un risque élevé pour les individus. Les ACTEURS ont alors l'obligation d'effectuer une étude d'impact préalablement à sa mise en place. Si l'analyse démontre un risque élevé pour les individus, les ACTEURS devront consulter les autorités de protection, qui pourront s'opposer au traitement. La même obligation s'impose dans le cas de profilage produisant des effets juridiques ou affectant significativement les personnes, ou encore de suivi à grande échelle d'une zone accessible au public.

Obligation de notification des violations de données

Les données personnelles doivent être traitées de manière à garantir une sécurité et une confidentialité appropriées.

Lorsqu'il constate une violation de données à caractère personnel, le responsable de traitement a 72 heures pour la notifier à l'autorité de contrôle. En cas de risque élevé, les personnes concernées devront également être informées dans les meilleurs délais.
De même, lorsque le sous-traitant a connaissance d'une violation de données, il le notifie au responsable de traitement dans les meilleurs délais.

5.3.6 Le transfert des données hors UE

En cas de besoin de transférer des données hors UE, les ACTEURS doivent s'assurer que les pays destinataires offrent un niveau de protection adéquat.

Sous l'égide de la Directive de 1995, 11 pays bénéficient d'une décision d'adéquation. En l'absence d'adéquation, le RGPD autorise les transferts internationaux à condition que des garanties suffisantes soient fournies par le pays tiers, que les individus disposent de droits effectifs relatifs au traitement et à la protection de leurs données et que des solutions juridiques existent en cas de manquement à ces droits.

Pour effectuer leurs transferts dans ce cas, les ACTEURS peuvent s'appuyer sur un certain nombre d'outils reconnus comme légitimes par le Règlement, parmi lesquels :

→ Mettre en place des « règles d'entreprise contraignantes » ;
→ Ou inclure au contrat des « clauses contractuelles types » approuvées.

Dans tous les cas, les personnes concernées devront avoir été informées du transfert et des garanties associées.

5.3.7 Principe « d'accountability »

Ce principe impose aux entreprises de prendre toutes les mesures pour assurer la conformité aux obligations du règlement, et d'être en mesure de démontrer cette conformité.

Cette démarche proactive dispense les entreprises des formalités préalables.

Les obligations induites sont :

→ Documentation des traitements via des registres adaptés ;
→ Mise en œuvre des mesures permettant de s'assurer et prouver que les traitements sont conformes à la réglementation ;
→ Désignation d'un DPO le cas échéant ;
→ Intégration du principe de protection des données dès la conception ;
→ Élaboration d'études d'impact pour certains traitements.

Les sanctions

Une palette de sanctions est prévue à l'encontre des ACTEURS, allant du simple avertissement, jusqu'à l'obligation d'effacer les données.

Les amendes peuvent s'élever, selon le type de manquement, jusqu'à 10 ou 20 millions d'euros, ou 2 % ou 4 % du CA mondial (le montant le plus élevé fait office de plafond).

Par ailleurs, toute personne ayant subi un dommage du fait de la violation du règlement, aura droit d'obtenir du responsable de traitement ou du sous-traitant réparation du préjudice.

5.4 La mise en conformité des entreprises

Les tâches pour la mise en conformité concernent les clauses contractuelles, les conditions générales et les politiques de confidentialité, mais également les procédures (collecte consentement, sécurité, registres…).
Chaque entreprise doit être en mesure de démontrer qu'elle est en conformité avec les exigences du règlement.

Les entreprises, pour se mettre en conformité, devront :

→ Revoir (ou mettre en place) leurs procédures techniques et administratives (sécurité des données, droits des personnes, mentions d'information…) ;
→ Documenter systématiquement l'ensemble des politiques, procédures et traitements (pour respecter l'obligation de démontrer la conformité au règlement) ;
→ Revoir les clauses de confidentialité ainsi que leurs conditions générales ;
→ Revoir les contrats les liant avec les responsables de traitements ou les sous-traitants ;
→ Revoir le cadre juridique des traitements comportant des transferts de données hors UE ;
→ Le plus souvent, désigner un DPO.

