MASTER EN SCIENCES INFORMATIQUES

Anonymisation des données
pourquoi et comment?

Burniaux, Francois-Xavier

Award date:
2013

Awarding institution:
Universite de Namur

Facultés Universitaires Notre-Dame de la Paix, Namur
Faculté d'Informatique
Année académique 2012–2013

Anonymisation des donnes: pourquoi et comment ?

François-Xavier Burniaux

Promoteur : _____ (Signature pour approbation du dépôt - REE art. 40)
Jean Noël Colin

Mémoire présenté en vue de l'obtention du grade de
Master en Sciences Informatiques.

Résumé - Summary

La problématique de la protection des données à caractère personnel fait l'objet de nombreux écrits. Chaque seconde, des données sont encodées, manipulées et conservées. Ces données, souvent à caractère personnel, se doivent d'être protégées. Régie par différents règlements, directives et lois, l'utilisation de ces données doit être protégée de différentes façons. Dans ce mémoire, nous allons nous intéresser à l'anonymisation des données pour la protection de celles-ci. Après un examen du cadre légal, nous allons définir la démarche à suivre pour anonymiser correctement les données. Ensuite, nous proposerons les différentes techniques pour y parvenir et nous les analyserons à partir d'une méthode principale, k-anonymity, et de plusieurs variantes de ce modèle dont les méthodes l-diversity et t-closeness. Enfin, après l'analyse des avantages et inconvénients de chacune de ces méthodes, nous exposerons les diverses techniques d'implémentation de la méthode k-anonymity, pour comprendre en profondeur son fonctionnement. Enfin, nous analyserons un cas pratique d'anonymisation, à savoir, un sondage effectué sur la mobilité en 2010 par le Groupe de Recherche de Transport de l'Université de Namur (GRT).

Mots Clés : Protection de la vie privée, données pesonnelles, anonymisation, pseudonymisation, k-anonymity, l-diversity, t-closeness, m-invariant.

The problem of the personal data protection is the object of numerous papers. Every second, data are encoded, treated and kept. These data, often personal, must be protected. Governed by various regulations, directives and laws, the use of these data can be protected in various ways. In this report, we are interested in the anonymisation of the data to protect them. After an examination of the legal frame, we will define the approach to follow to anonymise correctly the data. Then, we will propose the various methods to reach there and we will analyze them from a main method, k-anonymity, and from several variants of this model among which l-diversity and t-closeness. Finally, we will analyse a practical case of anonymization, namely a sounding made on the mobility in 2010 by the Group of Research for Transport of the University of Namur (GRT).

Keywords : Protection of life privacy, personal data, anonymisation, pseudonymisation, k-anonymity, l-diversity, t-closeness, m-invariant.

Avant-propos

Je tiens à remercier mon promoteur Monsieur Colin pour m'avoir proposé un sujet si passionnant ainsi que pour ses conseils avisés ; Monsieur Cornelis, pour m'avoir permis d'analyser le sondage du GRT ; mon épouse, Anne-Sophie, pour son soutien tout au long de la rédaction de mon mémoire et pour ses nombreuses relectures et mon fils, Louis, pour sa patience du haut de ses quatre mois.

Table des matières

Introduction

La protection de la vie privée est à un tournant de son histoire. Jamais nous n'avons eu autant de données à notre disposition. L'avènement des réseaux sociaux (comme Facebook, Twitter, Google plus, etc.) a complètement changé la donne sur la diffusion et l'utilisation des données personnelles. Maintenant, en quelques clics, nous pouvons accéder au profil d'un utilisateur pour autant qu'il n'ait pas bloqué celui-ci. Nombreuses sont les personnes qui ignorent les mécanismes de protection proposés tellement ces réseaux sont faits pour rendre nos données publiques. Impensable il y a quelques années, nous pouvons maintenant, en regroupant plusieurs données issues de différentes sphères - quelles soient professionnelles (Linked), sociales (Facebook, les blogs), corporatives (forums) ou encore musicales (Spotify, Deezer) - établir le portrait robot d'une personne sans jamais l'avoir vue. A tel point que les recruteurs, lorsqu'ils reçoivent le curriculum vitae d'un candidat, effectuent une recherche sur Google pour se faire une idée de celui-ci et déjà obtenir les réponses à certaines questions : est-il actif sur les réseaux sociaux ? Mène-t-il une vie saine ? Ecrit-il correctement ? Est-il joueur ?[1]. Autant de réponses que l'internet actuel apporte sans nécessiter le moindre effort de la part de l'utilisateur intéressé si ce n'est d'ouvrir son navigateur internet.

Dans de nombreux cas, ce danger provient évidemment de la possibilité de croiser et de confronter les données. Des données qui, pour la majorité, sont persistantes et alimentent les réseaux actuels comme nous le verrons par la suite. De plus, aucune information ne nous est donnée quant à la collecte, au traitement et à la diffusion de ces données. Il est en effet très rare que ces sites nous informent de leurs façons de procéder et du fait qu'ils archivent nos fichiers de données. Malheureusement, cela s'incrit dans la logique du fonctionnement de ces réseaux communautaires. En effet, l'essence même de ceux-ci est l'information fournie par leurs utilisateurs. Il serait inimaginable dès lors qu'il soit prévu que les données soient supprimées automatiquement.

Paradoxalement, le nombre de données personnelles explosent sur internet mais la sécurité ne suit pas, comme l'atteste la faille de sécurité dont a été victime la SNCB en décembre 2012. Grâce à une requête effectuée dans un moteur de recherche, un internaute est tombé sur les données personnelles (nom, prénom, code postal, ville, téléphone privé, téléphone public, e-mail, etc.) de plus

d'1 400 000 personnes, soit plus de 10% de la population de notre royaume[2]. Des chiffres hallucinants qui montrent à quel point la confidentialité des données et la protection de celles-ci sont prises à la légère. D'autant plus que le risque est grand de voir ces données distribuées et partagées à des fins commerciales ou autres. Dans ce cas-ci, des informations personnelles concernant des personnalités politiques, des membres de la Commission européenne, ainsi que des parlementaires ont été divulguées [3].

Malheureusement, ce danger n'est pas prêt de s'arrêter avec l'essor récent du cloud computing. En effet, de plus en plus de personnes font confiance à de grands noms de l'informatique comme Amazon, Dropbox, Google ou encore Microsoft pour stocker leurs données, persuadées de les mettre à l'abri de tout intrus. Or, de nombreuses failles de sécurité ont déjà été révélées. En 2011, Dropbox a laissé les comptes de plus de 25 millions de personnes accessibles pendant quatre heures. N'importe quel mot de passe fonctionnait ! Des dizaines de milliards de fichiers étaient donc totalement disponibles sans la moindre protection, sans que cela ne gène Dropbox. En effet, la société de San Fransico à minimisé cette faille de sécurité en affirmant que cela "n'avait touché que moins d'1% de ses utilisateurs" [4], mais cela représente quand même 200 000 comptes ! Quant aux conditions de protection et de stockage, les fournisseurs de cloud computing pratiquent la politique de l'autruche et se retranchent derrière des termes flous et souvent obscurs dans le but de gérer comme ils le veulent la protection des données. Dropbox, par exemple, argue dans ses règles de sécurité, que les données sont cryptées en AES-256. Nous pourrions croire qu'il s'agit d'une protection optimale mais si nous regardons plus en profondeur les conditions de sécurité, nous constatons que le personnel dispose de la clé de cryptage pour certains traitements [4].Il ne s'agit là que d'un exemple parmi tant d'autres.

Ces exemples mettent en exergue le problème actuel. Face aux flux de données, la sécurité de nos données personnelles n'est pas assurée. Pourtant, la détention et l'utilisation de données personnelles sont strictement réglémentées par la loi. Cependant, eu égard au coût engendré et au peu d'intérêt que cela représente pour les sociétés, la sécurité est négligée. L'objet de ce mémoire est donc de s'intéresser à l'anonymisation des données. Cette méthode permet de diffuser des données sans pouvoir en retrouver le propriétaire. Nous verrons que cette méthode est efficace mais qu'elle a ses limites et que l'utilisation de celle-ci doit faire préalablement l'objet d'une solide analyse sur les besoins, objectifs et exigences poursuivis. L'application de la méthode à un cas pratique sera riche en enseignements et montrera également les limites de l'exercice.

Ce mémoire sera structuré comme suit. Chapitre 1, nous replaçons la protection des données dans le contexte actuel, nous définissons de manière formelle ce que nous entendons par "traitement de données à caractère personnel" et dans quel cadre législatif ce traitement s'inscrit. Chapitre 2, nous propo-

sons les différentes étapes nécessaires à une anonymisation en bonne et due forme. Pour réaliser celle-ci, un ensemble de méthodes et de techniques sont proposées dans le chapitre 3. Ces méthodes sont classifiées en trois catégories en fonction des effets qu'elles produisent. Chapitre 4, nous expliquons les différents modèles d'anonymisation pour les données statiques et pour les données dynamiques. Pour chacun des modèles envisagés, nous définissons leurs techniques et nous l'exemplifions. Une analyse de leurs avantages et inconvénients est également faite. Pour le modèle k-anonymity, nous présentons les différents algorithmes qui mettent en place la principale méthode d'anonymisation. Chapitre 5, nous appliquons nos analyses théoriques à un cas pratique sur un sondage effectué en 2010 sur la mobilité. Enfin, nous tirons les conclusions de notre analyse et nous déterminons l'utilité de l'anonymisation pour la protection des données à caractére personnel.

Chapitre 1

La protection des données à caractère personnel

1.1 Une problématique actuelle

Suite à l'émergence des systèmes d'information et de communication, un nombre conséquent de données est encodé et automatisé. Par exemple, nous encodons un grand nombre de données (souvent inconsciemment) sur internet pour créer un compte e-mail, acheter un livre, commander un meuble, louer un film ou encore réserver nos vacances. Lors de chacune de ces actions, nous avons encodé des données personnelles qui nous caractérisent et qui nous identifient en tant qu'un seul et même individu. La protection de ces données est essentielle et nous devons avoir le contrôle de ces données. Malheureusement, ce n'est pas souvent le cas.

Google, par exemple, a récemment revu sa politique de protection de la vie privée[5], en synthétisant son règlement en une seule et même page. Google a profité de l'ambiguïté de la langue française et s'autorise à utiliser nos données personnelles pour l'ensemble des services qu'il propose. Sous le couvert d'une expérience utilisateur unique, il collecte, diffuse et utilise nos données personnelles comme bon lui semble à travers ses services. Cette modification, entrée en vigueur en mars 2012, a d'ailleurs fait débat[6].

Utilisés par 70% de la population, les réseaux sociaux sont également un réservoir d'informations personnelles important. L'utilisateur de Facebook, par exemple, peut créer un compte pour y mettre des commentaires, des photos, des vidéos, etc. En somme, des données qui le concernent et qui concernent son entourage. Toutes ces actions devraient donc être contrôlées par l'utilisateur. Pourtant, alors que Facebook possède plus d'un milliard d'utilisateurs [7], ce réseau ne permet pas que nous supprimions, par exemple, toutes les informations liées à notre compte ni même notre compte. Nous pouvons juste le désactiver, mais il est impossible d'avoir une suppression complète et rapide des données. Le prétexte avancé par les administrateurs de Facebook réside

dans le fait que, si l'utilisateur veut revenir, il retrouvera un environnement qui lui est familier et récupèrera ses contacts. Une photo uploadée sur le serveur et ensuite supprimée reste dans le serveur de Facebook à jamais même si le propriétaire fait de nombreuses réclamations afin de la supprimer[8]. Facebook va même plus loin en obligeant ses utilisateurs à créér un compte sous leur vrai nom plutôt que sous un pseudonyme. Facebook s'autorise à supprimer tout compte qui ne respecte pas cette règle. Contestée par le Conseil d'Etat, un juge allemand a quant à lui, déclaré que cette mesure était parfaitement légale puisque le siège social de Facebook se trouve en Irlande où la loi sur la protection des données est plus souple[9].

Ces exemples nous montrent que, sous le couvert de différents prétextes, nos données ne nous appartiennent plus et que nous devons, au contraire, offrir nos données pour obtenir des services. Les internautes belges ne s'y trompent pas. Selon l'eurobaromètre 74.3[10], 23 % des utilisateurs belges accordent leur confiance aux données personnelles déposées sur les réseaux sociaux, les moteurs de recherche et les services de messagerie. 23 % de ces mêmes utilisateurs se sentent également obligés de divulguer leurs informations pour obtenir un service, soit un utilisateur sur quatre [10]. Avec l'internet 2.0, la diffusion de l'information a complètement évolué. Les internautes créent et partagent du contenu. Ils sont au coeur du développement de l'information. Ils diffusent de nombreuses données sur leurs goûts, leurs envies, leurs habitudes. En somme, leur vie.

Une étude menée par la société ComScore[11] en décembre 2007 pour le New York Times et publiée en mars 2008 a révélé des chiffres édifiants : sur une période d'un mois, 336 milliards de transmissions de données ont été collectées à l'insu des utilisateurs par Yahoo, Google, Microsoft, AOL et MySpace. Yahoo arrive en tête avec une moyenne de 2520 données uniques collectées par visiteur, viennent ensuite MySpace avec 1229, AOL avec 610, Google avec 578 et Microsoft avec 355. Yahoo collecte donc par mois, 110 milliards de transmissions, soit 811 informations par jour par utilisateur qui surfe sur son site.

Il faut être conscient que toutes ces données présentes sur internet représentent une vraie valeur commerciale pour les entreprises. Les départements marketing de nombreuses sociétés seraient prêts à débourser des fortunes pour obtenir des renseignements sur leurs clients ou prospects. En effet, grâce aux données laissées par l'internaute, nous pouvons véritablement dresser son portrait. Une fois ces informations obtenues, une société pourrait très bien développer pour cet internaute un marketing ciblé en fonction de ses goûts et ce, à un degré de raffinement très précis. Les outils de Costumer Relationship Management (CRM) permettent par la suite de compléter parfaitement le profil de l'acheteur. Nos données personnelles sont donc recherchées et elles ont un coût comme l'atteste cette loi qui a été adoptée par les députés allemands en

juillet dernier. Ceux-ci ont voté "une loi autorisant les services municipaux à communiquer des données privées concernant leurs citoyens à des tiers, à des fins éventuellement commerciales"[12]. Autre exemple, Samarati[13] explique que la plupart des municipalités aux États-Unis vendent les registres de la population. Ces registres contiennent "les identités des individus avec leurs données démographiques ; comme le recensement local, les listes d'électeurs, les répertoires de la ville et d'agences de véhicules automobiles, les agences immobilières"[13] et bien d'autres informations encore.

Outre ces données récoltées (sans que nous le sachions), nous avons vu que nous encodions de nombreuses données pour divers services. Nous pourrions penser que ces données sont conservées de manière optimale et confidentielle. Malheureusement, l'histoire nous prouve que ce n'est pas le cas. En 2008, la Deutsche Telekom a avoué qu'une faille de sécurité avait permis d'avoir accès aux données personnelles (avec la carte bancaire) de 30 millions de comptes[14]. Il faut noter que ce n'est pas la seule société à avoir eu une faille de ce type. En 2009, suite à une faille de sécurité, le site d'Orange a permis à n'importe quel internaute de consulter quelque 400 000 fiches et ce, pendant plusieurs semaines. En modifiant simplement un numéro dans l'url, nous avions accès à la fiche des clients comprenant leurs coordonnées complètes, les offres souscrites, leurs codes d'accès ou encore leur numéro de téléphone[15].

La protection des données doit donc être mise en place afin de préserver le mieux possible l'intégrité et la diffusion de celles-ci. Dans une société où tout devient public, la mission n'est pas aisée mais des méthodes et processus ont été mis en place afin d'éviter que certaines données jugées sensibles soient divulguées à des personnes autres que celles qui ont reçu l'autorité pour les consulter. L'une de ces méthodes, l'anonymisation des données, est l'objet de ce mémoire. Nous nous intéresserons donc à cette méthode qui consiste à rendre les données d'un tableau anonymes. A travers différentes techniques, l'anonymisation garantit la protection des données de manière efficace et complexifie la recherche illégale d'informations sur ces données anonymisées. Cependant, comme nous le verrons par la suite, ces différentes méthodes d'anonymisation ne garantissent pas une protection optimale dans tous les cas.

Nos données doivent donc être anonymisées mais pas de n'importe quelle manière. En 2006, AOL décide d'anonymiser une liste de 20 millions de requêtes effectuées via son moteur de recherche[16]. Pour chaque requête, il remplace le nom du chercheur par un numéro. Cependant, cette méthode de protection s'est avérée beaucoup trop faible. En effet, après l'analyse des requêtes du numéro "4417749" portant sur "engourdissement du doigt", "homme célibataire 60 ans", "chien qui urine partout", un journaliste a pu retrouver en affinant ses recherches qu'il s'agissait de Thelma Arnold, 62 ans qui vit à Lilburn et possède trois chiens. Cet exemple montre bien que l'anonymisation ne

doit pas être prise à la légère et qu'il faut une véritable réflexion sur la nature des données, la méthode à choisir et les objectifs poursuivis. C'est pourquoi, dans ce mémoire, nous examinerons chacun de ces points afin d'apporter des solutions pour une anonymisation la plus efficiente possible.

1.2 Cadre législatif

Face au nombre croissant de données présentes et diffusées sur les différents systèmes d'information, divers règlements, directives et lois ont été mis en place dans de nombreux pays pour s'assurer que les données encodées par les utilisateurs soient correctement traitées par le destinataire. Afin d'avoir une vision complète du cadre législatif dans lequel s'inscrit le traitement de ces données, nous allons retracer de manière chronologique et dans les grandes lignes les différents textes juridiques belges et internationaux sur le sujet.

En septembre 1980, l'*Organisation for Economic Co-operation and Development* établit les lignes directrices régissant la protection de la vie privée et les flux transfrontaliers de données à caractère personnel[17]. Elle définit les données personnelles comme des données qui appartiennent à une identité ou à un individu identifiable. Ces données sont gérées par le contrôleur des données qui est soumis à la loi en vigueur. Elle entend par flux transfrontaliers de données, les flux de données qui passent d'un pays à un autre. Elle établit quelques principes qui seront détaillés plus loin : la collection des données, la qualité des données, l'utilisation de ces données et leurs limitations. Et surtout, elle établit les droits du citoyen à l'égard de ses données. Elle suggère également aux différents pays d'encadrer administrativement, légalement et à travers différentes procédures, la protection des données à caractère personnel.

Suite à l'émergence des nouveaux systèmes de communication et d'information, le Conseil de l'Europe ratifie le 28 janvier 1981, la convention 108 pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel[18]. Elle définit les bases de la protection de la vie privée à travers deux principes : (1)la minimisation des données personnelles et (2)la souveraineté sur les données personnelles[19]. En effet, les données automatisées doivent (1) "être adéquates, pertinentes et non excessives par rapport aux finalités pour lesquelles elles sont enregistrées"[18]. Dans le cadre d'une transaction bancaire, le site e-commerce n'a besoin de connaître que nos achats et notre adresse de livraison. Il n'a nullement besoin de savoir nos coordonnées bancaires ni l'adresse de facturation. Et réciproquement, le site de banque en ligne n'a pas besoin de savoir ce que l'on a acheté. Ces données doivent également avoir "une durée n'excédant pas celle nécessaire aux finalités"[18]. Ces données devraient donc être effacées une fois que l'utilisation de celles-ci n'est plus nécessaire. L'utilisateur qui a encodé ses données personnelles sur un site est le seul maître de ses données (2). Il bénéficie donc de garantie sur celles-ci.

Il doit savoir à quelles fins elles seront utilisées et doit connaître l'identité de celui qui les possède. Il pourra également à tout moment demander à les recevoir. Et si le destinataire des données ne respecte pas les deux principes vus ci-dessus, nous pouvons lui demander la rectification ou la suppression de ces données. Cependant, afin de respecter notre modèle démocratique actuel, ces dispositions peuvent être levées si elles concernent "la protection de la sécurité de l'État, la surêté publique, les intérêts monétaires de l'État ou la représentation des infractions pénales" mais également "la protection de la personne concernée et les droits et libertés d'autrui"[18].

En 1992, en Belgique, la loi relative à la protection de la vie privée à l'égard des traitements de données à caractère personnel est ratifiée. Il s'agit de la première loi belge traitant spécifiquement des données à caractère personnel. Communément appelée "loi vie privée", cette loi a été profondément modifiée suite à la directive internationale 95/46/CE du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation des données[20]. En effet, en 1995, suite à l'apparition d'internet dans tous les foyers et à la démocratisation de l'accès à l'information, il fallait une directive pour s'assurer "que les données à caractère personnel puissent circuler librement d'un État membre à l'autre, mais également que les droits fondamentaux des personnes soient sauvegardés"[20]. Cette directive dresse une liste complète et exhaustive du cadre dans lequel ces données doivent être traitées à travers les lois domestiques. C'est pourquoi la loi vie privée de 1992 a été mise à jour en 1998 et mise en exécution en 2003 par un arrêté royal[21].

Elle définit ce que nous entendons par "traitement" et "données à caractère personnel"[22]. Les données à caractère personnel sont "toute information concernant une personne physique identifiée ou identifiable"[22]. Une personne identifiable est "une personne qui peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs éléments spécifiques, propres à son identité physique, physiologique, psychique, économique, culturelle ou sociale"[22]. Le nom, le prénom, l'adresse postale, la date de naissance sont des exemples parmi d'autres de données à caractère personnel. Celles-ci font l'objet d'un traitement, c'est-à-dire de "toute opération ou ensemble d'opérations effectuées ou non à l'aide de procédés automatisés"[22]. Les opérations sont multiples et de différentes formes comme "la collecte, l'enregistrement, l'organisation, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, ainsi que le verrouillage, l'effacement ou la destruction de données à caractère personnel"[22].

Elle définit l'usage que nous pouvons faire de ces données. Elles doivent être, entre autres, traitées "loyalemenent et licitement"[22] mais peuvent éga-

lement faire l'objet d'une utilisation ultérieure pour des statistiques, des données historiques ou des recherches scientifiques. Cependant, les données collectées ont une temporalité. Elles ne peuvent dépasser un certain délai qui doit être compatible avec la finalité de la collection des données. De 1997 à 2006, d'autres directives ont été adoptées principalement sur les méthodes de communication[23, 24, 25].

Le 25 janvier 2012, la Commissaire et Vice-Présidente de la Commission européenne Viviane Reding a proposé un nouveau règlement en matière de protection des données[26]. Cette nouvelle règle a été motivée par deux raisons : l'ancienneté de la directive (aucune nouvelle directive importante depuis 17 ans) et la nécessité d'harmoniser cette directive à travers l'ensemble des Etats membres de l'Union européenne. Dans ce règlement, il est notamment prévu l'amélioration du droit du citoyen sur ses données en imposant plus de transparence en matière de traitement des données. Il est également prévu d'accorder au citoyen le droit à l'oubli (que ses données soient effacées du site). Il entend également conscientiser les acteurs et sociétés qui traitent les données. Enfin, il permet également d'alléger les mesures administratives pour la protection des données. Ainsi, l'obligation de déclaration pour le traitement des données est supprimée.

Cependant, cette déclaration a été accueillie de manière très mitigée par les différents États membres, réticents à l'idée d'une nouvelle directive. D'autant plus que la précédente directive est appliquée de manière différente selon les États. La Suède prône une application stricte de la directive tandis que le Royaume-Uni accorde plus de souplesse à certains points de la directive pour les industries. En Belgique, nous avons eu récemment une demande d'anonynimisation qui témoigne de l'application de cette directive. En effet, le 9 octobre 2012, le tribunal de première instance de Bruxelles s'est opposé à une demande d'anonymisation des données conformément à la disposition de la directive européenne. Les faits étaient les suivants : la partie demanderesse mentionnée dans plusieurs articles de presse pour des faits judiciaires passés a demandé à obtenir le droit à l'oubli : que ses données personnelles (nom et prénom) soient anonymisées. En effet, par les fonctions d'archivage des sites de presse spécialisés, nous pouvions retrouver aisément ses faits d'armes. Les sites de presse ont utilisé comme argument de défense le fait que ce n'était pas de leur ressort mais de celui des moteurs de recherche qui indexaient les pages de leurs sites. Le tribunal a reconnu que la diffusion des archives était bien sous l'application de la loi sur la protection des données à caractère personnel. Cependant, il a refusé de permettre l'anonymisation de ces données (pourtant suggérée par la directive) car les publications à des fins journalistiques ont une obligation d'informer les personnes. Dans ce cas-ci, ce jugement privilégie le droit à l'information au droit à l'oubli. Il s'agit d'un exemple typique de la latitude et du flou qui existent entre les directives et leurs applications[27].

Consciente de ces problèmes, Viviane Reding est revenue à la charge le 26 octobre 2012 en insistant sur la nécessité d'avoir une nouvelle directive mais en proposant plus de souplesse sur trois points : les PME, les actes délégués et les actes de mise en oeuvre et enfin, la flexibilité pour le secteur public. En effet, dans son précédent règlement, même si elle proposait déjà plus de souplesse pour les PME en ne les obligeant pas à nommer un responsable de la protection des données, elle obligeait les PME s'occupant du traitement des données à caractère personnel de respecter les mêmes règles que celle des grandes firmes. La Commission Européenne de Justice s'autorisait également à définir les formulaires de collecte d'informations à caractère personnel afin que ceux-ci soient uniformes et réglementaires. Viviane Reding a promis de réexaminer la cinquantaine d'actes proposés et de ne garder que les essentiels. Dans un premier temps, opposée à l'idée d'avoir un règlement pour le secteur public et un pour le secteur privé, elle propose d'adapter le règlement du privé au public. Comme l'a concédé Viviane Reding, parfois, nous sommes obligés d'avoir des règles spécifiques pour le secteur public, comme le cadastre. Comme vous avez pu le voir, de nombreux efforts ont donc été consentis par la Commissaire afin que ce règlement prenne enfin corps. Fin 2012, Viviane Redding a quitté la vice-présidence pour passer le relais à l'Irlandais Alan Shatter dont l'ambition reste la même que sa précédesseure : obtenir un accord politique en 2013 sur la protection des données à caractère personnel.

C'est dans cette optique que la Commission des libertés civiles, de la justice et des affaires intérieures a publié en janvier 2013 un projet de rapport sur la proposition de règlement du Parlement européen et du Conseil 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[28]. Ce rapport, communément appelé "rapport Albrecht" dû au nom de son rapporteur, Jan Philipp Albrecht, a suscité l'ire des archivistes et des généalogistes[29]. En effet, dans ce rapport, nous constatons un renforcement des mesures en matière de protection des données à caractère personnel. Tout devrait, en fait, être protégé. Les barrières pour lever cette protection seraient à la fois élevées et subjectives. Il est effectivement permis, dans ce nouveau rapport, que "toute recherche qu'elle soit menée par des universitaires ou des entreprises, y compris, par exemple, une étude de marché, pourrait servir de prétexte à la levée des protections prévues dans les autres parties présentes du règlement"[28]. Les exigences pour lever cette protection sont élevées et particulièrement arbitraires. Quel instrument va servir à mesurer la pertinence de la recherche, sa nécessité ? Une pétition a d'ailleurs vu le jour et compte déjà 40 000 signatures pour manifester contre ce rapport.

Si l'Europe se mobilise, nos voisins Français, ont également décidé de s'intéresser à la protection des données à caractère personnel. En effet, le gouvernement français a proposé à la fin de l'année 2012 un "Habeas Corpus

Numérique" pour assurer la protection des données à caractère personnel. Initié par la Ministre de l'économie numérique, Fleur Pellerin, il s'agirait "d'un projet de loi sur ces questions, sur un corpus de règles qui permettrait de garantir la protection des données personnelles et de la vie sur internet"[30]. Prévu dans le courant de l'année 2013, ce corpus sera réalisé en concertation avec le Ministre de la Justice et de l'Intérieur. Il faut savoir que ce corpus est né des suites des différents abus de grandes firmes de l'informatique comme Google ou Facebook sur les données à caractère personnel des utilisateurs (dont nous avons vu deux exemples ci-dessus). En octobre dernier, la Commission Nationale de l'Informatique et des Libertés (CNIL) avait d'ailleurs pointé du doigt les modifications récurrentes des règles de confidentialité de Facebook et ceci, sans en avoir averti les utilisateurs[31].

1.3 Conclusion

Comme nous l'avons vu précédemment, la protection des données est un sujet au coeur de l'actualité qui a fait et fait encore l'objet de nombreux débats et projets de règlement, directives et autres. Cependant, il n'existe pas une proposition qui fasse l'unanimité parmi tous les États membres (même si la directive de 1995 est la référence). Néanmoins, il faut saluer l'effort fait par le Gouvernement et le legislateur pour protéger nos données. Surtout en 2013, avec l'essor du cloud computing et la vague de données personnelles qui arrivent sans cesse sur internet, les méthodes de protection des données sont essentielles pour assurer leur intégrité et leur sauvegarde.

Chapitre 2

L'anonymisation des données : les données à protéger

L'examen du cadre législatif nous a montré que le traitement, la collecte et l'utilisation de nos données personnelles sont strictement réglementés et encadrés par les autorités compétentes, à la fois nationales et internationales. Grâce à ces réglementations, le législateur a posé les jalons de la protection des données à caractère personnel et nous ne pouvons que nous en féliciter. Cependant, l'établissement de ces directives, lois, règlements pose d'autres problèmes. En effet, les différentes procédures imposées pour protéger des données personnelles s'avèrent souvent lourdes, fastidieuses mais surtout coûteuses pour les sociétés. Afin de leur offrir une alternative, l'anonymisation des données a été introduite. Grâce à cette technique, les sociétés peuvent traiter des données sans devoir respecter la loi nationale sur la vie privée en vigueur (dans notre cas, la loi sur la vie privée de 1982). En rendant les données personnelles anonymes, le détenteur de ces données peut alors les utiliser à sa guise sans devoir respecter tel ou tel article législatif pour autant qu'aucune réidentification ne soit possible. Une recommandation du Conseil de l'Europe sur la protection des données à caractère personnel collectées et traitées à des fins statistiques donne une définition précise de l'anonymisation. L'anonymisation "consiste à supprimer les données d'identification afin que les données individuelles ne puissent plus être attribuées nommément aux diverses personnes concernées[...]. Le retrait des données d'identification ne met parfois pas totalement à l'abri d'une réidentification : le risque de dévoilement ne peut pas toujours être rigoureusement nul. Sans qu'on s'attache à définir un "risque acceptable", l'efficacité de l'anonymisation est d'une certaine relativité - nature des informations en cause, conditions de sécurité, date d'anonymisation, stade du traitement, etc."[32] Cette définition souligne deux points importants que nous développerons par la suite : le but de l'anonymisation, à savoir, l'impossibilité de trouver l'identité d'une personne et de découvrir ses données personnelles et la difficulté d'anonymiser les données et d'avoir un

résultat à 100% sécuritaire.

L'objet de ce chapitre est de proposer une démarche pour anonymiser de la meilleure façon qui soit ces données. En effet, nous avons vu qu'une mauvaise anonymisation pouvait avoir des conséquences catastrophiques sur les données présentes dans l'échantillon de données anonymisées. Pour réaliser cette démarche, nous avons examiné différents référentiels[33, 34, 35] qui définissent les bases d'une bonne anonymisation. Suite à leurs lectures, nous avons organisé ce chapitre de la manière suivante.

Section 1, nous définirons le contexte d'utilisation de l'anonymisation. Section 2, nous nous intéresserons à la nature des données à protéger. Des données, à la fois multiples et variées, qui se doivent d'être identifiées dans l'analyse et non lors de l'anonymisation. Section 3, nous définirons les objectifs et buts possibles poursuivis par l'anonymisation. Une étape cruciale qui définira de manière déterminante les formes d'anonymisation, au nombre de deux, qui seront examinées dans la section 4. Section 5, nous analyserons les diverses exigences demandées par le détenteur des données pour les protéger. Section 6, nous récapitulerons la démarche à suivre et nous tirerons les conclusions qui s'imposent.

2.1 Dans quel cadre utilise-t-on l'anonymisation ? L'anonymisation a priori ou a posteriori ?

De manière générale, l'anonymisation peut être utilisée sur n'importe quelle donnée personnelle qui doit être protégée. Cependant, à l'heure actuelle, nous retrouvons principalement l'anonymisation dans la recherche en santé, la planification des réseaux de santé et la surveillance de la santé publique mais également dans le secteur judiciaire, pour l'anonymisation des jugements même si cela reste, pour le moment, à l'état embryonnaire. En fait, l'anonymisation sera utilisée chaque fois que nous voudrons utiliser ces données à des fins statistiques et les divulguer à d'autres personnes. En effet, nous l'avons vu, la directive internationale impose de nombreuses contraintes pour utiliser ces données qui, au départ, ne nous appartiennent pas ; il faut le rappeler.

Cependant, avant d'entrer dans les étapes d'anonymisation, il faut faire la distinction entre deux types d'anonymisation : l'anonymisation a priori et l'anonymisation a posteriori. En effet, cette distinction n'existe pas dans la littérature consultée mais, à notre sens, celle-ci est nécessaire pour se rendre compte à quel point l'étape de collecte des données est cruciale dans la démarche d'anonymisation. En effet, c'est lors de la collecte des données que les données identifiantes sont réceptionnées, ensuite traitées et puis diffusées. Une intervention à la source du problème peut donc s'avérer utile dans le cas d'un traitement spécifique qui ne demanderait que quelques données et des données non identifiantes. De plus, nous l'avons vu dans le chapitre précédent,

certaines données sont collectées parfois sans raison valable et souvent de manière inutile pour l'organisme en question (mais destinées à d'autres sociétés sous-jacentes).

C'est pourquoi, pour limiter l'enregistrement massif de ces données, l'une des alternatives est d'effectuer une anonymisation a priori et de collecter directement de manière anonyme les informations et de n'inscrire aucune donnée directement identifiante. De cette manière, seules des données non identifiantes sont collectées. Grâce à ce processus, plus aucun autre traitement n'est nécessaire pour protéger ces données. En intervenant directement à la racine, nous éliminons déjà une grande partie du problème. Il faut savoir que la réalisation d'un formulaire anonyme est soumis à un règlement strict et doit se faire par des entreprises capables de pouvoir gérer et garantir cet anonymat. Le guide de la CNIL "Mesurer pour progresser vers l'égalité des chances"' donne d'ailleurs des grands principes sur la collecte de données de manière anonyme pour des enquêtes [36]. Premièrement, et de façon évidente, les données identifiantes ne doivent pas être collectées (nom, prénom, adresse postale ou électronique, date de naissance), de même que certaines données potentiellement identifiantes comme le poste du salarié interrogé, par exemple. Deuxièmement, aucun ordre ne doit être établi, de sorte que nous ne puissions pas retrouver ces données en croisant un autre tableau lié au premier. Troisièmement, en vue d'éviter de recouper certaines données entre elles, les propositions de réponses à certaines questions se doivent d'être vagues. Par exemple, si je recherche une information sur le salaire d'une personne, je dois lui proposer une tranche salariale (e.g. entre 20 000 et 30 000 euros) et non lui laisser un champ libre, pour qu'il inscrive de manière exacte son salaire. La CNIL précise également que si l'enquête se fait par internet ou par téléphone, de nouvelles mesures doivent être prises. Sur internet, le formulaire de statistiques doit déjà être préalablement rempli, la saisie des données identifiantes exclues et l'adresse IP de l'utilisateur non enregistrée. De plus, aucun historique de la session ne doit être gardé. En effet, le site web hébergeant le formulaire de statistiques ne peut permettre qu'un autre utilisateur arrive sur une session terminée, puisse revenir en arrière par le biais du navigateur internet et découvre alors les données que l'utilisateur précédent a inscrites. Par téléphone, l'anonymat est délicat dans la mesure où certaines données sont déjà collectées avant le questionnaire. Par exemple, le numéro de téléphone de la personne est connu ainsi que la liste nominative des salariés interrogés. Ceux-ci doivent donc être supprimés, de même que toute mention du nom et du prénom lors de la collecte des données.

Nous avons donc vu que selon la manière dont nous récoltons des données à des fins statistiques, que ce soit sur support papier, par internet ou par téléphone, la façon de procéder est différente et soumise à une réglementation et à un contrôle strict. Au vu de cette méthode, nous pourrions nous demander

pourquoi nous n'effectuons pas systématiquement une anonymisation a priori. Tout simplement parce qu'elle exclut directement un certain nombre de données qui pourraient se révéler utiles dans le cadre de telle ou telle analyse. En effet, il n'est pas rare d'avoir besoin de l'année de naissance ou de la localité d'une personne pour effectuer un sondage en particulier. L'anonymisation a priori trouve son utilité dans le cadre de statistiques portant sur un nombre restreint de données et où les données identifiantes ne sont pas nécessaires.

C'est pourquoi, dans ce chapitre, nous nous intéresserons à l'anonymisation la plus largement utilisée, l'anonymisation a posteriori. Il est évident que comme cette anonymisation s'effectue après la collecte des données, de nombreuses données sont encodées et sont donc potentiellement vulnérables. Nous allons donc définir dans cette seconde section un ensemble d'étapes allant de l'identification des données jusqu'à la forme d'anonymisation choisie afin de réaliser une anonymisation la plus efficiente et sécuritaire possible.

2.2 L'anonymisation en quatre étapes

2.2.1 Etape 1 : identifier la nature des données à anonymiser

La première étape avant d'anonymiser les données est d'analyser celles qui doivent l'être. En effet, les techniques employées seront différentes s'il s'agit de données statiques ou dynamiques, de données textuelles ou audiovisuelles. Dans le cadre de ce mémoire, nous nous sommes concentrés sur les données textuelles mais il faut savoir que des techniques d'anonymisation existent également pour les données multimédia (flouter l'image pour rendre la divulgation de l'identité impossible, réencoder du matériel audiovisuel en modifiant certaines données, etc.[34]).

Après avoir identifié la nature des données à analyser, il faut examiner quelles sont les données à anonymiser. S'agit-il de l'ensemble des données ou d'un échantillon de données ? Au sein de ces données, il faut examiner quelles sont les données identifiantes mais aussi les données critiques qui serviront à l'analyse. Il s'agit d'une étape déterminante car elle définira de manière formelle la façon de protéger et d'anonymiser les données. Une mauvaise analyse pourra, comme nous l'avons vu, avoir des conséquences non voulues ou, encore plus grave, procurer l'effet contraire, à savoir la divulgation de données potentiellement identifiantes. D'ailleurs, face à la vulnérabilité des données statistiques, l'Office National de Statistiques (ONS) a dû définir une série de guides et de méthodologies pour éviter de divulguer des informations trop sensibles à des fins statistiques[37].

Cette analyse des besoins doit également être envisagée en fonction de la (possible) menace. En effet, lors de la sélection des données, il faut mettre en balance, d'une part, la nécessité d'avoir cette donnée et, d'autre part, le risque

que cette donnée soit un jour dévoilée. Il s'agit donc de trouver le bon équilibre entre le besoin d'avoir cette donnée et la sécurité de celle-ci. Un exercice périlleux mais essentiel. Outre les données identifiantes, d'autres données doivent souvent être anonymisées. Malheureusement, il n'existe pas de liste exhaustive et, en fonction du contexte, certaines données seront dissimulées ou pas. Ces données peuvent concerner le statut professionnel d'un individu (profession exercée, rémunération), son statut social (réseaux d'ami, de parents, loisirs), son état de santé (données médicales), des caractéristiques propres à son physique (voix, visage, taille, corpulence), des caractéristiques culturelles (style d'alimentation, langue) ou encore l'appartenance religieuse (musulman, bouddhiste). La liste pourrait être encore longue mais stigmatise bien la première difficulté de l'anonymisation, la sélection des données à anonymiser. Outre celle-ci, le contexte d'utilisation, de diffusion et de disponibilité future de ces données représentent sans conteste la deuxième difficulté de l'anonymisation. En effet, il faut non seulement prendre en compte les données du tableau mais également le contexte de diffusion dans lequel ces données seront présentées. La diffusion de statistiques à un groupe de deux personnes ou sur internet, ne revêt pas la même complexité ni le même danger. De même, le danger ne sera pas le même s'il s'agit de données statiques ou de données mises à jour quotidiennement, de données disponibles pendant une durée limitée ou de données accessibles à vie.