5.5 Les craintes et les fausses idées

À l'heure de la rédaction de ces lignes, un travail conjoint entre la CNIL et les acteurs du marketing direct reste à faire pour l'application de la loi.
Par exemple, selon l'interprétation des textes, les courtiers pourraient être bridés par une obligation d'annoncer a priori aux contacts finaux, la liste exhaustive des partenaires avec qui ils pourront travailler (ce qui semble peu réaliste).

Ce travail est évidemment en cours.

La liste ci-dessous est extraite du document publié par le Sncd « *Règlement général sur la protection des données : idées reçues et foire aux questions* ».

Les fausses idées :

1. Le règlement ne concerne que les entreprises implantées en Europe : **FAUX**
Comme vu plus haut dans ce document, le règlement s'applique aux entreprises implantées dans les 28 États membres, mais également à toute entreprise dès lors qu'elle procède à un traitement de données lié à des personnes situées au sein de l'UE.

2. Le règlement ne concerne pas les entreprises traitant des données BtoB : **FAUX**
Le règlement ne fait pas de distinction entre les données BtoB et BtoC.

3. Le règlement ne concerne que le traitement automatisé des données : **FAUX**
Le règlement couvre les traitements de données automatisés ou non (traités par une personne).

4. Collecter le consentement des personnes pour le traitement de leur donnée est une obligation systématique : **FAUX**
Le règlement intègre une balance entre les intérêts légitimes du responsable de traitement, et les intérêts ou libertés et droits fondamentaux de la personne concernée. Dans tous les cas, le droit d'opposition s'applique.

5. L'obligation de désigner un DPO (Délégué à la protection des données) ne s'applique qu'aux grandes entreprises : **FAUX**
La désignation d'un DPO ne dépend pas de la taille, mais de la nature des activités.

Références

« Code Général de Déontologie de la Communication Directe » (SNCD 2011)
« Code Général de Déontologie de la Communication Directe Électronique » (SNCD 2005)
« Charte du développement responsable » (SNCD 2013)
« Règlement général sur la protection des données : idées reçues et foire aux questions » (SNCD 2017)
« Charte sur la Publicité Ciblée et la Protection des Internautes » (UFMD 2010)
« Guide des bonnes Pratiques concernant l'usage des Cookies Publicitaires » (UFMD 2010)

La Législation

- Loi Informatique et Libertés N°78-17 du 6 janvier 1978
- Modifiée par la loi N°2004-801 du 6 août 2004 (publiée au Journal Officiel le 7 août 2004)
- Loi pour la Confiance dans l'Économie Numérique N°2004-575 du 21 juin 2004 modifiée (publiée au Journal Officiel du 22 juin 2004)
- Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données
- Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique

6. Les référentiels

6.1 Définition

Il existe de nombreuses définitions de la notion de référentiel. Elles sont souvent liées à un domaine d'activité : mathématique, physique, informatique…
Nous considérons ici un référentiel comme étant un ensemble structuré de données de références : référentiels des Civilités, des Prénoms mais aussi des Communes, des Voies ou des Sociétés…

L'objectif de ce chapitre n'est pas de créer un annuaire des différents référentiels disponibles (qui ne saurait de toute façon être exhaustif) mais de préciser et décrire les apports des référentiels dans les traitements de Qualité de données.

6.2 Pourquoi

L'utilisation d'un ou plusieurs référentiels est fonction des objectifs que vous souhaitez atteindre. En effet si vous souhaitez géolocaliser vos adresses, le fichier BAN peut répondre à ce besoin, tout comme un référentiel de prénoms peut être utilisé pour définir le sexe de votre client.

Dans nos métiers, l'utilisation des référentiels répond bien souvent aux problématiques suivantes:

- Fiabiliser la donnée collectée (Hexavia, Hexacle…)
- Enrichir les points de contacts (cookies, emails, téléphone…) ;
- Enrichir votre connaissance client (Insee, référentiels privés…).

Les traitements de Qualité de données répondent à des objectifs de fiabilisation de l'information et d'enrichissement des données incomplètes. Les référentiels représentent donc la composante essentielle de la quasi-totalité des prestations de qualité de données. Ils permettent la fiabilisation (contrôle, validation, voire normalisation) de l'information traitée et constituent la source à partir de laquelle les données vont pouvoir être complétées.

6.3 Comment

L'utilisation des référentiels dans le domaine de la Qualité des données sera illustrée au travers des prestations suivantes :

- Normalisation des adresses postales (RNVP) ;
- Dédoublonnage/Déduplication ;
- Enrichissement.

La normalisation des adresses postales

Les traitements de normalisation des adresses postales mettent en œuvre deux natures distinctes de référentiels :

- Les référentiels qui vont permettre d'identifier les différents constituants d'une adresse. Ils sont principalement utilisés lors des phases de Restructuration et de Normalisation de l'adresse ;
- Les référentiels permettant de contrôler et valider l'existence et la cohérence des informations géographiques contenues dans les 3 voire 4 dernières lignes de l'adresse. Ils correspondent à la phase Validation postale.

La première catégorie de référentiels se révèle spécifique à chaque éditeur de solution logicielle, tant au niveau du contenu, plus ou moins complet, qu'au niveau de l'articulation des données entre elles, plus ou moins élaborée. Ils précisent pour chaque élément référencé les informations suivantes :

- Libellé ;
- Éventuelle abréviation (selon les différentes normes) ;
- Genre ;
- Nombre ;
- Type (mot simple, mot composé) ;
- Nature(s) (Article, Adjectif, Nom commun, Civilité, Prénom…) ;
- Indicateur précisant que l'élément peut éventuellement être lié au mot qui précède et/ou au mot qui suit (gestion des mots composés) ;
- …

Ces référentiels sont constitués à partir de données provenant d'organismes publics (table des prénoms de l'INSEE, liste des compléments de distribution et des types de voie de la norme AFNOR NF Z10-011…) et du vocabulaire courant (articles, adjectifs, nom commun…).

Le niveau de sophistication du référentiel matérialise souvent la compétence et l'expérience de l'éditeur. Il est ainsi possible de parler, pour les plus aboutis d'entre eux, de véritable grammaire intégrant des notions de syntaxe et de sémantique. Ces notions permettent, après une qualification de chacun des mots composant l'adresse, de les regrouper et d'identifier ainsi les différents constituants de l'adresse traitée. Chaque constituant se voit ensuite doté d'une fonction (complément de distribution, information voirie, mention de distribution spéciale…). Il ne reste plus, alors, qu'à (re)positionner chaque constituant sur les lignes adresse qui lui correspondent.

La deuxième catégorie de référentiels utilisés est produite par le Service National de l'Adresse (La Poste) et permet la validation et la normalisation des libellés de communes, de voies et de certains compléments de distribution. Ils permettent également de valider l'adéquation entre un code postal et une commune ou un code postal et la voie pour les communes disposant de plusieurs codes postaux. La gamme géographique des référentiels proposés par le SNA se compose des fichiers suivants :

Hexaposte	Fichier des codes postaux ménages et CEDEX, des communes, des lieux-dits référencés par La Poste et des bureaux CEDEX
Hexavia	Fichier des voies (et anciennes dénominations)
Hexaclé	Fichier des numéros dans la voie
Hexaligne3	Fichiers des compléments de distribution
Cedexa	Fichiers des adresses des entreprises bénéficiant d'un contrat CEDEX

Ces référentiels peuvent être complétés par certains fichiers proposés par l'INSEE tels que le fichier des Populations légales (nombre d'habitants d'une commune), le Code Officiel Géographique…

6.4 Différentes approches

Selon les utilisations souhaitées et les objectifs visés, vous utiliserez tel ou tel référentiel. La complexité reste dans la connaissance d'un référentiel et de ses limites et avantages.

Cependant nous pouvons faire une distinction entre les référentiels dits "classiques" et les alternatives qui sont apparues récemment.

Les classiques

Dans le traitement de l'adresse postale il existe les référentiels postaux qui nous permettent de valider une voie et de la normaliser selon les indications de La Poste française (hexavia, hexaclé…).

De même, les annuaires téléphoniques sont souvent utilisés pour de l'enrichissement en numéro de téléphone.

Et pour enrichir la relation client nous utilisons les données de l'Insee afin de scorer le niveau de richesse, un âge, le nombre d'enfants…

Les trois fournisseurs les plus courants en France ont été La Poste, France Télécom et l'Insee.