Avant de choisir l'une ou l'autre forme d'anonymisation, il faut donc définir les objectifs poursuivis par cette anonymisation en terme d'utilisation, de diffusion et de disponibilité des données afin de pouvoir en identifier les dangers et proposer une forme d'anonymisation, dans la mesure du possible, en adéquation avec les besoins de l'utilisateur afin d'éviter des erreurs irréparables.

2.2.2 Etape 2 : identifier les besoins de l'utilisateur des données

Il est essentiel de s'intéresser aux objectifs de l'anonymisation car ce sont eux qui définiront la manière dont on doit anonymiser les données. Il faut noter que l'anonymisation est victime d'un paradoxe. En effet, les utilisateurs des données veulent les données les plus complètes, les plus fournies mais veulent également que ces données soient les plus sûres possibles. Il est techniquement impossible de proposer un tableau de données le plus complet possible avec un niveau de protection maximale.

Il faut donc faire un choix en fonction des besoins et du niveau de sécurité souhaité. Nous l'avons vu précédemment, certaines données peuvent être sensibles pour un tableau et totalement anodines pour un autre. Nous avons catégorisé les besoins en trois sous-sections : les besoins en terme de qualité, d'intégrité des données, de confidentialité ou de diffusion des données et de disponibilité, mise à jour des données. Nous avons effectué ce choix car cette

découpe est logique et témoigne du nombre de besoins différents auxquels l'utilisateur de ces données pourrait être confronté.

2.2.2.1 La qualité et l'intégrité des données

Nous l'avons dit plus haut : dans un monde idéal l'utilisateur voudrait avoir un degré de qualité équivalent à ces données d'origine une fois l'anonymisation effectuée. Or, ce n'est malheureusement pas possible. Cependant, nous pouvons définir avec l'utilisateur quelles sont les données qui doivent être impérativement gardées telles quelles et celles qui peuvent être modifiées à travers différentes méthodes que nous verrons dans le chapitre suivant.

Cette modification devra également être encadrée et définie. En effet, l'utilisateur devra nous dire si cette donnée sera utilisée ou non. D'une part, si celle-ci est utilisée, l'utilisateur devra définir de quelle manière il a besoin de ces données. Par exemple, a-t-il besoin de ces données pour estimer le nombre de personnes habitant telle région ? Auquel cas, nous pourrions modifier l'adresse en ne gardant que la localité. Ou, au contraire, a-t-il juste besoin d'un ensemble de données et, dans ce cas, nous pourrions modifier les données présentes sans respecter la donnée originelle de sorte que les statistiques globales soient respectées. D'autre part, si celle-ci n'est pas utilisée, la suppression pourra être envisagée.

Dans le deuxième cas de figure (modification des données sans respecter la donnée originelle), nous touchons clairement à l'intégrité des données. En effet, les données présentes dans le tableau anonymisé ne seront plus intègres car nous aurons touché aux données originelles. Parfois, nous sommes obligés d'utiliser ces techniques tellement les données sont identifiantes et donc, compromettantes. Il est évident que plus nous touchons à l'intégrité de certaines données, plus le tableau perd en qualité mais aussi en fiabilité. De plus, l'usage de ces données se révèlera limité au seul but initialement prévu. Toute réutilisabilité à d'autres fins sera, dans la plupart du temps, compromise. Il faut également noter que l'intégrité doit être analysée sur les données mais également une fois l'anonymisation effectuée. En effet, nous devrions vérifier si après avoir anonymisé les données, nous ne rencontrons pas de problèmes de synonymie ou d'homonymie. Par exemple, confondre les données d'un patient avec un patient précédent possédant le même nom.

En conclusion, des données trop généralisées seront très robustes du point de vue sécuritaire mais du point de vue qualitatif et informatif, presque nulles. Le but des données statistiques étant de donner un maximun d'informations, il faut trouver un juste équilibre pour, à la fois, avoir un bon niveau de protection et une lecture ainsi qu'une analyse des données permises, suffisamment profondes

2.2.2.2 La confidentialité ou la diffusion des données

Une autre question à se poser dans le cadre de l'anonymisation est de savoir si les données seront diffusées à grande échelle ou bien si elles feront l'objet d'un accès privilégié. En effet, des données fortement exposées seront plus sujettes aux attaques et devront faire preuve d'une robustesse plus accrue aux attaques que des données dont l'accès est limité. Dans le cadre d'une grande diffusion de données anonymisées, il faudra examiner les données présentes aux alentours afin de pouvoir estimer les risques d'inférence (que nous examinerons plus bas) mais aussi encadrer leur diffusion. Nous pourrions nous demander, au vu des risques encourus, l'utilité de diffuser ces données à large échelle mais cela s'explique par l'utilisation que nous voulons faire de ces données. Par exemple, en France, dans le domaine de la santé, suite à la loi N° 2004-810 du 13 août 2004 [38], le département de la santé a instauré un dossier médical personnel. Ce dossier a été mis en place afin d'offrir un véritable "curriculum vitae" de la santé d'un patient aux différents corps médicaux pour suivre un patient et pouvoir accéder directement aux données personnelles de celui-ci. Disponible sur internet (dmp.gouv.fr), cet outil de centralisation des données est un véritable réservoir d'informations sur un individu. En effet, celui-ci peut être consulté pour une hospitalisation, une demande d'adoption ou simplement un avis sur un enfant ou sur un adulte.

Avec un nombre de données importantes présentes sur ce portail, l'exigence de sécurité et d'anonymisation des données est très forte. En effet, si nous accédons au profil d'un patient, nous pourrions connaître son état de santé, ses différentes hospitalisations, ses pathologies, etc. Toutes des données qui, si elles ne tombent pas entre de bonnes mains, pourraient s'avérer être une arme dangereuse.

Bien évidemment, l'avantage de cette technique de centralisation et de diffusion, est de permettre le rassemblement de toutes les informations des patients en un seul et même endroit. Le suivi et le traitement du patient ne peuvent qu'en être améliorés. Malheureusement, cet outil, même s'il est anonymisé, revêt de par son caractère centralisé, un réél danger. En effet, plutôt que d'aller chercher et recouper les informations trouvées à différents endroits, l'intrus sait qu'il doit se rendre sur tel site pour obtenir un maximun d'informations. Il faut donc prendre conscience que, dans ce cadre-ci, même si le suivi des patients peut être amélioré, sa vie privée pourrait en faire les frais.

Outre ce danger, la centralisation des données de manière anonyme, pose également un problème au niveau éthique. En effet, selon le serment d'hypocrate, "admis dans l'intimité des personnes, je tairai les secrets qui me sont confiés"[39]. Par le développement de cette plateforme, tous les médecins pourraient, par exemple, avoir accès à des données d'une personne internée dans un asile. Le serment serait alors violé, car il ne s'agit plus que du médecin

traitant qui est au courant du cas de son patient, mais de l'ensemble du corps médical qui pourrait, potentiellement, avoir accès aux données le concernant.

A l'inverse, si nous restreignons l'accès aux données à un petit nombre de personnes, nous nous privons de l'utilisation future de nos données à d'autres fins. Par exemple, si nous disposons d'un tableau de données de personnes jugées et condammées, ces données une fois anonymisées pourraient être utilisées à des fins d'analyse sociétale ou comportementale.

La diffusion des données doit être longuement étudiée, pas seulement du point de vue du traitement que nous voulons effectuer mais également du potentiel d'analyse que ces données représentent. Malheureusement, nous devrons encore faire une fois un compromis entre, d'une part, la diffusion de l'information et, d'autre part, le danger par rapport à cette large diffusion.

2.2.2.3 La disponibilité et la mise à jour des données

Une fois le cadre d'utilisation des données établi, l'utilisateur de celles-ci devra définir pendant combien de temps ces données seront disponibles. Il est évident que le niveau de sécurité variera en fonction de la durée. Parfois, la durée sera définie dès le départ mais sera modifiée en cours de route. Il faudra alors réexaminer les différentes étapes afin de voir quelles sont les conséquences au niveau de la sécurité de nos données. Dans la même optique, la mise à jour des données doit être prise en compte. En effet, si le tableau de données évolue avec le temps, nous permettons à l'utilisateur d'avoir des données différentes pour la même personne. En publiant différents tableaux de données, nous donnons une opportunité à l'intrus de pouvoir effectuer des comparaisons en fonction des personnes présentes et puis absentes du tableau, des données modifiées, etc. Souvent négligé, cet aspect devra être pris en compte.

2.2.3 Etape 3 : choisir une forme d'anonymisation

Après avoir défini les données à anonymiser, précisé les besoins attendus par l'utilisateur de ces données ; le choix d'une forme d'anonymisation s'impose. Cette anonymisation peut revêtir trois formes différentes : l'anonymisation irréversible, l'anonymisation réversible et l'anonymisation inversible[40]. L'objectif de ces trois formes est le même : à savoir, rendre les données anonymes mais la manière d'utiliser ou de réutiliser ces données sera différente. Nous adopterons donc une forme par rapport à l'autre en fonction des besoins. Nous pourrions nous poser la question de la présence d'une forme d'anonymisation réversible dans ce cadre-ci d'analyse. En effet, alors que le but de l'anonymisation est de rendre impossible la réidentification, pourquoi la permettre dans ce cadre et l'inclure comme une forme d'anonymisation. Gilles Trouessin rejoint cet avis et estime quant à lui qu'il ne s'agit pas à proprement parler "d'une forme d'anonymisation"[41]. Dans ce mémoire, nous la

considérons pourtant comme telle car il s'agit avant tout d'une méthode qui rendra anonyme les données d'un tableau. De plus, à une fin d'étude et de comparaison, il est très intéressant de pouvoir confronter les deux méthodes afin d'en faire ressortir les avantages et inconvénients.

2.2.3.1 L'anonymisation irréversible

L'anonymisation irréversible est la méthode d'anonymisation par excellence. Elle consiste à rendre anonyme les données et ne permet pas de revenir aux données originelles. Nous verrons d'ailleurs que dans l'analyse des techniques étudiées, plus de 90% d'entres elles sont de cette forme. L'anonymisation irréversible se distingue donc de l'anonymisation réversible sur deux points : le caractére définitif de l'anonymisation et l'impossibilité de retrouver les données originelles. Par exemple, l'institut national de veille sanitaire a mis en place une application informatique pour la surveillance épidémiologique des maladies infectieuses à déclaration obligatoire comme le sida. Le but de cette application est de disposer à tout moment des informations sur "les maladies qui nécessitent une intervention urgente locale, nationale ou internationale et les maladies dont la surveillance est nécessaire à la conduite et à l'évaluation de la politique de santé publique" [42]. Elle a soumis son application à la CNIL qui a conseillé l'utilisation d'une anonymisation irréversible et double.

Cependant, même s'il s'agit de la technique la plus usitée, celle-ci peut être à l'origine de problèmes non décelés immédiatement. Par exemple, dans le cadre de l'échantillonnage de tissus et de données, les laboratoires anonymisent les données de façon irréversible[43]. Nous pourrions penser qu'il s'agit d'une bonne technique mais qu'en est-il des données des utilisateurs ? Par exemple, si Bob décide de donner des échantillons de tissus et que nous anonymisons son nom de manière irréversible, Bob n'a plus aucun moyen de retrouver des informations sur ses tissus. Plus grave, Bob n'a aucun moyen de contrôle sur ses données et ne sait donc pas à quelles fins elles seront utilisées. Ceci pose un problème d'éthique. En effet, les personnes qui mettent leurs échantillons et données au service de la science, spécifient dans quel cadre ils doivent être utilisés ou non. Malheureusement, comme ce procédé anonymise les échantillons de manière irréversible, nous n'avons aucune certitude quant à l'utilisation conforme de ces échantillons. Un autre problème que soulève l'irreversibilité, c'est l'incapacité d'être au courant des résultats des recherches et donc, d'en profiter.

C'est pourquoi, à des fins épidémilogiques, une autre méthode a été proposée, spécialement dans le domaine médical : l'anonymisation inversible, que nous allons exposer plus loin.

2.2.3.2 L'anonymisation réversible

L'anonymisation réversible signifie que nous pouvons, après avoir anonymisé une donnée, la récupérer et la désanonymiser. Cette méthode peut s'avérer utile dans le cadre de recherches sur des maladies à risque ou n'ayant actuellement pas de traitement efficace. En effet, les données de personnes atteintes du cancer peuvent être anonymisées afin d'effectuer une recherche sur cette maladie. En cas de résultat positif, il faudra alors désanonymiser les données afin d'administrer le traitement trouvé aux différents patients atteints de cette maladie.

Cette méthode peut donc s'avérer fort utile, notamment dans le domaine de la santé, mais pose évidemment un problème important. En effet, si nous pouvons anonymiser et désanonymiser, nous devons disposer d'une clé pour faire ce changement. Le choix du responsable de la clé devra alors être étudié. En effet, il faut être sûr que le dépositaire de la clé soit intègre et ne le transmette pas de manière abusive. La CNIL propose d'ailleurs de mettre en place "des mesures organisationnelles" comme diviser la clé en trois et la confier à trois personnes différentes[35]. De sorte que la réunion de deux des trois personnes permette de désanonymiser les données.

2.2.3.3 L'anonymisation inversible

La troisième forme d'anonymisation, l'anonymisation inversible, est à la croisée des chemins entre l'anonymisation irréversible et l'anonymisation reversible. En effet, il s'agit d'une part d'une sorte d'anonymisation irréversible, car le but premier est d'anonymiser des données à des fins de recherches mais d'autre part, il s'agit également d'une forme d'anonymisation réversible dans le sens où nous pouvons désanonymiser les données suite à des résultats importants mais uniquement sur demande exceptionnelle et uniquement pour les personnes habilitées à connaître ces données. Cette forme d'anonymisation est principalement utilisée dans le cadre médical. En effet, certaines recherches menées par différents médecins ou effectuées sur différents patients sont anonymisées de manière inversible. Cette technique consiste à remplacer les données nominales d'une personne par un pseudonyme. Par exemple, suite aux recherches fructueuses de Docteur Bob sur une maladie, nous avons décidé d'anonymiser de façon inversible ces recherches. Le nom de Docteur Bob se trouvant alors remplacé par un pseudonyme. Pour rétablir la véritable identité de Bob, seule "une autorité habilitée à évaluer les comportements des médecins pourrait rétablir les identités réelles des médecins" comme le confirme Gilles Trouessin[44]. Les personnes habilitées peuvent être des médecins conseils, des médecins inspecteurs, des médecins traitants mais ne doivent le faire que quand c'est nécessaire. Cette méthode d'anonymisation est également appelée "pseudonymisation" dans la littérature consultée.

Il faut noter que l'un des dangers de l'utilisation d'un pseudonyme provient de son utilisation. En effet, si celui-ci est utilisé de manière universelle, un recoupement des données par inférence sera toujours possible. Nous devons donc prendre en compte toutes ces données lorsque nous effectuons cette forme d'anonymisation. La différence principale entre l'anonymisation réversible et l'inversible est le caractère exceptionnel de la démarche de réversion. En effet, dans le cadre de l'anonymisation réversible, la réversion est permise et même encouragée. Dans le cadre de l'inversibilité, la réversion est considérée comme une exception et est très encadrée.

2.2.4 Etape 4 : satisfaire les exigences d'anonymisation

Nous l'avons vu, chaque forme d'anonymisation représente des dangers particuliers et aucune de ces formes n'est meilleure par rapport aux autres. Il s'agit juste de les utiliser à bon escient en fonction d'abord des besoins de l'utilisateur des données et, ensuite de la sécurité que le propriétaire de ces données est en droit d'exiger de la part du détenteur de ces données. Nous avonc identifié quatre exigences qui nous semblent fondamentales à prendre en compte en fonction de la forme utilisée : la robustesse à la reversion mais également à l'inférence, la non-chaînabilité et la non-obervabilité des données. Nous allons examiner chacune d'entre elles en détail et nous agrémenterons nos explications d'exemples explicites sur ce sujet.

2.2.4.1 La robustesse à la réversion

Dans les deux formes d'anonymisation (réversible ou inversible), nous pouvons retourner aux données originales. Une robustesse très forte est donc demandée pour que nous soyons certains que ces données ne puissent être reliées aux données nominatives de la personne. En effet, il faut que l'algorithme de cryptage soit suffisament robuste pour ne pas que nous retrouvions la clé de cryptage. Cette remarque vaut également pour le pseudonyme utilisé dans le cas d'une recherche épidémiologique. En effet, si la procédure de pseudonymisation n'a pas été suffisamment robuste, nous pourrions retrouver facilement l'identité réelle d'une personne simplement parce que le pseudonyme est peu sécuritaire ou trop souvent utilisé dans d'autres domaines. Par exemple, si nous utilisons le pseudonyme Fra, alors que nous nous appelons François, le risque est élévé de retrouver mon identité. De même, si nous utilisons trop souvent un type de pseudonyme, nous pourrions être retrouvé par croisement des données.

2.2.4.2 La robustesse à l'inférence

La principale qualité attendue de l'anonymisation est la robustesse des données face aux risques d'inférence ou de réversion des données (dans le cas

d'une anonymisation irréversible). L'inférence est une méthode qui consiste, à partir de différentes données, à retrouver des informations non dévoilées par simple déduction. Elle représente donc un des risques majeurs pour la confidentialité des données d'autant plus qu'elle peut revêtir différentes formes. Gilles Troussien et co distinguent quatre formes d'inférence qu'il faut prendre en compte : l'inférence déductive, l'inférence inductive, l'inférence abductive et l'inférence adductive ou probabiliste[41].

L'inférence déductive consiste à deviner un fait non mentionné suite à un enchainement d'événements. Par exemple, Bob va au tribunal, et peu de temps après, Bob va en prison. Nous pouvons en déduire que Bob a été condamné et qu'il était coupable. Contrairement à l'inférence déductive, l'inférence inductive ne se base pas sur un cas particulier mais sur une analyse globale. Par exemple, il est de notoriété publique que les japonais sont des gens stressés et donc plus sujets à des maladies cardiaques. L'inférence abductive consiste à établir à partir de faits une règle, un modèle et à définir des scénarios types. Par exemple, les policiers se rendent souvent au domicile de Bob, nous pourrions donc en déduire qu'il a des problèmes avec la justice. En établissant des règles, nous restreignons le champ des possibilités et nous éliminons les solutions improbables. Dans cet exemple, nous avons exclu la possibilité qu'il ait un(e) ami(e) travaillant pour la police. L'inférence adductive ou probabiliste consiste à obtenir la confirmation de "données sensibles déduites" à partir d'informations accessibles. Par exemple, Bob travaille dans une société d'informatique, nous savons qu'il est architecte et, vu son âge, nous arrivons à estimer son salaire en fonction du barème en vigueur dans le secteur.

Ces différentes formes d'inférence se basent donc sur des données connues qui servent ensuite de base à l'inférence. Mais comment ces données sont connues de l'intrus ? Selon Bruce Schneier, spécialiste en sécurité, il existe six types de données différentes présentes sur internet : les données de service, les données divulguées, les données confiées, les données fortuites, les données comportementales et les données dérivées [45]. Cette taxonomie est très intéressante car elle permet de se rendre compte à quel point le nombre de données à la disposition de l'intrus est grand et le risque d'inférence très élevé. Nous allons donc expliquer de manière succincte ces divers types de données. Les données de service sont les données que nous encodons lors de l'inscription à un réseau social. Cela peut aller de l'âge jusqu'à la carte de crédit. Les données divulguées sont les données que nous transmettons par l'intermédiaire d'un article, d'un blog, d'un site, d'un message, d'un commentaire ou de toute autre forme d'informations partagées. Les données confiées sont similaires aux données divulguées, à la différence que nous n'avons pas le contrôle sur celles-ci. Dans ce cas de figure, il s'agit par exemple de commentaires déposés sur la page d'une autre personne, d'une réaction à un article. Le créateur du message n'en a plus le contrôle une fois celui-ci déposé sur la

page de quelqu'un d'autre. Les données fortuites représentent les informations que les autres personnes déposent sur vous. Il s'agit évidemment du type de données le plus difficile à contrôler. Le tag, par exemple, sur une photo d'une personne est autorisé par Facebook, sans que la personne en question donne son accord pour être identifiée sur cette photo. Les données comportementales nous échappent également. En effet, il s'agit de données collectées en fonction de nos habitudes sur tel ou tel site. Par exemple, si Bob va sur un site de jeux vidéo et qu'il joue à des jeux orientés stratégie, nous pouvons en déduire qu'il aime ce style de jeux et lui proposer dans une newsletter, d'autres jeux de la même veine. Nous pouvons également analyser son parcours sur le mois, pour voir à quel moment il joue, pendant combien de temps. Nous pourrions ainsi déduire son emploi du temps, définir si Bob a un emploi ou pas, etc. Les données dérivées sont les données que nous pouvons déduire en analysant l'ensemble des données à notre disposition. Par exemple, si dans ses amis Facebook, Bob a 80% de ses amis qui jouent au tennis, nous pourrions en déduire que Bob joue probablement au tennis. Une avalanche de données très utile pour l'intrus et qui ne doivent pas être négligées dans le cadre de l'anonymisation.

2.2.4.3 La non-chaînabilité

La non-chaînabilité des données exprime le fait que deux opérations faites par une même personne ne peuvent être reliées entre elles. Nous ne pouvons pas faire le lien, par exemple, si un homme qui a mal au dos et va sur un site pour commander des calmants pour son dos sur un autre site pour s'acheter un matelas optimisé pour sa colonne vertébrale. Ces données doivent donc non seulement être anonymisées mais rendre l'enchaînement des actions impossible à retracer. Afin d'en comprendre l'importance, nous allons l'exemplifier à partir d'un cas concret dont les nouvelles modalités sont d'ailleurs très récentes (décembre 2012).

En 2009, la RATP et la STIF décident de remplacer la carte Orange par la carte Navigo. Le pass "Navigo" était une carte à puce qui contenait le nom, le prénom, l'adresse postale et une photo d'identité de la personne qui a souscrit à cet abonnement. De plus, par le biais d'une puce RFID, elle-même associée à un numéro d'identifiant, la RATP recensait les trajets effectués par l'utilisateur. Évidemment, l'utilisation des transports en commun est rendu plus agréable pour l'utilisateur[46]. En effet, celui-ci n'a plus qu'à scanner son pass sur une borne pour se rendre à la destination qu'il souhaite mais à quel prix ? En effet, de par la liaison entre le numéro d'identifiant et les données personnelles d'un individu, la RATP collectait de manière non anonyme des informations sur les habitudes des navetteurs. Quel train prenait-il le plus ? Vers quelle destination ? Combien de trajets effectue-t-il par mois ? Des questions auxquelles les données collectées permettaient d'apporter des réponses

fiables mais qui portent gravement atteinte à la vie privée. De facto, il s'agit d'un exemple de chaînabilité. En effet, si Bob prend le métro tous les jours (sauf le week-end) à la bouche de métro Abbesses, à Paris, pour se rendre à l'arrêt Argentine, également à Paris, nous pouvons en déduire que Bob travaille sur Paris et que son travail se situe dans une zone proche de l'arrêt en question. En effet, c'est grâce aux informations obtenues sur les différents jours que nous avons pu en déduire qu'il prenait le train. Si nous n'avions obtenu l'information sur son trajet que sur un jour, nous n'aurions pu établir les mêmes constatations avec certitude. C'est pourquoi la CNIL, lorsqu'elle a examiné le pass Navigo "avait rappelé que les usagers des transports publics ont le droit de voyager de manière anonyme[...]. En effet, avec le pass "Navigo", les données de validation (date, heure et lieu de passage) sont associées au numéro d'abonné ce qui les rendent nominatives. Les données sont conservées durant 48 heures à des fins de lutte contre la fraude"[47]. Suite à ces remarques, la RATP et la STIP ont proposé un nouveau pass "Navigo Découverte" dans lequel les données de validation étaient encore enregistrées mais cette fois, de manière anonyme. Cela permet de ne plus pouvoir relier un individu à un trajet effectué sur ces lignes. Néammoins, la CNIL a regretté que l'achat d'un pass "Navigo Découverte" soit payant et qu'il n'était pas valable pour "les tarifs réduits ou sociaux"[48].

2.2.4.4 La non-observabilité

La non-observatibilité est l'impossibilité de pouvoir observer si une opération est en cours. Cette dernière exigence prend tout son sens, par exemple, dans le milieu bancaire. En effet, il faut absolument que pendant la transaction aucune personne ne soit au courant de ce que nous sommes en train de faire et de qui nous sommes. Il faut, en effet, que nous puissions agir de manière anonyme, sans pouvoir être pisté ni suivi. Il faudra donc veiller à ne laisser aucune trace de notre passage pendant l'opération afin de permettre d'effectuer des opérations de manière anonyme.

2.3 Conclusion

Tout au long de ce chapitre, nous avons détaillé les démarches à suivre pour effectuer l'anonymisation des données demandées par le détenteur de celles-ci. Nous avons proposé une démarche en quatre étapes et insisté sur le fait que celle-ci devait s'effectuer pas à pas et de manière rigoureuse. Car, une fois la méthode choisie, il est très difficile de faire marche arrière. Mais surtout, il ne faut pas oublier que chacune des étapes est liée aux autres. En effet, une durée courte couplée à une disponiblité restreinte des données représentera une exigence en sécurité plus faible qu'une durée illimitée et une disponibilité

à grande échelle. De plus, dans ce deuxième cas, la risque d'inférence sera plus grand.

De manière générale et afin de bien comprendre l'enchaînement des actions, nous pouvons dire qu'une généralisation forte, une disponibilité courte, une diffusion restreinte offriront une sécurité maximale mais un degré de qualité, d'utilisation et d'informations faible. A l'inverse, une faible généralisation, une grande disponibilité, une large diffusion offriront une sécurité minimale mais un degré de qualité, d'utilisation et d'informations plus grand. Le paradoxe que nous avions évoqué plus haut prend ici tout son sens et il faut toujours garder à l'esprit qu'un équilibre doit être trouvé entre la sécurité d'une part et les données (dans toutes leurs formes d'utilisation possibles) d'autre part.

Chapitre 3

L'anonymisation des données : les techniques

Pour effectuer cette démarche, nous avons de nombreuses techniques d'anonymisation à notre disposition et nous allons les examiner en détail car ces techniques seront utilisées par la suite dans le cadre des modèles que nous avons analysés. Ce chapitre est inspiré du référentiel sur l'anonymisation[33] et du guide réalisé par l'autorité indépendante anglaise spécialisée dans la protection des données, l'ICO (Information Commissioner's Office)[34], qui recensent les différents moyens d'anonymiser les données. En fonction des effets des techniques sur les données, nous avons décidé de les diviser en trois parties : les techniques qui appauvrissent les données, celles qui les dégradent et celles qui ne les dégradent pas.

Au sein de ces parties, pour chacune des techniques, nous les analyserons comme suit : nous définirons la technique, ensuite nous l'exemplifierons pour enfin examiner les avantages et inconvénients de celle-ci. Cette approche nous permettra d'avoir déjà une bonne idée de la technique utilisée en fonction du contexte rencontré. De plus, ces techniques seront employées dans le dernier chapitre, consacré au cas pratique. Une bonne compréhension de ces techniques est donc nécessaire avant d'être en mesure de l'aborder.

3.1 Les techniques d'appauvrissement des données

Les techniques d'appauvrissement des données sont les techniques les plus faciles à mettre en place. Il s'agit de prendre le tableau de données qui sera mis à disposition de l'intrus et de se mettre à la place de celui-ci pour examiner quelles sont les données qui pourraient être esseulées ou directement identifiables. L'appauvrissement consiste dès lors à convertir ou supprimer une ou plusieurs données afin de rendre les données plus difficiles à identifier. L'appauvrissement est très utile car il permet de satisfaire un très grand nombre

de modèles d'anonymisation. Trois techniques permettent de mettre en place ce système : la suppression des données (globale ou locale), la généralisation globale et le masquage des données.

3.1.1 La suppression des données

Par définition, la suppression des données consiste à enlever des données ou une partie des données jugées compromettantes pour l'anonymisation correcte de la base de données. Cependant, le choix des données à supprimer n'est pas évident. En effet, il n'existe pas un modèle, une technique, qui permet de définir dans un tableau de données quelles sont les données à garder et celles à supprimer. La difficulté de cette technique réside donc dans le choix de la ou les données à supprimer. Il est certain que certaines données sont évidentes et doivent donc être supprimées comme les noms et prénoms. Pour les autres données, il faut appliquer la suppression au cas par cas en fonction du contexte d'utilisation et de l'objectif poursuivi par l'anonymisation.

N°	Age	Sexe	Code postal	Salaire annuel
1	**	*	****	******
2	<20	F	[6000-8000]	Élevé
3	>=30	F	[1000-2000]	Manquant
4	>=30	M	[1000-2000]	Faible
5	>50	M	[5000-5500]	Élevé
6	>50	M	[5000-5500]	Moyen

FIGURE 3.1 – *Tableau montrant un exemple de suppression en ligne et par élément*

La suppression des données peut se faire de deux manières : la première manière est de supprimer une colonne dans un tableau, c'est-à-dire toutes les données de celui-ci. Celle-ci peut, par exemple, être utilisée dans le cadre où ces données pourraient être trop identifiantes et dévoiler des informations qui pourraient être compromettantes et utiles pour la personne mal intentionnée. L'autre manière consiste à supprimer une ligne entière du tableau jugée trop identifiante par rapport aux autres données reprises. Par exemple, si nous avons un tableau de statistiques où nous n'avons que des femmes hormis un homme, nous devons supprimer la ligne concernant l'homme car cette information sera tout de suite trouvée par l'attaquant. La suppression peut également se faire localement dans le tableau de données. Par exemple, si l'attaquant cherche un homme d'une cinquantaine d'années et que dans le tableau, dans la colonne "âge", une seule cellule contient la donnée généralisée "50-55", cette donnée se doit d'être supprimée. La donnée supprimée sera remplacée par un texte indiquant "manquant". Évidemment, ce genre de procédé ne fonctionne que si le nombre de données est important. En effet, si

dans un tableau de données, il n'y a que des femmes et juste un homme, et que nous mentionnons dans la colonne "genre", "manquant" pour l'homme ; l'intrus aura 50% de chance de ne pas se tromper. L'un des inconvénients de la technique de suppression des données, c'est qu'elle dénature le contenu de la table et pose de facto un problème de réemploi dans d'autres domaines où les données auraient pu être sollicitées. Néammoins, cette technique est couramment utilisée, comme chez Bouygues Telecom[49].

3.1.2 La généralisation globale

N°	Age	Sexe	Code postal	Salaire annuel
1	<20	F	[6000-8000]	Faible
2	<20	F	[6000-8000]	Élevé
3	>=30	F	[1000-2000]	Moyen
4	>=30	M	[1000-2000]	Faible
5	>50	M	[5000-5500]	Élevé
6	>50	M	[5000-5500]	Moyen

FIGURE 3.2 – *Tableau généralisé de données*

La généralisation globale est une technique largement utilisée par de nombreux modèles que nous allons voir par la suite. Il s'agit de généraliser des données afin de rendre difficile la recherche d'une donnée et de proposer à l'intrus plusieurs choix possibles. Il est évident que plus le nombre de lignes similaires est important dans un tableau et plus la recherche de la part de l'intrus sera difficile. La généralisation peut se faire de différentes manières.

Premièrement, nous pouvons généraliser en plaçant la donnée à travers un écart. Par exemple, si une personne a 20 ans, nous mettrions dans sa colonne à l'emplacement de son âge, [20-25] ans. Si une autre personne a 22 ans, elle pourrait donc également se retrouver dans la même classification et donc, au lieu de retrouver directement la personne, l'attaquant aurait un doute puisque deux personnes figureraient dans le même écart. Deuxièmement, nous pouvons généraliser en bornant les données. Nous classifions alors l'âge en mettant des bornes inférieures, supérieures ou égales pour chacune des personnes, ce qui permettra également de regrouper les données. Dans notre tableau ci-dessous, une personne de 19 ans se trouvera dans la borne des <20 ans. Troisièmement, nous pouvons généraliser en textualisant des données chiffrées par des estimations. Par exemple, nous pouvons définir des catégories pour les salaires et estimer que si une personne gagne 30 000 euros, elle a un salaire faible et que si elle gagne 80 000 euros, elle a un salaire élevé. Cependant, il existe certaines catégories de données qu'il est impossible de généraliser comme le sexe d'une personne par exemple. Il faut noter que la généralisation doit être

appliquée avec modération car elle diminue la qualité des données et une trop forte généralisation pourrait rendre les données totalement inexploitables. Ci-dessus un tableau avec les variables "Age"', "Sexe", "Code postal" et "Salaire annuel" généralisées.

3.1.3 Le masquage des données

Le masquage des données est à la croisée de la généralisation et de la suppression. Cette technique consiste à masquer une partie des données tout en laissant l'autre partie dévoilée. Ce processus est souvent utilisé pour les données bancaires où nous masquons la majorité des chiffres sauf les quatre derniers, par exemple, XXXX-XXXX-3125. Nous utiliserons également cette technique dans la méthode d'anonymisation principale pour les codes postaux.

3.2 Les techniques de dégradation des données

Dans les techniques d'appauvrissement, nous avons vu que les données présentes dans le tableau une fois anonymisé étaient moins qualitatives et moins précises que les données originales. Cependant, aucune donnée n'a subi de détérioration. En effet, soit la donnée était absente, généralisée ou masquée mais aucune donnée présente n'avait été dégradée. A l'opposé, la deuxième technique que nous allons exposer dégrade les données afin de donner à l'intrus des informations incorrectes par rapport aux données qu'il recherche. Non seulement la qualité du tableau de données va en subir les conséquences mais également la véracité de celui-ci. Pour bien comprendre la différence entre ces deux techniques (appauvrissement et dégradation), nous allons l'exemplifier à travers le tableau de données de la figure 10. Si nous voulons anonymiser ce tableau par appauvrissement, nous pouvons utiliser l'une des techniques ou la combinaison de plusieurs techniques vues précédemment et les appliquer au tableau. Suite à ce traitement, nous obtiendrons un tableau qui aura par exemple perdu certaines données ou alors certaines données auront perdu en qualité. Cependant, les données présentes dans ce nouveau tableau représenteront l'exact reflet des données originelles (à l'exception des données supprimées). Par contre, si nous voulons anonymiser ce tableau par dégradation, nous allons toucher à l'intégrité des données et donc il risque d'y avoir un écart important entre la donnée originale et la nouvelle donnée anonymisée.

Ces techniques de dégradation doivent donc être utilisées avec une extrême prudence. Une mauvaise anonymisation pourrait avoir des conséquences désastreuses. Onze techniques de dégradation ont été recensées. Celles-ci peuvent être regroupées en deux types : les méthodes qui échangent les données (le data swapping, la "Post-Randomization", le rééchantillonnage) et les méthodes qui modifient les données du tableau (la micro-aggrégation, l'ajout du bruit, le

décalage, le viellissement, la génération et le remplacement de données, le chiffrement des données, la fonction de hachage, la concaténation et l'obfuscation).

3.2.1 Le data swapping

Le data swapping est une technique qui consiste à échanger les valeurs qui se trouvent dans les tables afin de rendre difficile leur identification[50]. Les valeurs qui sont échangées sont appelées "les attributs échangés", il peut s'agir de n'importe quelle valeur qui pose problème dans le tableau. Pour échanger une donnée, nous devons tenir compte des "attributs de contrainte". Il s'agit d'attributs auxiliaires à la donnée échangée mais similaires ou différents aux attributs auxiliaires de la donnée à échanger. Ceux-ci sont définis en fonction de l'objectif poursuivi. Le nombre de valeurs échangées dans le tableau fait l'objet d'un pourcentage appelé "taux d'échange" et noté "r". Celui-ci est souvent de l'ordre de 1 à 10%[34]. Par exemple, si nous examinons le tableau ci-dessous, nous constatons que deux données sont uniques dans le tableau, l'âge des lignes 1 [23-27] et 10 [56-60].

N°	Age	Sexe	Code postal	Salaire annuel	Dépense annuelle
1	[23-27]	F	6254	30 000	20 000
2	[28-35]	F	6040	32 000	10 000
3	[28-35]	F	1000	38 000	5 000
4	[36-40]	M	1024	40 000	15 000
5	[36-40]	M	6520	42 000	30 000
6	[41-50]	M	5245	43 000	6000
7	[41-50]	F	1000	38 000	5 000
8	[51-55]	M	1024	40 000	15 000
9	[51-55]	M	6520	42 000	30 000
10	[56-60]	F	5245	43 000	6000

FIGURE 3.3 – *Tableau de dix salaires*

N°	Age	Sexe	Code postal	Salaire annuel	Dépense annuelle
1	[56-60]	F	6254	30 000	20 000
10	[23-27]	F	5245	43 000	6000

FIGURE 3.4 – *Echantillon du tableau avec les deux données échangées*

Ces deux lignes représentent donc un danger puisqu'elles sont aisément identifiables. Nous allons appliquer la méthode du data swapping pour échanger ces données esseulées. L'attribut d'échange sera donc l'âge et l'attribut de contrainte sera le sexe de la personne. La seule donnée échangée sera donc

l'âge, le reste des données restant dans leur pristine état. En intervertissant ces données, nous perturbons la table et nous rendons la recherche par inférence fausse. En effet, si Bob sait qu'Alice a 21 ans, il saura qu'elle est en ligne 4 et qu'elle gagne un salaire faible et dépense beaucoup alors qu'en fait, elle a un gros salaire et dépense peu. En détournant les données, l'attaquant pourrait cibler la bonne personne mais tomber sur une ligne composée de données éronées. Le même problème se pose sur l'utilisation de ces données à des fins statistiques ou autres. En effet, en fonction des données échangées, l'examen du tableau des données donnera des réponses correctes et d'autres fausses. En effet, si un organisme de statistiques décide d'estimer la moyenne des salaires des personnes habitant dans telle région, les données seront correctes de même que si une analyse est effectuée sur la moyenne salariale en fonction des hommes et des femmes. Cependant, une analyse par âge de la moyenne salariale sera tronquée puisque les données ont été échangées et que dans notre cas de figure, une femme âgée gagne moins qu'une femme plus jeune alors que c'est l'inverse selon les statistiques originales[51].

Les attributs échangés doivent donc être choisis en fonction des objectifs poursuivis par l'anonymisation et en fonction de leur représentation globale dans le tableau de données. En effet, nous pourrions avoir le même cas de figure que celui exemplifié mais avec des salaires et revenus identiques. Le data swapping de ces deux données, dans ce cas-là, serait alors inefficace. L'inconvénient du data swapping est que la réutilisation des données à d'autres fins est de manière évidente très compliquée de par la présence d'informations fausées dans le tableau.

3.2.2 La technique de "Post Randomisation" (PRAM)

La technique de "Post-Randomisation" est une technique probabiliste qui consiste à perturber les données[52] en utilisant une matrice de Markov. Nous commencons par définir quelle variable du tableau nous voulons pertuber, ensuite nous définissons pour cette variable toutes les catégories possibles qu'elle peut avoir. Enfin, nous effectuons la permutation des données en appliquant une probabilité préalablement définie. Pour bien le comprendre, nous allons nous appuyer sur un exemple très clair de l'ICO que nous avons modifié pour les besoins de notre étude.[34].

Par exemple, dans le cadre de la figure 3.5, nous constatons que nous avons autant de personnes masculines que féminines. De manière évidente, nous allons choisir la catégorie du sexe car c'est cette catégorie qui nous permettra ensuite de perturber ce tableau afin de fausser certaines données.