Cependant, suite à l'arrêt des monopoles et à l'obligation de fournir des données au public (Open Data), l'intérêt et le poids de ces référentiels ont fortement bougé. L'annuaire France Télécom n'est plus aussi exhaustif que par le passé, il faut réussir à compiler des données venant des différents acteurs comme Free, SFR, Orange, Bouygues… pour constituer une référentiel aussi large qu'avant.

Aussi nous voyons apparaître d'autre types de référentiels afin de combler les lacunes des classiques.

Les alternatives

La notion d'alternatives à l'utilisation de référentiels peut être perçue à différents niveaux.
Le premier consiste à envisager des traitements de qualité de données sans utilisation de référentiels.

Or ne pas utiliser de référentiels reviendrait à limiter les traitements de Qualité de données aux seuls aspects algorithmiques que sont le contrôle du format et de la longueur et, lorsque que l'information contient dans sa structure une clé de contrôle (clé RIB pour un compte bancaire), la validation de sa cohérence sans pouvoir, toutefois, garantir son existence réelle.

Par exemple, une chaîne de caractères de 10 chiffres commençant par un zéro correspond syntaxiquement à un numéro de téléphone. Sans confrontation de ce numéro à un référentiel Annuaire, il est impossible d'affirmer que ce numéro est bien affecté.

Il n'existe donc pas d'alternative crédible à l'utilisation de référentiels hormis l'utilisation d'autres référentiels. Les seules variations envisageables concernent la façon de les consulter. Cet aspect est développé ci-dessous.

L'open data met à disposition un grand nombre de données et notamment des données issues des référentiels postaux (BAN et BANO) gratuitement. Alors pourquoi ne pas les utiliser pour remplacer les "classiques" ?

Ces référentiels permettent souvent d'obtenir des résultats similaires aux classiques et cela gratuitement (fichier Siret passé dans le domaine de l'open data).

D'autres référentiels apparaissent afin de proposer des données nécessaires à la communication. Ces référentiels sont pour la plupart des référentiels privés constitués à partir d'agrégation de sources différentes (base mutualisée), permettant ainsi d'enrichir votre base avec des données de contacts supplémentaires et/ou des comportements d'achats spécifiques.

L'avantage de ces référentiels est que tout le monde peut les constituer dans la limite des réglementations, et ainsi gérer les sources et les mises à jour.

Prospectives

Aujourd'hui, l'évolution imposée par le législateur permet d'accéder à des référentiels libres (Open Data). Les gouvernements mettent à disposition de tous des données "publiques" qui ouvrent des perspectives nouvelles notamment dans le traitement de la donnée.

La question posée serait de savoir comment substituer aux référentiels historiques de nouveaux fichiers issus de l'Open Data.

Principalement concernés par cette approche, les référentiels géographiques proposés par le Service National de l'Adresse qui peuvent potentiellement être concurrencés par la Base Adresse Nationale (BAN), émanation du travail collaboratif de La Poste, de l'IGN, des collectivités, des communes et des SDIS (Service Départemental d'Incendie et de Secours). A ce jour, il apparaît clairement que la **BAN** ne peut pas encore être utilisée pour des traitements de normalisation des adresses postales (*les libellés voies peuvent différer de ceux des référentiels dits "officiels" ; les informations CEDEX ne figurent pas dans la **BAN** ; et, plus grave, le lieu-dit (ligne 5 de l'adresse) est également absent*).

En revanche, la BAN précisant les coordonnées Lambert (X,Y), la longitude et la latitude de chaque numéro dans la voie constitue une source très crédible pour les traitements de géolocalisation et/ou de géocodage (enrichissement en code IRIS, en Carreaux) avec comme limite de ne couvrir que la France métropolitaine et les Départements d'Outre-mer. La mise à jour hebdomadaire des informations contenues dans la BAN constitue également un point positif.

Le mode de diffusion de la BAN (licence spécifique « gratuite de repartage » obligeant, notamment, à retourner tout enrichissement aux concédants) représente un frein à une utilisation professionnelle de ce référentiel.

De même qu'il existe des référentiels privés comme celui de Google permettant de géolocaliser en (X,Y) ou le référentiel TomTom etc. ces référentiels construits pour des fins de navigation pourraient être utilisés pour du traitement de données.

Nous pourrions envisager d'autres modes comme interroger directement un référentiel complément délocalisé auprès d'une institution, d'une société privée ou même du consommateur. On parlerait alors de datasharing.

En Suède le gouvernement met à la disposition d'acteurs privés le référentiel de tous les Suédois, moyennant finance. Ce qui permet de vérifier l'utilisation qui en est faite et ainsi de maîtriser les droits de chacun.

En revanche, la **BAN** précisant les coordonnées Lambert (X,Y), la longitude et la latitude de chaque numéro dans la voie constitue une source très crédible pour les traitements de géolocalisation et/ou de géocodage (enrichissement en code IRIS, en Carreaux) avec comme limite de ne couvrir que la France métropolitaine et les Départements d'Outre-mer. La mise à jour hebdomadaire des informations contenues dans la **BAN** constitue également un point positif.

Le mode de diffusion de la BAN (licence spécifique « gratuite de repartage » obligeant, notamment, à retourner tout enrichissement aux concédants) représente un frein à une utilisation professionnelle de ce référentiel.

De même qu'il existe des référentiels privés comme celui de Google permettant de géolocaliser en (X,Y) ou le référentiel TomTom etc. ces référentiels construits pour des fins de navigation pourraient être utilisés pour du traitement de données.

Nous pourrions envisager d'autres modes comme interroger directement un référentiel complément délocalisé auprès d'une institution, d'une société privée ou même du consommateur. On parlerait alors de datasharing.

En Suède le gouvernement met à la disposition d'acteurs privés le référentiel de tous les Suédois, moyennant finance. Ce qui permet de vérifier l'utilisation qui en est faite et ainsi de maîtriser les droits de chacun.

6.5 Les dangers, les craintes et les fausses idées

Exhaustivité des référentiels

Pour pouvoir parler de référentiel, celui-ci doit être exhaustif et mis à jour.
Or aujourd'hui qui détient un référentiel exhaustif ? mis à part peut-être l'Etat ?

Quid des mises à jours ?

La mise à jour d'un référentiel nécessite un travail de chaque instant. Cela sous-entend des coûts non négligeables. Si l'accès à ces référentiels devient gratuit (siret) comment financer les coûts inhérents à ces mises à jour ?

La facilité d'accès à des référentiels gratuits signifie-t-elle une qualité des données optimale ? Nous pourrions voir apparaître une baisse de qualité importante des traitements réalisés, creusant encore le risque de non-conformité réglementaire.

La BAN est plus complète que les référentiels du SNA. Est-elle plus exhaustive ?
Oui et non !

Concernant la BAN
La BAN contient 27 millions d'entrées, mais… L'objet de la BAN n'est pas de référencer des adresses postales, mais, plus largement, de réaliser une cartographie géographique.

Nous y trouvons des adresses non habitées (points cadastraux, cimetières, monuments), par conséquent des numéros de voies inexistants postalement et avec une structure différente (absence du lieu-dit).
A ce jour, la BAN n'est pas adaptée à un traitement postal.

D'une manière plus large, un référentiel existe pour répondre à une finalité. Attention à ne pas faire de raccourcis malheureux entre vos besoins et la description de ceux-ci.

7. Glossaire

RNVP

(Restructuration, Normalisation et Validation Postale)
La RNVP est la terminologie d'usage pour parler de correction informatisée d'adresses postales.
En France, la RNVP s'appuie sur : la norme AFNOR NF Z 10-011 du 19 janvier 2013 (règles d'écriture et présentation d'une adresse postale) pour la partie "RN" ; ainsi que sur des référentiels postaux (pour la partie "VP").
Bien qu'en théorie l'origine des référentiels ne soit pas imposé (pourvu qu'ils aient une exhaustivité satisfaisante), en pratique, il est difficile de ne pas s'appuyer sur les référentiels fournis par le SNA (Service National de l'Adresse, filiale de LA POSTE).
Une RNVP est un traitement très complexe. Les logiciels de RNVP font l'objet (en France) d'une homologation par le SNA.

Déduplication

La déduplication consiste à confronter plusieurs sources de données (fichiers ou flux), afin d'en déterminer les contacts communs.
La déduplication fait appel à des technologies pointues, permettant de gérer les écarts phonétiques ou typographiques.
La déduplication peut être « batch » (traitement de fichiers) ou « online » (c'est-à-dire recherche de préexistence dans une base de données).