Pour ce faire, nous allons supposer que la variable "Sexe" contient deux catégories : une catégorie masculine (1) et une catégorie féminine (2). Nous allons mettre ces catégories dans une matrice de Markov 2x2 où nous allons attribuer des probabilités pour les transitions selon la méthode de PRAM. Par

N°	Age	Genre	Code postal	Revenu
1	24	F	6040	20 000
2	27	M	6509	30 000
3	38	M	7040	40 000
4	39	F	8090	20 000
5	41	F	6000	60 000
6	53	M	5000	70 000

FIGURE 3.5 – *Tableau de six Belges*

exemple, nous pourrions dire que la probabilité qu'un homme soit un homme et qu'une femme soit une femme est de 90% et la probabilité qu'un homme soit une femme et qu'une femme soit un homme est de 10%. Avec ces probabilités, nous obtenons le nouveau tableau de la figure 3.6 où deux personnes ont eu le genre modifié comme l'atteste les lignes 1 et 3.

N°	Age	Genre	Code postal	Revenu
1	24	M	6040	20 000
2	27	M	6509	30 000
3	38	F	7040	40 000
4	39	F	8090	20 000
5	41	F	6000	60 000
6	53	M	5000	70 000

FIGURE 3.6 – *Tableau de six Belges post-randomizé*

Avec cette technique, nous constatons que la répartition homme-femme est respectée mais que nous ne sommes plus certain de la véracité de ces données. Il est en effet difficile pour l'intrus de trouver une donnée exacte même s'il arrive à retrouver les probabilités définies par la méthode de PRAM.

3.2.3 Le rééchantillonnage

Le rééchantillonnage est une technique qui consiste à échanger des données mais, cette fois-ci, par rapport à un échantillon plus large. Cette méthode fonctionne comme suit : premièrement, nous choisissons la variable que nous voulons modifier, ensuite nous examinons la répartition de celle-ci sur le tableau de données. Par exemple, sur le tableau ci-dessous, la donnée "revenu" est celle que nous voulons pertuber. En faisant la somme des revenus de chaque mois, nous constatons que pour les six personnes sur deux mois (janvier : 267 000, février : 132 000), cela représente 399 900 euros. C'est cette répartition que nous devrons respecter lors du rééchantillonnage des données. Par une méthode de distribution statistique, nous allons intervetir les données

du mois de janvier et de février tout en respectant la répartition globale des deux mois.

N°	Age	Sexe	Région	Revenu : janvier	Revenu : février
1	[56-60]	F	Bruxelles-Capitale	30 000	15 000
2	[23-27]	M	Flandres	45 000	30 000
3	[56-60]	F	Wallonie	20 000	20 000
4	[23-27]	M	Bruxelles-Capitale	60 000	25 000
5	[56-60]	F	Flandres	82 000	15 000
6	[23-27]	M	Wallonie	30 000	27 000

FIGURE 3.7 – *Tableau des revenus de six travailleurs indépendants*

N°	Age	Sexe	Région	Revenu : janvier	Revenu : février
1	[56-60]	F	Bruxelles-Capitale	30 000	45 000
2	[23-27]	M	Flandres	25 000	60 000
3	[56-60]	F	Wallonie	30 000	20 000
4	[23-27]	M	Bruxelles-Capitale	20 000	30 000
5	[56-60]	F	Flandres	27 000	15 000
6	[23-27]	M	Wallonie	15 000	82 000

FIGURE 3.8 – *Tableau rééchantillonné des revenus de six travailleurs indépendants*

Sur le tableau rééchantillonné, nous constatons que les revenus du mois de janvier et de février sont différents mais que la somme des deux mois est la même (399 900 euros). Même si elle respecte cette répartition, il faut noter que les moyennes mensuelles sont, par ce changement, inexactes. En effet, sur ce deuxième tableau, la somme du mois de janvier est de 147 000 au lieu de 267 000 dans le tableau original et pour le mois de février, 252 000 contre 132 000. Nous constatons que le rapport est inversé et qu'après rééchantillonnage, si nous réalisons des statistiques sur les revenus mensuels, celles-ci se révèleraient inexactes. Par contre, si les statistiques visent la rémunération globale des employés sur les deux derniers mois afin d'effectuer des moyennes, ces données rééchantillonnées seront parfaitement utilisables.

3.2.4 La micro-aggrégation

La méthode consiste à aggréger certaines données et à répartir celles-ci dans n-partitions pour rendre plus ardue la détection par l'attaquant. Cette méthode part du principe que certaines données pourraient être connues par l'attaquant. Imaginons que Bob veut sortir avec Alice. Lorsqu'ils ont parlé salaire, Alice lui a dit combien elle gagnait. Avant de sortir avec elle, Bob veut s'assurer qu'elle n'est pas trop dépencière et tombe sur le tableau de la figure

ci-dessous. Il constate tout de suite qu'Alice se trouve sur la première ligne et qu'elle ne fait aucune économie. Il y a donc un risque de divulgation de l'information. La micro-aggrégation va résoudre ce problème. En effet, cette technique va d'abord diviser les tuples en n-partitions. Dans cet exemple-ci, une partition correcte serait de diviser le tableau en 2-partitions. La première composée des trois premières lignes et la seconde, des trois dernières. Comme la donnée potentiellement identifiante pour Bob est le salaire, nous allons faire la moyenne des trois données de chacune des partitions et remplacer ces données par la moyenne de celles-ci.

N°	Age	Sexe	Code postal	Salaire annuel	Dépense annuelle
1	23	F	6254	30 000	20 000
2	24	F	6040	32 000	10 000
3	27	F	1000	38 000	5 000
4	40	M	1024	40 000	15 000
5	47	M	6520	42 000	30 000
6	52	M	5245	43 000	6000

FIGURE 3.9 – *Tableau de salaires*

Nous obtenons donc le tableau ci-dessous avec, dans la colonne salaire, deux salaires différents pour toute la table : 33 333 et 41 000. Lorsque Bob examinera le tableau micro-aggrégé, il ne saura plus si Alice est dépensière ou non. La migro-aggrégation est efficace sur un petit nombre de données. En effet, dans l'exemple, nous avons vu que la donnée "salaire" avait été modifiée et donc tronquée mais de manière légère. La différence entre la donnée initale et la donnée micro-aggrégée n'étant pas trop grande, la qualité des données n'est pas trop dégradée. Mais pour un grand nombre de données, il pourrait y avoir des disparités beaucoup plus grandes.

Par rapport à la généralisation, la micro-aggrégation détourne les données et donc les modifie. Dans la généralisation, nous avons une perte de granularité de l'information mais l'information présente reste exacte. Dans la micro-aggrégation, nous avons, dans l'exemple du salaire, une moyenne, et donc une information incorrecte. Cependant, l'un des avantages de cette méthode est qu'une fois la moyenne effectuée, il est impossible de pouvoir retrouver avec exactitude le salaire d'une personne.

3.2.5 L'ajout de bruit

La technique de l'ajout de bruit s'applique uniquement aux données numériques. Elle consiste à modifier ces données en fonction d'un nombre aléatoire e, avec une moyenne à zero et en fonction de la variance. Une forte variance diminuera la qualité de la donnée et, de fait, sa précision mais l'information ne sera pas exacte pour l'attaquant. Et vice-versa.

N°	Age	Sexe	Code postal	Salaire annuel	Dépense annuelle
1	23	F	6254	33 333	20 000
2	24	F	6040	33 333	10 000
3	27	F	1000	33 333	5 000
4	40	M	1024	41 000	15 000
5	47	M	6520	41 000	30 000
6	52	M	5245	41 000	6000

FIGURE 3.10 – *Tableau de salaires micro-aggregés*

Soit le tableau de données de la figure 2.8 et les données suivantes, nous obtenons le tableau avec du bruit suivant :

N°	Age	Genre	Code postal	Revenu
1	22	F	6224	20 000
2	26	M	6000	22 000
3	27	F	5203	32 000
4	35	M	1000	31 500
5	36	F	2506	68 000
6	40	M	1060	28 000

FIGURE 3.11 – *Tableau de six Belges avec leur salaires*

N°	Revenu	Valeur aléatoire E	E x 1000 (A)	Revenu + (A)
1	20 000	-0.171932015	-172	19 828
2	22 000	1.862281351	1826	23 862
3	32 000	0.959896624	960	32 960
4	31 500	-2.543129085	-2543	28 957
5	68 000	-1.049088496	-1049	66 951
6	28 000	-0.308324388	-308	27 692

FIGURE 3.12 – *Tableau avec du bruit de six Belges avec leur salaires*

Les données situées dans Revenu + (A) se retrouveront donc dans un nouveau tableau à la place des revenus initiaux. Cependant, nous pouvons constater dans l'exemple que l'écart n'est pas fort conséquent. Si l'attaquant cherche le salaire d'une personne, il pourrait savoir s'il a un gros, un moyen ou un petit salaire. Pour ce faire, il suffit qu'il ait une idée du salaire pour trouver la donnée qu'il veut ou qu'il soit lui-même dans le tableau et puisse alors établir une base de référence.

Aussi appelée "variance", cette technique introduit donc un paramètre qui va modifier la donnée originale. En effet, cette technique prend la donnée et la fait varier d'un certain niveau à un autre. Par exemple, pour le salaire, nous

pourrions établir une variance de -10% à +25% d'un salaire estimé à 30 000 euros.

3.2.6 Le décalage

Le décalage peut également être envisagé. Il s'agit de décaler les lettres, à l'image du code de César, en établissant une table de translation qui effectue le lien entre la donnée d'entrée et la donnée décalée. Évidemment, l'inconvénient de cette méthode est que si l'intrus tombe sur la table de translation, il n'aura aucun mal à faire la correspondance et à trouver la donnée qu'il convoite. En effet, cette méthode est vulnérable principalement à trois types d'attaques : l'attaque par analyse fréquentielle, qui consiste à reproduire deux graphiques, un graphique avec les différentes lettres et leur pourcentage d'apparition dans le texte codé, et un autre graphique avec le texte à déchiffrer, pour examiner les décalages ; l'attaque par brute force, qui consiste à essayer toutes les combinaisons possibles (surtout efficace si le nombre de décalages est limité) et l'attaque par recherche du mot probable, qui consiste à supposer qu'un mot que nous connaissons se retrouve dans le message crypté. Le mécanisme de décalage doit donc être évolué car de nombreux mécanismes de décalage ont déjà été percés[53].

3.2.7 Le viellissement

Le vieillissement des données consiste à modifier des données en les vieillissant ou en les rajeunissant. Cette méthode est surtout utilisée pour les dates. Par exemple, si nous avons dans notre tableau des données concernant des mineurs, nous pourrions vieillir leur âge ou modifier leur date de naissance afin que nous ne puissions plus savoir qu'ils ont moins de 18 ans. L'inconvénient de cette technique est qu'elle ne considére que la donnée en tant que telle et non le contexte dans lequel la donnée de la personne s'inscrit. En effet, comme le souligne l'AFCDP, "une utilisation non maîtrisée de la technique du viellissement donne quelques fois des résultats étonnants : un PDG âgé de 4 ans ou un collaborateur décédé avant d'avoir rejoint l'entreprise" [33]. Il faudra donc définir de manière rigoureuse le viellisement à effectuer en fonction du cadre dans lequel il est utilisé.

3.2.8 La génération et le remplacement de données

La génération de données est une technique qui consiste à créer, après extraction de données existantes, un tableau de données fictives qui reprend les mêmes caractéristiques à partir de règles de génération. Le remplacement par des données fictives est une technique similaire, à la différence que nous remplaçons directement dans le tableau des informations confidentielles par des

données fictives. Cette génération peut être automatique ou manuelle et peut concerner toute une série de données comme le code postal, l'âge, l'adresse, etc. Cette technique est souvent utilisée pour les jeux de tests[54] afin de ne pas divulguer des données confidentielles pour vérifier que telle méthode d'anonymisation fonctionne. Néammoins, l'inconvénient de cette technique est qu'il faut avoir analysé le tableau en profondeur pour établir les règles de génération. Ce processus peut se réveler long et fastidieux.

3.2.9 Le chiffrement des données

Une méthode classique et employée dans de nombreux services est le chiffrement des données. Il s'agit de crypter les données pour les rendre inviolables. Il existe deux types de chiffrement : le chiffrement symétrique et le chiffrement asymétrique.

Le chiffrement symétrique consiste à utiliser une clé de chiffrement pour lire un message. Par exemple, Bob envoie un tableau de données à Alice. Pour qu'Alice puisse lire le tableau, elle a besoin d'une clé. Cette clé est donc envoyée par un canal sécurisé afin qu'Alice puisse lire les données en question. La difficulté réside ici dans le canal de diffusion utilisé pour transmettre la clé et son niveau de sécurité. Il est évident qu'une transmission de la clé par e-mail représenterait un réel danger.

Le chiffrement asymétrique permet de résoudre le problème de la transmission de la clé en introduisant deux clés : une clé publique et une clé privée. Dans notre exemple, Bob enverrait alors un tableau de données en le chiffrant avec la clé publique. Pour qu'Alice puisse lire le fichier, elle devra utiliser sa clé privée.

Cependant, cette technique pose deux problèmes : la gestion de la clé de chiffrement et la lourdeur du processus. En effet, dans le cadre de la cryptographie des données, le processus est long et fastidieux. Nous devons utiliser une clé de cryptage à chaque fois que nous voulons travailler sur ces données. L'emploi des données ne peut donc être immédiat. De plus, la distribution de ces données à des fins d'analyse est également limitée. En effet,nous ne pouvons concevoir de diffuser cette clé à un grand nombre. Cela aurait pour conséquence de rendre plus fragile ce mécanisme de protection. C'est pourquoi, dans les méthodes exposées, aucune n'utilise ces techniques de cryptographie.

3.2.10 La fonction de hachage

Le calcul d'empreintes (hash) est une technique rapide et efficace. Elle consiste à prendre une ou des données nominatives en entrée et, grâce à une fonction de hachage, à produire une valeur numérique incompréhensible par l'utilisateur lambda en sortie. Dans l'exemple que nous avons plus haut, la CNIL avait recommandé d'utiliser une anonymisation irréversible pour le trai-

tement des données sur les maladies infectieuses. La méthode prônée a été celle du hachage. En effet, "à partir de la première lettre du nom de la personne, de son prénom, de sa date de naissance et de son sexe, un numéro d'anonymat sous forme d'une chaîne de 16 caractères en majuscule" a été réalisée[55]. Cette méthode a d'ailleurs été couplée à une autre méthode que nous avons vue, la chiffrement des données, "à partir de l'identifiant et d'une clé secrète détenue par l'INVS"[55]. Grâce à cette double anonymisation, aucun lien ne pourra être fait entre la donnée et l'identifiant. En effet, d'après la CNIL, la combinaison de ces deux méthodes permet de "satisfaire aux exigences de confidentialité dans la mesure où elle s'accompagne de mesures de sécurité efficaces"[55].

La CNIL prône cette méthode car elle représente trois avantages importants que nous allons reprendre ci-dessous. Premièrement, le caractère réellement irréversible de la méthode en soulignant la "quasi-impossiblité mathématique de retrouver à partir du résultat final les données" hachées [55]. Deuxièmement, le risque de collission s'avère très faible. En effet, le risque d'une collission "entre deux des six milliards d'habitants sur terre serait inférieur à un sur plusieurs milliards de milliards"[55]. Troisièment, les algorithmes proposés pour le hachage sont performants. Néammoins, la CNIL souligne que cette méthode reste vulnérable à un type d'attaque : l'attaque par dictionnaire. L'attaque par dictionnaire consiste à s'octroyer l'identité d'une personne, à appliquer la fonction de hachage et à se présenter comme telle personne pour obtenir des données personnelles de l'individu que nous prétendons être. Pour contrer ce genre d'attaques, la CNIL propose "d'intégrer une clé secrète qui permet de vérifier lors de la présentation ultérieure d'une valeur de hachage que celle-ci est bien authentique"[55]. Pour ce faire, elle propose de réaliser une clé mais de la diviser en cinq fragments. Chacune de ces parties contenant des informations redondantes. Ces cinq parties seront ensuite distribuées à cinq personnes différentes. Afin d'éviter les problèmes de disponibilité et les éventuelles absences, la réunion de seulement trois parties de la clé permettra de reconstituer la clé d'origine. Et, afin d'avoir une sécurité encore plus optimale, la CNIL propose d'effectuer cette fonction de hachage deux fois et avec des clés différentes[55].

3.2.11 La concaténation

La concaténation permet de regrouper plusieurs données en une seule afin de rendre la lecture de ces données incompréhensible pour l'intrus. La concaténation est principalement utilisée pour les jeux de tests.

3.2.12 L'obfuscation

L'obfuscation est une technique qui consiste à ajouter des éléments, à créer du bruit dans les données afin de noyer les vraies données au sein des données fictives. Grâce à l'ajout de données fictives, nous ne touchons pas aux données d'origine mais nous rendons la recherche plus compliquée pour l'intrus. Cette technique permet de rajouter de l'incertitude dans la table en incluant de fausses données dans celle-ci. Pour prendre un cas pratique, en février 2008, le site note2be.com ouvre ses portes. Ce site propose à n'importe quel étudiant de coter son ou ses professeur(s)[56]. Dénoncé par la CNIL car portant atteinte à la protection des données personnelles[57], ce site a fait l'objet d'une pratique d'obfuscation par les internautes scandalisés par l'existence de celui-ci. En effet, comme vous pouvez le voir sur l'illustration ci-dessous le nom et le prénom de la personne et le lieu où il exerce son métier étaient affichés sur le site. Autant de données qui, comme nous l'avons vu, pourraient servir de base pour une attaque par inférence. Les internautes ont créé de faux profils de professeurs sur ce site, utilisant des noms de personnalités fictives ou décédées comme Karl Marx ou Achille Talon afin de décrédibiliser ce site et d'enfouir la véritable information au sein d'informations incorrectes ou improbables. En mars 2003, le site a été obligé d'enlever les données nominatives des professeurs suite à une décision de justice[58].

FIGURE 3.13 – Capture d'écran du site internet note2be avec les données personnelles des professeurs

.

Cet exemple nous montre bien l'utilité de cette technique, à savoir rendre pour l'intrus une donnée véritable aussi incertaine qu'une donnée fictive. L'inconvénient est également évident : en ajoutant des données fictives aux tableaux de données, nous rendons les statistiques et analyses sur notre tableau partiellement faussées. Le niveau d'incorrection étant proportionnel aux types de statistiques demandées.

3.3 Les techniques non-dégradantes

Face à la dégradation des données, deux autres techniques ont été proposées qui, à l'inverse des techniques dévellopées ci-avant, ne touchent pas aux données mais soit les échantillonnent (en enlevant juste les éléments identifiants) soit les réagencent en fonction des besoins. Ces techniques, bien qu'ambitieuses, sont très délicates et malheureusement, perméables aux diverses attaques que nous verrons par la suite.

3.3.1 L'échantillonnage

L'échantillonnage est une technique qui est possible uniquement si nous avons un grand nombre de données à notre disposition. Une méthode simple est donc de prendre un échantillon de ces données et d'en supprimer les données identifiantes. La sélection des données à echantillonner peut se faire soit de manière aléatoire soit par un échantillonnage probabiliste comme celui de Bernouilli (basé sur l'échantillonnage de poisson)[34]. L'échantillonnage reste néammoins dangereux car comme il s'agit d'une sélection aléatoire, nous pourrions prendre des échantillons de données dont une donnée serait très particulière et donc identifiable. Imaginons que nous échantillonnons des données médicales, sur le tableau de données, nous retrouvons différentes maladies communes et une rare. En effet, une personne dans le tableau est atteinte de progeria, une maladie génétique très rare qui provoque le viellessiment accéléré de la personne. Avec trois cas en France et vingt-cinq en Europe[59], l'intrus pourra facilement confondre plusieurs données et retrouver la personne atteinte de cette maladie. Un cas réel témoigne d'ailleurs de cette problématique.

En 2007, la société Nefix, spécialisée dans le cinéma, décide de publier une base de donnée de 100 000 millions d'avis de 480 000 utilisateurs sur 18 000 films différents[60]. Les données contenaient juste la date, le film, l'avis et le groupe ID. Les données identifiantes ayant été supprimées. Fier de leur système, elle a mis au défi quiconque pourrait deviner de quelles personnes il s'agissait. Et le défi a été relevé, puisque l'étudiant Arvind Narayanan et le professeur Vitaly Shmatikov ont dévoilé qu'ils avaient pu retrouver deux individus au sein de cette liste. Ces mêmes chercheurs ont également rappelé

que supprimer uniquement les nom et prénom n'était en aucun cas une preuve d'anonymisation[61].

3.3.2 La tabulation de données

La technique de "cross-tabulation data" s'intéresse aux variables multiples dans un tableau. Cette technique utilise les données pour refaire un tableau différent en condensant les données et en éliminant les lignes afin que nous ne puissions plus identifier un individu directement.

Le revers de la médaille de cette technique est que si nous n'avons aucune donnée disponible, nous donnons une information à l'attaquant. Par exemple, sur le tableau 3.14 qui reprend les grades des étudiants ayant passé l'examen de sécurité, nous voyons que nous pourrions les organiser d'une manière différente comme sur le tableau 3.15.

N°	Age	Grade "examen sécurité"
1	M	Distinction
2	M	Grande distinction
3	F	Satisfaction
4	M	Satisfaction
5	F	Distinction
6	M	Satisfaction
7	F	Insufissant
8	F	Distinction

FIGURE 3.14 – *Tableau de six étudiants en sécurité informatique*

Genre	I	S	D	GD
Homme	0	2	1	1
Femme	1	1	1	0
Total	1	3	2	1

FIGURE 3.15 – *Tableau des résultats de l'examen en sécurité informatique par genre*

Cette organisation donne une autre lecture du tableau et concentre les données différemment. Cependant, elle peut donner d'autres révélations à l'attaquant. En effet, au vu du tableau, nous pouvons être sûrs de deux faits : aucun étudiant masculin n'a raté l'examen de sécurité informatique et aucune étudiante n'a réussi l'examen avec grande distinction. Il s'agit d'un exemple typique qui illustre le danger d'une anonymisation incorrecte qui, avec des ambitions louables, pourrait dévoiler plus d'informations qu'elle n'en cache.

3.4 Conclusion

Suite à l'examen de ces différentes techniques, nous avons pu constater les avantages et les inconvénients de chacune d'entre elles. Nous avons vu qu'il n'y avait pas de technique meilleure ou moins bonne qu'une autre mais que chacune d'elle trouvait son utilité en fonction du contexte et des données à anonymiser. De plus, cette analyse nous a permis de constater que la majorité des techniques étaient irréversibles.

En effet, si nous examinons le tableau de la figure 3.16 récapitulant les techniques classifiées selon les catégories et les formes d'anonymisation (à l'exception de la forme d'anonymisation inversible trop spécifique au secteur d'activité médicale), nous constatons que la seule méthode réversible disponible est le chiffrement. Ce tableau est révélateur de deux constatations : premièrement, le choix en matière d'anonymisation réversible est pour l'heure limité. Deuxièmement, nous constatons que peu importe la méthode utilisée (appauvrissante, dégrandante ou non), celle-ci s'avère être irréversible.

Il faut également noter que ces techniques peuvent être appliquées de concert. Nous pourrions en effet avoir une combinaison de méthodes appauvrissantes, dégradantes et non dégradantes pour satisfaire une demande d'anonymisation bien spécifique. Afin de nous en rendre compte, nous allons examiner un modèle d'anonymisation, k-anonimity et ses dérivés, pour montrer l'utilisation combinée de plusieurs techniques d'anonymisation. Ce modèle nous permettra également de mettre en évidence certains problèmes ou contraintes non perceptibles à la lecture des techniques précédemment analysées.

Technique	Réversible	Irréversible
Les techniques d'appauvrissement des données		
La suppression des données		X
La généralisation globale		X
Le masquage des données		X
Les techniques de dégradation des données		
Le data-swapping		X
Le rééchantillonnage		X
Le Post-Ramdomization (PRAM)		X
La migro-aggrégation		X
Le décalage		X
Le viellissement		X
La génération des données		X
Le chiffrement	X	
La fonction de hachage		X
La concaténation		X
Les techniques non-dégradantes		X
L'échantillonnage		X
La tabulation des données		X

FIGURE 3.16 – *Tableau des techniques avec leurs formes d'anonymisation*

Chapitre 4

Les modèles d'anonymisation

Nous avons identifié une technique d'anonymisation principale, k-anonymity et diverses variantes que nous allons exposer de la manière suivante. Tout d'abord, nous définissons chacune des méthodes de manière formelle, ensuite nous expliquons cette définition à l'aide d'un exemple et enfin, nous présentons les avantages et inconvénients de chacune de ces méthodes.

4.1 K-Anonymity

4.1.1 Définition

Face aux problèmes évoqués ci-dessus, Sweeney a proposé une méthode largement adoptée maintenant afin de rendre les données d'une table anonyme[62]. Une table respecte les principes de k-anonymity s'il existe au moins k-1 fois la même classe d'équivalence. C'est sur cette méthode que vont se baser les autres modèles d'anonymisation (dont une liste non exhaustive est proposée par [63]). Afin de comprendre cette méthode, nous allons nous appuyer sur les différentes définitions établies par Sweeney [64].

Données à caractère personnel					
Nom	**Prénom**	**Naissance**	**Nationalité**	**Sexe**	**Code postal**
Dupond	Georges	1982	Belge	Masculin	6223
Henderson	Stephanie	1986	Belge	Féminin	6041
Dellapiana	Vanessa	1978	Italien	Féminin	60944
Nakamura	Hiroshi	1965	Japonais	Masculin	98033
Proto	Sylvio	1968	Belge	Masculin	9800
Scifo	Enzo	1978	Italien	Masculin	60944

FIGURE 4.1 – *Tableau de données à caractère personnel*

Définition 1 (attributs) *Soit $B(A_1, ..., A_n)$ une table avec un nombre fini de tuples. Le nombre fini d'attributs de B est $B(A_1, ..., A_n)$ d'après[64].*

Parmi les attributs, il faut en distinguer trois : les attributs identifiants, les attributs non sensibles et les attributs sensibles. Les attributs identifiants sont, par exemple, le nom, le prénom ou encore l'adresse postale. Comme il s'agit d'informations trop évidentes pour être laissées, ces attributs sont éliminés. Dans un tableau k-anonymisé, il ne reste donc que les attributs sensibles et les attributs non sensibles. Les attributs non sensibles sont les quasis-identifiants et les attributs sensibles sont les données qui ne peuvent être dévoilées sous aucun pretexte (santé, information judiciaire, salaire, etc.). Dans la figure 3, la colonne "casier judiciaire" est une colonne d'attributs sensibles. En effet, il s'agit d'attributs qui ne doivent pas être révélés à n'importe quel individu. Ce sont des données confidentielles.

Définition 2 (quasi-identifiants) *Soit une population d'entité U, une table d'entité spécifique $T(A_1, ..., A_n)$, $f_c : U \to T$ et $f_g : T \to U'$, où $U \subseteq U'$. Un quasi-identifiant de T, écrit Q_r, est un ensemble d'attributs $(A_1, ..., A_n) \subseteq (A_1, ..., A_n)$ où : $\exists p_i \in U$ tels que $f_g(f_c(p_i)[Q_t]) = p_i$ d'après[64].*

Les quasi-identifiants sont toutes les données qui permettent de retrouver un individu à partir de ceux-ci. Sweeney, par exemple, nous explique que selon une étude, 87% de la population américaine possède des caractéristiques qui les rendent uniques simplement grâce à leur code postal, leur sexe et leur date de naissance.[62]. En effet, nous constatons que, malheureusement, même en supprimant les attributs identifiants, nous pouvons retrouver les informations convoitées simplement par *linking* (inférence). Par exemple, si nous examinons la ligne 1 de la figure 1, nous voyons que Georges Dupond est un homme belge qui habite au 6223 et est né en 1982. Supposons maintenant que l'intrus recherche justement cet individu et veut savoir s'il a un casier judiciaire. Grâce aux données qu'il a collectées dans le tableau de la figure 3, il n'aura aucun mal à trouver l'information qu'il souhaite dans un tableau anonymisé.

Quasi-identifiants			
Date de naissance	**Nationalité**	**Sexe**	**Code postal**
1982	Belge	Masculin	6223
1986	Belge	Féminin	6041
1978	Italien	Féminin	60944
1965	Japonais	Masculin	98033
1968	Belge	Masculin	9800
1978	Italien	Masculin	60944

FIGURE 4.2 – *Données quasi-identifiantes*

Définition 3 (k-anonymity) *Soit $RT(A_1, ..., A_n)$ une table et QI_{RT} un quasi-identifiant qui lui est associé. RT satisfait à la règle de k-anonymity si et seulement si chaque séquence de valeurs dans $RT[QI_{RT}]$ apparaît avec moins de k occurences dans $RT[QI_{RT}]$ d'après[64].*

Age	Nationalité	Code postal	Casier judiciaire
26	Belge	6041	Meurtre
24	Belge	6024	Meurtre
18	Italien	5396	Meurtre
39	Japonais	98033	Vol
35	Belge	98000	Possession de stupéfiants
15	Italien	5345	Rien
14	Belge	5346	Vol à l'étalage
37	Américain	98011	Vol
25	Français	6040	Meurtre

FIGURE 4.3 – *Données du commissariat de police*

4.1.2 Technique de k-anonymisation

Pour arriver à rendre une table k-anonyme, Sweeney utilise deux techniques que nous avons précédemment vues : une technique de généralisation des données et de suppression de certains tuples[65],[13]. Nous allons l'expliciter ci-dessous.

Définition 4 (généralisation) *Soit $T_i(A_1, ..., A_n)$ et $T_j(A_1, ..., A_n)$ deux tables avec le même nombre d'attributs. T_j est une généralisation de T_i, écrit $T_i <= T_j$ si : 1.$|T_i| = |T_j|$. 2.$\forall A_z \in A_1, ..., A_n$: $dom(A_z, T_i) <=_D dom(A_z, T_j)$ 3. Nous pouvons définir une trajectoire bijective entre T_i et T_j qui associe chaque tuple $t_i \in T_i$ avec un tuple $t_j \in T_j$ tel que $t_i[A_z] <=_v t_j[A_z]$ pour tout $A_z \in A_1, ..., A_n$ d'après[64]*

Un attribut quasi-identifiant peut être généralisé. Si nous prenons un code postal, par exemple, 6224, nous pouvons le généraliser en enlevant le chiffre le plus à droite, comme suit : 622* et ainsi de suite, jusqu'à arriver à la généralisation la plus complète 6***. Nous pouvons également généraliser une date de naissance en inscrivant juste l'année de naissance ou encore généraliser une ville en indiquant dans quelle province elle se situe.

Chaque attribut fait partie d'un domaine composé d'*integer*, par exemple, pour un code postal ou de *string* pour la province de la ville. Pour définir une relation entre deux domaines, Sweeney propose la formule suivante : $D_i <= D_j$[64]. Cela signifie que D_i est une généralisation de D_j. Il définit deux

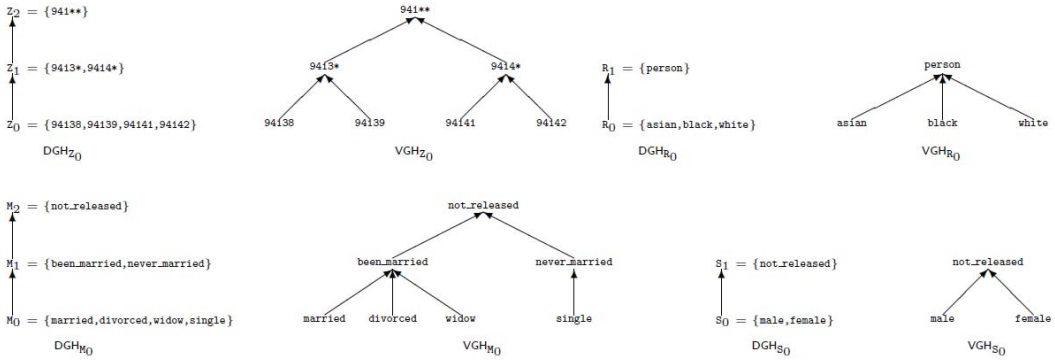


FIGURE 4.4 – Exemple de généralisation DGH_d d'après [13]

conditions à cette généralisation. Premièrement, les attributs peuvent être généralisés jusqu'à arriver à un seul domaine de généralisation (par exemple, 6***) et sont ordonnés. Deuxièmement, cet ordre est hiérarchique et est appelé "*domain generalization hierarchy*" noté DGH_d. Chaque domaine hiérarchique de généralisation contient des valeurs hiérarchiques de généralisation. Celles-ci peuvent être représentées sous la forme d'un arbre dont le sommet est la valeur maximale du domaine et les feuilles, les valeurs suivantes avec un degré de généralisation en moins en fonction des niveaux. Sur la figure 4, l'ensemble R_0 contient toutes les races présentes dans le tableau (*asian*, *black*, *white*) et l'ensemble plus général R_1 les regroupe sous le terme générique de "person". R_1 est une généralisation de R_0.

Une fois la relation et la hiérarchie d'un domaine établies, nous pouvons définir une stratégie de généralisation. Par exemple, si nous prenons un tuple d'attributs (code postal, date de naissance), nous voyons sur la figure 4.4 qu'il y a trois chemins qui partent des attributs non généraux vers le domaine de généralisation le plus complet (6***, 1***). Nous avons donc le choix entre trois stratégies à adopter.

La généralisation permet de rendre les données plus difficiles à trouver mais cela se fait aux dépens de la qualité et de la pertinence des données disponibles. Si nous généralisons certaines données, nous ne pourrons plus les utiliser après pour d'autres traitements. La généralisation doit donc être la plus minimale possible afin de généraliser uniquement ce qui est nécessaire. Nous en reparlerons à la section 5 lors de l'implémentation algorithmique de cette méthode.

4.1.3 Avantages et inconvénients de cette méthode

L'avantage de cette technique est qu'elle permet de protéger les données contre les attaques par inférence et qu'elle les garde intègres. Cependant, la

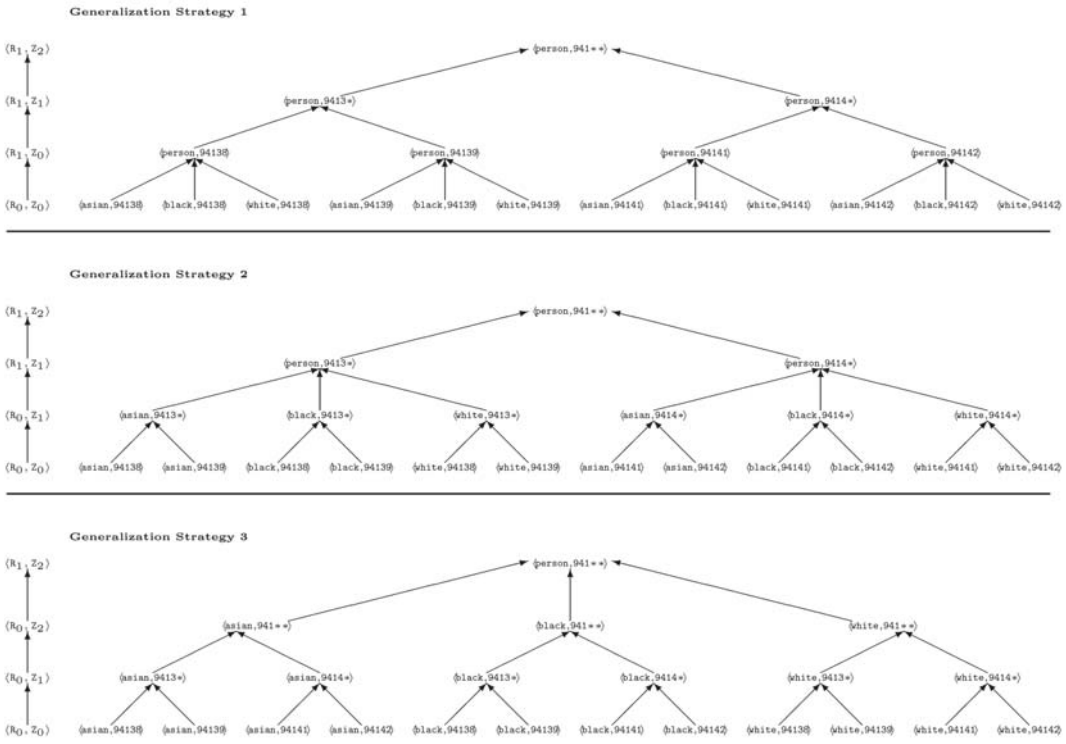


FIGURE 4.5 – Exemple de domaine de généralisation DGH_d d'après [13]

méthode k-anonymity présente des défauts. Elle ne sait pas résister à cinq types d'attaques : les *matching attacks*, les *complementary attacks*, les *temporal attacks*, les *homogeneous attacks* et les *background knowledge attacks*. Nous allons les examiner une par une.

4.1.3.1 Matching attack

Les *matching attacks* consistent à trouver les données confidentielles convoitées par l'intrus en partant du principe que ces données anonymisées sont dans le même ordre que les données normales. Une solution intuitive est de mettre ces données anonymisées dans un ordre aléatoire, mais cela ne suffit pas pour empêcher d'autres types d'attaques.

4.1.3.2 Complementary attack

La complémentarité de deux tables anonymisées s'avère suffisante pour pouvoir trouver la table d'origine. La solution pour éviter ce problème est

d'examiner en premier lieu toute les tables anonymisées existantes avant d'anonymiser une nouvelle table.

4.1.3.3 Temporel attack

L'anonymisation des données évolue dans le temps. Il y a des données nouvelles qui entrent et d'autres qui sont supprimées ou mises à jour. Cette évolution des données pose un souci en terme de protection des données. En effet, l'intrus peut, à l'aide de deux tables généralisées prises à différents moments, retrouver les données originales.

4.1.3.4 Homogeneous attack

Soit l'exemple suivant : Alice rencontre Bob dans la rue. C'est le coup de foudre. Alice et Bob s'échangent alors leur numéro de téléphone. Lors de sa conversation avec Bob, Alice apprend que Bob a 29 ans, qu'il est belge et qu'il habite à Gosselies (6041). Ayant eu de mauvaises expériences par le passé, elle décide de vérifier si son futur petit copain n'a pas de casier judiciaire. Lors de ses recherches, elle tombe sur la table 3-anonyme suivante de la police de Charleroi. Avec les informations dont elle dispose, trois choix s'offrent à elle : le 7, 8, 9. Or, dans la colonne casier judiciaire, nous constatons qu'ils ont chacun été condamné pour meurtre. Il ne fait donc aucun doute que Bob a un casier et qu'il a commis un meurtre.

N°	Age	Nationalité	Code postal	Casier judiciaire
1	<20	*	53**	Meurtre
2	<20	*	53**	Rien
3	<20	*	53**	Vol à l'étalage
4	>30	*	980**	Vol
5	>30	*	980**	Possession de stupéfiants
6	>30	*	980**	Vol
7	2*	*	60**	Meurtre
8	2*	*	60**	Meurtre
9	2*	*	60**	Meurtre

FIGURE 4.6 – *Données du commissariat de Police 3-anonymisée*

Cet exemple nous montre les limites de cette méthode car, en anonymisant certaines données, nous ne sommes pas à l'abri d'un manque de diversité des données sensibles. C'est logique car la méthode k-anonymy se concentre sur les attributs qu'elle a généralisés et non sur les attributs sensibles.

4.1.3.5 Background knowledge attack

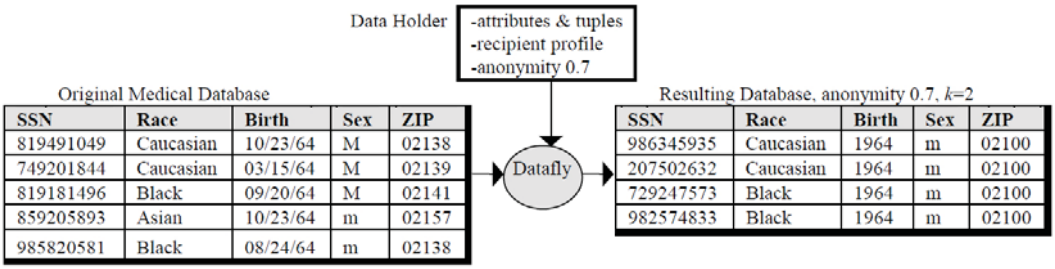
La difficulté pour l'anonymisation des données est que nous ne savons pas ce que l'autre sait. Nous n'avons donc pas une vue complète sur le comportement et les actions de l'intrus. Or, c'est justement le *background* d'une personne qui lui permettra de trouver les données et les informations qu'il recherche. L'intrus peut donc aisément grâce à ce *background* trouver les informations qu'il désire. Par exemple, l'intrus veut savoir quel type de maladie a son voisin. Il sait qu'il est grec et qu'il a 29 ans. Imaginons que l'intrus tombe sur un tableau 3-anonymisé dans lequel il a plusieurs choix possibles. Trois choix s'offrent à lui : problèmes cardio-vasculaires, cancer ou arythmie. Comme il sait que les personnes méditerranéennes n'ont presque jamais de problèmes au coeur (grâce à leur cuisine à l'huile d'olive), il en déduit que son voisin a probablement un cancer.

4.1.4 Implémentation algorithmique des méthodes

Afin de comprendre la manière d'implémenter un modèle, nous avons décidé, comme il s'agit de la méthode principale et uniquement pour celle-ci, d'analyser quelques algorithmes qui mettent en place de différentes manières les techniques de généralisation et de suppression. L'étude algorithmique est nécessaire car elle permet d'être conscient des difficultés d'application du modèle théorique mais également de ses points forts. Dans le cadre de cette analyse, nous en avons selectionné six qui sont les plus connus et les plus intéressants à examiner.

4.1.4.1 Le système Datafly

Le système Datafly a été mis en place par Sweeney[66] pour permettre, de manière automatique, par généralisation, substitution et suppression de transformer les données présentes dans un tableau en un ensemble de données qui permet de diminuer les attaques par *linking* et par *matching*. Le système est présenté à la figure 10.



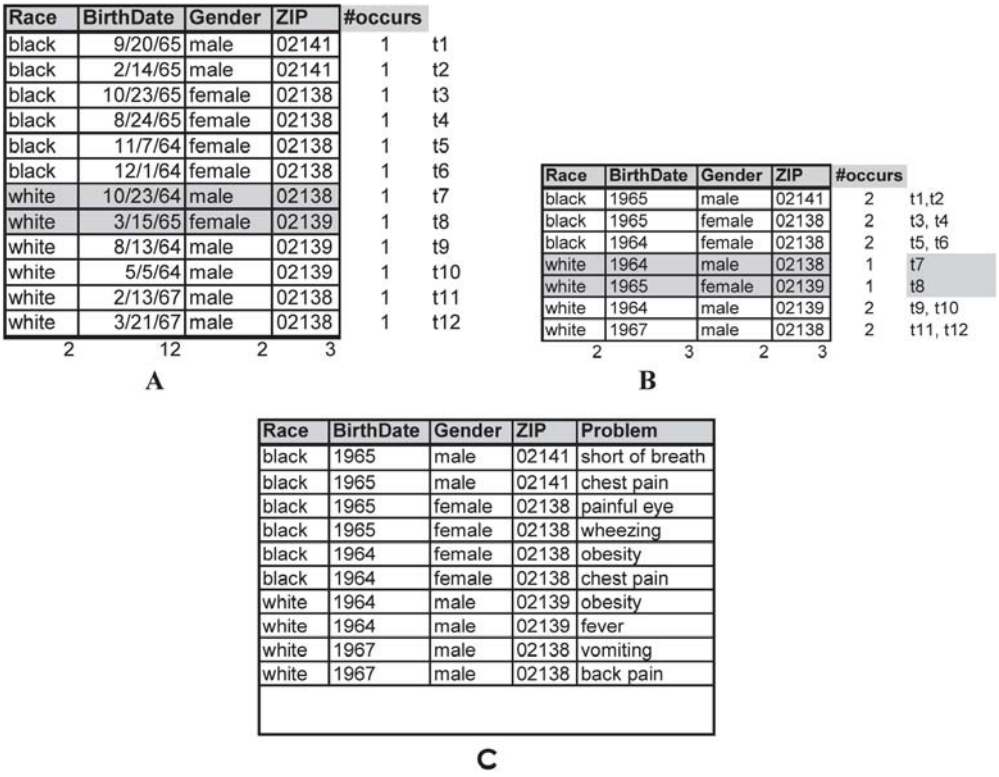
SSN	Race	Birth	Sex	ZIP
819491049	Caucasian	10/23/64	M	02138
749201844	Caucasian	03/15/64	M	02139
819181496	Black	09/20/64	M	02141
859205893	Asian	10/23/64	m	02157
985820581	Black	08/24/64	m	02138

SSN	Race	Birth	Sex	ZIP
986345935	Caucasian	1964	m	02100
207502632	Caucasian	1964	m	02100
729247573	Black	1964	m	02100
982574833	Black	1964	m	02100

FIGURE 4.7 – Illustration du système Datafly d'après [66]

À gauche, nous avons un extrait d'une base de données médicales et au milieu, le Datafly qui prend en entrée le tableau de données médicales suivant les exigences du détenteur des données (*data holder*). Le résultat est le tableau de droite k-anonymity.

Avant d'examiner la méthode, nous pouvons déjà constater que (1) la ligne 5 contenant la race *Asian* a été supprimée, (2) les dates ont été généralisées à l'année de naissance, (3) les numéros de sécurité sociale ont été remplacés par des données alternatives et (4) les codes postaux ont été modifiés en mettant les deux derniers chiffres à 0. Le tableau de droite est donc 2-anonymisé.



Race	BirthDate	Gender	ZIP	#occurs	
black	9/20/65	male	02141	1	t1
black	2/14/65	male	02141	1	t2
black	10/23/65	female	02138	1	t3
black	8/24/65	female	02138	1	t4
black	11/7/64	female	02138	1	t5
black	12/1/64	female	02138	1	t6
white	10/23/64	male	02138	1	t7
white	3/15/65	female	02139	1	t8
white	8/13/64	male	02139	1	t9
white	5/5/64	male	02139	1	t10
white	2/13/67	male	02138	1	t11
white	3/21/67	male	02138	1	t12
2	12	2	3		

A

Race	BirthDate	Gender	ZIP	#occurs	
black	1965	male	02141	2	t1,t2
black	1965	female	02138	2	t3, t4
black	1964	female	02138	2	t5, t6
white	1964	male	02138	1	t7
white	1965	female	02139	1	t8
white	1964	male	02139	2	t9, t10
white	1967	male	02138	2	t11, t12
2	3	2	3		

B

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

C

FIGURE 4.8 – Les différentes étapes de l'algorithme Datafly d'après [66]

Dans le système Datafly, c'est le détenteur des données qui va définir les différentes opérations à effectuer sur les attributs du tableau. Celles-si peuvent se résumer en cinq opérations spécifiques : il établit quels sont les attributs et les tuples qui doivent être traités, il les regroupe en sous-ensembles de quasi-identifiants, pour chaque attribut dans chaque quasi-identifiant, il leur assigne un poids de 0 à 1. 0 signifie peu probable et 1 signifie fort probable. Ensuite, le détenteur des données spécifie un degré de k-anonymity et enfin, il spécifie pour chaque attribut s'il veut qu'il soit fortement modifié ou pas du tout. 0

signifie à ne pas modifier et 1 à généraliser fortement. [65]. Pour l'exemple ci-dessous, nous supposons ici que le degré de k-anonymity est de 2.

Une fois les conditions d'initialisation établies, l'algorithme peut être résumé en cinq étapes : (1) l'algorithme calcule le nombre d'attributs différents dans chaque ensemble quasi-identifiant et le nombre d'occurrences de chaque séquence. A la figure suivante, nous constatons que le nombre d'occurrences est de 1 pour chaque séquence et que le nombre d'attributs différents est le suivant : 2 pour la race, 12 pour la date de naissance, 2 pour le sexe et 3 pour le code postal. (2) Tant qu'il y a un nombre d'attributs distincts supérieur à k, il faut généraliser. Sur notre figure, la date de naissance contient donc le plus grand nombre de valeurs distinctes. C'est ce groupe d'attributs qui va être généralisé. Suite à la généralisation de la date de naissance en année, nous constatons sur la figure 4.9 qu'il n'y a plus que 3 attributs différents pour la date de naissance et que le nombre d'occurences est égal à 2, excepté pour t7,t8. (3) Ces deux tuples sont supprimés car ils ne sont pas égaux à k-anonymity. (4) Cette suppression doit respecter les principes de k-anonymity. Enfin, (5) le système produit la table généralisée que nous pouvons voir à la figure suivante[65].

L'avantage de cet algorithme est qu'il satisfait toujours le principe de k-anonymity. Par contre, il ne garantit pas une généralisation minimale. Le désavantage est qu'il généralise "toutes les valeurs associées à un attribut" [65] et que la valeur du tuple qui ne respecte pas le principe de k-anonymity est supprimée.

4.1.4.2 Le système Mu-Argus

Cette méthode a été inventée par Hundepool et Willenborg[67]. Le fonctionnement est similaire à celui de Datafly. Le détenteur des données spécifie une valeur de k-anonymity et définit quel attribut est sensible à travers quatre valeurs : "*not identifying*", "*most identifying*" , "*more identifying*" et "*identifying*". Le système fonctionne comme suit : (1) en premier lieu, il vérifie le nombre d'attributs distincts et le nombre d'occurrences de chaque séquence. Ensuite, (2) les valeurs de chaque attribut sont généralisées afin de satisfaire la k-anonymity. Puis, (3) le programme va essayer plusieurs combinaisons de deux ou trois valeurs pour voir lesquelles posent problème. Il place ces combinaisons problématiques dans un *outlier*. Après, (4) et (5) il choisit les attributs à généraliser parmi les *outliers*. Enfin, (6) il supprime les valeurs dans chaque combinaison d'*outliers*.

Soit le tableau de quasi-identifiants de la figure ci-dessous et les valeurs des attributs suivants : race <*identifying*>, birth <*most identifying*>, sexe <*more identifying*> et code postal <*more identifying*>. Nous constatons que la combininaison de deux attributs *most* et *more* (à savoir la date de naissance et le code postal) montre que le tuple t8 ne satisfait pas le principe de k-anonymity.

Il est donc placé dans les *outliers*. Sur la figure, dans la colonne *outliers*, nous voyons aux lignes 4 et 5 toutes les combinaisons de quasi-identifiants qui ne satisfont pas aux principes de k-anonymity.

Birth	ZIP	occurs	sid	outliers
1965	02141	2	{t1,t2}	{}
1965	02138	2	{t3,t4}	{}
1964	02138	3	{t5,t6,t7}	{}
1965	02139	1	{t8}	{}
1964	02139	2	{t9,t10}	{}
1967	02138	2	{t11,t12}	{}

V

Race	Birth	Sex	ZIP	occurs	sid	outliers
black	1965	male	02141	2	{t1,t2}	{}
black	1965	female	02138	2	{t3,t4}	{}
black	1964	female	02138	2	{t5,t6}	{}
white	1964	male	02138	1	{t7}	{}
white	1965	female	02139	1	{t8}	{{birth,zip}}
white	1964	male	02139	2	{t9,t10}	{}
white	1967	male	02138	2	{t11,t12}	{}

freq

FIGURE 4.9 – Test sur les combinaisons et les fréquences de Most x More d'après [67]

Race	Birth	Sex	ZIP	occurs	sid	outliers
black	1965	male	02141	2	{t1,t2}	{}
black	1965	female	02138	2	{t3,t4}	{}
black	1964	female	02138	2	{t5,t6}	{}
white	1964	male	02138	1	{t7}	{{**birth**,sex,zip}, {race,**birth**,zip}}
white	1965	female	02139	1	{t8}	{{**birth**,zip}, {**sex**,zip}, {**birth**,**sex**,zip}, {race,**birth**,**sex**}, {race,**birth**,zip}, {race,**sex**}, {race,**birth**}}
white	1964	male	02139	2	{t9,t10}	{}
white	1967	male	02138	2	{t11,t12}	{}

FIGURE 4.10 – Tableau des fréquences avant suppression d'après [67]

Un défaut de cette technique soulevé par [65] est qu'elle n'examine que deux ou trois combinaisons et pas toutes les combinaisons possibles. Or, il se pourrait qu'il y ait quatre ou plusieurs combinaisons. De plus, cette technique ne garantit pas la k-anonymity au contraire du système Datafly. L'exemple de la figure 4.10 l'atteste. Elle ne permet pas de résister aux attaques par *linking* et *matching*.

4.1.4.3 Algorithme de généralistion minimale

Après le système Datafly, Samaraty et Sweeney ont proposé un algorithme mininal[64]. Le but de cet algorithme est de trouver un nombre k-minimal de généralisations. Il définit *le k-minimal generalization* comme une table k-

id	Race	BirthDate	Gender	ZIP
t1	black	1965	male	02141
t2	black	1965	male	02141
t3	black	1965	female	02138
t4	black	1965	female	02138
t5	black	1964	female	02138
t6	black	1964	female	02138
t7	white		male	02138
t8	white			02139
t9	white	1964	male	02139
t10	white	1964	male	02139
t11	white	1967	male	02138
t12	white	1967	male	02138

MT

id	Race	BirthDate	Gender	ZIP
t1	black	1965	male	02141
t2	black	1965	male	02141
t3	black	1965	female	02138
t4	black	1965	female	02138
t5	black	1964	female	02138
t6	black	1964	female	02138
t7	white	1964	male	02138
t8	white		female	02139
t9	white	1964	male	02139
t10	white	1964	male	02139
t11	white	1967	male	02138
t12	white	1967	male	02138

MT actual

FIGURE 4.11 – Tableau final obtenu après l'exécution de l'algorithme μ-Argus d'après [67]

Race:R_0	ZIP:Z_0
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142

PT

Race:R_1	ZIP:Z_0
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142

$GT_{[1,0]}$

Race:R_1	ZIP:Z_1
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*
person	9414*

$GT_{[1,1]}$

Race:R_0	ZIP:Z_1
asian	9413*
asian	9413*
asian	9414*
asian	9414*
black	9413*
black	9413*
black	9414*
black	9414*
white	9413*
white	9413*
white	9414*
white	9414*

$GT_{[0,1]}$

Race:R_0	ZIP:Z_2
asian	941**
asian	941**
asian	941**
asian	941**
black	941**
black	941**
black	941**
black	941**
white	941**
white	941**
white	941**
white	941**

$GT_{[0,2]}$

Race:R_1	ZIP:Z_2
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**
person	941**

$GT_{[1,2]}$

FIGURE 4.12 – Plusieurs tables k-anonymisées d'après [64]

anonymisée qui n'est pas une généralisation d'une autre table k-anonymisée. Il établit également un seuil maximum de suppression des données, intitulé MaxSup [64]. Pour une même table de données, plusieurs tables généralisées sont donc possibles afin de satisfaire la k-anonymity. Par contre, une question se pose : jusqu'où généraliser ? Car si cette méthode est efficace, elle touche quand même à la qualité des données de la table. Une table généralisée est par essence moins qualitative et moins informative que la table d'origine. Il faut donc généraliser ce qui doit l'être, sans plus. L'algorithme de généralisation minimale vérifie les différentes généralisations possibles d'une même table et

elle vérifie chacune des tables généralisées entre elles afin d'examiner laquelle peut être reconstruite à partir d'une autre. Au fur et à mesure de sa progression, elle ne va garder que la table optimale. Un exemple complet de la démarche est proposé dans l'article[64].

4.1.4.4 K-optimise

FIGURE 4.13 – Index de l'algorithme K-Optimise d'après [68]

Agrawal et Bayardo ont proposé un autre type d'algorithme intitulé *K-Optimise*[68]. Cet algorithme se base sur l'établissement d'une structure et d'un index (composé d'un integer). Par exemple, si nous regardons la figure ci-dessus, nous constatons qu'il y a deux types de quasi-identifants : la race et le code postal. Dans chaque quasi-identifiant, il y a plusieurs catégories d'attributs. Pour la race, il y a trois catégories constituées d'intervales : [Asian], [Black], [White] et pour le code postal, quatre intervalles : [941038], [94139], [94141] et [94142]. Si nous suivons l'ordre hiérarchique suivant, race puis code postal, nous pouvons mettre en-dessous de ces différentes catégories un numéro de gauche à droite. Ainsi les catégories race sont numérotées de 1 à 3 et les catégories code postal sont numérotées de 4 à 5. Première constatation : les valeurs les plus faibles de chaque catégorie peuvent être supprimées car elles seront comprises dans la généralisation. Dans notre exemple, il s'agit du 1 [Asian] et du 4 [94138].

Après avoir établi les différents principes, l'algorithme *K-Optimise* construit un *set enumeration tree* à l'aide des index des ensembles. La racine est l'ensemble vide. Pour établir quelle est la meilleure stratégie de généralisation, nous comparons à chaque noeud la valeur de ce noeud par rapport au meilleur coût trouvé jusqu'alors. Si c'est la valeur du noeud actuel qui est la meilleure alors elle devient le meilleur coût et remplace l'ancienne valeur[51]. Mais si c'est le contraire, alors nous ne parcourons plus les fils du noeud puisque nous allons simplement enlever ce noeud de notre recherche puisqu'il n'est pas optimal. Pour un exemple du dévellopement de l'arbre, nous vous invitons à lire l'article de Agrawal et Bayardo[68].

4.1.4.5 Incognito

Une autre technique basée sur l'algorithme *Breadth-first search*, Incognito, a été développé par LeFevre, DeWitt et Ramakrishan [69, 70]. Soit une table

privée donnée, Incognito va générer toute les tables *k-minimal* possibles.
La méthode de cet algorithme est la suivante : (1) il examine chaque attribut pour vérifier si celui-ci respecte le principe de k-anonymity. Si ce n'est pas le cas, il ne généralise pas ces attributs. (2) Ensuite, il combine les différents attributs restants et effectue cette vérification sur les paires d'attributs. (3) Enfin, il effectue cette même opération sur les triples et ainsi de suite, jusqu'à avoir parcouru l'ensemble des quasi-identifiants.

Pour comprendre cette méthode, nous allons nous appuyer sur l'exemple proposé par Ciriani et al.[51] qui, dans la figure 4.15, expose les différentes étapes de l'algorithme.

Race	**Sexe**	**Statut marital**
Asiatique	Féminin	Divorcée
Asiatique	Féminin	Divorcée
Asiatique	Féminin	Mariée
Asiatique	Masculin	Mariée
Asiatique	Masculin	Mariée
Noir	Féminin	Célibataire
Noir	Féminin	Célibataire
Blanc	Féminin	Célibataire
Blanc	Féminin	Veuve

FIGURE 4.14 – Table privée contenant des données médicales d'après [51]

Soit la figure ci-dessus, les quasi-identifiants (race, sexe et statut marital) et un tableau devant être 2-anonymisé, nous constatons que les quasi-identifiants race et sexe respectent ce principe mais pas le quasi-identifiant statut marital. Lors de la première itération, nous pouvons voir sur la figure 10 à droite qu'Incognito va supprimer M_0. Ensuite, lors de la deuxième itération, Incognito examine les couples (race, sexe), (race, statut marital) et (sexe, statut marital). Le premier satisfait les principes de 2-anonymity. L'examen du second et du troisième permet de voir que certains couples ne le respectent pas. Ils sont donc supprimés. Il s'agit des couples (R_0, M_0) et (R_1,M_0) pour (race, statut marital) et des couples (S_0, M_0) et (S_1, M_0) pour (sexe, statut marital). Enfin, la troisième itération effectue le même contrôle mais sur les trois identifiants. Nous pouvons voir le résultat sur la figure 10. Il faut noter que dans l'exemple que nous avons pris, nous voyons tout de suite que la ligne qui pose problème dans la généralisation est la dernière. En effet, elle contient une valeur dans le statut marital différente de toutes les autres.

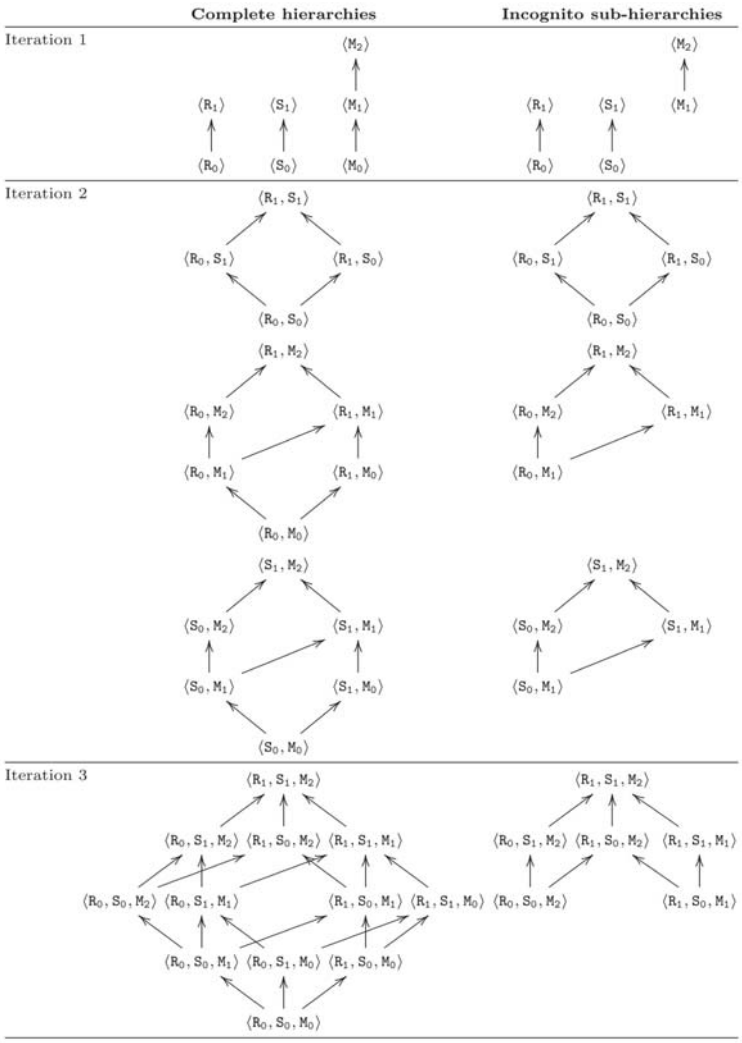


FIGURE 4.15 – Différentes étapes de l'algorithme Incognito d'après [69]

Marital_status	*Sex*	*Hours*	*#tuples (Hyp. values)*
divorced	M	35	2 (0Y, 2N)
divorced	M	40	17 (16Y, 1N)
divorced	F	35	2 (0Y, 2N)
married	M	35	10 (8Y, 2N)
married	F	50	9 (2Y, 7N)
single	M	40	26 (6Y, 20N)

FIGURE 4.16 – Différentes étapes de l'algorithme Incognito d'après [69]

4.1.4.6 Mondrian Multidimensional

Les créateurs de l'Incognito ont également développé un algorithme qui représente la généralisation de manière spaciale soit en plusieurs dimensions[71]. La table privée est représentée sous la forme d'un graphe.

La méthode de l'algorithme est la suivante : (1) premièrement, nous définissons la dimension du graphe. Si nous prenons un couple d'identifiants, le graphe aura deux dimensions. Ensuite, nous représentons les différentes catégories de chaque attribut sur les ordonnées et les abscisses comme sur la figure 4.17. Ensuite, nous plaçons les différents points sur le graphe en fonction des relations établies entre les couples. Enfin, nous marquons le nombre d'occurrences à côté du point. Par exemple, pour le couple de quasi-identifiants suivant (statut marital, sexe) et pour un degré de 10-anonymity, l'algorithme le présente sous la forme d'un graphique *2-dimensional* avec en ordonnée, les deux attributs M et F et en abscisse les différents attributs du statut marital (single, married, divorced). Ceux-ci sont représentés par des points avec le nombre d'occurrences de chaque tuple. Le point 19 représente donc 19 tuples de personnes masculines et divorcées dans le tableau. (2) Deuxièmement, nous pouvons alors commencer à partitionner le graphe en différentes régions. Par exemple, sur la figure 4.17b, nous partitionnons la catégorie statut marital en deux. En effet, le graphe fait la distinction entre *divorced*, *married* et *single*. Ensuite, sur la figure 4.17c, nous avons partitionné la catégorie sexe en deux : soit masculin, soit féminin. Enfin, la dernière partition possible est celle de diviser au sein de la catégorie statut marital, *divorced* et *married*. La figure 4.17d nous donne donc le résultat final de l'algorithme.

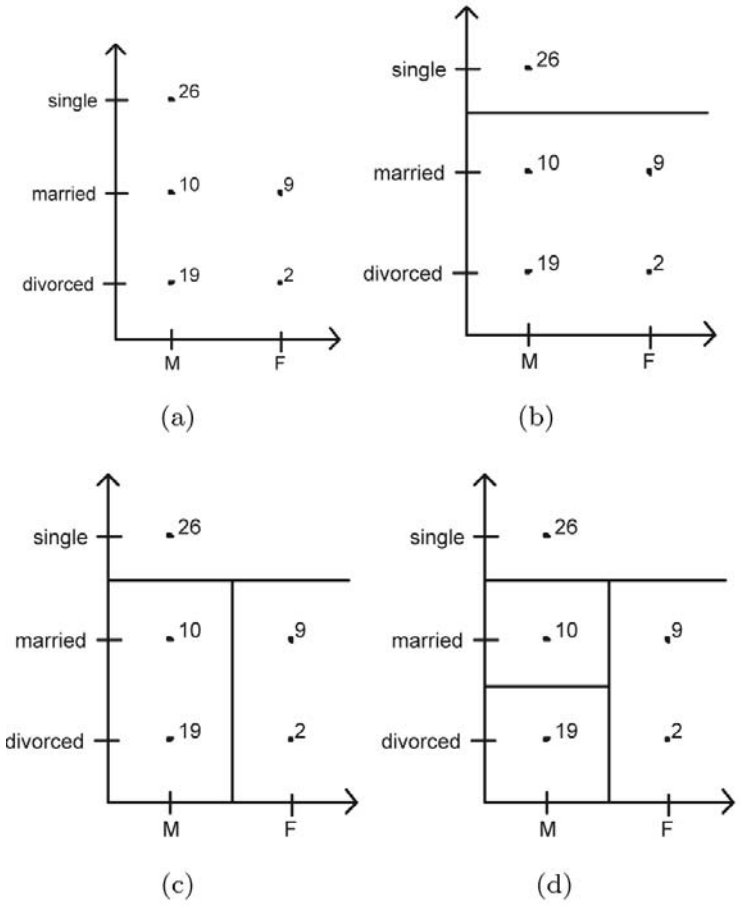


FIGURE 4.17 – Différentes étapes de l'algorithme de Mondrian d'après [71]

4.2 L-diversity

4.2.1 Définition

La méthode l-diversity a été introduite par Machanavajjhala et al.[72] pour résister aux *homogeneous attacks* et aux *background knowledge attacks*. Cette méthode établit une technique pour que les valeurs sensibles soient les mieux représentées dans la table afin d'éviter que ce genre d'attaques ne réussise.

Définition 5 (l-diversity) *Une colonne q∗ est l-diverse si elle contient au moins l "bien représentées" valeurs des attributs sensibles S. Une table est l-diverse si chaque q∗-colonne est l-diverse.*

4.2.2 Technique de l-diversity

Même s'il est impossible de savoir complètement ce que l'adversaire sait, les auteurs nous font remarquer que nous pouvons utiliser certaines données à notre disposition [72] : (1) l'adveraire a accès à la table publique rendue anonyme (et il sait que cette table est une généralisation), (2) il peut savoir que celui qu'il recherche est dans la table et (3) il a une connaissance des attributs sensibles et non sensibles de la population. Ainsi, nous constatons que, par exemple, Alice peut connaître des informations sur Bob dû à deux choses : le manque de diversité des données sensibles et le degré de background knowledge de l'intrus.

Pour résoudre ces deux problèmes, les auteurs se sont basés sur la technique probabiliste d'inférence de Bayes[72] pour construire la méthode l-diversity, qui consiste, nous l'avons vu plus haut, à avoir un nombre "bien-représenté" de valeurs différentes dans les données sensibles.

Afin de définir le terme "bien représenté" (*well represented*), les auteurs proposent trois interpretations de ce terme : distinct l-diversity, entropy l-diversity et recursive l-diversity. Nous allons les exposer briévement.

4.2.2.1 Distinct l-diversity

Une première manière de définir ce terme, intuitive, est de dire qu'il faut au moins l valeurs sensibles distinctes au sein de chaque classe d'équivalence.

4.2.2.2 Entropy l-diversity

Une deuxième manière de le faire est d'utiliser la notion d'entropie. Elle vérifie non seulement que chaque classe d'équivalence a des données sensibles différentes mais aussi que ces données sont correctement réparties dans l'ensemble des classes d'équivalence[72]. Cela signifie que l'entropie de distribution des données sensibles est égale à log (l)[72]. Évidemment, si les données sensibles ne sont pas assez diversifiées, l'entropie n'en sera que plus faible.

4.2.2.3 Recursive l-diversity

Une troisième manière est de vérifier que les valeurs les plus fréquentes ne soient pas trop représentées.

Définition 6 (recursive) $r_1 < c(r_l + r_{l+1} + ... + r_m)$.

4.2.3 Avantages et inconvénients de cette méthode

L'avantage de cette méthode est qu'elle offre une meilleure résistance aux *homogeneous attacks* et rend plus complexe la *background knowledge attack*.

N°	Age	Nationalité	Code postal	Casier judiciaire
1	<30	*	[5000-98000]	Meurtre
2	<30	*	[5000-98000]	Rien
3	<30	*	[5000-98000]	Vol à l'étalage
4	>30	*	[5000-98000]	Vol
5	>30	*	[5000-98000]	Possession de stupéfiants
6	>30	*	[5000-98000]	Vol
7	<30	*	[5000-98000]	Meurtre
8	<30	*	[5000-98000]	Meurtre
9	<30	*	[5000-98000]	Meurtre

FIGURE 4.18 – *Données du commissariat de Police 3-diversity*

En effet, grâce à cette méthode, Alice ne sait plus savoir avec exactitude si Bob a un casier ou pas. Avec ce modèle, nous avons maintenant un tableau avec 3-diversifiées valeurs sensibles : meurtre, vol à l'étalage ou pas de casier. Par contre, elle n'est pas à l'abri des *similarity attacks* et des *skewness attacks*.

4.2.3.1 Similarity attack

Supposons qu'Alice décide de sortir avec Bob. Bob est malade, mais ne veut pas dire à Alice ce qu'il a. Alice, curieuse, décide de chercher quelle maladie a son petit ami. Elle tombe sur une table 3-anonyme et 3-diversifiée de l'hôpital. Avec les données qu'elle a en sa possession, elle arrive à affiner sa recherche pour n'avoir plus que trois choix possibles comme le montre la figure 8. Nous constatons que, malgré la diversification des données (arythmie, cardiomyopathie, maladie coronarienne), Alice sait que Bob a une maladie du coeur. Nous constatons que les méthodes l-diversity et k-anonymity sont plus efficaces plus le k et le l sont grands.

N°	Age	Maladie
1	<20	Obésité
2	<20	Diabète
3	<20	Hypertension artérielle
4	>30	Asthme
5	>30	Maladie d'Alzheimer
6	>30	Sclérose latérale amyothrophique
7	2*	Arythmie
8	2*	Cardiomyopathie
9	2*	Maladie coronarienne

FIGURE 4.19 – *Données de l'hôpital 3-anonymity et 3-diversity*

4.2.3.2 Skewness attack

Un autre type d'attaque est la *skewness attack* dite "attaque disymétrique". Par exemple, si nous prenons deux attributs sensibles dans un tableau : HIV positif et HIV négatif et que le niveau de répartition de ces attributs dans un tableau est le suivant : 1% de HIV positifs et 99% de HIV négatifs. Supposons également que nous avons une classe d'équivalence qui contient un nombre égal d'enregistrements positifs et négatifs. La méthode l-diversity ne va pas faire la différence entre cette classe d'équivalence (49 HIV positifs + 1 HIV négatif) et celle-ci (1 HIV positif + 49 HIV négatifs). Si nous supposons qu'il y a 20000 enregistrements, pour avoir un 2-diversity, il devrait y avoir au plus 20000*1% = 200 classes d'équivalence.

4.2.4 Les variantes de cette méthode

Comme nous l'avons vu plus haut, la méthode l-diversity ne permet pas de faire de distinction entre les ressemblances au sein des valeurs sensibles. Les données doivent être différentes mais l'appartenance de ces données à un groupe commun n'est pas prise en compte. Pour remedier a cette vulnérabilité, Qian wang et Xiangling shi ont dévellopée la méthode (α, d) diversity en utilisant la même méthode mais en lui ajoutant un facteur supplémentaire [73]. Yunli Wang et al.[74] ont également proposé une amélioration de cette méthode intitulé L-SR diversity.

4.3 T-closeness

Li et al.[75] proposent une autre solution pour mieux répartir les attributs sensibles afin qu'ils soient plus proches non seulement dans chaque classe d'équivalence mais également dans la table tout entière.

4.3.1 Définition

Définition 7 (t-closeness) *Une classe d'équivalence est dite "t-closeness" si la distance entre la distribution des attributs sensibles dans la classe et la distribution des attributs dans toute la table n'est pas plus grande que le seuil t. Une table t-closeness est une table où chaque classe d'équivalence est t-closeness d'après [75].*

4.3.2 Technique de t-closeness

Pour ce faire, t-closeness utilise l'*Earth Mover's Distance* (EMD). Cette technique est "basée sur la quantité de travail minimale que l'on doit faire pour transformer une distribution en une autre en déplaçant la masse de distribution entre chacune d'elles"[75]. Les auteurs calculent l'EMD selon le type

d'attributs[75]. Pour les attributs sensibles numériques, l'EMD est calculée sur base du nombre de valeurs entre le nombre total. Pour les attributs sensibles de catégorie, elle est évaluée soit par *equal distance* (la distance entre deux valeurs de catégorie d'attributs est définie à 1) ou par *hierarchical distance* (la distance entre deux valeurs se base sur les niveaux d'écart entre les valeurs). Par exemple, sur la figure 4.20, la première classe d'équivalence

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

A

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

B

FIGURE 4.20 – Tableau 3-diversity et 3-anonymity / Tableau 0.167 t-closeness pour le salaire d'après [75]

contient dans la colonne de l'attribut sensible salaire, trois salaires différents : 3k, 4k, 5k. Ce principe respecte donc bien la règle de l-diversity. Cependant, cette classification n'est pas optimale et pose problème. En effet, si l'intrus recherche le salaire d'une personne qui est dans la première classe d'équivalence. Il saura qu'il gagne entre 3K et 5k. L'écart est trop faible. Si nous examinons la deuxième classe d'équivalence, nous constatons que l'écart est plus grand et donc meilleur, entre 6k et 11k. Afin de rendre cette table t-closeness, la méthode va analyser les différents écarts entre les classes d'équivalence par rapport à l'ensemble des valeurs. Puis, elle va généraliser la table afin de rendre l'écart plus important entre les attributs numériques sensibles de chaque classe d'équivalence. Sur la figure 9b, nous voyons que l'écart pour la première classe est de 6, pour la deuxième de 5 et pour la dernière de 6.

4.3.3 Avantages et inconvénients de cette méthode

Même si en théorie, cette méthode présente une meilleure résistance aux différentes attaques, aucune implémentation n'a encore été proposée pour cette méhode. Par contre, l'un des inconvénients de cette technique est que plus le t est petit, plus les données sont endommagées. De plus, le temps de calcul sera également augmenté.

4.4 (a,k) anonymity

L'a,k anonymity proposée par Wong et al[76] est une méthode qui prend en compte les spécificités des données sensibles.

4.4.1 Définition

Définition 8 (α, k-anonymity) *Soit un ensemble de données D, un ensemble d'attributs Q et une valeur sensible s dans la domaine des attributs $S \notin Q$. Soit (E, s) un ensemble de tuples dans chaque classe d'équivalence E contenant s issu de S et α, le seuil qui est compris entre $0 < \alpha < 1$. L'ensemble des données D est alpha-k anonymisé si la fréquence relative des s dans chaque classe d'équivalence est plus petit ou égal au seuil α. La table α, k-anonymisé doit être $|(E, s)|/|E| <= \alpha$ pour toutes les classes d'équivalence E.*

Par exemple, si nous examinons la figure 4.21, nous constatons que dans la colonne des maladies, certaines sont plus graves que d'autres. C'est sur la donnée la plus sensible (ici, le sida) que la méthode va se baser. Pour ce faire, nous devons utiliser la définition établie par Wong et al[76].

4.4.2 Technique d'α, k-anonymity

Nous allons expliquer la méthode à travers l'exemple du tableau de la figure 4.21. Tout d'abord, nous examinons le nombre de classes d'équivalence qu'il y a dans le tableau. En l'occurrence, nous en avons trois : {1,2,3,4}, {5,6,7,8},{9,10,11,12}. Considérant la donnée sensible "Sida", nous avons dans la première classe d'équivalence, une seule fois le sida, ce qui est traduit par $|(E, s)|/|E| <= 0.25$, dans la deuxième classe d'équivalence, cette maladie apparaît deux fois $|(E, s)|/|E| <= 0.50$ et dans la dernière classe d'équivalence, aucune fois $|(E, s)|/|E| <= 0$. Nous constatons que la classe d'équivalence est $|(E, s)|/|E| <= 0.50$.

4.4.3 Avantages et inconvénients de cette méthode

Comme nous l'avons vu, la généralisation est efficace. Elle permet de semer le doute chez l'attaquant qui recherche une donnée précise et est donc susceptible de lui faire perdre beaucoup de temps. Cependant, en généralisant, notre table de données devient de moins en moins précise et qualitative. De plus, si les données sensibles ne sont pas variées, cette méthode ne pourra être efficace.

N°	Age	Maladie
1	<20	Rhume
2	<20	Fièvre
3	<20	Migraine
4	<20	Sida
5	>30	Sida
6	>30	Rhume
7	>30	Rhumatisme
8	>30	Sida
9	2*	Problème de dos
10	2*	Migraine
11	2*	Fièvre
12	2*	Mal de gorge

FIGURE 4.21 – *Données de l'hopital 0,5 d'alpha et 4-anonymity*

4.5 Anatomy

C'est pourquoi, Xiakui & Tao[77] a proposé une autre méthode basée sur la k-anonymity, l'anatomy. Il s'agit de réaliser un tableau de quasi-identifiants et un tableau de valeurs sensibles qui sont séparés par une valeur numérique pour protéger les données.

4.5.1 Définition

Définition 9 (l-diverse partition) *Une partition avec m groupes quasi-identifiants est l-diverse si chaque groupe quasi-identifiant* $QI_j(1 <= j <= m)$ *satisfait aux conditions suivantes. Soit v la valeur* A^s *la plus fréquente dans* QI_j *et* $c_j(v)$*, le nombre de tuples* $t \in QI_j$ *avec* $t[d+1] = v$ *alors* $c_j(v)|QI_j| <= 1/l$ *d'après[77].*

Par exemple, sur le tableau des données judiciaires de la page suivante, nous constatons que le groupe ID 1, 2, 3 constitue des partitions 2-diverse.

Définition 10 (anatomy) *Soit une partition l-diverse avec m-groupes de quasi-identifiants, anatomy produit un tableau de quasi-identifiant (QIT) et un tableau de données sensibles (ST) comme suit. Le groupe quasi-identifiant aura le schéma suivant :*
$(A_1^{qi}, A_2^{qi}, ..., A_d^{qi}$*, Groupe de QI) Pour chaque groupe quasi-identifiant,* $QI_j(1 <= j <= m)$ *et chaque tuple* $t \in QI_j, QIT$ *est un tuple de la forme :* $(t[1], t[2], ..., t[d], j)$*. Le tableau de données sensibles aura le schéma suivant :*
$(Groupe - ID, A^s$*, Nombre de fois) d'après[77].*

4.5.2 Technique de anatomy

Nous allons expliquer cette méthode à travers un exemple. Soit le tableau de la figure 4.22.