Dédoublonnage

Le dédoublonnage consiste à rechercher des doublons au sein d'un même fichier (ou base de données).
Le dédoublonnage fait appel aux même techniques que la déduplication.

Over-kill

Un traitement de déduplication restitue des « doublons certains », des « doublons douteux » et des « uniques ».
Les outils de déduplication permettent généralement de déterminer le seuil de détermination d'un doublon.
Un réglage « over-kill » (qui « sur-élimine » des doublons) tend à considérer les doublons douteux comme doublons avérés.
Ce type de réglage est plutôt préféré lors des campagnes de prospection.

Under-kill

Un traitement de déduplication restitue des « doublons certains », des « doublons douteux » et des « uniques ».
Les outils de déduplication permettent généralement de déterminer le seuil de détermination d'un doublon.
Un réglage « under-kill » (qui « sous-élimine » des doublons) tend à considérer les doublons douteux comme des uniques.
Ce type de réglage est plutôt préféré pour les nettoyages de bases clients.

Doublon

Un doublon (ou groupe de doublons) est constitué de plusieurs contacts redondants (par exemple, « Martine Michu » présente plusieurs fois à la même adresse). Ce groupe est constitué :

- d'un 1er de groupe (qui sera le contact a conserver) ;
- d'un (ou de plusieurs) suivant(s) de groupe(s) (qui seront les contacts à supprimer, et pour lesquels les historiques seront à fusionner avec celui du 1er de groupe).

Doublon douteux

Les écarts de rédaction des contacts et adresses (faute phonétique ou typographique, omission d'éléments…) font que la certitude d'un groupe de doublons est variable d'un groupe à l'autre.
En deçà d'un certain seuil, un groupe peut être considéré comme douteux.
Exemple

Madame Marie ROBERT	Monsieur Robert MARIE
Résidence du Peuple Belge	12 Avenue du Peuple Belge
59800 LILLE	59000 LILLE

Les outils de déduplication permettent (ou devraient permettre) de choisir de les conserver ou non.

Priorité de déduplication (ou de dédoublonnage)

Un groupe de doublons est constitué d'un :

- 1er de groupe (généralement considéré comme celui à conserver) ;
- Suivant(s) de groupe (considéré comme « en trop »).

Le choix du 1er de groupe est généralement paramétrable (priorité selon le canal d'origine, priorité selon la qualité du contact, priorité selon un score prédéfini…).

Match-code (ou phonème ou clé de déduplication)

Que ce soit pour la RNVP ou la déduplication, les algorithmes de recherches sont tolérants aux écarts, que ceux-ci soient phonétiques ou typographiques.
Un match-code est une représentation phonétique d'un élément (nom de famille, voie, commune, etc.).

Erreur typographique

Une erreur typographique correspond à une faute de frappe.
Nous y trouvons :

1. les inversions de caractères :	« rue de Phil**da**elphie »
2. le doublement de caractères :	« rue de Phil**aa**delphie »
3. l'omission d'un caractère :	« rue de piladelphie »
4. la scission d'un mot :	« rue de Phil adelphie »
5. la fusion de 2 mots :	« rue dePhiladelphie »

Parmi les algorithmes, nous avons :

- Algorithme de distance Levenshtein (traite les cas 2, 3) ;
- Algorithme de distance Damerau-Levenshtein (traite les cas 1, 2, 3).

CNIL

La Commission Nationale de l'Informatique et des Libertés est une autorité administrative indépendante française. Le rôle de la CNIL est de veiller à ce que l'informatique ne porte pas atteinte, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques.

CIL

Le correspondant informatique et libertés ou CIL se positionne en intermédiaire entre le responsable des traitement des données concernées (l'entreprise dans un contexte marketing) et la CNIL.
Il est responsable :

- de la création et de la mise à jour d'une liste des traitements effectués ;
- de la publicité de cette liste ;
- d'une fonction de conseil et de recommandation auprès des responsables des traitements ;
- de l'intermédiation CNIL / structure.