N°	Age	Sexe	Code postal	Casier judiciaire
1	21	M	6224	Abus de confiance
2	22	M	5020	Extorsion
3	23	M	1024	Extorsion
4	24	M	4520	Abus de confiance
5	32	F	6000	Home jacking
6	33	F	5620	Contrefaçon
7	33	F	3000	Home jacking
8	37	F	6050	Espionnage
9	40	F	6000	Corruption
10	41	F	5620	Harcèlement sexuel
11	42	F	3000	Corruption
12	43	F	6050	Home jacking

FIGURE 4.22 – *Tableau de données judiciaires*

Un rapide constat sur ce tableau nous permet de voir que, même en généralisant ce tableau sur le sexe et le code postal, les données sensibles ne se trouveraient pas correctement diversifiées. Pour ce faire, il suffit d'examiner les quatres premières lignes. Même la méthode l-diversity se révèlera inefficace. Car, dans le premier cas, nous aurons 50% de chance de deviner que l'homme que nous recherchons a été condammé pour abus de confiance ou pour extorsion. Et pour le deuxième groupe, si nous recherchons une femme, il y a 50% de chance qu'elle soit responsable d'un home jacking. Idem pour le troisième groupe avec la corruption.

La construction d'un tableau respectant le principe d'anatomy se réalise en trois étapes : premièrement, nous divisons le tableau de données en groupe de données quasi-identifiantes. Dans le cas de l'exemple envisagé plus haut, nous pouvons effectuer une division du groupe en deux. Dans le premier groupe, nous prenons les quatre premiers tuples, les quatre suivants dans le deuxième groupe et les quatre derniers dans le troisième groupe. Deuxièmement, nous créons un tableau en remplaçant la colonne des données sensibles par une colonne intitulée "Groupe ID" dans lequel nous numérotons pour chaque tuple le groupe auquel il appartient. Ce qui nous donne le tableau de la figure 4.23.

Troisièmement, nous créons un deuxième tableau intitulé "Tableau de données sensibles" dans lequel pour chaque groupe ID, nous retrouvons les données sensibles correspondantes ainsi que le nombre de fois où nous les avons retrouvées dans le tableau original.

La méthode anatomy s'articule autour de deux tableaux : un tableau de

N°	Age	Sexe	Code postal	Groupe ID
1	21	M	6224	1
2	22	M	5020	1
3	23	M	1024	1
4	24	M	4520	1
5	32	F	6000	2
6	33	F	5620	2
7	33	F	3000	2
8	37	F	6050	2
9	40	F	6000	3
10	41	F	5620	3
11	42	F	3000	3
12	43	F	6050	3

FIGURE 4.23 – *Anatomy : tableau QIT*

Groupe ID	Casier judiciaire	Nombre de fois
1	Abus de confiance	2
1	Extorsion	2
2	Home jacking	2
2	Contrefaçon	1
2	Espionnage	1
3	Corruption	2
3	Harcèlement sexuel	1
3	Home jacking	1

FIGURE 4.24 – *Anatomy : Tableau des données sensibles*

données excluant les données sensibles et un tableau les contenant. L'encapsulation des données sensibles dans le groupe permet une meilleure protection de ces données. En effet, contrairement aux autres méthodes, nous ne pouvons accéder directement aux données sensibles. Ici, nous devons avoir les deux tableaux.

4.5.3 Avantages et inconvénients de cette méthode

Malgré cette nouvelle protection, nous pouvons constater que le résultat est le même que la méthode l-diversity. Vu le manque de données sensibles, nous retrouvons plusieurs fois des données sensibles au sein de chaque groupe. De plus, il ne faut pas oublier que si l'intrus tombe sur le tableau de données sensibles et le tableau QIT, il pourrait faire le lien entre les deux tableaux.

4.6 Le modèle d'anonymisation km

Le modèle dévellopé par Terrovitis & co[78] s'intéresse à appliquer l'anonymisation des données sur des bases de données transactionnelles (par exemple, un tableau de données recensant les articles achetés dans un supermarché par les clients, par jour). Le but étant de transformer une base de données A en une base de données A'. Comme dans d'autres modèles, nous sommes obligés de tenir compte des connaissances que l'adversaire a potentiellement sur une personne figurant dans cette base de données. A priori, il est impossible de pouvoir définir ce degré de connaissance. Ce modèle se propose donc "de définir un modèle générique pour la vie privée, qui garantit que la connaissance des adversaires est limitée à un niveau, exprimé comme un paramètre du modèle"[78]. Ce paramètre est nommé "m". Cela signifie que l'attaquant connaît maximun "m" données dans la base de données lui permettant de le relier à l'attaquant. Il est donc impératif que "pour n'importe quel ensemble de m ou moins de données, il y ait au moins "k" transactions, qui contiennent cet ensemble, dans la base de données publiée A"[78].

Les auteurs soulignent également la différence par rapport aux données anonymisées dans le cadre des modèles k-anonymity. Pour les bases de données transactionnelles, il n'existe pas de quasi-identifiants ni de données sensibles. En effet, nous pourrions très bien avoir une base de données avec un ensemble de données non-idendifiantes et reprendre ces données qui pourraient servir de quasi-identifiant dans un autre cas de figure. Il n'est pas possible de les catégoriser car leur nature peut être changeante en fonction de la transaction. De plus, la taille et la longueur des bases de données n'est pas fixe.

4.6.1 Définition

Définition 11 (km-anonymisation) *Soit une base de données D, aucun attaquant qui a un background knowledge jusqu'à m articles de la transaction t E D peut utiliser ces articles pour identifier moins que les k tuples de D. En d'autres termes, n'importe quelle requête de taille m ou moins de l'attaquant devrait renvoyer rien ou plus que des réponses K d'après [78].*

4.6.2 Technique d'anonymisation km

Le principe est le suivant : prendre une base de donnée transactionnelle qui ne respecte pas l'exigence de km anonymisation et pour chaque article se trouvant dans la base de données, les généraliser jusqu'à atteindre une base de données transactionnelles respectant le modèle km anonymisation. Par exemple, si nous avons dans notre base de données des articles intitulés "lait demi-écrémé, lait entier, lait écrémé", le modèle va les généraliser en "lait" ou encore "produits laitiers".

Pour ce faire, le modèle va s'appuyer sur une généralisation hiérarchique. Le but étant que les produits possédant certaines spécificités mais d'ensemble commun, se retrouvent dans le même domaine de généralisation. Par exemple, "pain", "biscotte", "céréales" se retrouvent dans "produits céréaliers".

Si nous prenons le tableau de la figure 4.25, nous constatons que nous avons un tableau avec des articles commandés dans un magasin par différentes personnes. Chaque ligne est différente, ce qui permet à l'attaquant de retrouver facilement les données d'une personne. Supposons que nous voyons par la fenêtre mon voisin avec un paquet de courses avec du lait et du fromage, je peux connaître la suite des aliments qu'il a acheté si le magasin diffuse la liste des transactions effectuées par jour. Nous pouvons également voir que ce tableau ne respecte pas le principe de k-anonymity.

Transaction	**Produits**
1	Lait, pain, céréales
2	Fromage, pain
3	Fromage, pain, céréales
4	Lait, fromage, céréales

FIGURE 4.25 – *Base de données transactionnelles des articles commandés en magasin par différentes personnes*

Si nous appliquons la généralisation prônée par ce modèle pour un tableau 22 anonymity, nous arrivons au tableau suivant. Nous constatons déjà que la généralisation sur les produits laitiers rend la divulgation des données d'achats beaucoup plus difficile. Maintenant, nous devons regarder si ce tableau respecte le principe de km-anonymity. Pour ce faire, nous allons examiner deux articles issus du tableau non généralisé pour voir si nous le retrouvons moins de k fois dans le tableau généralisé. Si nous prenons le couple (fromage, pain), nous constatons que nous le retrouvons dans trois lignes. M est donc supérieur à K, ce qui rend donc la tâche de l'attaquant beaucoup plus difficile.

Transaction	**Produits**
1	Produit laitier, pain, céréales
2	Produit laitier, pain
3	Produit laitier, pain, céréales
4	Produit laitier, céréales

FIGURE 4.26 – *Base de données transactionnelles avec les articles km anonymisées*

Il faut également noter qu'une généralisation à outrance n'est pas à conseiller comme le témoigne l'exemple suivant. Si nous généralisons le lait, pain, snickers, pomme à articles de magasin, nous respectons parfaitement le principe mentionné ci-dessous. Cependant, nous perdrions alors un grand nombre d'in-

formations. La généralisation doit être envisagée selon les objectifs poursuivis par les utilisateurs actuels ou futurs des données.

4.6.3 Avantages et inconvénients de cette méthode

Comme nous l'avions vu dans les techniques d'appauvrissement, la technique de généralisation est utilisée dans le cadre de ce modèle. La généralisation est donc adaptée au facteur "m" de connaissance de l'individu. L'inconvénient de cette technique est que nous ne savons pas ce que l'intrus sait de la personne. La généralisation doit donc être effectuée selon ce degré de connaissance, inconnu, et donc calculé de manière hypothétique.

4.7 M-invariance

Précédemment, les méthodes que nous avons analysées s'intéressaient à l'anonymisation pour un tableau. Ici, cette méthode s'assure que les données sensibles au sein des classes d'équivalence soient correctement réparties au sein des différents tableaux de données publiés à des temps j différents.

Afin de comprendre ces méthodes, nous devons d'abord reprendre les différentes définitions utiles à la compréhension de celles-ci.

4.7.1 Définitions

Définition 12 (groupe QI) *Pour une table de données T(j), le groupe QI (ou partition) est composé d'un sous-ensemble disjoint de tuples dans T(j), dont l'union égale T(j). d'après [79]*

Définition 13 (la signature) *Soit QI*, le groupe QI de la table anonymisée T*(j) pour chaque j de 1 à n. La signature de QI est un ensemble de valeurs sensibles distinctes dans QI*, d'après [79]*

Définition 14 (m-unique) *Une table généralisée T*(j) est m-unique, si chaque groupe QI dans T*(j) contient au moins m-tuples, et que tous ces tuples dans le groupe ont des valeurs sensibles différentes, d'après [79]*

Définition 15 (m-invariant) *Une séquence de relations publiées T*(1),..., T*(n) (où n >=1) est m-invariant si et seulement si : //1. T*(j) est m-unique pour tout j de 1 à n //2. Pour chaque tuple t avec une durée de vie [x,y], t.QI*(x), t.QI*(x+1),...,t.QI*(y) ont la même signature, ou t.QI*(j) montre que le groupe QI contient t au temps j (appartenant à [x,y]), d'après [75]*

4.7.2 Technique de m-invariant

La méthode m-invariance s'intéresse ici aux attaques possibles par comparaison de deux tableaux de données. En effet, si Bob se retrouve dans le premier tableau avec une classe d'équivalence, par exemple de deux catégories (meurtre, vol) et que dans le deuxième tableau, nous retrouvons une autre classe d'équivalence avec une des catégories déjà présente précédemment (meurtre, délit de fuite), l'attaquant pourrait aisément faire l'intersection entre les deux et deviner que Bob a été condamné pour meurtre. Pour éviter ce type d'attaque, les auteurs de la méthode m-invariance introduisent une donnée inexistante dans le tableau 4.27. Il s'agit d'une donnée "contrefaite". Nous rajoutons dans le tableau une ligne de données fictives (c1, [40-49], [50k-69k], excès de vitesse) qui permet de bloquer ce type d'attaque. Le but de cette manipulation est d'éviter que l'attaquant n'associe plusieurs valeurs similaires à un seul individu. Même si cela permet d'éviter cette attaque, une autre type d'attaque peut se présenter "attaque par valeur d'équivalence". En effet, en comparant le tableau 1 avec le tableau 3 anonymisé par la méthode m-invariance, nous constatons que Bob et Greg ont la même maladie, à savoir, un rhume. Ce procédé pourrait continuer encore et encore.

Nous l'avons vu, la connaissance de l'adversaire pour mettre à jour les données sensibles est un atout non négligeable et est impossible à quantifier. Cependant, nous pouvons nous prémunir de certains risques. En effet, lorsque nous avons des données sensibles et que nous effectuons fréquemment des mises à jour, nous risquons de dévoiler sans le savoir de précieuses informations à l'attaquant. Par exemple, imaginons que Bob ait été condamné pour harcèlement sexuel. Il a purgé sa peine et décide maintenant de retrouver un emploi. Entre-temps, la police publie plusieurs statistiques sur les personnes ayant enfreint la justice. Un premier tableau de données voit le jour en juin, l'autre en juillet avec des mises à jour. En juillet, Bob rencontre un employeur qui est fortement intéréssé par son profil. Avec les données en sa possession (Bob, 26 ans, habite à Charleroi), il décide d'aller examiner de plus près son profil. Lors de sa recherche, il tombe sur les deux tableaux généralisés ci-dessous.

Nous constatons qu'avec la connaissance du futur employeur et les deux tableaux ci-dessous, il arrive à la conclusion que dans le premier tableau, Bob a un casier judiciaire et qu'il a été condamné soit pour harcèlement sexuel soit pour vol à l'étalage. A ce stade-ci, il a un doute à hauteur de 50%. Mais quand il examine le deuxième tableau actualisé où il sait que figure de nouveau Bob, il constate qu'il y a encore la donnée sensible "harcèlement sexuel" mais plus la donnée "Vol à l'étalage". Il est donc certain que Bob a été condamné pour harcèlement sexuel et il est alors fort probable que Bob ne soit pas engagé dans cette société. La méthode m-invariance propose d'ajouter des données contrefaites dans le tableau. Soit le tableau d'après.

Groupe ID	Age	Code postal	Casier judiciaire
1	[25,30]	5000-6000	Harcèlement sexuel
1	[25,30]	5000-6000	Vol à l'étalage
2	[31-37]	6001-7000	Piratage
2	[31-37]	6001-7000	Vol à la tire
3	[38-42]	7001-8000	Piratage
3	[38-42]	7001-8000	Trafic de stupéfiants
4	[43-48]	8001-9000	Viol
4	[43-48]	8001-9000	Homicide
5	[50-55]	9001-10000	Faux-monnayage
5	[50-55]	9001-10000	Extorsion

FIGURE 4.27 – *Tableau de données de juin généralisé*

Groupe ID	Age	Code postal	Casier judiciaire
1	[25,27]	3000-6000	Harcèlement sexuel
1	[25,27]	3000-6000	Recel
2	[30-35]	6001-6582	Piratage
2	[30-35]	6001-6582	Hold up
3	[36-40]	7001-7500	Trafic d'influence
3	[36-40]	7001-7500	Trafic de stupéfiants
4	[42-48]	8001-9000	Empoisonnement
4	[42-48]	8001-9000	Homicide
5	[50-65]	9001-10000	Trafic de drogue
5	[50-65]	9001-10000	Extorsion

FIGURE 4.28 – *Tableau de données de juillet généralisé*

Groupe ID	Age	Code postal	Casier judiciaire
1	[25,27]	3000-6000	Harcèlement sexuel
c1	[25,27]	3000-6000	Vol à l'étalage
2	[30-35]	6001-6582	Piratage
2	[30-35]	6001-6582	Hold up
3	[36-40]	7001-7500	Piratage
c2	[36-40]	7001-7500	Trafic de stupéfiants
4	[42-48]	8001-9000	Empoisonnement
4	[42-48]	8001-9000	Homicide
5	[50-65]	9001-10000	Trafic de drogue
5	[50-65]	9001-10000	Extorsion

FIGURE 4.29 – *Tableau de données de juillet m-invariant*

Groupe ID	Nombre de fois
1	1
3	1

FIGURE 4.30 – *Tableau des données contrefaites*

Grâce à cette méthode, nous constatons que l'introduction de données contrefaites dans le tableau de juillet permet de respecter la répartition des données sensibles par rapport au premier tableau de juin. En effet, dans la classe d'équivalence de Bob, nous retrouvons les mêmes données sensibles : harcèlement sexuel et vol à l'étalage. La mise à jour du tableau, sous cette forme, ne permettra donc pas à l'intrus d'obtenir des informations supplémentaires sur ces deux tableaux.

4.7.3 Avantages et inconvénients de cette méthode

Suite à l'examen de cette méthode, nous avons constaté qu'elle était résistante aux attaques par valeur d'association. En effet, grâce à la méthode m-invariance, nous offusquons les données en ajoutant de l'incertitude avec les données contrefaites. Par contre, il faut également prendre en compte le fait que l'attaquant pourrait lui-même faire partie du tableau de données et donc deviner en fonction de son propre cas, les valeurs sensibles des personnes ciblées mais également d'autres personnes de la base par déduction.

4.8 Conclusion

Dans ce chapitre, nous avons premièrement examiné en détail, la méthode d'anonymisation principale, k-anonymity, qui, malgré quelques faiblesses constitue un bon point de départ pour une anonymisation. Deuxièmement, nous avons analysé la méthode l-diversity qui, tout en se basant sur la précédente, s'est intéressée à donner de la diversité aux différents attributs sensibles dans chaque classe d'équivalence. Troisièmement, nous avons étudié le modèle t-closeness, qui a approfondi le modèle l-diversity jusqu' à offrir une diversité dans toute la table. Quatrièmement, nous avons examiné d'autres méthodes pour les bases de données transactionnelles mais aussi pour les données mises à jours. Nous avons constaté que ces différents modèles utilisaient de nombreuses techniques exposées dans le chapitre 3 et que cette combinaison pouvait être efficace dans certains cas de figure. L'implémentation algorithmique de la méthode principale nous a permis de dresser une liste non exhaustive de divers algorithmes. Nous avons constaté que certains algorithmes étaient optimaux et d'autres complets. Il en ressort néammoins qu'aucun modèle n'a pu assurer une fiabilité à 100% et reste vulnérable à différentes attaques.

Chapitre 5

Analyse d'un cas pratique

5.1 Le sondage sur la mobilité

Après avoir défini une démarche pour anonymiser correctement les données voulues et avoir expliqué les techniques à mettre en oeuvre pour arriver au résultat voulu, il est intéressant d'allier la théorie à la pratique en l'appliquant à un cas concret. Nous avons donc décidé, sur les conseils de Monsieur Colin et grâce à l'aimable autorisation de Monsieur Cornellis, d'analyser les données publiées de manière anonyme suite à un sondage sur la mobilité des Belges effectué en 2011 et réalisé par le Groupe de Recherche sur les Transports (GRT) de l'Université de Namur en collaboration avec les chercheurs de l'IMOB (Université d'Hasselt) et du CES (Facultés Saint-Louis). Ce sondage fait suite à un premier sondage sur le même sujet réalisé en 1999. La collecte des données a été lancée en décembre 2009 et s'est terminée en décembre 2010. Il a été réalisé sur 8532 ménages (soit 15821 personnes âgées de six ans et plus).

5.1.1 La préparation du sondage

Avant de réaliser ce sondage, l'équipe des chercheurs s'est adressée au registre national afin d'obtenir les coordonnées d'un certain nombre de personnes. Cette demande a été motivée par le besoin d'obtenir des informations sur les déplacements des belges et sur leur utilisation des transports en commun en Belgique. Après avoir étudié la question, le registre national a autorisé le prélèvement au sein de sa base de données d'un échantillon de personnes suivant les critères demandés par le groupe de recherche. Il faut noter que c'est le registre national qui est en charge du tirage et de la transmission des données. Les chercheurs ont donc demandé un tirage aléatoire par province et par ménage d'une personne (taux de réponse le plus bas). Ils ont également spécifié le tirage et les caractéristiques de celui-ci. La Commission de la vie privée a ensuite examiné ces données et rendu un avis portant le numéro 005014032. Malheureusement, lors de la rédaction de ce mémoire, l'avis n'était

plus disponible sur le site de la Commision de la vie privée.

Il est intéressant de noter qu'à ce stade les personnes tirées au sort n'ont pas donné leur accord pour participer au sondage. Pourtant, la Commission de la vie privée a permis que ce tirage soit effectué. En effet, si avant d'effectuer le sondage, il fallait contacter les 11 008 000 de Belges pour leur demander leur autorisation, cela ne serait pas possible. Cependant, nous constatons que le registre national divulgue nos données sans que nous en soyons avertis. Nous constatons donc que nos données personnelles n'ont plus réellement une valeur "personnelle" puisque c'est le registre national avec la Commission de la vie privée qui décident ce qui est bon ou pas pour nos données.

Une fois les coordonnées personnelles des individus tirées au sort collectées, le groupe de recherche de transport a sous-traité l'envoi des fichiers à une société tierce. La société tierce a donc reçu également les donnnées personnelles des individus sélectionnés. Celle-ci a envoyé 15 jours avant le contact pour le sondage une lettre expliquant que la personne avait été sélectionnée pour réaliser un sondage sur la mobilité. Sur le papier, il était notamment noté qu'il n'y avait aucune obligation légale d'y répondre. Il faut souligner cette initiative qui montre que même si le registre national a divulgué nos données personnelles sans notre accord, le groupe de recherche y accorde une importance et préfère prévenir les tirés au sort. Cependant, il faut noter que, même si la personne n'était pas intéressée suite au premier courrier, elle a quand même reçu la lettre. Dans cette lettre, le sous-traitant mentionnait également le jour où la personne serait contactée. Si personne ne répondait, une relance téléphonique était de mise quelques jours après le premier contact. Pour les personnes dont le numéro de téléphone était correcte et existant, le sondage a été réalisé par téléphone. Pour les autres, un courrier avec les questions leur a été envoyé. A charge pour eux de le renvoyer par la poste, le port étant payé par le destinataire.

5.1.2 La collecte des données

Pour réaliser ce sondage, deux formulaires ont été envoyés : un formulaire sur le ménage et un formulaire individuel pour chaque personne du ménage agée de six ans et plus. Afin de pouvoir comprendre la méthode d'anonymisation effectuée sur ces données, nous allons analyser ces deux formulaires pour constater quelles sont les données collectées à la source. Ces deux formulaires ont été joints en annexe afin de permettre aux lecteurs d'avoir une vision de la manière dont les données ont été collectées (champs libres, tranches de réponses prédéfinies, etc.).

5.1.2.1 Questionnaire du ménage

Sur la première page du formulaire, le GRT spécifie directement que "les renseignements que vous nous communiquerez seront traités de manière anonyme et conformément à la déclaration à la Commission de la protection de la vie privée" et renvoie à la page 8 pour plus d'informations sur le sujet. De manière claire et précise, elle met en confiance l'utilisateur sur l'encodage futur des données de ce sondage.

Premièrement, la personne de contact est invitée à reconstituer son ménage. Pour ce faire, pour chaque personne du ménage, elle doit remplir son prénom, son année de naissance, son sexe, sa nationnalité, sa relation par rapport à lui, son diplôme ou certificat le plus élevé obtenu et son statut professionnel actuel.

Deuxièmement, elle doit remplir le type et le nombre de véhicules présents chez elle (vélo d'enfant, vélo d'adulte, cyclomoteur, moto, voiture, camionnette ou autres véhicules). Ensuite, pour chaque véhicule motorisé ou deux roues, la personne de contact doit remplir les spécificités. Les données suivantes sont demandées : marque, modèle, type de véhicule (pour les véhicules motorisés), cylindrée, année d'achat, nature de l'achat (neuf, occasion, etc.), nombre de kilomètres (si plus d'un an), type de carburant, l'endroit de stationnement du véhicule et l'utilisateur principal du véhicule (avec mention de son nom).

Troisièmement, des informations sur l'habitation et le quatrier du ménage sont collectées. Les données demandées concernent le type d'habitation (villa, appartement) si la personne en est locatataire ou propriétaire, sur le nombre de voitures dans le garage, la présence d'un parking avoisinant, la nature du parking. Après avoir collecté le nombre de vélos dans le ménage, l'avis du ménage est demandé sur l'utilisation qui est faite des transports en commun.

Enfin, pour une recherche à des fins économiques, la personne de contact est invitée à cocher le revenu net de son ménage. A la fin du questionnaire, en bas de la page, les différentes mesures de protection des données sont rappelées.

5.1.2.2 Questionnaire individuel

Le questionnaire individuel est composé de trois catégories : une première catégorie sur les habitudes de mobilité de la personne sondée, une deuxième catégorie sur une journée type de déplacement et une troisième catégorie consacrée à l'avis de la personne sondée sur la mobilité.

La première catégorie est elle-même subdivisée en trois parties (que nous nous devons de présenter car les informations récoltées dans celles-ci nous seront utiles) : une partie sur les usages des différents modes de déplacement, les déplacements longue distance et les déplacements domicile-travail et vice-versa. Dans la première partie, chaque personne du ménage est invitée à répondre notamment aux informations suivantes : les modes de déplacement

utilisés (marche, vélo, etc.) et la fréquence à laquelle ils sont utilisés, le nombre de fois où elle utilise les transports en commun, la nature de l'abonnement, la possesion d'un permis de conduire, etc. Dans la deuxième partie, les données sur les déplacements à l'étranger sont collectées. Des informations sur le dernier voyage de plus de 100 kilomètres sont demandées. Le lieu de départ (ville, pays), le lieu d'arrivée (ville, pays) ainsi que le mode de transport sont demandés, le nombre de nuits ainsi que la raison du voyage. Enfin, dans la troisième partie, chaque personne est invitée à mentioner son logement (avec l'adresse complète). Il est également demandé où la personne travaille (à son domicile, dans un lieu fixe, etc.). Elle doit également encoder l'adresse de son lieu de travail ou d'étude (s'il est étudiant), le nombre de jours de travail par semaine et les différents modes de transport qu'elle utilise pour se rendre au travail/lieu d'étude. Pour chacun de ces modes, le nombre de kilomètres "aller-simple" est demandé. Si la personne posséde deux travails, ou un travail et est encore aux études, une deuxième colonne est proposée pour remplir d'autres données. Ensuite, des questions sont posées sur la distance par rapport aux transports en commun (arrêt de bus, métro, train, etc.), sur la présence d'un parking à proximité, sur la difficulté de trouver une place, l'activité réalisée pendant le déplacement vers son domicile/lieu de travail. Une fois ces données collectées, des données spécifiques aux travailleurs sont demandées. Elles concernent les frais de déplacement, si la personne a un travail qui demande des déplacements, le mode privilégié de déplacement vers son lieu de travail, si la personne effectue du co-voiturage. Afin de spécifier la nature du travail, la personne doit encoder des données sur le nombre de jours de travail, la période ainsi que le nombre d'heures prestées par semaine ainsi que le secteur d'activité.

La deuxième partie demande à chaque individu de retracer une journée type de déplacement. La personne sondée doit encoder pour chaque déplacement son lieu de départ et d'arrivée, l'heure de départ et d'arrivée, le temps et la distance du trajet ainsi que le mode de transport utilisé et le motif du déplacement (aller à la maison, aller au travail, etc.). A cette fin, dix emplacements sont réservés pour indiquer chacun des déplacements sur une journée. Deux autres questions concernant les déplacements nombreux ou au contraire, absents sont posées.

La troisième et dernière partie concerne la récolte d'opinions sur la mobilité. Cette partie est reservée aux personnes de plus de 16 ans. Dans cette partie, des questions sur les transports en commun sont posées aux personnes sondées sur leur qualité, sur des propositions d'améliorations, sur ce qu'il faudrait améliorer, etc.

Enfin, comme pour le questionnaire sur le ménage, celui-ci se termine avec un rappel de la manière dont les données collectées sont conformes au respect de la vie privée.

5.1.3 Le traitement des données

La collecte des données a donc duré un an. Les sondages effectués par téléphone ont directement été encodés dans un tableau Excel et transmis au fur et à mesure au GRT. Pour les sondages par courrier, les données ont été recopiées manuellement dans ce même tableau Excel. Après avoir collecté l'ensemble des données, le sous-traitant a transmis un fichier définitif au GRT. Après avoir reçu le fichier, le GRT a constaté qu'il comportait de nombreuses erreurs de saisie suite à quelques vérifications automatiques (par exemple, si le trajet dure deux minutes et qu'il a fait 100 kilomètre ou inversément). Le GRT a donc dû réexaminer les courriers reçus par le sous-traitant pour être sûr de la véracité des informations présentes dans le fichier. Le GRT a d'ailleurs mentionné sur son site Beldam que "les résultats nécessitent un minimum de prudence dans leurs interprétations". Suite aux erreurs d'encodage du sous-traitant, le traitement des données reçues a duré presque quatre mois, de décembre à avril 2011.

5.1.4 L'usage de ces données

Une fois les données collectées et traitées, le GRT a organisé une conférence de presse fin de l'année 2011 pour présenter les résultats de son sondage. Comme le GRT avait toutes les données à sa disposition, il a pu réaliser des analyses précises, rigoureuses et complexes sur les habitudes en terme de mobilité : sur l'utilisation des transports en commun, du vélo, sur le covoiturage ou encore la marche à pied. Malheureusement, comme ces données contiennent de nombreuses données identifiantes, celles-ci ne peuvent être partagées à d'autres organismes. C'est pourquoi le GRT a décidé d'anonymiser ces données en accord avec les règles établies par la Commission de la vie privée. Un rapport complet de plus de trois cent pages est disponible sur leur site afin d'examiner les conclusions que l'organisme a pû tirer de cette enquête.

5.2 La diffusion du sondage de manière anonyme

5.2.1 La technique d'anonymisation

Lorsque nous examinons les deux questionnaires, nous constatons qu'une première anonymisation a priori a été effectuée. En effet, le nom de famille et la date de naissance n'ont pas été demandés. Il s'agit déjà de deux données fortement identifiantes volontairement écartées de la recherche. Ce qui est une bonne chose. En effet, seuls le prénom et l'année de naissance sont demandés. Cependant, certaines données identifiantes ont quand même été encodées comme l'adresse de départ pour les voyages à l'étranger ou encore les différentes adresses pour retracer le chemin d'une personne sur une jour-

née. Autant de données qui doivent être anonymisées. C'est pourquoi une deuxième anonymisation, cette fois, a posteriori, a été effectuée. Pour ce faire, le GRT a utilisé deux techniques que nous avons présentées dans le chapitre 3 : la généralisation et la suppression. Ceci afin de permettre de diffuser ce sondage aux personnes désirant étudier ces données, tout en protégeant les données personnelles des personnes sondées. En effet, les données personnelles telles que les différentes adresses ont été supprimées. Seule l'adresse restait visible, généralisée à la localité. L'anonymisation a été établie pour permettre à n'importe quel organisme désireux d'étudier la mobilité et les habitudes des navetteurs, de pouvoir disposer d'une base de données solide pour en tirer certaines conclusions.

5.2.2 L'obtention des données anonymisées

Pour pouvoir obtenir ces données, il faut se rendre sur le site de Beldam et télécharger un formulaire à remplir et à retourner au GRT. Dans ce formulaire, il est inscrit une série de règles qu'il faut respecter suite à l'obtention des données. Celles-ci doivent être utilisées en respectant l'objectif premier du sondage, la mobilité. Elles ne doivent pas être utilisées à des fins commerciales, ni diffusées ni transférées ni vendues à des tiers et ce, même au sein de l'organisme en charge du projet utilisant les données ciblées. Quant à la confidentialité des données, elle est laissée au bon soin du nouveau détenteur des données qui est tenu d'en assurer la plus stricte sécurité. De plus, une description du projet de recherche visant à utiliser ces données doit être spécifiée de manière précise et rigoureuse. Après avoir rempli ce formulaire et l'avoir transmis par courrier à l'organisme, le GRT renvoie à l'adresse mentionnée dans la lettre, un CD-ROM avec les données anonymisées.

Pour obtenir le CD de données, nous avons dû, conformément à la procédure télécharger un formulaire disponible sur le site de Beldam.be en langue française ou néerlandaise. Dans ce formulaire, nous avons dû indiquer nos nom et prénom ainsi que le cadre d'activité dans lequel s'inscrivent nos recherches. Le formulaire spécifiant que cela ne peut se faire que dans le "cadre des recherches, des exploitations décrites en annexe à ce document et ce, en accord avec les termes de ce document". Ensuite, Beldam rappelle quelques principes qui s'inscrivent dans le cadre des prescrits de la Commission de la protection de la vie privée (malheureusement, le lien vers le pdf n'est plus accessible, celui-ci nous redirigeant vers une erreur 404).

Premièrement, le but poursuivi par la recherche doit être d'avoir "une meilleure connaissance de la mobilité en Belgique". Deuxièmement, les données doivent être scrupuleusement utilisées, elles ne peuvent être ni transférées ni vendues ni utilisées à des fins commerciales. De plus, elles doivent être détruites une fois la démarche effectuée. Beldam rappelant également que la confidentialité de ces données (assurées par celle-ci) doit également l'être par

le nouveau détenteur des données, allant même jusqu'à rappeler que "tous les résultats fournis, diffusés ou publiés seront agrégés à un niveau tel que l'anonymat des répondants soit garanti". Beldam demande d'annexer au formulaire une description du projet, du but et de la manière d'exploiter ces données.

Nous avons donc adressé le formulaire dûment complété ainsi qu'une description de notre projet, à savoir l'étude de l'anonymisation effectuée sur ce sondage. Deux jours après, nous avons reçu le disque de données avec la mention "confidentiel" qui contenait une archive zip. Celle-ci était munie d'un mot de passe, de sorte que, si le CD s'égare dans les postes ou autres organismes d'envoi, celui-ci ne puisse être lisible. Une protection simple et efficace. Cependant, en fonction du mot de passe choisi, il est possible, grâce à des logiciels gratuits de le découvrir. Une fois le CD reçu, le GRT envoie par e-mail à l'utilisateur des données le mot de passe pour ouvrir l'archive.

Lors de la décompression de l'archive, nous obtenons seize fichiers. Quatre tableaux Excel et douze fichiers pdfs. Ces quatre tableaux Excel reprennent les données anonymisées du sondage. Le tableau "HH" reprend les données des différents membres du ménage, le tableau "IND" reprend les informations sur les différents individus, le tableau "TRIP" reprend la journée référence demandée dans le questionnaire individuel et, pour finir, le tableau "VEH" reprend toutes les informations sur les véhicules des personnes sondées. Chaque tableau pouvant être relié à l'autre par un numéro d'identifiant unique par personne. Pour chacun de ces tableaux, nous avons des fichiers pdfs qui donnent d'une part, une description des colonnes (les fichiers "label") et, d'autre part, un autre fichier intitulé "format" qui fait correspondre les numéros à la valeur réelle dans ce sondage (les fichiers "format"). Ceux-ci sont disponibles aussi bien en français qu'en anglais.

5.2.3 Une anonymisation trop générale ?

Lors de l'interview que nous avons réalisée de Monsieur Cornelis, coordinateur de ce sondage sur la mobilité, il nous a confirmé que les données avaient été anonymisées et que l'anonymisation avait été validée par la Commission de la vie privée. Celle-ci lui ayant demandé de retirer le prénom de la personne, mais aussi les informations sur les adresses de domicile et de destination. La date de naissance a été généralisée à l'année de naissance, de même que le domicile a été généralisé au code postal. Le GRT s'est donc exécuté et, à la lecture des quatre différents tableaux, nous constatons que ces mesures ont été appliquées à la lettre.

Néanmoins, Monsieur Cornelis m'a fait part des inconvénients de cette anonymisation. En effet, comme le but de ce sondage est d'obtenir le plus d'informations possibles sur les habitudes des citoyens en matière de mobilité, il aimerait pouvoir utiliser en profondeur ces données. Or, par la généralisation, il perd de facto la possibilité d'offrir à d'autres organismes des recherches

avancées en matière de mobilité. Pour bien comprendre ce problème, prenons l'exemple suivant. Bob se rend tous les jours à son travail à Namur depuis Charleroi. Pour ce faire, il prend sa voiture et ensuite, va voir Alice, sa fiancée à Bouges. Si un organisme veut pouvoir offrir une alternative à la voiture à Bob, celui-ci doit connaître son adresse de départ, l'adresse de son boulot ainsi que de sa fiancée, Alice. Or, par la généralisation, l'organisme sait juste qu'il part de telle commune pour arriver à telle autre. Cet organisme ne peut donc offrir d'itinéraire alternatif en transport en commun à Bob. En effet, qu'il parte d'un point ou d'un autre de Charleroi, l'offre de transport en commun peut s'avérer diamétralement différente.

Cela témoigne encore une fois des désavantages de cette méthode et de la difficulté évoquée dans le premier chapitre de trouver un compromis entre qualité et sécurité. En l'occurrence, au profit de la sécurité, de nombreuses informations ont dû être dégradées et le tableau de données a donc perdu en qualité. Le champ d'exploitation de celui-ci se retrouvant, de facto, plus restreint.

Pour preuve, les résultats de ce sondage anonymysés n'ont été demandés que par moins d'une dizaine de personnes comme nous l'a mentionné Monsieur Cornellis. Pourtant, malgré la généralisation, ce tableau regorge d'informations intéressantes et même parfois un peu trop identifiantes comme nous allons le voir par la suite.

5.3 Analyse de la robustesse de l'anonymisation du sondage sur la mobilité

Comme nous l'avons proposé à Monsieur Cornellis, avant d'examiner la possibilité de proposer une autre méthode d'anonymisation moins dégradante pour ces données, nous allons en vérifier la robustesse afin de voir si cette technique d'anonymisation rempli bien son rôle premier, à savoir l'impossibilité de retrouver l'identité d'une personne. Pour vérifier la sécurité du tableau de données et la protection accordée par celle-ci grâce à l'anonymisation, nous avons décidé de nous mettre dans la peau d'un intrus. Nous avons donc décidé de trouver une personne dans ces données afin de montrer que l'anonymisation n'avait pas été effectuée de manière correcte et que les données des personnes présentes dans le tableau pouvaient être facilement révélées en divulguant leurs nom et prénom. Pour ce faire, nous avons utilisé les techniques d'inférence que nous avons évoquées au chapitre 3 sur l'anonymisation des données. Nous avons structuré notre démarche comme suit. Premièrement, nous identifions quelles sont le ou les données identifiantes. Deuxièmemment, une fois la personne choisie, nous collectons toutes les données sur celle-ci. Troisièmement, nous établissons un profil type et nous essayons de croiser les données. Quatrièmement, nous analysons le profil établi et nous en tirons les

conclusions qui s'imposent.

Face au flux de données présentes dans les tableaux, nous avons dû faire un choix parmi la ligne et donc, la personne (ou le ménage). Afin d'effectuer ce choix, nous avons regardé le tableau d'une manière différente, en nous focalisant sur le moindre élément discriminant, différents des autres et donc potentiellement identifiant. Lors de notre examen, nous avons identifié une colonne contenant des valeurs très discrimantes. Intitulée "iq25a" dans le tableau "IND", cette colonne représentait le type de profession pratiquée par les personnes ayant de nombreux déplacements comme nous l'avons vu dans la présentation du questionnaire individuel. Dans cette colonne, nous avons retrouvé de nombreuses professions comme dentiste, médecin, kiné, avocat, etc. Nous avions alors trouvé l'élément qui différencie les personnes les unes des autres. En effet, par la profession exercée par la personne choisie, nous réduisions déjà fortement notre champ de recherche. Notre choix s'est porté sur la personne portant l'identifiant "565172-1" et exercant le métier de "notaire" mais il faut noter que nous aurions très bien pû prendre une autre profession. Cependant, nous avons pris cette personne-ci car le nombre de notaires est restreint en Belgique. De plus, comme les études sont attribuées par concours, nous pensions qu'il était potentiellement possible d'obtenir des informations complémentaires que nous aurions pu croiser (ce qui n'a pas été le cas).

Une fois la profession choisie et donc la personne, nous avons examiné les autres données le concernant. Nous avons constaté que c'était la personne de contact d'une famille de cinq personnes. En analysant le tableau de données IND, nous avons pû obtenir les informations suivantes : la personne de contact est un homme né en 1957, de nationalité belge, titulaire d'un diplôme universitaire et qui exerce une profession libérale. Sa conjointe ou épouse est une femme de nationalité belge, née en 1972, titulaire d'un diplôme universitaire et employée. Le couple a trois enfants : deux filles et un fils. Leur première fille est née en 1991, de nationalité belge, et est étudiante en secondaire générale. La dernière fille est née en 1998, de nationalité belge et est écolière en primaire. Leur fils est né en 1992, de nationalité belge, et est étudiant en secondaire générale.

De plus, nous savons que cette famille habite à 7080. Une rapide visite sur "code postale.be" nous indique qu'il s'agit de Frameries. A la lecture de ces lignes, nous constatons déjà que nous obtenons beaucoup d'informations sur la composition du ménage. Nous verrons d'ailleurs que certaines de ces données nous seront utiles dans le cadre de la découverte de l'identité de la personne. Après avoir examiné ces informations, nous avons décidé de nous tourner vers les autres données du tableau, à savoir les données concernant les déplacements. Nous avons constaté que, dans son formulaire, la personne de contact stipulait qu'il faisait de la marche cinq fois par semaine et qu'il se situait à un kilomètre de son lieu de travail. Grâce à ces informations, nous

avons pu aisément deviner qu'il travaillait dans sa région. En combinant cette dernière donnée et les autres, nous nous retrouvions déjà avec une série de données importantes et potentiellement identifiantes : un notaire de 56 ans, marié à une femme de 41 ans et père de famille de 3 enfants, tous encore à l'école. De plus, grâce à la distance parcourue, nous savons que son cabinet de notaire se situe à Frameries.

Nous avons donc effectué une recherche dans Google mentionnant la phrase suivante "Notaire Frameries 1957", le moteur de recherche nous a alors renvoyé sur de nombreuses pages dont une qui a attiré notre attention. Intitulée " Étude Notaire Vilain à Frameries", cette page nous renvoyait vers l'url "www.notaire.be". Nous nous sommes donc rendus sur ce site qui recense toutes les études notariales du pays et ce, de manière exacte et exhaustive. Pour chaque étude, le site "notaire.be" affiche les coordonnées de l'étude ainsi que la liste des collaborateurs (nom, prénom, adresse e-mail et parfois téléphone). Sur ce site, dans le moteur de recherche proposé à droite, nous pouvions rechercher un notaire par nom ou par code postal. Nous avons donc tapé "7080" et nous avons obtenu deux réponses : l'étude du Notaire Vilain et l'étude du Notaire Raucent. Après un clic sur le premier cité, nous obtenons l'adresse du cabinet d'étude ainsi que l'année d'obtention de celle-ci. Pour le notaire Vilain, il est mentionné qu'il a obtenu l'étude le 8 Janvier 1976. Or, notre notaire est né en 1957, il était donc impossible qu'à 19 ans, il ait obtenu ce cabinet car il faut six ans d'étude pour obtenir le statut de licencié en notariat. Nous avons exclu de manière logique ce premier notaire. Nous sommes donc revenus à la page de sélection et nous avons cliqué sur l'étude du Notaire Paul Raucent. Lors de l'examen de l'obtention de l'étude, nous avons constaté que celui-ci l'obtient le 25 Mai 1993. Un rapide calcul nous permet de déduire que notre notaire aurait pu obtenir celle-ci à l'âge de 36 ans. Une hypothèse qui semble parfaitement plausible. Néammoins, et même s'il s'agit du seul notaire probable sur Frameries, nous ne pouvions nous arrêter à ces comparaisons. Nous voulions avoir la certitude qu'il s'agissait bien de l'homme recherché.

En effet, pour le moment, notre chemin vers la désanonymisation nous a conduit vers ce notaire, mais sans certitude aucune, juste par pure déduction logique et non par preuve affirmative. Pour démontrer qu'il s'agissait bien du notaire en question, nous avions besoin d'une preuve tangible, irréfutable, témoin du lien entre ces deux personnes afin de montrer qu'elles ne font qu'une. Il fallait donc essayer de faire correspondre l'année de naissance de la personne recherchée avec celle du notaire en question. Malheureusement, lors de nos recherches, nous n'avons obtenu aucune donnée supplémentaire sur cette personne. En effet, la recherche sur Google s'est avérée infructueuse. Aucune information ne remontait sur son année de naissance. Nous avons également obtenu les mêmes échecs sur les réseaux communautaires. Le profil Facebook

de Paul Raucent est vide, il ne possède pas de compte Twitter ni de profil Linked et ne participe à aucune activité extra-professionnelle qui pourrait figurer sur internet. En fait, nous avons constaté que le notaire a tout fait pour que sa vie privée le reste, et nous n'arrivons pas à réussi à obtenir la moindre photo, information sur lui en dehors des informations professionnelles présentes sur "notaire.be". La recherche se trouvait dans l'impasse. Sans la correspondance avec l'année de naissance, il était impossible d'affirmer que cette personne était bien le notaire de Frameries, Paul Raucent.

Mais en relisant les informations à notre disposition, nous avons constaté que le site "notaire.be" donnait également d'autres informations complémentaires. En effet, pour chaque étude de notaire, nous pouvons consulter l'historique des différents notaires ayant tenu cette étude. Nous avons donc pensé à retrouver l'acte de passation de l'étude entre André Lemaître, le précédent notaire et Paul Raucent mais nos recherches se sont avérées également infructueuses jusqu'à ce que nous prêtions attention à la constitution de cette étude.

Lors de notre recherche sur Google, nous avons constaté qu'il était indiqué SPRL Étude de Notaire Paul Raucent. Comme toute constitution d'une SPRL se doit d'être publiée au Moniteur Belge, nous nous sommes rendus sur le site du Moniteur Belge, dans sa section base de données, nous avons ensuite tapé le nom de la personne, Paul Raucent, quatorze résultats se sont alors affichés. Sur ces quatorze résultats, nous sommes tombés sur un rapport de la constitution de la SPRL. En annexe, nous avons pu lire la photocopie de toute la publication. L'acte de constitution nous a permis de constater que Paul Raucent était né à Leuven en 1957, comme la personne que nous recherchons. Avec cette correspondance, nos certitudes se confirment quant à l'identité probable de cette personne.

Cependant, afin de montrer que le tableau de données anonymisées recèle un nombre important de données identifiantes, nous avons continué notre recherche afin d'obtenir d'autres facteurs de ressemblance. Dans ce même extrait du Moniteur Belge, nous avons également d'autres informations comme le statut marital de Paul Raucent. En effet, le Moniteur Belge stipule qu'il est marié à Hélène Baudoux. Cette nouvelle information nous confirme que cette personne vit en couple comme la personne que nous recherchons. Un rapide retour sur la liste des collaborateurs de l'étude Paul Raucent sur le site "notaire.be" nous confirme que son épouse travaille bien avec lui, puisqu'elle est mentionnée dans cette liste comme Hélène Baudoux, licenciée en droit et en notariat. De plus, via le Moniteur Belge nous apprenons que celle-ci occupe la fonction de clerc de notaire. Pour obtenir la confirmation que ces données sont les mêmes que celles de notre couple, nous avons examiné le profil de cette personne. Certes, nous n'avons pas l'âge de la personne mais nous avons d'autres données identifiantes. En effet, nous savons qu'elle travaille comme

employée pour son mari. Information qui est confirmée par le type d'emploi que l'épouse de la personne que nous recherchons occupe. De plus, nous savons qu'elle travaille avec son mari. Un rapide examen de son trajet nous confirme que c'est bien le cas. En effet, elle ne met qu'un kilomètre pour aller à son travail et son travail se situe à Frameries, tout comme notre notaire. Autant d'informations qui nous permettent d'augmenter nos certitudes de manière importante.

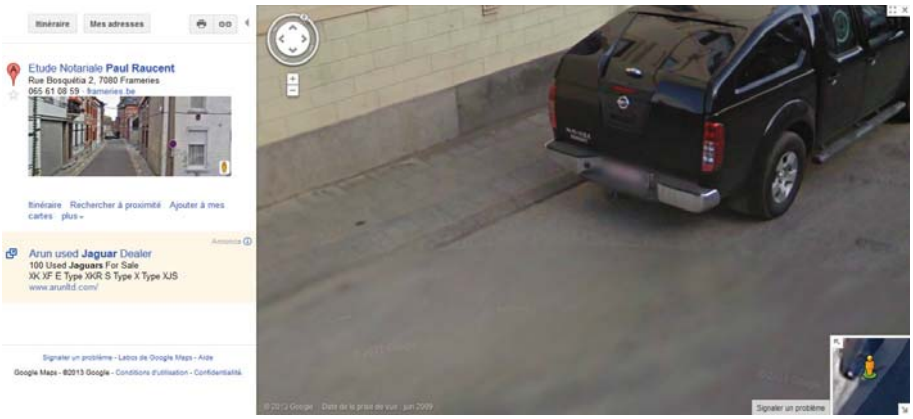


FIGURE 5.1 – Photographie de la Nissan Navara devant l'étude du notaire Paul Raucent réalisée par Google.

A ce stade, nous avions déjà beaucoup de données identiques : même âge, même fonction, même localité de travail, même fonction de son épouse. Rien qu'avec ces informations, notre pourcentage d'inférence s'élevait à plus de 99%. Cependant, nous voulions prouver que nous pouvions encore aller plus loin dans la recherche de l'information concernant une personne.

Suite à l'examen du tableau de données "VH", nous pouvions voir quel véhicule les personnes sondées utilisaient. Dans le cadre de notre recherche, nous avons donc examiné les véhicules de la personne recherchée. Nous avons constaté que le ménage disposait de trois véhicules : une Nissan Navara, une Nissan Qashquai et une Jaguar X-drive. Pour pouvoir prouver que l'une de ces voitures appartient bien à Paul Raucent, nous avons utilisé Google Maps et Google Street View. En effet, ces deux sites vont nous être d'une grande utilité. Nous avons alors tapé dans le moteur de recherche de Google Maps l'adresse de l'étude notariale Paul Raucent. Une fois arrivés sur son adresse, nous avons utilisé Google Street View pour examiner les voitures à proximité de l'étude. Nous avons alors constaté, comme vous pouvez le voir sur la figure 5.1, qu'une Nissan Navara se trouvait juste devant le cabinet lors de la prise de vue de Google Street View. Selon le tableau anonymisé, cette Nissan Navara est un 4X4 de 2600 cylindrés et qui a été acheté en 2010 et construite en 2007. Afin d'être sûrs que cette voiture corresponde bien au modèle de 2007, nous

avons effectué une recherche sur Google pour trouver des modèles similaires. Par comparaison, sur simple photo, nous avons en effet pu constater qu'il s'agit bien d'une Nissan Navara de 2007. Cependant, nous avons constaté que la date à laquelle Google Street View a pris la photo est de 2009. Or, comme la personne a spécifié qu'elle avait acheté sa voiture en 2010, nous ne pouvons affirmer qu'il s'agit de sa voiture. En effet, trois cas de figures sont possibles : soit il ne s'agit pas de la voiture de la personne recherchée, soit il s'agit bien de sa voiture mais qu'elle s'est trompé d'un an dans la date d'achat de son véhicule, soit la date de prise de vue de Google est erronée. Autant de possibilités qui nous renforcent dans l'idée qu'il s'agit bien de la personne concernée.

L'analyse du compte Facebook de Paul Raucent nous a également permis de confirmer, mais de manière moins formelle, l'identité de la personne recherchée. En effet, comme nous l'avions dit, sur son compte Facebook, nous n'obtenons aucune information sur lui. Cependant, nous avons la possibilité de consulter la liste de ses amis. Et c'est justement cette liste qui nous donnera des informations. En effet, dans celle-ci, nous constatons qu'il est ami avec de nombreuses filles et garçons dont le nom de famille est Raucent. En cliquant sur chacun des profils, nous sommes tombés sur le profil de Pauline Raucent, dont le profil était également fermé mais dont l'une des activités indiquaient "Les belles chansons des filles 3E :) par Lena Bengui.". Sachant que le sujet que nous recherchons a une fille née en 1998, il pourrait très bien s'agir de sa fille de 15 ans. Ces données ne nous permettent pas de conclure qu'il s'agit bien de la fille de la personne qui figure dans le ménage avec certitude mais le fait qu'il existe de nombreuses personnes dans ses amis pouvant être considérés comme son fils et ses filles est lourd de sens couplé avec les éléments préalablement cités.

5.4 Propositions d'amélioration

Nous l'avons constaté, le tableau de données anonymisées est dans un état insuffisant pour assurer une protection optimale. En effet, en quelques heures et en choissisant les bons outils, nous avons pu montrer qu'une personne anonymisée pouvait être inférée avec une certitude de l'ordre de 99%. La robustesse du tableau doit être revue d'autant plus que nous avons utilisé des outils gratuits et disponibles en ligne. Il faut garder à l'esprit que l'intrus, s'il veut percer les secrets d'une table, utilisera de nombreux moyens et stratagèmes pour y arriver. Nous l'avons dit, lors de l'examen préalable, le nombre de données présent dans le tableau est trop important. En effet, avec l'ensemble de ces données, nous offrons un nombre incroyable de possibilités d'inférences à la fois sur la profession, le type de véhicule, le nombre de personnes dans le foyer, la localisation, le dernier voyage, etc.

Notre recommandation serait premièrement d'appliquer des techniques d'anonymisation sur certaines données qui n'ont pas ou très peu été anonymisées. L'adresse de la personne ne devrait pas être généralisée à la commune car c'est trop réstrictif (comme le cas de notre notaire à Frameries) mais devrait être masquée de sorte, que par exemple, pour Frameries, nous ayons 7****.

Deuxièmement, les mentions des professions doivent être supprimées ou généralisées par type (profession libérale, employé, etc.). En effet, certaines professions comme notaire, croque-mort sont trop peu nombreuses pour se retrouver isolées. En effet, comme nous l'avons vu avec le modèle k-anonymity, il faut examiner le tableau dans son ensemble pour voir les risques d'inférence. Or, dans ce tableau-ci, il y en a énormément.

Troisièmement, la localisation de départ et d'arrivée pour les voyages à l'étranger doit également être généralisée ou masquée. Quatrièmement, il faut analyser les véhicules présents dans le tableau afin d'éliminer ceux qui pourraient se révéler très identifiants. En effet, lors d'une recherche d'une marque de véhicule inconnue, nous avons constaté qu'il s'agissait d'un tracteur et nous avons alors pu en déduire que la personne sondée était un agriculteur.

De manière générale, ce tableau doit être revu en profondeur et dans son intégralité. Non seulement tableau par tableau mais également de manière globale, en examinant toutes les combinaisons possibles de données. Pour ce faire, une démarche en quatre étapes, expliquées dans le chapitre 3, est nécessaire pour assurer de manière efficace la protection des données. En effet, les failles de sécurité que nous avons pu relever à travers notre exemple du notaire montre à quel point les données initiales ont été sous-estimées et doivent donc faire l'objet d'une analyse détaillée.

5.5 Conclusion

Lors de l'analyse du sondage sur la mobilité, nous avons constaté qu'un nombre important de données étaient collectées. Deux étapes d'anonymisation ont été utilisées : l'une a priori, l'autre a posteriori. Il en ressort que, malgré cette double anonymisation, le tableau de données se révèle vulnérable. En quelques heures, nous avons pu retrouver avec une certitude à 99% les données d'une personne. Par ce fait, nous montrons que le citoyen qui a rempli ces données en étant sûr qu'elles seraient anonymisées est floué. En effet, une fois l'identité de la personne inférée, nous avons, grâce à ce tableau du sondage, un nombre important de données qui sont personnelles. Par exemple, nous savons que Paul Raucent a une Jaguar. Une donnée intéressante pour un voleur. Nous avons des données sur son salaire, ses dernières vacances, la composition de sa famille, son niveau d'étude, sa journée type, son parc de voitures, etc. Autant de données identifiantes qui doivent être protégées sous peine de dévoiler la vie privée d'un individu qui se croyait pourtant à l'abri de ce genre de divulgation.

Conclusion

L'examen du mécanisme d'anonymisation a été révélateur de plusieurs points critiques sur le traitement des données à caractère personnel.

Premièrement, dans la majorité des cas, les données à caractère personnel ne sont pas protégées. Nous l'avons vu avec le problème de la SCNB. Les données n'étaient pas anonymisées et traînaient sur leurs serveurs. Or, lorsque ces données sont collectées, la loi impose que celles-ci soient protégées. Ce décalage s'explique malheureusement par le coût engendré par ce type de protection, surtout lorsque le nombre de données n'est pas élevé, comme le soulignait les PME suite à la proposition de la nouvelle directive de Viviane Reding.

Deuxièmement, lorsque ces données sont protégées, elles ne le sont pas suffisamment de sorte que celle-ci peuvent être découvertes par simple inférence (déductive, abductive, etc.). Le cas pratique du sondage sur la mobilité en est une preuve flagrante. A juste titre, le GRT s'est appuyé sur la Commission de la vie privée pour savoir de quelle manière ces données devaient être exploitées. De par son accord sur ces données anonymisées, le GRT les a alors autorisées à la diffusion en toute bonne foi et persuadé que ces données ne pourraient se révéler identifiantes. Or, nous avons constaté que ce n'était pas le cas. Dans ce cadre, nous avons donc vu que deux acteurs étaient trompés : l'organisme de recherche (le GRT) et le citoyen ayant rempli le formulaire. En effet, celui-ci à la lecture du formulaire et avant le remplissage, retrouvait sur papier certaines certitudes comme l'utilisation de manière anonyme de ses données.

Troisièmement, pour protéger les données personnelles, nous avons constaté qu'il n'existe aucune technique optimale mais que l'utilisation d'un mécanisme d'anonymisation doit utiliser une ou kkkplusieurs techniques combinées pour arriver au résultat souhaité. Ces techniques doivent être envisagées en fonction des besoins, des objectifs et des exigences de l'utilisateur. Le modèle k-anonymity nous a d'ailleurs permis d'en montrer un exemple concret. Différentes failles de sécurité ont alors été révélées. Cela nous a permis de nous rendre compte de la difficulté d'obtenir une protection des données parfaite. Même les améliorations de ce modèle n'ont pu proposer une solution totalement sécuritaire.

L'analyse de l'anonymisation nous a permis de constater sa nécessité, de plus en plus grandissante, suite à l'afflux de données personnelles mais également sa difficulté. En partant d'un point de vue théorique et en analysant un cas pratique, nous avons démontré qu'il fallait suivre une démarche structurée pour effectuer une anonymisation correcte sur les données collectées. En effet, une analyse des données identifiantes et des recoupements possibles entre les données auraient pu éviter de retrouver le nom du notaire figurant dans les données personnelles anonymisées du sondage sur la mobilité.

Enfin, il ne faut pas sous-estimer la prolixité de nos données (souvent, nos données peuvent en dire plus que nous le pensons), le contexte dans lequel nos données sont diffusées et l'éco-système entourant nos données (existe-il des annuaires sur les notaires, avocats, comédiens ?). En effet, de par la volonté de centraliser toute une série de données pour notre facilité, l'internet d'aujourd'hui permet facilement de combiner plusieurs données a priori non identifiantes mais dont la réunion permet de le devenir. De plus, ce danger n'est malheureusement pas contrôlable car nous n'avons pas le contrôle sur toutes les données nous concernant sur internet. De par notre fonction, nos amis, notre parcours, de nombreuses données sont diffusées sans que notre accord en soit explicitement demandé. Une grande prudence doit donc être appliquée à toutes les données que nous diffusons, aux données dont nous autorisons la diffusion et au traitement de celles-ci. Car avec ces données, c'est notre vie qui pourrait se retrouver à notre insu dévoilée sur n'importe quel site et ce, sans notre accord. Notre vie privée s'en trouverait alors bafouée, contrairement aux lois en vigueur et à notre propre volonté.

Bibliographie

[1] L'Express. Les chasseurs de têtes vont à la pêche sur le net, Consulté le 11 avril 2013. http://www.lexpress.fr/actualite/high-tech/les-chasseurs-de-tetes-vont-a-la-peche-sur-le-net_735441.html. (Cité en page vi.)

[2] Mailys Charlier. La sncb divulgue les coordonnées privées de ses clients sur internet, Consulté le 11 avril 2013. http://www.lesoir.be/143379/article/actualite/belgique/2012-12-24/sncb-divulgue-coordonnées-privées-ses-clients-sur-internet. (Cité en page vii.)

[3] Claude de Decker. Sncb : les données personnelles de magnette rendues publiques, Consulté le 11 avril 2013. http://www.lesoir.be/145627/article/actualite/belgique/2012-12-28/sncb-données-personnelles-magnette-rendues-publiques. (Cité en page vii.)

[4] Emilien Ercolani. Une faille de sécurité dropbox laisse les comptes accessibles, Consulté le 12 décembre 2012. http://www.linformaticien.com/actualites/id/20972/une-faille-de-securite-dropbox-laisse-les-comptes-accessibles.aspx. (Cité en page vii.)

[5] Google. Des règles de confidentialité unifiées pour une expérience google unique, Consulté le 7 février 2012. http://www.google.com/policies/?hl=fr. (Cité en page 1.)

[6] Le Monde. Google invité à expliquer sa charte de confidentialité au congrès américain, Consulté le 7 février 2012. http://www.lemonde.fr/technologies/article/2012/01/31/google-invite-a-expliquer-sa-charte-de-confidentialite-au-congres-americain_1636698_651865.html. (Cité en page 1.)

[7] Le Monde. Facebook franchit la barre du milliard d'utilisateurs, Consulté le 10 décembre 2012. http://www.lemonde.fr/technologies/article/2012/10/04/facebook-franchit-la-barre-du-milliard-d-utilisateurs_1770255_651865.html. (Cité en page 1.)

[8] Zdnet. Facebook : des photos supprimées peuvent rester en ligne plusieurs mois, Consulté le 5 février 2012. http:

//www.zdnet.fr/actualites/facebook-des-photos-supprimees-peuvent-rester-en-ligne-plusieurs-mois-39755440.htm. (Cité en page 2.)

[9] Le Monde. Interdiction des pseudonymes : victoire judiciaire pour facebook en allemagne, Consulté le 19 février 2013. http://www.lemonde.fr/technologies/article/2013/02/15/interdiction-des-pseudonymes-victoire-judiciaire-pour-facebook-en-allemagne_1833410_651865.html. (Cité en page 2.)

[10] Commission Européenne. Attitudes on data protection and electronic identity in the european union, Consulté le 5 février 2012. http://ec.europa.eu/public_opinion/archives/eb_special_359_340_en.htm. (Cité en page 2.)

[11] New York Times. To aim ads, web is keeping closer eye on you, Consulté le 10 février 2013. http://www.nytimes.com/2008/03/10/technology/10privacy.html?pagewanted=1&_r=2&hp&. (Cité en page 2.)

[12] Le Huffington Post. Euro 2012 : Les députés allemands ont profité du match de foot pour faire passer une loi sur les données personnelles, Consulté le 16 février 2012. http://www.huffingtonpost.fr/2012/07/09/euro-2012-deputes-allemand-foot-euro-match-loi-donnee-personnelle_n_1658471.html?utm_hp_ref=france. (Cité en page 3.)

[13] P. Samarati. Protecting respondents identities in microdata release. *Knowledge and Data Engineering, IEEE Transactions on*, 13(6) :1010–1027, 2001. (Cité en pages 3, 44, 45 et 46.)

[14] 01Net. Les données de 30 millions de clients de deutsche telekom en accès libre, Consulté le 10 février 2013. http://www.01net.com/editorial/393163/les-donnees-de-30-millions-de-clients-de-deutsche-telekom-en-acces-libre. (Cité en page 3.)

[15] L'Express. Orange a laissé un accès libre à 400.000 fiches de ses clients, Consulté le 10 février 2013. http://lexpansion.lexpress.fr/high-tech/orange-a-laisse-un-acces-libre-a-400-000-fiches-de-ses-clients_171365.html. (Cité en page 3.)

[16] New York Times. A face is exposed for aol searcher no. 4417749, Consulté le 10 février 2013. http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all. (Cité en page 3.)

[17] Organisation for Economic Co-operation and Development(OECD). Recommendation of the council concerning guidelines governing the protection of privacy and transborder flows of personal data, 23 septembre 1980. (Cité en page 4.)

[18] Etats membres du Conseil de l'Europe. Convention 108 pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, 28 janvier 1981. (Cité en pages 4 et 5.)

[19] C. Nicomette V. et Roy M. Deswarte, Y. et Aguilar-Melchor. Protection de la vie privée sur internet. *Revue de l'Électricité et de l'Électronique (REE*, (9) :65–74, 2006. (Cité en page 4.)

[20] Directive 95/46/ce du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation des données. (Cité en page 5.)

[21] Arrêté royal du 13 février 2001 portant exécution de la loi du 8 décembre 1992 relative à la protection de la vie privée à l'égard des traitements de données à caractère personnel. (Cité en page 5.)

[22] Loi relative à la protection de la vie privée à l'égard des traitements de données à caractère personnel, 8 décembre 1992. (Cité en page 5.)

[23] Directive 97/66/ce du parlement européen et du conseil du 15 décembre 1997 concernant le traitement des données à caractère personnel et la protection de la vie privée dans le secteur des télécommunications. (Cité en page 6.)

[24] Directive 2002/58/ce du parlement européen et du conseil du 12 juillet 2002 concernant le traitement des données à caractère personnel et la protection de la vie privée dans le secteur des communications électroniques (directive vie privée et communications électroniques). (Cité en page 6.)

[25] Directive 2006/24/ce du parlement européen et du conseil du 15 mars 2006 sur la conservation de données générées ou traitées dans le cadre de la fourniture de services de communications électroniques accessibles au public ou de réseaux publics de communications, et modifiant la directive 2002/58/ce,. (Cité en page 6.)

[26] Commission Européenne. Proposal for a regulation of the european parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (general data protection regulation, 25 janvier 2012. (Cité en page 6.)

[27] Elise Defreyne et Quentin Van Enis. La loi belge sur la protection des données à caractère personnel et l'anonymisation des archives de presse en ligne, Consulté le 27 avril 2013. `http://e-watchdog.overblog.com/la-loi-belge-sur-la-protection-des-donnees-a-caractere-personnel-et-l-anonymisation-des-archives-de-presse-en-ligne`. (Cité en page 6.)

[28] Projet de rapport sur la proposition de règlement du parlement européen et du conseil 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. (Cité en page 7.)

[29] ActuaLitte. Données personnelles, la recherche perd-elle la mémoire ?, Consulté le 7 avril 2013. `http://www.actualitte.com/`

patrimoine/les-archivistes-en-lutte-pour-la-memoire-de-l-europe-41518.htm. (Cité en page 7.)

[30] AFP. Un projet de loi pour protéger vos données sur le net prêt pour 2013, Consulté le 10 janvier 2013. http://www.01net.com/editorial/578143/un-projet-de-loi-pour-proteger-vos-donnees-sur-le-net-pret-pour-2013/. (Cité en page 8.)

[31] CNIL. Les conclusions de la cnil sur le bug facebook, Consulté le 03 janvier 2013. http://www.cnil.fr/la-cnil/actualite/article/article/les-conclusions-de-la-cnil-sur-le-bug-facebook/. (Cité en page 8.)

[32] Conseil de l'Europe. Comité des ministres. Recommandation rec(97)18 sur la protection des données à caractère personnel, collectées et traitées à des fins statistiques, 30 septembre 1997. (Cité en page 9.)

[33] Belleil A. Réferentiel afcdp des dispositifs d'anonymisation, 28 mai 2008. http://www.afcdp.net/IMG/pdf/AFCDP_Referentiel_Anonymisation_080522-3.pdf. (Cité en pages 10, 24 et 34.)

[34] ICO. Anonymisation : managing data protection risk code of practice, 2012. http://www.ico.org.uk/Global/-/media/documents/library/Data_Protection/Practical_application/anonymisation_code.ashx. (Cité en pages 10, 12, 24, 28, 29 et 38.)

[35] CNIL. Guide : La sécurité des données personnelles, 2010. http://www.cnil.fr/fileadmin/documents/Guides_pratiques/Guide_securite-VD.pdf. (Cité en pages 10 et 18.)

[36] CNIL. Mesurer pour progresser vers lŠégalité des chances, Consulté le 10 décembre 2012. http://www.defenseurdesdroits.fr/sites/default/files/upload/promotion_de_legalite/progress/fiches/ldd_cnil_interactif.pdf. (Cité en page 11.)

[37] Office National de Statistique. Guidance and methodology, Consulté le 17 février 2012. http://www.ons.gov.uk/ons/guide-method/index.html. (Cité en page 12.)

[38] Loi numéro 2004-810 du 13 août 2004 relative à l'assurance maladie. (Cité en page 15.)

[39] Ludwig Edelstein. *The Hippocratic Oath : Text, Translation and Interpretation, Supplement to the History of Medicine*. The Johns Hopkins Press, 1943. (Cité en page 15.)

[40] Y. et Trouessin G. et Cordonnier E. El Kalam, AA et Deswarte. Personal data anonymization for security and privacy in collaborative environments. In *Collaborative Technologies and Systems, 2005. Proceedings of the 2005 International Symposium on*, pages 56–61. IEEE, 2005. (Cité en page 16.)

[41] Trouessin G. Sécurité et intimité des données à caractère personnel. *La lettre d'ADELI*, 44 :35–44, 2001. (Cité en pages 16 et 20.)

[42] Académie des sciences morales et politiques. Groupe d'études société d'information et vie privée, Consulté le 10 février 2012. `http://www.asmp.fr/travaux/gpw/internetvieprivee/rapport3/chapitr7.pdf`. (Cité en page 17.)

[43] B. Elger. La protection de la personnalité et des données : l'anonymisation irréversible comme dilemme éthique. *Bulletin des médecins suisses*, 45 :2510–2512, 2006. (Cité en page 17.)

[44] Yves et Trouessin Gilles et Cordonnier Emmanuel El Kalam, Anas Abou et Deswarte. Une démarche méthodologique pour l'anonymisation de données personnelles sensibles. In *Actes du symposium SSTIC04*, 2004. (Cité en page 18.)

[45] Bruce Schneier. A revised taxonomy of social networking data, Consulté le 15 octobre 2012. `http://www.schneier.com/blog/archives/2010/08/a_taxonomy_of_s_1.html`. (Cité en page 20.)

[46] Michel Arnaud. Le passe navigo anonyme revisité, Consulté le 14 avril 2013. `http://www.lecreis.org/colloquescreis/2010/PasseNavigoanonymeMichelArnaud.pdf`. (Cité en page 21.)

[47] CNIL. 2ème contrôle des passes anonymes navigo, Consulté le 14 avril 2013. `http://www.cnil.fr/la-cnil/actu-cnil/article/article/2eme-controle-des-passes-anonymes-navigo/`. (Cité en page 22.)

[48] CNIL. Testing de la cnil auprès de la ratp : l'exercice du droit des usagers à se déplacer anonymement n'est pas garanti, Consulté le 14 avril 2013. `http://www.cnil.fr/la-cnil/actu-cnil/article/article/testing-de-la-cnil-aupres-de-la-ratp-lexercice-du-droit-des-usagers-a-se-deplacer-anonymement/`. (Cité en page 22.)

[49] Patrick Chambet Jean-Luc Lambert. L'anonymisation de données en masse chez bouygues telecom, Consulté le 16 février 2012. `http://www.ossir.org/jssi/jssi2011/1B.pdf`. (Cité en page 26.)

[50] W.E. et al. Kim, J.J. et Winkler. Masking microdata files. In *Proceedings of the Survey Research Methods Section, American Statistical Association*. Citeseer, 1995. (Cité en page 28.)

[51] S. et Foresti S. et Samarati P. Ciriani, V. et Capitani di Vimercati. κ-anonymity. *Secure Data Management in Decentralized Systems*, pages 323–353, 2007. (Non cité.) ? ? ? ?

[52] Peter et Willenborg LCRJ et De Wolf Peter-Paul Gouweleeuw, JM et Kooiman. Post randomisation for statistical disclosure control : Theory

and implementation. *Journal of official statistics*, 14 :463–478, 1998. (Cité en page 29.)

[53] Friedrich L.Bauer. *Decrypted Secrets : Methods and Maxims of Cryptology*. Springer, 2010. (Cité en page 34.)

[54] REVER S.A. Jeux de données anonymisées, Consulté le 16 avril 2013. http://www.rever.eu/rever_new/fr/content/test-data-management. (Cité en page 35.)

[55] CNIL. L'état des lieux en matière de procédés d'anonymisation, Consulté le 16 avril 2013. http://www.cnil.fr/la-cnil/actu-cnil/article/article/letat-des-lieux-en-matiere-de-procedes-danonymisation. (Cité en page 36.)

[56] Le Monde. Note2be : les élèves évaluent les professeurs, Consulté le 15 avril 2014. http://www.lemonde.fr/societe/article/2008/02/20/note2be-les-eleves-evaluent-les-professeurs_1013612_3224.html. (Cité en page 37.)

[57] Le Monde. La cnil juge le site de notation des professeurs note2be.com "illégitime", Consulté le 15 avril 2014. http://www.lemonde.fr/societe/article/2008/03/06/la-cnil-juge-le-site-de-notation-des-professeurs-note2be-com-illegitime_1019770_3224.html. (Cité en page 37.)

[58] Philippe Crouzillacq. La justice dit non aux noms des professeurs sur note2be, Consulté le 15 avril 2014. http://www.01net.com/editorial/372605/la-justice-dit-non-aux-noms-des-professeurs-sur-note2be/. (Cité en page 37.)

[59] Institut national de la santé et de la recherche médicale. Progeria (syndrome de hutchinson-gilford), Consulté le 17 février 2012. http://www.inserm.fr/thematiques/genetique-genomique-et-bioinformatique/dossiers-d-information/progeria-syndrome-de-hutchinson-gilford. (Cité en page 38.)

[60] Robert Lemos. Researchers reverse netflix anonymization, Consulté le 16 février 2012. http://www.securityfocus.com/news/11497. (Cité en page 38.)

[61] Vitaly Narayanan, Arvind et Shmatikov. Robust de-anonymization of large sparse datasets. In *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125. IEEE, 2008. (Cité en page 39.)

[62] L. Sweeney. k-anonymity : A model for protecting privacy. *International Journal on Uncertainty Fuzziness and Knowledgebased Systems*, 10(5) :557–570, 2002. (Cité en pages 42 et 43.)

[63] Y. et Le J. Luo, Y. et Zhao. A survey on the privacy preserving algorithm of association rule mining. In *Electronic Commerce and Security, 2009.*

ISECS'09. Second International Symposium on, volume 1, pages 241–245. IEEE, 2009. (Cité en page 42.)

[64] L. Samarati, P. et Sweeney. Protecting privacy when disclosing information : k-anonymity and its enforcement through generalization and suppression. Technical report, Technical report, SRI International, 1998. (Cité en pages 42, 43, 44, 51, 52 et 53.)

[65] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems*, 10(5) :571–588, 2002. (Cité en pages 44, 50 et 51.)

[66] L. Sweeney. Guaranteeing anonymity when sharing medical data, the datafly system. In *Proceedings of the AMIA Annual Fall Symposium*, page 51. American Medical Informatics Association, 1997. (Cité en pages 48 et 49.)

[67] S. Hundepool, A. et Netherlands. The argus-software. *Monographs of official statistics*, page 347, 2004. (Cité en pages 50, 51 et 52.)

[68] R. Bayardo, R.J. et Agrawal. Data privacy through optimal k-anonymization. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 217–228. IEEE, 2005. (Cité en page 53.)

[69] D.J. et Ramakrishnan R. LeFevre, K. et DeWitt. Incognito : Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60. ACM, 2005. (Cité en pages 53 et 55.)

[70] D.J. et Ramakrishnan R. LeFevre, K. et DeWitt. Workload-aware anonymization techniques for large-scale datasets. *ACM Transactions on Database Systems (TODS)*, 33(3) :17, 2008. (Cité en page 53.)

[71] D.J. et Ramakrishnan R. LeFevre, K. et DeWitt. Mondrian multidimensional k-anonymity. In *Data Engineering, 2006. ICDE'06. Proceedings of the 22nd International Conference on*, pages 25–25. Ieee, 2006. (Cité en pages 56 et 57.)

[72] D. et Gehrke J. et Venkitasubramaniam M. Machanavajjhala, A. et Kifer. l-diversity : Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1) :3, 2007. (Cité en pages 57 et 58.)

[73] Xiangling Wang, Qian et Shi. (a, d)-diversity : Privacy protection based on l-diversity. In *Software Engineering, 2009. WCSE'09. WRI World Congress on*, volume 3, pages 367–372. IEEE, 2009. (Cité en page 60.)

[74] Y. et Geng L. et Liu H. Wang, Y. et Cui. A new perspective of privacy protection : Unique distinct l-sr diversity. In *Privacy Security and Trust*

(PST), 2010 Eighth Annual International Conference on, pages 110–117. IEEE, 2010. (Cité en page 60.)

[75] T. et Venkatasubramanian S. Li, N. et Li. t-closeness : Privacy beyond k-anonymity and l-diversity. In *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*, pages 106–115. IEEE, 2007. (Cité en pages 60, 61 et 68.)

[76] J. et Fu A.W.C. et Wang K. Wong, R.C.W. et Li. (α, k)-anonymity : an enhanced k-anonymity model for privacy preserving data publishing. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 754–759. ACM, 2006. (Cité en page 62.)

[77] Y. Xiao, X. et Tao. Anatomy : Simple and effective privacy preservation. In *Proceedings of the 32nd international conference on Very large data bases*, pages 139–150. VLDB Endowment, 2006. (Cité en page 63.)

[78] N. et Kalnis P. Terrovitis, M. et Mamoulis. Privacy-preserving anonymization of set-valued data. *Proceedings of the VLDB Endowment*, 1(1) :115–125, 2008. (Cité en page 66.)

[79] Y. Xiao, X. et Tao. M-invariance : towards privacy preserving republication of dynamic datasets. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 689–700. ACM, 2007. (Cité en page 68.)

Annexe

QUESTIONNAIRE DU MÉNAGE

À REMPLIR PAR LA PERSONNE DE CONTACT

Ce questionnaire s'adresse à la personne de contact du ménage.

Nous vous remercions d'avance:

- de remplir ce questionnaire;
- de remplir un questionnaire individuel;
- de veiller à ce que chaque personne de votre ménage (à partir de 6 ans) remplisse un questionnaire individuel. Il est utile pour cela de lire les instructions aux pages 10 à 12 du questionnaire individuel avant le jour de référence;
- de renvoyer tous les questionnaires complétés dans l'enveloppe fournie, le plus rapidement possible après les avoir complétés. Un timbre n'est pas nécessaire (port payé par le destinataire).

Comment remplir le questionnaire ?
Il y a trois manières de répondre aux questions:

	cocher un(des) cercle(s) pour choisir votre(vos) réponse(s) dans une liste ou un tableau;
	écrire un nombre dans une case. Si votre réponse est "aucun" ou "ne s'applique pas", alors indiquer 0;
.....	écrire votre réponse en IMPRIMÉ sur des pointillés.

Votre participation est précieuse.

Les renseignements que vous nous communiquerez seront traités de manière anonyme (*).

Si vous souhaitez plus d'informations ou de l'aide pour remplir le questionnaire,

appelez gratuitement le **0800/80.770**
du lundi au vendredi
de 9h00 à 20h00.

Vragenlijsten in het Nederlands zijn beschikbaar op aanvraag.
Fragebögen auf Deutsch sind auf Anfrage erhältlich.

(*) conformément à la déclaration à la Commission de la protection de la vie privée, figurant en page 8.

LES MEMBRES DE VOTRE MENAGE

1 Dans le tableau ci-dessous, remplissez une colonne pour chaque membre de votre ménage (sans aucune limite d'âge).
Par ménage, nous entendons les personnes, de votre famille ou non, qui habitent, même partiellement, le même logement et partagent le même budget.