Dans le cadre du RGPD, **le CIL est remplacé, depuis mai 2018, par le Délégué à la Protection des Données (DPO).**

DPO

Le Délégué à la Protection des Données est un poste ou une fonction qui est obligatoire dans de nombreuses entreprises dans le cadre de l'application du règlement européen de protection des données personnelles (mai 2018). Un délégué est obligatoire pour les entreprises publiques et pour les entreprises faisant un usage intensif des données personnelles.
Il est le garant du respect des textes juridiques relatifs au traitement des données personnelles et à leur sécurité. Il peut être considéré comme l'héritier du CIL.

Pseudonymisation des données

On parle de pseudonymisation de données dès lors que celles-ci ne peuvent plus être rapportées à l'individu qu'elles concernent, sauf recours à des informations additionnelles.
Cette notion est apportée par la nouvelle réglementation européenne.
La pseudonymisation réduit les risques liés à la protection des données.

Principe « d'accountability »

Ce principe est apporté par la nouvelle réglementation européenne. Il encourage les entreprises à prendre toutes les mesures pour assurer la conformité aux obligations du règlement, et être en mesure de démontrer cette conformité.

RGPD (ou GDPR)

Règlement Général sur la Protection des Données (ou ***General Data Protection Regulation***).
L'Union européenne a adopté en 2016 ce nouveau règlement, applicable à compter du 25 mai 2018.
Ce règlement introduit de nouvelles obligations pour les entreprises et de nouveaux droits pour les citoyens européens. Les entreprises doivent modifier les procédures et les démarches en matière de protection de la vie privée et des données à caractère personnel.
Pour plus de détails, le lecteur pourra se référer au paragraphe « Déontologie » du présent document.

IRIS

(ancienne dénomination : IRIS2000)
Les communes de plus de 10000 habitants (et certaines communes de plus de 5000) sont découpées en zones géographiques cohérentes (quartiers). Leur population évolue entre 1800 et 5000 habitants.
Un IRIS est identifié par un **Code IRIS** (5 digits alphanumériques) composé du code INSEE de la commune historique de rattachement (qui peut être différent de la commune actuelle), ainsi que d'un suffixe de 4 digits.
A ce jour, la France est découpée en environ 16500 IRIS (sur les communes les plus importantes).
Le code IRIS permet de récupérer des données statistiques sur la population le composant.

Carreaux, Rectangles

L'histoire des carreaux remonte à 2006 (au moins pour ce qui est relatif à une utilisation systématique).
Ils sont issus d'un rapprochement effectué par l'INSEE entre les données fiscales et des informations géographiques.
La France métropolitaine ainsi que la Réunion et la Martinique sont découpés en carreaux de 200 m de côté.
Si ceux-ci sont considérés trop peu peuplés (moins de 11 ménages) pour respecter l'anonymisation des données, ils sont alors regroupés en rectangles.
Les variables statistiques sont au nombre de 18 (caractéristiques des ménages, habitations, PCS et revenus).

Internet des Objets (IdO) ou Internet of Things (IoT)

Baptisé Web 3.0, l'Internet des objets connectés représente les échanges d'informations et de données provenant de dispositifs du monde réel avec le réseau Internet.

Coordonnées X,Y (ou Coordonnées géographiques)

Par coordonnées géographiques d'un lieu sur la Terre, on entend un système de deux coordonnées qui sont le plus souvent : la latitude, la longitude (dans nos métiers traitant des coordonnées de lieu, l'altitude est oubliée, considérée comme étant la surface terrestre).
Il existe pléthore de systèmes de projection utilisés dans le monde. Les plus connus sont : WGS84 (coordonnées GPS), Lambert, Mercator.

Remerciements

Le Sncd remercie les adhérents membres de la commission Data & Technologies qui se sont particulièrement investis dans l'élaboration et la production de ce Livre blanc :

- **Goulven Aubrée (Edgewhere),**
- **Christophe Blin (76310),**
- **Guy Cals (Amabis),**
- **Marc de Fougerolles (Data Company).**

76310
VOS DONNÉES SONT SACRÉES

amabis
Donnons du sens à la data client

edgewhere

DATA company

sncd
Data Marketing
Industrie

68 boulevard Saint Marcel - 75005 PARIS

01 55 43 06 11

www.sncd.org

info@sncd.org

www.linkedin.com/company/sncd

twitter.com/Sncdmulticanal