	Personne 1	Personne 2	Personne 3
Prénom	.....	.....	.....
Année de naissance	☐☐☐☐	☐☐☐☐	☐☐☐☐
Sexe	1 ○ masculin 2 ○ féminin	1 ○ masculin 2 ○ féminin	1 ○ masculin 2 ○ féminin
Nationalité	.....	.....	.....
Relation par rapport à la personne de contact *1 seule réponse possible*	1 ○ personne de contact 2 ○ époux(se), compagnon(agne) 3 ○ enfant (même adresse officielle ou autre adresse officielle) 4 ○ autre	1 ○ personne de contact 2 ○ époux(se), compagnon(agne) 3 ○ enfant (même adresse officielle ou autre adresse officielle) 4 ○ autre	1 ○ personne de contact 2 ○ époux(se), compagnon(agne) 3 ○ enfant (même adresse officielle ou autre adresse officielle) 4 ○ autre
Diplôme ou certificat le plus élevé obtenu à ce jour *1 seule réponse possible*	1 ○ pas concerné(e) car moins de 12 ans 2 ○ aucun diplôme 3 ○ primaire 4 ○ secondaire général (inférieur ou supérieur) 5 ○ secondaire technique (inférieur ou supérieur) 6 ○ secondaire professionnel (inférieur ou supérieur) 7 ○ secondaire enseignement spécial 8 ○ supérieur non universitaire, 2 à 3 ans 9 ○ supérieur non universitaire, 4 à 5 ans 10 ○ universitaire	1 ○ pas concerné(e) car moins de 12 ans 2 ○ aucun diplôme 3 ○ primaire 4 ○ secondaire général (inférieur ou supérieur) 5 ○ secondaire technique (inférieur ou supérieur) 6 ○ secondaire professionnel (inférieur ou supérieur) 7 ○ secondaire enseignement spécial 8 ○ supérieur non universitaire, 2 à 3 ans 9 ○ supérieur non universitaire, 4 à 5 ans 10 ○ universitaire	1 ○ pas concerné(e) car moins de 12 ans 2 ○ aucun diplôme 3 ○ primaire 4 ○ secondaire général (inférieur ou supérieur) 5 ○ secondaire technique (inférieur ou supérieur) 6 ○ secondaire professionnel (inférieur ou supérieur) 7 ○ secondaire enseignement spécial 8 ○ supérieur non universitaire, 2 à 3 ans 9 ○ supérieur non universitaire, 4 à 5 ans 10 ○ universitaire
Statut professionnel à ce jour *Si vous en avez plusieurs, citez votre statut professionnel principal.* *1 seule réponse possible*	1 ○ enfant non scolarisé(e) 2 ○ écolier, étudiant(e) 3 ○ femme/homme au foyer 4 ○ chercheur(se) d'emploi 5 ○ (pré)pensionné(e) 6 ○ invalide 7 ○ ouvrier(ère) 8 ○ cadre 9 ○ employé(e) 10 ○ indépendant(e) 11 ○ profession libérale 12 ○ enseignant(e) 13 ○ agriculteur(trice) 14 ○ autre (précisez):	1 ○ enfant non scolarisé(e) 2 ○ écolier, étudiant(e) 3 ○ femme/homme au foyer 4 ○ chercheur(se) d'emploi 5 ○ (pré)pensionné(e) 6 ○ invalide 7 ○ ouvrier(ère) 8 ○ cadre 9 ○ employé(e) 10 ○ indépendant(e) 11 ○ profession libérale 12 ○ enseignant(e) 13 ○ agriculteur(trice) 14 ○ autre (précisez):	1 ○ enfant non scolarisé(e) 2 ○ écolier, étudiant(e) 3 ○ femme/homme au foyer 4 ○ chercheur(se) d'emploi 5 ○ (pré)pensionné(e) 6 ○ invalide 7 ○ ouvrier(ère) 8 ○ cadre 9 ○ employé(e) 10 ○ indépendant(e) 11 ○ profession libérale 12 ○ enseignant(e) 13 ○ agriculteur(trice) 14 ○ autre (précisez): ....

Si votre ménage comporte **plus de 6 personnes**, vous pouvez demander une page supplémentaire au numéro gratuit indiqué en première page, ou en télécharger une à l'adresse Internet suivante : www.beldam.be

	Personne 4	Personne 5	Personne 6
Prénom	.....	.....	.....
Année de naissance	☐☐☐☐	☐☐☐☐	☐☐☐☐
Sexe	1 ○ masculin 2 ○ féminin	1 ○ masculin 2 ○ féminin	1 ○ masculin 2 ○ féminin
Nationalité	.....	.....	.....
Relation par rapport à la personne de contact *1 seule réponse possible*	1 ○ personne de contact 2 ○ époux(se), compagnon(agne) 3 ○ enfant (même adresse officielle ou autre adresse officielle) 4 ○ autre	1 ○ personne de contact 2 ○ époux(se), compagnon(agne) 3 ○ enfant (même adresse officielle ou autre adresse officielle) 4 ○ autre	1 ○ personne de contact 2 ○ époux(se), compagnon(agne) 3 ○ enfant (même adresse officielle ou autre adresse officielle) 4 ○ autre
Diplôme ou certificat le plus élevé obtenu à ce jour *1 seule réponse possible*	1 ○ pas concerné(e) car moins de 12 ans 2 ○ aucun diplôme 3 ○ primaire 4 ○ secondaire général (inférieur ou supérieur) 5 ○ secondaire technique (inférieur ou supérieur) 6 ○ secondaire professionnel (inférieur ou supérieur) 7 ○ secondaire enseignement spécial 8 ○ supérieur non universitaire, 2 à 3 ans 9 ○ supérieur non universitaire, 4 à 5 ans 10 ○ universitaire	1 ○ pas concerné(e) car moins de 12 ans 2 ○ aucun diplôme 3 ○ primaire 4 ○ secondaire général (inférieur ou supérieur) 5 ○ secondaire technique (inférieur ou supérieur) 6 ○ secondaire professionnel (inférieur ou supérieur) 7 ○ secondaire enseignement spécial 8 ○ supérieur non universitaire, 2 à 3 ans 9 ○ supérieur non universitaire, 4 à 5 ans 10 ○ universitaire	1 ○ pas concerné(e) car moins de 12 ans 2 ○ aucun diplôme 3 ○ primaire 4 ○ secondaire général (inférieur ou supérieur) 5 ○ secondaire technique (inférieur ou supérieur) 6 ○ secondaire professionnel (inférieur ou supérieur) 7 ○ secondaire enseignement spécial 8 ○ supérieur non universitaire, 2 à 3 ans 9 ○ supérieur non universitaire, 4 à 5 ans 10 ○ universitaire
Statut professionnel à ce jour *Si vous en avez plusieurs, citez votre statut professionnel principal.* *1 seule réponse possible*	1 ○ enfant non scolarisé(e) 2 ○ écolier, étudiant(e) 3 ○ femme/homme au foyer 4 ○ chercheur(se) d'emploi 5 ○ (pré)pensionné(e) 6 ○ invalide 7 ○ ouvrier(ère) 8 ○ cadre 9 ○ employé(e) 10 ○ indépendant(e) 11 ○ profession libérale 12 ○ enseignant(e) 13 ○ agriculteur(trice) 14 ○ autre (précisez):	1 ○ enfant non scolarisé(e) 2 ○ écolier, étudiant(e) 3 ○ femme/homme au foyer 4 ○ chercheur(se) d'emploi 5 ○ (pré)pensionné(e) 6 ○ invalide 7 ○ ouvrier(ère) 8 ○ cadre 9 ○ employé(e) 10 ○ indépendant(e) 11 ○ profession libérale 12 ○ enseignant(e) 13 ○ agriculteur(trice) 14 ○ autre (précisez):	1 ○ enfant non scolarisé(e) 2 ○ écolier, étudiant(e) 3 ○ femme/homme au foyer 4 ○ chercheur(se) d'emploi 5 ○ (pré)pensionné(e) 6 ○ invalide 7 ○ ouvrier(ère) 8 ○ cadre 9 ○ employé(e) 10 ○ indépendant(e) 11 ○ profession libérale 12 ○ enseignant(e) 13 ○ agriculteur(trice) 14 ○ autre (précisez): ....

LES VÉHICULES AU SEIN DE VOTRE MÉNAGE

2 Pour chaque type de véhicule proposé ci-contre, combien sont présents dans votre ménage?
Nous sommes intéressés par tous les véhicules <u>en ordre de fonctionnement</u> dont votre ménage dispose pour un usage privé, donc aussi par les véhicules de société s'ils peuvent être utilisés en dehors du contexte du travail.

Veuillez remplir le tableau ci-dessous en cochant, pour chaque type de véhicule, le cercle qui correspond au nombre de véhicules. Une seule réponse par ligne.

	Nombre de véhicules					
Type de véhicule	0	1	2	3	4	5 et plus
Vélo d'enfant	○ 0	○ 1	○ 2	○ 3	○ 4	○ 5
Vélo d'adulte	○ 0	○ 1	○ 2	○ 3	○ 4	○ 5
Cyclomoteur (moins de 50 cc)	○ 0	○ 1	○ 2	○ 3	○ 4	○ 5
Moto (50 cc ou plus)	○ 0	○ 1	○ 2	○ 3	○ 4	○ 5
Voiture (y compris monospace et 4x4, max. 9 personnes)	○ 0	○ 1	○ 2	○ 3	○ 4	○ 5
Camionnette, pick-up	○ 0	○ 1	○ 2	○ 3	○ 4	○ 5
Autre (précisez):.	○ 0	○ 1	○ 2	○ 3	○ 4	○ 5

3a Dans le tableau ci-dessous, remplissez une <u>colonne</u> pour chaque <u>vélomoteur ou moto</u> que vous avez cité à la question 2.
- Si aucun vélomoteur ou moto n'est présent dans votre ménage, allez à la question 3b ci-contre.
- S'il y a plus de trois vélomoteurs ou motos présents dans votre ménage, ne remplissez le tableau que pour les trois vélomoteurs ou motos les plus utilisés.
- Si nécessaire, vous trouverez les réponses à certaines questions sur les documents du véhicule.

	Cyclomoteur/moto 1	**Cyclomoteur/moto 2**	**Cyclomoteur/moto 3**
Marque *par ex.: "Honda"*	. . . . .	. . . . .	. . . . .
Modèle *par ex.: "Wallaroo"*	. . . . .	. . . . .	. . . . .
Cylindrée et/ou Puissance si électrique	□□□□ cc □□□ kW	□□□□ cc □□□ kW	□□□□ cc □□□ kW
Année d'achat	□□□□	□□□□	□□□□
Année de construction	□□□□	□□□□	□□□□
Si vous avez le véhicule depuis plus d'un an: **nombre de km parcourus par an**	□□□□□□ km / an	□□□□□□ km / an	□□□□□□ km / an
Kilométrage actuel	□□□□□□ km	□□□□□□ km	□□□□□□ km

3b Dans le tableau ci-dessous, remplissez une <u>colonne</u> pour chaque <u>voiture, camionnette ou autre véhicule motorisé</u> que vous avez cité à la question 2. *- Si aucune voiture, camionnette ou autre véhicule motorisé n'est présent, allez à la question 4 à la page suivante.*
- S'il y a plus de trois voitures, camionnettes ou autres véhicules motorisés présents dans votre ménage, ne remplissez le tableau que pour les trois véhicules les plus utilisés.
- Si nécessaire, vous trouverez les réponses à certaines questions sur les documents du véhicule.

	Véhicule 1	Véhicule 2	Véhicule 3
Marque *ex.: "Volkswagen"*	. . . . .	. . . . .	. . . . .
Modèle *ex.: "Golf"*	. . . . .	. . . . .	. . . . .
Type de véhicule *Une seule réponse possible*	1 ○ voiture (y compris monospace et 4x4) 2 ○ camionnette 3 ○ autre(précisez): …….....	1 ○ voiture (y compris monospace et 4x4) 2 ○ camionnette 3 ○ autre(précisez): …….....	1 ○ voiture (y compris monospace et 4x4) 2 ○ camionnette 3 ○ autre(précisez): ……..... .....
Cylindrée	☐☐☐☐ cc	☐☐☐☐ cc	☐☐☐☐ cc
Année d'achat (ou mise à disposition)	☐☐☐☐	☐☐☐☐	☐☐☐☐
Comment avez-vous pris possession du véhicule? *Une seule réponse possible*	1 ○ acheté neuf 2 ○ acheté d'occasion 3 ○ véhicule de société 4 ○ autre	1 ○ acheté neuf 2 ○ acheté d'occasion 3 ○ véhicule de société 4 ○ autre	1 ○ acheté neuf 2 ○ acheté d'occasion 3 ○ véhicule de société 4 ○ autre
Année de construction	☐☐☐☐	☐☐☐☐	☐☐☐☐
Si vous avez le véhicule depuis plus d'un an: **nombre de km parcourus par an**	☐☐☐☐☐☐ km / an	☐☐☐☐☐☐ km / an	☐☐☐☐☐☐ km / an
Kilométrage actuel	☐☐☐☐☐☐ km	☐☐☐☐☐☐ km	☐☐☐☐☐☐ km
Carburant *Une seule réponse possible*	1 ○ essence sans plomb 2 ○ diesel/gasoil 3 ○ LPG 4 ○ hybride 5 ○ autre	1 ○ essence sans plomb 2 ○ diesel/gasoil 3 ○ LPG 4 ○ hybride 5 ○ autre	1 ○ essence sans plomb 2 ○ diesel/gasoil 3 ○ LPG 4 ○ hybride 5 ○ autre
La nuit, où stationne le plus souvent ce véhicule ? *Une seule réponse possible*	1 ○ dans un garage, box, ou un autre emplacement réservé, situé à ☐☐ minutes à pied du logement 2 ○ dans la rue 3 ○ autre : ……	1 ○ dans un garage, box, ou un autre emplacement réservé, situé à ☐☐ minutes à pied du logement 2 ○ dans la rue 3 ○ autre : ……	1 ○ dans un garage, box, ou un autre emplacement réservé, situé à ☐☐ minutes à pied du logement 2 ○ dans la rue 3 ○ autre : ……
Utilisateur principal, c'est-à-dire la personne qui effectue le plus de km avec le véhicule? *Une seule réponse possible*	Prénom: ..... Il s'agit de : A ○ la personne n° ☐ du ménage (indiquez le numéro de la colonne de la personne à la question 1) B ○ une autre personne n'appartenant pas au ménage	Prénom: ..... Il s'agit de : A ○ la personne n° ☐ du ménage (indiquez le numéro de la colonne de la personne à la question 1) B ○ une autre personne n'appartenant pas au ménage	Prénom: ..... Il s'agit de : A ○ la personne n° ☐ du ménage (indiquez le numéro de la colonne de la personne à la question 1) B ○ une autre personne n'appartenant pas au ménage

VOTRE HABITATION ET VOTRE QUARTIER

4 Vous habitez… *1 seule réponse possible*

- 1 ◯ un appartement ou un studio.
- 2 ◯ une maison individuelle mitoyenne des 2 côtés (2 façades).
- 3 ◯ une maison individuelle jumelée (3 façades).
- 4 ◯ une maison individuelle séparée (4 façades).
- 5 ◯ autre

5 Votre ménage est-il propriétaire ou locataire du logement où vous vivez? *Une seule réponse possible*

- 1 ◯ propriétaire
- 2 ◯ locataire
- 3 ◯ autre

6 Combien de voitures pouvez-vous garer dans un garage ou un emplacement de parking privé (c'est-à-dire pas sur la voie publique) dont vous disposez ? *Dans le cas où votre garage ou votre emplacement de parking est utilisé à d'autres fins que pour garer une ou plusieurs voitures, combien de voitures pourrait-il normalement accueillir?*

0 ◯ 0 1 ◯ 1 2 ◯ 2 3 ◯ 3 ou plus

7 Sur votre lieu de résidence, ou dans ses environs proches, trouver un emplacement de parking gratuit en rue pour une voiture ……

- 1 ◯ ne pose pas de problème.
- 2 ◯ pose quelques difficultés.
- 3 ◯ pose beaucoup de difficultés.

8 Sur votre lieu de résidence, ou dans ses environs proches, le parking en rue est …

- 1 ◯ gratuit et à durée illimitée. → *allez à la question 10*
- 2 ◯ gratuit et à durée limitée (zone bleue par exemple).
- 3 ◯ payant.

9 Si le parking en rue est payant ou à durée limitée, quel(s) type(s) de cartes avez-vous pour (au moins) une voiture de votre ménage ? *(cochez la(les) carte(s) dont vous disposez)*

- 1 ◯ carte de riverain
- 2 ◯ carte « handicapé »
- 3 ◯ sans objet, je n'ai pas de voiture
- 4 ◯ autre (précisez) : ……..
- 5 ◯ aucune carte

10 Combien de vélos pouvez-vous abriter chez vous ou dans les communs ? *Une seule réponse possible*

0 ◯ 0 1 ◯ 1 2 ◯ 2 3 ◯ 3 ou plus 4 ◯ ne sait pas

11 Concernant les transports en commun, quel est votre degré de satisfaction au niveau des points suivants ?

	très satisfait	satisfait	insatisfait	très insatisfait	sans avis
fréquence depuis votre domicile	1	2	3	4	5
fréquence par rapport à vos destinations habituelles	1	2	3	4	5
ponctualité	1	2	3	4	5
prix	1	2	3	4	5
sécurité	1	2	3	4	5
autre :	1	2	3	4	5

12 Afin de lier les résultats de cette enquête à des données économiques, nous souhaiterions avoir une idée du revenu net de votre ménage le mois dernier.

Pour ce faire, totalisez les revenus professionnels nets (c'est-à-dire ce que chaque personne de votre ménage reçoit effectivement chaque mois) mais aussi les autres revenus comme les allocations familiales, les allocations de chômage, les aides sociales, les pensions, les primes, les revenus immobiliers, mobiliers, commerciaux, etc.

Compte tenu de tout cela, merci d'indiquer à quelle catégorie appartient le revenu net de votre ménage du mois dernier.

Une seule réponse possible

1 ○ 0 à 499 euros par mois
2 ○ 500 à 999 euros par mois
3 ○ 1000 à 1499 euros par mois
4 ○ 1500 à 1999 euro par mois
5 ○ 2000 à 2499 euros par mois
6 ○ 2500 à 2999 euros par mois
7 ○ 3000 à 3999 euros par mois
8 ○ 4000 à 4999 euros par mois
9 ○ 5000 à 9999 euros par mois
10 ○ plus de 10000 euros par mois

NOUS VOUS REMERCIONS POUR VOTRE AIMABLE COLLABORATION

N'OUBLIEZ PAS…

- **de remplir un questionnaire individuel;**
- **de veiller à ce que chaque personne de votre ménage (à partir de 6 ans) remplisse un questionnaire individuel;**
- **de renvoyer tous les questionnaires complétés, dans l'enveloppe fournie, le plus rapidement possible après les avoir complétés. Un timbre n'est pas nécessaire (port payé par le destinataire).**

Si vous avez des questions, des remarques ou des suggestions à propos de ce questionnaire ou de notre enquête en général, veuillez les écrire ci-dessous.

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

En application de l'article 4 de la loi du 8 décembre 1992 relative à la protection de la vie privée à l'égard des traitements de données à caractère personnel, nous devons vous informer des éléments suivants:

1. Responsable du traitement : le Service Public Fédéral Mobilité et Transports, Rue du Progrès, 56 à 1210 Bruxelles.
2. Finalité du traitement : Les données recueillies seront utilisées pour l'enquête nationale sur la mobilité des ménages.
3. Destinataires des données : les données anonymisées seront utilisées par différents Services Publics et groupes de recherche scientifique.
4. Les ménages tirés au sort ne sont pas dans l'obligation de répondre.
5. Toute personne concernée par les données a le droit d'y accéder et a le droit de demander leur rectification (uniquement possible durant le déroulement de l'enquête, avant anonymisation des données).
6. Des renseignements complémentaires peuvent être obtenus dans le registre tenu par la Commission de la protection de la vie privée : http://www.privacycommission.be (numéro d'identification du traitement: : VT 005014032).

QUESTIONNAIRE INDIVIDUEL

À REMPLIR PAR CHAQUE PERSONNE DU MÉNAGE À PARTIR DE 6 ANS
(LES PARENTS PEUVENT AU BESOIN AIDER LEURS ENFANTS À RÉPONDRE)

Qui doit remplir le questionnaire individuel ?
Ce questionnaire doit être rempli par toutes les personnes du ménage, âgées de 6 ans et plus.

Pour pouvoir relier votre questionnaire à celui de votre ménage, veuillez indiquer :

votre prénom: .....

votre numéro à la question 1 du questionnaire du ménage (numéro de la colonne complétée) : ☐

De quoi se compose le questionnaire ?
Le questionnaire se compose de trois parties :
- Partie 1 (p.2 à 9) : vos habitudes en matière de déplacements,
- Partie 2 (p.10 à 18) : les déplacements que vous effectuez un **jour de référence**. Ce jour de référence est le même pour chaque personne de votre ménage, soit le

- Partie 3 (p. 19 à 20): vos opinions en matière de mobilité.

Que faire avec le questionnaire ?
Avant le jour de référence, lire les indications et les questions de la partie 2 (p.10 à 18) ; cela vous aidera à répondre ensuite aux questions. Vous pouvez déjà répondre à la partie 1.

Après le jour de référence, remplir le plus rapidement possible ce questionnaire et veiller à le renvoyer dès que possible, avec les autres questionnaires de votre ménage, dans l'enveloppe fournie. Inutile d'y apposer un timbre, le port est payé par le destinataire.

Comment remplir le questionnaire ?
Il y a trois manières de répondre aux questions:

5 ○	cocher un(des) cercle(s) pour choisir votre(vos) réponse(s) dans une liste ou un tableau;
☐☐ min	écrire un nombre dans une case. Si votre réponse est "aucun" ou "ne s'applique pas", alors indiquer 0;
.....	écrire votre réponse en IMPRIMÉ sur des pointillés.

Votre participation est précieuse.

Les renseignements que vous nous communiquerez seront traités de manière anonyme (*).

Si vous souhaitez plus d'informations ou de l'aide pour remplir le questionnaire,

appelez gratuitement le **0800/80.770**
du lundi au vendredi
de 9h00 à 20h00.

Vragenlijsten in het Nederlands zijn beschikbaar op aanvraag.
Fragebögen auf Deutsch sind auf Anfrage erhältlich.

(*) conformément à la déclaration à la Commission de la protection de la vie privée, figurant en page 20.

PARTIE 1 : VOS HABITUDES EN MATIÈRE DE DÉPLACEMENTS

VOTRE USAGE DES DIFFÉRENTS MODES DE DÉPLACEMENT

1a A quelle fréquence avez-vous utilisé les modes de déplacement ci-dessous au cours des 12 derniers mois, que ce soit en Belgique ou à l'étranger et quelle que soit la raison (promenades comprises) **?** *Veuillez remplir le tableau ci-dessous en cochant, pour chaque mode de déplacement, le cercle qui correspond à votre choix. Une seule réponse par ligne.*

Modes de déplacement	au moins cinq jours par semaine	un à quelques jours par semaine	un à quelques jours par mois	un à quelques jours par an	jamais
marche (plus de 10 minutes)	1	2	3	4	5
vélo	1	2	3	4	5
cyclomoteur, moto	1	2	3	4	5
transport public	1	2	3	4	5
taxi	1	2	3	4	5
voiture en tant que conducteur	1	2	3	4	5
voiture en tant que passager	1	2	3	4	5
avion	1	2	3	4	5

1b Habituellement, à quelle fréquence avez-vous recours aux services des sociétés de transport ci-dessous (un aller-retour=deux fois) ? *Veuillez remplir le tableau ci-dessous en cochant, pour chaque société de transport, le cercle qui correspond à votre choix. Une seule réponse par ligne.*

Sociétés de Transport	10 fois par semaine ou +	4 à 8 fois par semaine	2 fois par semaine	4 à 6 fois par mois	2 fois par mois	moins de 2 fois par mois	jamais
SNCB	1	2	3	4	5	6	7
De Lijn	1	2	3	4	5	6	7
STIB	1	2	3	4	5	6	7
TEC	1	2	3	4	5	6	7
Cambio	1	2	3	4	5	6	7

2 A quelle(s) réduction(s) de tarifs ou mesure(s) de gratuité pour les transports publics avez-vous droit, même si vous ne l'(les) utilisez pas? *Une seule réponse par ligne.*

	oui	non	ne sait pas
enfant, jeune ou scolaire	1	2	3
senior	1	2	3
famille nombreuse	1	2	3
OMNIO / BIM (ex-VIPO : Veuf, Invalide, Pensionné, Orphelin)	1	2	3
bénéficiaire de RIS (revenu d'intégration sociale) ou ERIS	1	2	3
employé(e) d'une société de transport public, de Belgacom, de la Poste, ...	1	2	3
autre (précisez):	1	2	3

3a En ce moment, avez-vous un abonnement nominatif pour les transports publics?

1 ◯ non → *Allez à la question 3b*

2 ◯ oui → **Dans le tableau ci-dessous, remplissez une <u>ligne</u> pour chaque <u>abonnement</u> que vous possédez en ce moment.**

	Société de transport public	**Nombre de voyages** *Une seule réponse possible*	**Durée de l'abonnement** *Une seule réponse possible*
Abonnement 1 ***Si** abonnement **combiné**, cochez les différentes sociétés de transport comprises dans cet abonnement.*	1 ◯ De Lijn 2 ◯ TEC 3 ◯ STIB 4 ◯ SNCB	<u>Uniquement pour **SNCB**</u> 1 ◯ nombre limité (ex: Railflex, Campus) 2 ◯ nombre illimité	1 ◯ 1 semaine 2 ◯ 2 semaines 3 ◯ 1 mois 4 ◯ 3 mois 5 ◯ 1 an 6 ◯ autre (précisez):.....
Abonnement 2 ***Si** abonnement **combiné**, cochez les différentes sociétés de transport comprises dans cet abonnement.*	1 ◯ De Lijn 2 ◯ TEC 3 ◯ STIB 4 ◯ SNCB	<u>Uniquement pour **SNCB**</u> 1 ◯ nombre limité (ex: Railflex, Campus) 2 ◯ nombre illimité	1 ◯ 1 semaine 2 ◯ 2 semaines 3 ◯ 1 mois 4 ◯ 3 mois 5 ◯ 1 an 6 ◯ autre (précisez):.....

3b En ce moment, détenez-vous d'autres titres de transport (non nominatifs) ou carte de fidélité pour les transports publics ?

1 ◯ non → *Allez à la question 4a*

2 ◯ oui → **Cochez les titres de transports et carte de fidélité que vous détenez en ce moment.** *Plusieurs réponses possibles*

<u>Titres de transport</u>

A ◯ go pass / rail pass

B ◯ key card

C ◯ carte SNCB 10 voyages à destination fixe

D ◯ carte TEC multi-voyages

E ◯ carte STIB 1 ou plusieurs voyage(s) ou jour(s) (en carte papier ou contrat chargé sur votre carte MOBIB)

F ◯ carte JUMP 1, 5 ou 10 voyage(s) ou 1 jour (en carte papier, valable chez les 4 opérateurs dans Bruxelles)

G ◯ carte De Lijn multi-voyages (Lijnkaart)

<u>Carte de fidélité</u>

H ◯ carte de réduction 50% SNCB (payante et valable 1 an)

4a Avez-vous un permis de conduire? *Y compris les permis cyclomoteur et moto.*

1 ◯ Non, mais en apprentissage → *Allez à la question 5 à la page suivante*

2 ◯ Non → *Allez à la question 5 à la page suivante*

3 ◯ Oui → **Quel(s) type(s) de permis avez-vous?** *Plusieurs réponses possibles*

1 ◯ A3 - cyclomoteur

2 ◯ A2 - moto jusqu'à 400 cc

3 ◯ A1 - moto de plus de 400 cc

4 ◯ B - voiture et camionnette

5 ◯ C - camion

6 ◯ D - bus

7 ◯ E - véhicules des catégories B, C ou D avec une grande remorque

4b Si vous avez un permis B (voitures et camionnettes), depuis quelle année avez-vous ce permis ?

Depuis ☐☐☐☐

5 Avez-vous, pour des raisons physiques, des difficultés pour utiliser certains modes de déplacement, par exemple en raison d'un handicap permanent, de votre âge, etc.?
Veuillez remplir le tableau ci-dessous en cochant, sur chaque ligne, le cercle qui correspond à votre choix.

	... ne me pose aucun problème.	... m'est possible mais avec difficultés.	... m'est seulement possible avec des facilités d'accès ou des équipements particuliers.	... m'est impossible.
Marcher ...	1	2	3	4
Faire du vélo ...	1	2	3	4
Monter et descendre de voiture ...	1	2	3	4
Conduire une voiture ...	1	2	3	4
Accéder aux arrêts de tram ou bus ...	1	2	3	4
Accéder aux gares, quais et arrêts de train ou métro ...	1	2	3	4
Monter et descendre du train, bus, tram et métro ...	1	2	3	4
Autre (précisez): …….. **…….....**	1	2	3	4

6 Dans <u>la semaine écoulée</u>, avez-vous utilisé l'une ou plusieurs des sources d'information suivantes pour préparer ou mener à bien l'un ou plusieurs de vos déplacements (suivre un itinéraire inconnu, connaître la durée d'un retard éventuel, changer de mode de déplacement en fonction de la circulation, …) ? *Plusieurs réponses possibles.*

1 Carte routière ou plan de ville sur papier
2 Site web des sociétés de transport (train, tram, bus, avion, …)
3 Carte, plan ou service de calcul d'itinéraires sur internet (Mappy, Google Maps, …)
4 Infotrafic à la radio ou à la télévision
5 Service d'information téléphonique (Mobiligne, Touring Mobilis…)
6 GPS
7 Amis, collègues, famille
8 Service d'information à un arrêt ou à la gare (écran, affiche, guichet…)
9 Autre (précisez) : ……..

VOS DÉPLACEMENTS LONGUE DISTANCE

7a Pendant les 12 derniers mois, combien de voyages de plus de 100 km (distance aller) avez-vous effectués (excepté vos navettes quotidiennes)?

0 0 → *allez à la question 8 page suivante*

1 1 2 2 3 3 4 4 5 entre 5 et 9 6 10 ou plus

7b Parmi ces voyages de plus de 100 km, combien étaient des voyages dont la destination se situait à l'étranger ?

0 0 → *allez à la question 8 page suivante*

1 1 2 2 3 3 4 4 5 entre 5 et 9 6 10 ou plus

7c Décrivez ci-dessous votre dernier voyage de plus de 100 km (distance aller) vers l'étranger. Si vous avez effectué un circuit, ne notez comme lieu d'arrivée, que la destination principale de votre circuit.

Lieu de départ : Ville / commune : …….

Pays : BELGIQUE

Lieu d'arrivée : Ville / commune : …….

Pays : …….

Quel **mode de déplacement principal** avez-vous utilisé pour vous y rendre ?

Par mode de déplacement principal, nous entendons celui avec lequel vous avez parcouru la distance la plus longue.

Par exemple, si vous faites 20 km en bus, puis 300 km en train, cochez seulement la case "train". Une seule réponse possible.

1 vélo
2 moto
3 train
4 autocar
5 voiture
6 avion
7 autre (précisez): .....

Lors de ce dernier voyage, combien de **nuits** avez-vous passées hors de chez vous ?

☐☐ nuits

Quel était **le motif principal** de ce dernier voyage ? *Une seule réponse possible*

1 des raisons professionnelles
2 rendre visite à la famille ou à des amis
3 sports, loisirs, vacances
4 voyage scolaire
5 autre (précisez): . . . . . .

VOS DÉPLACEMENTS DOMICILE-TRAVAIL OU DOMICILE-ÉCOLE

8 Avez-vous un logement (autre que votre domicile) que vous occupez au moins 3 jours par semaine ?
Ex : kot, logement pour le travail durant la semaine

1 ○ oui → Adresse : Rue : ……..…….. N°……
Code postal : ☐☐☐☐ Localité : ……

2 ○ non

9 Exercez-vous un(des) emploi(s) / profession(s) ? Etes-vous écolier/étudiant? *Plusieurs réponses possibles.*

1 ○ Oui, j'exerce un(des) emploi(s). → *continuez ci-dessous à la question 10a*

2 ○ Oui, je suis écolier / étudiant. → *continuez ci-dessous à la question 10a*

3 ○ Non → *allez aux instructions de la partie 2, à la page 10*

10a Avez-vous un lieu de travail ou d'études fixe, c'est-à-dire où vous devez vous rendre au moins 2 jours par semaine, autre que votre domicile ? *Une seule réponse possible.*

1 ○ Non, car je travaille à domicile → *Allez à la question 16*

2 ○ Non, car le lieu de mon activité n'est pas fixe → *Allez à la question 14*

3 ○ Oui, j'ai un (ou plusieurs) lieu(x) de travail/étude fixe(s) (autre que mon habitation) → *Continuez ci-dessous à la question 10b*

10b Dans le tableau ci-dessous, complétez une colonne par lieu fixe de travail ou d'études.

LIEU DE TRAVAIL OU D'ÉTUDES FIXE FREQUENTE AU MOINS 2 JOURS /SEMAINE

Rue et numéro :……
Code postal : ☐☐☐☐ Localité : ……

Combien de jours par semaine : ☐ jours par semaine

Comment vous y rendez-vous habituellement ?
Entourez *les modes de déplacement successifs que vous utilisez et, pour chacun d'eux, indiquez le nombre de kilomètres (aller simple). Complétez uniquement le nombre d'étapes nécessaire (maximum 5).*

Etapes	Entourez un mode par étape				Distance (aller simple)
D'abord: mode 1	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		
Puis : mode 2	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		
Puis : mode 3	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		
Puis : mode 4	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		
Enfin : mode 5	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		

EVENTUELLEMENT : AUTRE LIEU DE TRAVAIL OU D'ÉTUDES FIXE FREQUENTE AU MOINS 2 JOURS / SEMAINE

Rue et numéro :……
Code postal : ☐☐☐☐ Localité : ……

Combien de jours par semaine : ☐ jours par semaine

Comment vous y rendez-vous habituellement ?
Entourez *les modes de déplacement successifs que vous utilisez et, pour chacun d'eux, indiquez le nombre de kilomètres (aller simple). Complétez uniquement le nombre d'étapes nécessaire (maximum 5).*

Etapes	Entourez un mode par étape				Distance (aller simple)
D'abord: mode 1	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		
Puis : mode 2	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		
Puis : mode 3	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		
Puis : mode 4	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		
Enfin : mode 5	auto conducteur	à pied	train	bus	☐☐☐ km
	auto passager	cyclo/moto	tram	autre :	
	taxi	à vélo	métro		

Si vous exercez un emploi et suivez aussi des cours, ne considérez que votre emploi pour les questions suivantes (11 à 13). Si vous avez plusieurs emplois, considérez votre emploi principal. Si vous avez 2 mi-temps, choisissez parmi ceux-ci.

11 Donnez pour chaque type d'arrêt de transport public repris ci-dessous, la distance entre votre lieu de travail ou d'études et l'arrêt le plus proche. *Une seule réponse possible par colonne.*

Arrêt de bus	Gare	Arrêt de tram	Station de métro
1 0 à 249 m	1 0 à 249 m	1 0 à 249 m	1 0 à 249 m
2 250 à 499 m	2 250 à 499 m	2 250 à 499 m	2 250 à 499 m
3 500 à 999 m	3 500 à 999 m	3 500 à 999 m	3 500 à 999 m
4 1 à 2 km	4 1 à 2 km	4 1 à 2 km	4 1 à 2 km
5 2 à 5 km	5 2 à 5 km	5 2 à 5 km	5 2 à 5 km
6 plus de 5 km	6 plus de 5 km	6 plus de 5 km	6 plus de 5 km
7 ne sait pas	7 ne sait pas	7 ne sait pas	7 ne sait pas

12 Sur votre lieu de travail ou d'études ou dans ses environs, disposez-vous (ou pourriez-vous disposer) **d'un garage ou d'un emplacement de parking privé** (c'est-à-dire pas sur la voie publique) **pour une voiture?**

1 Oui, il y a un parking privé gratuit.
2 Oui, il y a un parking privé payant.
3 Non.

13 D'après vous, trouver un emplacement de parking public pour une voiture sur votre lieu de travail ou d'études ou dans ses environs... *Une seule réponse possible.*

1 ne pose pas de problème.
2 pose quelques difficultés.
3 pose beaucoup de difficultés.

14 Quelles activités réalisez-vous pendant vos déplacements vers votre lieu de travail ou d'école ?
Veuillez remplir le tableau ci-dessous en cochant, pour chaque mode de déplacement qu'il vous arrive d'utiliser pour vos navettes domicile-travail ou domicile-école, la(les) activité(s) réalisée(s). Plusieurs réponses possibles par colonne.

Quand je me déplace… / *il m'arrive souvent de...*	à pied	en bus	en tram/ métro	en train	en voiture conducteur	en voiture passager
rêver, me reposer, dormir.	1	1	1	1	1	1
discuter avec d'autres personnes	2	2	2	2	2	2
lire.	3	3	3	3	3	3
travailler.	4	4	4	4	4	4
jouer.	5	5	5	5	5	5
téléphoner.	6	6	6	6	6	6
envoyer des messages (sms).	7	7	7	7	7	7
écouter la radio / de la musique.	8	8	8	8	8	8
regarder des films.	9	9	9	9	9	9
autre	10	10	10	10	10	10
Précisez :	.....	.....	.....	.....	.....	.....

Cette partie ne concerne que les personnes qui exercent un emploi.
Si vous n'exercez pas d'emploi, allez à la partie 2 à la page 10.

15 Vos frais de déplacements entre votre habitation et votre lieu de travail sont-ils remboursés ou payés par votre employeur (ou par un système tiers-payant) ? *Une seule réponse possible.*

1 ○ Oui, j'ai une voiture de société.
2 ○ Oui, mon employeur paye (ou me rembourse) **partiellement** mes frais de déplacements (en transports en commun, à vélo, en voiture personnelle …).
3 ○ Oui, mon employeur paye (ou me rembourse) **totalement** mes frais de déplacements (en transports en commun, à vélo, en voiture personnelle…).
4 ○ Non, mais le transport est assuré par mon employeur (bus de ramassage, etc.).
5 ○ Non, je ne suis pas du tout remboursé pour mes frais de déplacements.
6 ○ Je suis indépendant et mes frais de déplacement sont pris en compte dans mes frais réels.

16 Dans l'exercice de votre profession, avez-vous besoin de vous déplacer? *Une seule réponse possible.*

1 ○ Jamais → *Allez à la question 18 à la page suivante*
2 ○ Occasionnellement → *Continuez ci-dessous à la question 17*
3 ○ Très souvent → *Continuez ci-dessous à la question 17*

17 Quel(s) mode(s) de déplacement utilisez-vous alors pour ces déplacements professionnels?
Une réponse par ligne.

Mode(s) de déplacement utilisé(s)	Régulière-ment	De temps en temps	Jamais
à pied	1 ○	2 ○	3 ○
vélo	1 ○	2 ○	3 ○
cyclomoteur/moto	1 ○	2 ○	3 ○
train	1 ○	2 ○	3 ○
bus	1 ○	2 ○	3 ○
tram	1 ○	2 ○	3 ○
métro	1 ○	2 ○	3 ○
taxi	1 ○	2 ○	3 ○
véhicule de société comme conducteur	1 ○	2 ○	3 ○
véhicule de société comme passager	1 ○	2 ○	3 ○
véhicule personnel comme conducteur	1 ○	2 ○	3 ○
véhicule personnel comme passager	1 ○	2 ○	3 ○
transport organisé par la société	1 ○	2 ○	3 ○
avion	1 ○	2 ○	3 ○
autre (précisez):	1 ○	2 ○	3 ○

18 A quelle fréquence faites-vous du covoiturage (conducteur et/ou passager) pour vous rendre sur votre lieu de travail ?

Le covoiturage est un mode de déplacement où plusieurs personnes utilisent une seule voiture pour faire le même trajet ou presque.

Complétez chaque ligne du tableau ci-dessous :

Covoiturage...

	A **quelle fréquence** ?					Avec **combien de personnes** (le plus souvent) ?
	3 fois ou plus par sem	1 à 2 fois par sem.	au - 1 fois par mois	au - 1 fois par an	jamais	
a. avec un ou plusieurs **membres du ménage**	1	2	3	4	5	☐☐ pers.
b. avec une ou plusieurs personnes de la **même entreprise**	1	2	3	4	5	☐☐ pers.
c. avec une ou plusieurs personnes **d'autres entreprises**	1	2	3	4	5	☐☐ pers.

Si vous ne covoiturez pas avec des personnes extérieures au ménage, seriez-vous prêt à l'envisager, si l'opportunité se présente ?

1 ◯ Oui, uniquement comme passager.

2 ◯ Oui, comme conducteur et passager.

3 ◯ Non. Pourquoi ? : ……

19 Vous travaillez habituellement (c'est-à-dire pour plus de ¾ de votre temps de travail) **...** *Une seule réponse possible.*

1 ◯ durant la journée.

2 ◯ durant la nuit.

3 ◯ en pause, sans service de nuit.

4 ◯ en pause, avec service de nuit.

5 ◯ aucun de ces cas ; dispositions de travail (précisez) : ……

20 Vous avez ... *Une seule réponse possible.*

1 ◯ des heures de travail identiques chaque jour, fixées par votre employeur.

2 ◯ des heures de travail identiques chaque jour, fixées par vous-même.

3 ◯ des heures de travail variables, fixées par votre employeur.

4 ◯ des heures de travail variables, fixées par vous-même.

5 ◯ autre : ……

21 Si vous êtes salarié, votre temps de travail global est équivalent à… *Une seule réponse possible.*

1 ◯ moins qu'un mi-temps.

2 ◯ un mi-temps.

3 ◯ entre un mi-temps et un temps plein.

4 ◯ un temps plein.

22 Combien d'heures travaillez-vous par semaine habituellement ?

☐☐ heures / semaine

23 Dans quel secteur travaillez-vous ? *Si vous exercez plusieurs professions, considérez la principale. Une seule réponse possible.*

1 ◯ secteur privé

2 ◯ secteur (para)public

3 ◯ secteur associatif

PARTIE 2: VOS DÉPLACEMENTS LE …… (LE JOUR DE RÉFÉRENCE)

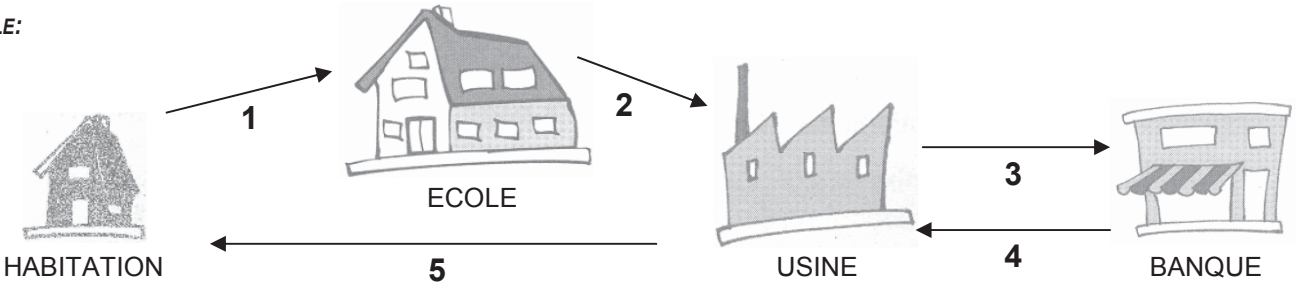
DE 4H DU MATIN À 4H LE LENDEMAIN MATIN

QU'APPELONS-NOUS UN DÉPLACEMENT?

Quand vous sortez pour aller quelque part, vous faites un déplacement. Voici quelques exemples de déplacements: aller au magasin, aller à l'école, aller au travail, aller chercher quelqu'un, aller chez le médecin, aller rendre visite à quelqu'un, aller promener le chien, aller boire un verre, ... Pour effectuer ces déplacements, vous utilisez un (des) mode(s) de déplacement. Vous vous déplacez à pied, à vélo, à moto, en train, en bus, en tram, en métro, en taxi, en voiture, ...

VOICI QUELQUES INDICATIONS QUI VOUS PERMETTRONT DE BIEN RENDRE COMPTE DE VOS DÉPLACEMENTS

EXEMPLE:



Déplacement 1	Déplacement 2	Déplacement 3	Déplacement 4	Déplacement 5
Déposer quelqu'un	Aller travailler	Services	Aller travailler	Rentrer à la maison

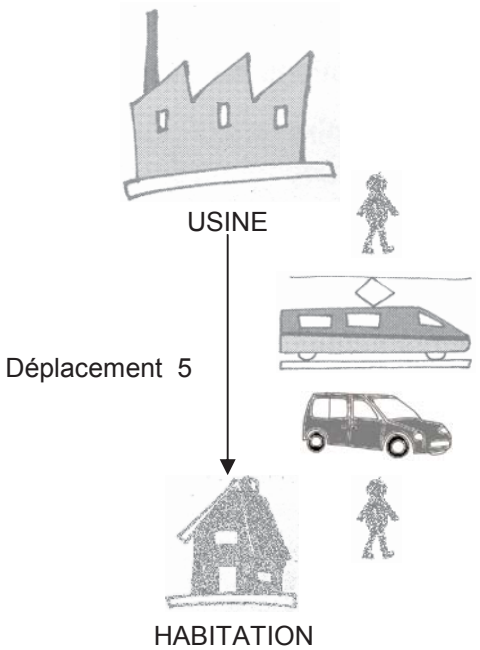
Quand vous sortez pour aller quelque part, vous faites un déplacement.

Si, en allant quelque part, vous vous arrêtez en chemin, vous devez considérer chaque déplacement séparément: avant l'arrêt et après l'arrêt. Par exemple, si en allant au travail, vous conduisez vos enfants à l'école, vous devez considérer deux déplacements: de votre habitation à l'école, puis de l'école à votre travail. (Ex: déplacements 1 et 2)

N'oubliez pas vos petits déplacements (aller acheter le journal ou chercher de l'argent au distributeur). (Ex: déplacement 3)

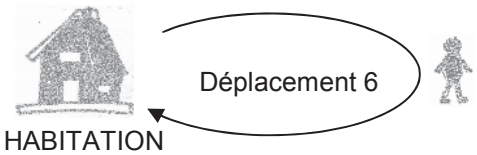
Le retour au point de départ est considéré comme un nouveau déplacement. (Ex: déplacement 4)

Un déplacement peut être réalisé à l'aide de plusieurs modes de déplacement consécutifs. Donnez, par déplacement, tous les modes de déplacement consécutifs que vous avez utilisés. (Ex : déplacement 5)



D'ABORD :	5 min **à pied**		km 500 m	
PUIS (1) :	45 min		km 50 km	m
auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	
PUIS (2) :	12 min		km 10 km	m
auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	
PUIS (3) :	min		km	m
auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	
ENFIN :	4 min **à pied**		km 400 m	

Se promener, sortir le chien ou faire une balade à vélo sans destination précise sont aussi considérés comme des déplacements. (Ex: déplacement 6)



Si vous n'exercez pas d'emploi, passez directement à la question 26 à la fin de la page.

24 Le jour de référence, avez-vous travaillé depuis votre domicile (télétravail) ?

1 ◯ non

2 ◯ oui → **Combien d'heures ?** ☐☐ **h**.

25 Exercez-vous une profession qui nécessite de nombreux déplacements quotidiens ?
Exemples : facteur, livreur, chauffeur, représentant de commerce, médecin durant ses visites à domicile, conducteur de tram, employé d'intercommunale (eau, gaz, électricité), etc.

1 ◯ non → *Allez à la question 26 (bas de la page)*

2 ◯ oui → **25a Quelle est cette profession ? ……..**

Comment indiquer ces déplacements professionnels (nombreux et aux caractéristiques similaires) dans les pages qui suivent ?

Une **série de déplacements professionnels**, que nous appellerons dans la suite **« tournée »**, peut être indiquée en une seule fois (dans une seule colonne), si le nombre de déplacements professionnels successifs est égal à 3 ou plus. Voici les indications qui vous permettront de noter ces déplacements :

- dans la case « destination », l'adresse de destination du dernier déplacement professionnel de la tournée, c'est-à-dire le dernier endroit (où vous vous êtes rendu pour raison professionnelle) avant votre retour au domicile, avant d'aller faire une course, etc. (tout déplacement autre que professionnel). N'incluez pas le retour au domicile dans la tournée mais indiquez le dans la colonne suivante avec le motif "aller à la maison" ;
- comme raison principale, cochez « pour le travail », et notez le nombre total de déplacements professionnels successifs dans la case « nombre si tournée » ;
- comme heure de départ, indiquez l'heure de départ du premier déplacement professionnel de la tournée ;
- notez le ou les modes utilisés pour ces déplacements, avec pour chaque mode, la durée totale de trajet et la distance totale parcourue ;
- et comme heure d'arrivée, indiquez l'heure d'arrivée du dernier déplacement professionnel de la tournée.

Exemple : *un livreur quitte son domicile pour se rendre à un dépôt de marchandise (déplacement 1), où il charge sa voiture avant d'effectuer 15 livraisons (déplacement 2). En fin de journée, il va faire des courses (déplacement 3) avant de retourner à son domicile (déplacement 4) sans repasser par le dépôt.*

*Le **déplacement 1** aura comme motif « aller travailler » et sera complété selon les modalités définies dans la page précédente.*

*Pour le **déplacement 2**, dans la case « destination », l'adresse de la dernière livraison sera indiquée.*

Le motif sera « se déplacer dans le cadre du travail », et dans les cases « nombre si tournée », le nombre de déplacements effectués sera noté (15 dans ce cas).

Heure de départ : 10h00

Modes de déplacement, distance et durée : voiture conducteur, 320 minutes (total du temps passé en voiture), 280 km (distance totale parcourue pour les 15 livraisons), 30 minutes de recherche d'une place de parking (somme des durées pour chercher un parking), et le véhicule 2 du ménage sera entouré, car c'est avec ce véhicule que le livreur aura effectué ses livraisons.

Heure d'arrivée : 18h00 (fin de la tournée des livraisons)

*Le **déplacement 3** aura comme motif « faire des courses » et décrira le déplacement entre le dernier lieu de livraison et le magasin visité.*

*Le **déplacement 4** aura comme motif « aller à la maison ».*

26 Dans le tableau suivant, remplissez une colonne pour chaque déplacement que vous avez effectué le jour de référence.
N'oubliez pas d'indiquer tous les déplacements effectués à pied, de toujours considérer un déplacement retour comme un nouveau déplacement, et d'indiquer le dernier déplacement que vous avez effectué le jour de référence (votre retour à la maison par exemple).

Merci de lire les instructions ci-dessous et d'indiquer vos déplacements dans l'ordre à partir de la page suivante.

Questions

D'où êtes-vous parti ?
Vous ne devez indiquer que le point de départ de votre premier déplacement (là où vous étiez le jour de référence à 4h du matin).

Où êtes-vous allé? Remplissez aussi précisément que possible. Si vous ne connaissez pas le nom de la rue, donnez le nom de l'endroit, du quartier, de l'entreprise, ... où vous alliez.

Quelle était la raison principale du déplacement?
Une seule réponse possible.

A quelle heure êtes-vous parti?

Comment y êtes-vous allé ?
Quels modes de déplacement successifs avez-vous utilisés?

Si vous avez utilisé plusieurs modes de déplacement (marche comprise), décomposez votre déplacement en plusieurs étapes.

Complétez uniquement le nombre d'étapes nécessaire (max. 5).

Par étape, indiquez la durée du trajet, la distance parcourue et entourez le mode de déplacement.

Pour les train, tram, bus et métro, comptez le temps d'attente dans le temps du trajet

N'oubliez pas d'indiquer tous les trajets effectués à pied.

Si vous avez effectué une partie de ce déplacement en **voiture**,
- estimez le temps qu'il vous a fallu pour trouver une place de parking

- si vous avez utilisé une **voiture du ménage**, entourez le numéro qui figure au-dessus de la description de ce véhicule dans le questionnaire ménage.

A quelle heure êtes-vous arrivé?

Pendant le déplacement, étiez-vous accompagné, d'enfant(s) en bas âge ou d'autres personnes ? Etiez-vous chargé de courses ou de bagages ?

Exemple

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: RUE NEUVE N° : .....
Localité : BRUXELLES Code postal : 1000

Raison principale 1 seule réponse
1 ○ déposer/chercher quelqu'un
2 ○ aller à la maison
3 ○ aller travailler
4 ○ pour le travail (si tournée, nombre : ☐☐ déplacements)
5 ○ suivre un cours (école, ...)
6 ○ prendre un repas à l'extérieur
7 ☒ faire des courses/du shopping
8 ○ services (médecin, banque, ...)
9 ○ rendre visite à la famille ou à des amis
10 ○ se promener, faire un tour
11 ○ loisirs, sports, culture
12 ○ autre (précisez): .....

Heure de départ : 14 h 50 min (si après midi : 13h, 14h, ...)

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD :	☐☐☐ min **à pied**			☐☐☐ km	☐☐☐ m
PUIS (1) :	10 min			5 km	☐☐☐ m
(auto conducteur)	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB		
taxi	à vélo	métro	bus TEC		
PUIS (2) :	30 min			50 km	☐☐☐ m
auto conducteur	à pied	(train)	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB		
taxi	à vélo	métro	bus TEC		
PUIS (3) :	5 min			1 km	500 m
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	(bus STIB)		
taxi	à vélo	métro	bus TEC		
ENFIN :	4 min **à pied**			☐☐☐ km	400 m

- Recherche d'une place de parking: 0 min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) :

véhicule 1	(véhicule 2)	véhicule 3

Heure d'arrivée : 15 h 40 min

Etiez-vous accompagné … ?
1 ○ d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2 ☒ d'autres personnes → combien ? 1 personnes
3 ○ d'animaux
4 ○ de courses / bagage

Merci d'indiquer dans l'ordre les déplacements que vous avez effectués le jour de référence à partir de 4h du matin jusqu'à 4h le lendemain.

Déplacement 1

Point de départ

Rue: …… N° : ……

Localité : …… Code postal : ☐☐☐☐

Destination Pays (si hors Belgique): ……

Rue: …… N° : ……

Localité : …… Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. déposer/chercher quelqu'un
2. aller à la maison
3. aller travailler
4. pour le travail (si tournée, nombre : ☐☐ déplacements)
5. suivre un cours (école, …)
6. prendre un repas à l'extérieur
7. faire des courses/du shopping
8. services (médecin, banque, …)
9. rendre visite à la famille ou à des amis
10. se promener, faire un tour
11. loisirs, sports, culture
12. autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

PUIS (1) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	……
taxi	à vélo	métro	bus TEC	

PUIS (2) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	……
taxi	à vélo	métro	bus TEC	

PUIS (3) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	……
taxi	à vélo	métro	bus TEC	

ENFIN : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. d'autres personnes → combien ? ☐☐ personnes
3. d'animaux
4. de courses / bagage

Déplacement 2

Point de départ = destination précédente

Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?

NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): ……

Rue: …… N° : ……

Localité : …… Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. déposer/chercher quelqu'un
2. aller à la maison
3. aller travailler
4. pour le travail (si tournée, nombre : ☐☐ déplacements)
5. suivre un cours (école, …)
6. prendre un repas à l'extérieur
7. faire des courses/du shopping
8. services (médecin, banque, …)
9. rendre visite à la famille ou à des amis
10. se promener, faire un tour
11. loisirs, sports, culture
12. autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

PUIS (1) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	……
taxi	à vélo	métro	bus TEC	

PUIS (2) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	……
taxi	à vélo	métro	bus TEC	

PUIS (3) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	……
taxi	à vélo	métro	bus TEC	

ENFIN : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. d'autres personnes → combien ? ☐☐ personnes
3. d'animaux
4. de courses / bagage

Le nombre de déplacements par jour est une donnée importante. N'oubliez aucun déplacement : les retours au domicile, les petits déplacements ou les brefs arrêts pour déposer quelqu'un ou acheter un journal, etc.

Déplacement 3

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …….....…… N° : ….....
Localité : ……..…….. Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. déposer/chercher quelqu'un
2. aller à la maison
3. aller travailler
4. pour le travail (si tournée, nombre : ☐☐ déplacements)
5. suivre un cours (école, …)
6. prendre un repas à l'extérieur
7. faire des courses/du shopping
8. services (médecin, banque, …)
9. rendre visite à la famille ou à des amis
10. se promener, faire un tour
11. loisirs, sports, culture
12. autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

PUIS (1) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (2) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (3) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

ENFIN : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. d'autres personnes → combien ? ☐☐ personnes
3. d'animaux
4. de courses / bagage

Déplacement 4

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …….....…… N° : ….....
Localité : ……..…….. Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. déposer/chercher quelqu'un
2. aller à la maison
3. aller travailler
4. pour le travail (si tournée, nombre : ☐☐ déplacements)
5. suivre un cours (école, …)
6. prendre un repas à l'extérieur
7. faire des courses/du shopping
8. services (médecin, banque, …)
9. rendre visite à la famille ou à des amis
10. se promener, faire un tour
11. loisirs, sports, culture
12. autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

PUIS (1) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (2) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (3) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

ENFIN : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. d'autres personnes → combien ? ☐☐ personnes
3. d'animaux
4. de courses / bagage

Le nombre de déplacements par jour est une donnée importante. N'oubliez aucun déplacement : les retours au domicile, les petits déplacements ou les brefs arrêts pour déposer quelqu'un ou acheter un journal, etc.

Déplacement 5

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …….....…… N° : ….....
Localité : …….....…… Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. ◯ déposer/chercher quelqu'un
2. ◯ aller à la maison
3. ◯ aller travailler
4. ◯ pour le travail (si tournée, nombre : ☐☐ déplacements)
5. ◯ suivre un cours (école, …)
6. ◯ prendre un repas à l'extérieur
7. ◯ faire des courses/du shopping
8. ◯ services (médecin, banque, …)
9. ◯ rendre visite à la famille ou à des amis
10. ◯ se promener, faire un tour
11. ◯ loisirs, sports, culture
12. ◯ autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD :	☐☐☐ min **à pied**		☐☐☐ km	☐☐☐ m	
PUIS (1) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB	.....	
taxi	à vélo	métro	bus TEC		
PUIS (2) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB	.....	
taxi	à vélo	métro	bus TEC		
PUIS (3) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB	.....	
taxi	à vélo	métro	bus TEC		
ENFIN :	☐☐☐ min **à pied**		☐☐☐ km	☐☐☐ m	

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. ◯ d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. ◯ d'autres personnes → combien ? ☐☐ personnes
3. ◯ d'animaux
4. ◯ de courses / bagage

Déplacement 6

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …….....…… N° : ….....
Localité : …….....…… Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. ◯ déposer/chercher quelqu'un
2. ◯ aller à la maison
3. ◯ aller travailler
4. ◯ pour le travail (si tournée, nombre : ☐☐ déplacements)
5. ◯ suivre un cours (école, …)
6. ◯ prendre un repas à l'extérieur
7. ◯ faire des courses/du shopping
8. ◯ services (médecin, banque, …)
9. ◯ rendre visite à la famille ou à des amis
10. ◯ se promener, faire un tour
11. ◯ loisirs, sports, culture
12. ◯ autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD :	☐☐☐ min **à pied**		☐☐☐ km	☐☐☐ m	
PUIS (1) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB	.....	
taxi	à vélo	métro	bus TEC		
PUIS (2) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB	.....	
taxi	à vélo	métro	bus TEC		
PUIS (3) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB	.....	
taxi	à vélo	métro	bus TEC		
ENFIN :	☐☐☐ min **à pied**		☐☐☐ km	☐☐☐ m	

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. ◯ d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. ◯ d'autres personnes → combien ? ☐☐ personnes
3. ◯ d'animaux
4. ◯ de courses / bagage

Le nombre de déplacements par jour est une donnée importante. N'oubliez aucun déplacement : les retours au domicile, les petits déplacements ou les brefs arrêts pour déposer quelqu'un ou acheter un journal, etc.

Déplacement 7

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …….....…… N° : ….....
Localité : ……..…….…… Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. ◯ déposer/chercher quelqu'un
2. ◯ aller à la maison
3. ◯ aller travailler
4. ◯ pour le travail (si tournée, nombre : ☐☐ déplacements)
5. ◯ suivre un cours (école, …)
6. ◯ prendre un repas à l'extérieur
7. ◯ faire des courses/du shopping
8. ◯ services (médecin, banque, …)
9. ◯ rendre visite à la famille ou à des amis
10. ◯ se promener, faire un tour
11. ◯ loisirs, sports, culture
12. ◯ autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

PUIS (1) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (2) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (3) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

ENFIN : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. ◯ d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. ◯ d'autres personnes → combien ? ☐☐ personnes
3. ◯ d'animaux
4. ◯ de courses / bagage

Déplacement 8

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …….....…… N° : ….....
Localité : ……..…….…… Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. ◯ déposer/chercher quelqu'un
2. ◯ aller à la maison
3. ◯ aller travailler
4. ◯ pour le travail (si tournée, nombre : ☐☐ déplacements)
5. ◯ suivre un cours (école, …)
6. ◯ prendre un repas à l'extérieur
7. ◯ faire des courses/du shopping
8. ◯ services (médecin, banque, …)
9. ◯ rendre visite à la famille ou à des amis
10. ◯ se promener, faire un tour
11. ◯ loisirs, sports, culture
12. ◯ autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

PUIS (1) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (2) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (3) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

ENFIN : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. ◯ d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. ◯ d'autres personnes → combien ? ☐☐ personnes
3. ◯ d'animaux
4. ◯ de courses / bagage

Le nombre de déplacements par jour est une donnée importante. N'oubliez aucun déplacement : les retours au domicile, les petits déplacements ou les brefs arrêts pour déposer quelqu'un ou acheter un journal, etc.

Déplacement 9

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …… N° : ….....
Localité : ……..…… Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. déposer/chercher quelqu'un
2. aller à la maison
3. aller travailler
4. pour le travail (si tournée, nombre : ☐☐ déplacements)
5. suivre un cours (école, …)
6. prendre un repas à l'extérieur
7. faire des courses/du shopping
8. services (médecin, banque, …)
9. rendre visite à la famille ou à des amis
10. se promener, faire un tour
11. loisirs, sports, culture
12. autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

PUIS (1) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (2) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (3) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

ENFIN : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. d'autres personnes → combien ? ☐☐ personnes
3. d'animaux
4. de courses / bagage

Déplacement 10

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …… N° : ….....
Localité : ……..…… Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. déposer/chercher quelqu'un
2. aller à la maison
3. aller travailler
4. pour le travail (si tournée, nombre : ☐☐ déplacements)
5. suivre un cours (école, …)
6. prendre un repas à l'extérieur
7. faire des courses/du shopping
8. services (médecin, banque, …)
9. rendre visite à la famille ou à des amis
10. se promener, faire un tour
11. loisirs, sports, culture
12. autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

PUIS (1) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (2) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

PUIS (3) : ☐☐☐ min ☐☐☐ km ☐☐☐ m

auto conducteur	à pied	train	bus De Lijn	autre :
auto passager	cyclo/moto	tram	bus STIB	.....
taxi	à vélo	métro	bus TEC	

ENFIN : ☐☐☐ min **à pied** ☐☐☐ km ☐☐☐ m

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. d'autres personnes → combien ? ☐☐ personnes
3. d'animaux
4. de courses / bagage

Le nombre de déplacements par jour est une donnée importante. N'oubliez aucun déplacement : les retours au domicile, les petits déplacements ou les brefs arrêts pour déposer quelqu'un ou acheter un journal, etc.

Déplacement 11

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …….....…… N° : ….....
Localité : ……..…….. Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. déposer/chercher quelqu'un
2. aller à la maison
3. aller travailler
4. pour le travail (si tournée, nombre : ☐☐ déplacements)
5. suivre un cours (école, …)
6. prendre un repas à l'extérieur
7. faire des courses/du shopping
8. services (médecin, banque, …)
9. rendre visite à la famille ou à des amis
10. se promener, faire un tour
11. loisirs, sports, culture
12. autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD :	☐☐☐ min **à pied**		☐☐☐ km	☐☐☐ m	
PUIS (1) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB		
taxi	à vélo	métro	bus TEC		
PUIS (2) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB		
taxi	à vélo	métro	bus TEC		
PUIS (3) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB		
taxi	à vélo	métro	bus TEC		
ENFIN :	☐☐☐ min **à pied**		☐☐☐ km	☐☐☐ m	

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. d'autres personnes → combien ? ☐☐ personnes
3. d'animaux
4. de courses / bagage

Déplacement 12

Point de départ = destination précédente
Quand et pourquoi avez-vous quitté cet endroit ? Pour quelle destination ?
NB: indiquez un éventuel voyage retour comme un nouveau déplacement.

Destination Pays (si hors Belgique): .....
Rue: …….....…… N° : ….....
Localité : ……..…….. Code postal : ☐☐☐☐

Raison principale *1 seule réponse*

1. déposer/chercher quelqu'un
2. aller à la maison
3. aller travailler
4. pour le travail (si tournée, nombre : ☐☐ déplacements)
5. suivre un cours (école, …)
6. prendre un repas à l'extérieur
7. faire des courses/du shopping
8. services (médecin, banque, …)
9. rendre visite à la famille ou à des amis
10. se promener, faire un tour
11. loisirs, sports, culture
12. autre (précisez): ……..

Heure de départ : ☐☐ h ☐☐ min — si après midi : 13h, 14h, ...

Pour chaque étape du déplacement, entourez le mode de transport et indiquez les **durées** et **distances** correspondantes.

D'ABORD :	☐☐☐ min **à pied**		☐☐☐ km	☐☐☐ m	
PUIS (1) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB		
taxi	à vélo	métro	bus TEC		
PUIS (2) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB		
taxi	à vélo	métro	bus TEC		
PUIS (3) :	☐☐☐ min		☐☐☐ km	☐☐☐ m	
auto conducteur	à pied	train	bus De Lijn	autre :	
auto passager	cyclo/moto	tram	bus STIB		
taxi	à vélo	métro	bus TEC		
ENFIN :	☐☐☐ min **à pied**		☐☐☐ km	☐☐☐ m	

- Recherche d'une place de parking: ☐☐ min

- Si vous avez utilisé une **voiture du ménage**, entourez le numéro (cf. questionnaire ménage) : véhicule 1 | véhicule 2 | véhicule 3

Heure d'arrivée : ☐☐ h ☐☐ min

Etiez-vous accompagné … ?

1. d'enfants de moins de 6 ans → combien ? ☐☐ enfants
2. d'autres personnes → combien ? ☐☐ personnes
3. d'animaux
4. de courses / bagage

Le nombre de déplacements par jour est une donnée importante. N'oubliez aucun déplacement : les retours au domicile, les petits déplacements ou les brefs arrêts pour déposer quelqu'un ou acheter un journal, etc.

26b Si vous avez effectué plus de 12 déplacements le jour de référence, combien de déplacements supplémentaires avez-vous faits ?

☐☐ déplacements supplémentaires

27 Si vous n'avez réalisé aucun déplacement le jour de référence, pourquoi n'avez-vous pas fait de déplacements ce jour-là ? *Plusieurs réponses possibles.*

1 ○ travail à la maison ou études à la maison
2 ○ obligations à la maison (garder quelqu'un, attendre le plombier, ...)
3 ○ handicap permanent, maladie de longue durée
4 ○ maladie de courte durée
5 ○ la météo
6 ○ pas d'activités à réaliser à l'extérieur, congé
7 ○ absence de moyens de transport
8 ○ autre (précisez): .....

PARTIE 3 : OPINIONS

Cette partie ne s'adresse qu'aux personnes de 16 ans et plus

28 Voici une liste de critères de qualité que doivent rencontrer les transports publics. Pouvez-vous pointer les trois critères les plus importants à vos yeux ? *Indiquez dans les trois cases ci-dessous les lettres correspondant aux 3 critères que vous jugez les plus importants, par ordre d'importance.*

	+++	++	+
Critères les plus importants :	☐	☐	☐

A La fréquence de passage
B La vitesse (facilité de circulation)
C La ponctualité
D L'information en cas de situation perturbée
E Le confort (pouvoir être assis)
F Le prix
G La sécurité dans les véhicules
H La sécurité dans les stations / aux arrêts
I La propreté
J La desserte du quartier
K La structure du réseau (correspondances faciles, nombre de lignes)

29 Selon vous, quelles mesures concrètes pourraient le plus inciter certains conducteurs de voiture à réduire l'utilisation de ce mode de déplacement pour alléger la pression du trafic automobile ? *Indiquez dans les quatre cases ci-dessous les lettres correspondant aux 4 mesures que vous jugez les plus pertinentes, par ordre d'importance.*

	++++	+++	++	+
mesures les plus appropriées :	☐	☐	☐	☐

A Une diminution significative de la durée de déplacement en transports en commun
B Une augmentation sensible du confort dans les transports en commun (places assises garanties)
C Une meilleure fiabilité des horaires des transports en commun (ponctualité, qualité de l'information, garantie des correspondances)
D Une offre plus importante des transports en commun (dessertes à fréquence de passage et amplitude accrues)
E Une dégradation significative des conditions de circulation et de parcage (embouteillage, mesures de péage, environnement, prix du carburant, absence de parking hors park&ride, …)
F Une amélioration de l'accessibilité des arrêts d'autobus et des gares (facilités de parcage, dispositifs d'accueil de vélos…)
G Une amélioration du confort aux arrêts d'autobus et dans les gares (abri des intempéries, propreté, information, …)
H Un aménagement des routes plus approprié aux vélos
I Une amélioration de l'aménagement des trottoirs
J Autre (précisez) : ……

30 En zone rurale, parmi les mesures suivantes, quelles sont celles qui vous paraissent les plus appropriées pour améliorer la mobilité ? *Indiquez dans les quatre cases ci-dessous les lettres correspondant aux 4 mesures que vous jugez les plus pertinentes, par ordre d'importance.*

	++++	+++	++	+
mesures les plus appropriées :	☐	☐	☐	☐

A Le développement de bus à la demande (réservation depuis son domicile auprès d'une centrale d'appel)

B Le développement de bus locaux (horaires et itinéraires propres à chaque commune)

C Le développement de lignes d'autobus régionales express (avec un seul arrêt par commune traversée)

D L'aménagement de dispositifs d'accueil de vélos aux gares ou arrêts d'autobus importants

E L'encouragement au covoiturage

F L'amélioration du confort aux arrêts d'autobus (abri des intempéries, propreté, information, …)

G Le développement de taxis de type social pour de courtes distances (association, A.S.B.L., C.P.A.S, ...)

H L'amélioration du réseau routier

I Autre (précisez) : ……..

31 Si vous avez des suggestions d'autres mesures en matière de mobilité en général, notez les ci-dessous.

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

NOUS VOUS REMERCIONS POUR VOTRE AIMABLE COLLABORATION

N'OUBLIEZ PAS…

…de renvoyer ce questionnaire complété, le plus rapidement possible, avec les autres questionnaires de votre ménage dans l'enveloppe fournie. Inutile d'y apposer un timbre, le port est payé par le destinataire.

Si vous avez des questions, des remarques ou des suggestions à propos de ce questionnaire ou de notre enquête en général, veuillez les écrire ci-dessous

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

En application de l'article 4 de la loi du 8 décembre 1992 relative à la protection de la vie privée à l'égard des traitements de données à caractère personnel, nous devons vous informer des éléments suivants:

1. Responsable du traitement : le Service Public Fédéral Mobilité et Transports, Rue du Progrès, 56 à 1210 Bruxelles.

2. Finalité du traitement : Les données recueillies seront utilisées pour l'enquête nationale sur la mobilité des ménages.

3. Destinataires des données : les données anonymisées seront utilisées par différents Services Publics et groupes de recherche scientifique.

4. Les ménages tirés au sort ne sont pas dans l'obligation de répondre.

5. Toute personne concernée par les données a le droit d'y accéder et a le droit de demander leur rectification (uniquement possible durant le déroulement de l'enquête, avant anonymisation des données).

6. Des renseignements complémentaires peuvent être obtenus dans le registre tenu par la Commission de la protection de la vie privée http://www.privacycommission.be (numéro d'identification du traitement:: VT 005014032).

Déposés ~~en même temps~~ :
- une expédition de l'acte
- les statuts coordonnés.

Déposé, 8 avril 2002.

3 151,95 T.V.A. 21 % 31,91 183,86 EUR

(56738)

N. 20020418 — 302

OM SAIRAM

Société coopérative à responsabilité illimitée

Rue Porte basse, 6 à 6900 Marche en Famenne

Marche en Famenne 20 270

464 291 488

Modification

AGE du 6 mars 2002

A partir d'aujourd'hui, SAWHNEY Ram Baldev démissionne de son poste de gérant. L'assemblée accepte sa démission et lui donne décharge de son mandat.

Est nommé gérant Monsieur SAWHNEY Neteen demeurant au N° 9 de la rue Lantigni à 6940 (Durbuy) Barvaux.
Monsieur SAWHNEY Ram Baldev et Madame SRESHEHTA Luxmi cessent d'être associés et cèdent toutes leurs parts sociales à Monsieur SAWHNEY Neteen qui accepte.

La nouvelle répartition du capital est désormais la suivante :

1 – SAWHNEY Robine	20 parts sociales
2 – SAWHNEY Rawine	20 parts sociales
3 – SAWHNEY Neteen	140 parts sociales
4 – SAWHNEY Pummy	20 parts sociales

(Signé) Sawhney, Neteen,
gérant.

Déposé, 8 avril 2002.

1 50,65 T.V.A. 21 % 10,64 61,29 EUR

(56739)

N. 20020418 — 303

«ETUDE NOTARIALE PAUL RAUCENT"

Société civile sous forme de société privée à responsabilité limitée

/080 FRAMERIES, rue du Onze Novembre, 7

CONSTITUTION

D'un acte reçu par Maître VILAIN Franz, notaire de résidence à Frameries, en date du vingt-six mars deux mille deux, enregistré à Colfontaine le vingt huit mars suivant, volume 517 folio 67 numéro 5, trois rôles, sans renvoi, au droit de nonante-trois euros, signé par le Receveur : B.JANSSEN.
Il résulte que :
Monsieur Paul Marie ghislain RAUCENT, Notaire, né à leuven le deux mai mil neuf cent cinquante-sept, époux de madame Hélène Francine Ferdinande Albert Ghislaine BAUDOUX, clerc de notaire, avec laquelle il demeure à Frameries, section de Frameries, rue du Onze Novembre,7.
A constitué une société civile sous forme de société privée à responsabilité limitée aux caractéristiques suivantes :

La société civile adopte la forme d'une société privée à responsabilité limitée.

Elle est dénommée « ETUDE NOTARIALE PAUL RAUCENT ».

Le siège social est établi à 7080 Frameries, rue du Onze Novembre, 7.

Il peut être transféré partout, dans les limites de l'obligation légale de résidence du notaire-associé, à toute autre adresse, par simple décision de la gérance à publier aux annexes au Moniteur belge.

La société est constituée pour une durée illimitée ; elle peut être dissoute par décision de l'assemblée générale délibérant comme en matière de modification aux statuts.

La société n'est pas dissoute par la mort, l'interdiction, la faillite ou la déconfiture d'un associé.

Le capital social a été fixé lors de la constitution à dix-huit mille six cent euros divisé en cent quatre-vingt-six parts sociales sans valeur nominale.
Chacune des parts est entièrement libérée;
Les fonds affectés à la libération de l'apport en numéraire ont été déposés à un compte spécial ouvert au nom de la société.

La gérance de la société ne peut être exercée que par un notaire ou notaire-associé.

En cas de décès ou d'empêchement de celui-ci, l'administration de la société peut être confiée à un autre notaire ou notaire-associé, désigné par le Président de la Chambre des Notaires ou à son défaut, son Vice-Président, à la requête de toute personne intéressée.

Le gérant peut accomplir tous les actes nécessaires ou utiles à l'accomplissement de l'objet social, sauf ceux que la loi réserve à l'assemblée générale ; le gérant représente seul la société à l'égard des tiers ainsi qu'en justice, soit en demandant, soit en défendant.

Dans ses rapports avec les tiers, le gérant peut, sous sa responsabilité, conférer des pouvoirs spéciaux à des mandataires de son choix.

Aussi longtemps que la société ne compte qu'un associé, ce dernier exerce les pouvoirs dévolus à l'assemblée générale, dans les conditions prévues par la loi.

En dehors de cette hypothèse, les associés se réunissent en assemblée générale pour délibérer sur tous les objets qui intéressent la société ou qui ne rentrent pas dans les pouvoirs d'administration de la gérance.

L'assemblée générale ordinaire se tient le dernier jeudi du mois de mai de chaque année à quatorze heures, au lieu désigné dans la convocation. Si ce jour est férié, l'assemblée est remise au plus prochain jour ouvrable.

L'exercice social commence le premier janvier et finit le trente et un décembre.

1°- Gérance :

Le comparant agissant en qualité d'assemblée générale appelle à la fonction de gérant, pour une durée indéterminée, Monsieur Paul RAUCENT prénommé qui accepte.

2°- Premier exercice social et assemblée générale ordinaire :

Le premier exercice social débutera dès le jour du dépôt au Greffe du Tribunal de Commerce d'un extrait du présent acte constitutif et se clôturera le trente et un décembre deux mille deux. La première assemblée générale aura donc lieu en deux mille trois.

3°- Commissaire :

Le comparant déclare que, d'après ses estimations, la société répondra, pour son premier exercice, aux critères légaux qui la dispensent de nommer un ou plusieurs commissaires.

4°- Effets suspensifs de l'article « 2 » du Code des Sociétés – reprise par la société des engagements du notaire pendant la période intermédiaire :

Le comparant déclare savoir que la société ne sera dotée de la personnalité juridique que du jour du dépôt au Greffe du Tribunal de Commerce d'un extrait du présent acte constitutif.

En conséquence, le comparant déclare s'autoriser à souscrire, pendant la période séparant la signature du présent acte et le dit dépôt au Greffe, pour le compte de la société en formation, tous les actes et engagements nécessaires à la réalisation de l'objet social de la future société.

Dès ledit dépôt au Greffe, ces actes et engagements pour compte de la société en formation seront réputés avoir été souscrits dès l'origine par cette société.

Pour extrait analytique :

(Signé) Franz Vilain,
notaire.

Déposé en même temps : une expédition de l'acte constitutif du 26 mars 2002 et une attestation bancaire.

Déposé à Mons, 8 avril 2002 (A/11470).

2 101,30 T.V.A. 21 % 21,27 122,57 EUR

(56740)

N. 20020418 — 304

SOCOBINCHE

Société anonyme

Rue des Chartriers, 33
7000 Mons

Mons - 125276

BE-446.587.604

NOMINATIONS DES ADMINISTRATEURS

Extrait du procès-verbal de l'assemblée générale extra-ordinaire du 18 janvier 2002.
L'assemblée accepte la démission de Mr Félix PALOMAR MARTINEZ de son poste d'administrateur à dater de ce jour.
Elle nomme, à l'unanimité des voix et pour une période de six ans, les administrateurs suivants :
Monsieur Luciano DI PASQUALE, domicilié à 7012 Jemappes, avenue Champ de Bataille, 81,
Monsieur Romolo PANDOLFI, domicilié à 7012 Jemappes, Cité Morette, 23,
Monsieur Antonio DI PASQUALE, domicilié à 59300 Valenciennes, Place du Marché aux Herbes, 8.
Monsieur Luciano DI PASQUALE est nommé au poste d'administrateur-délégué.
Les mandats sont exercés à titre gratuit, sauf décision contraire prise lors d'une prochaine assemblée.

(Signé) Di Pasquale,
administrateur délégué.

Déposé, 8 avril 2002.

1 50,65 T.V.A. 21 % 10,64 61,29 EUR

(56741)

N. 20020418 — 305

TIME POWER

Société Privée à responsabilité limitée

Rue Pré Sabot 7a
7190 ECAUSSINNES

RC Mons n° 136.873

TVA N° 460.507.696

Changement de siège social et de gérance

L'Assemblée Générale extraordinaire de la sprl TIME POWER a acté les modifications suivantes:

Le siège social de la Société est transféré 173 Rue de l'Avedelle à 7190 Marche-Lez-Ecaussinnes

Monsieur Bailly Olivier est nommé gérant de la Société en lieu et place de Madame Pirotte Valérie

Madame Roger Isabelle est nommée associée non active et sans aucune rémunération dans la société

Fait à Ecaussines, le 05 avril 2002.

(Signé) Bailly, Olivier,
gérant.

Déposé, 8 avril 2002.

1 50,65 T.V.A. 21 % 10,64 61,29 EUR

(56742)

N. 20020418 — 306

FRA-CLA

Société Privée A Responsabilité Limitée

Rue du Fisch Club 6
7000 Mons

Mons 141042

BE 547.947.391