DOCTRINE

Entrepôts de données et vie privée

Nathalie Lefever et Yves Poullet

Les entrepôts de données sont des types particuliers de bases de données. Des données en provenance de bases de données variées sont copiées dans l'entrepôt de données de sorte que des requêtes puissent être exécutées sans toucher à l'exécution ou à la stabilité des systèmes de production. Les entrepôts de données sont donc aptes à recueillir un très grand nombre d'informations, y compris des données à caractère personnel, et sont appropriés pour plusieurs différents usages de ce qu'il est convenu d'appeler le « data mining » ou forage de données. En particulier, les puissances de calcul utilisées par de telles bases de données permettent d'établir des corrélations entre des données au départ de nature fort différente et sans lien logique apparent, ce qui permettra dans un second temps de les appliquer à des personnes physiques présentant ce type de corrélations. Ainsi, le consommateur disposant d'un véhicule de telle marque, faisant ses courses le soir et achetant au minimum cinq litres de bière est à 80 % de chance intéressé par un voyage à Miami ou un fraudeur au fisc.

Leur existence cause des soucis sur la sécurité et l'utilisation légitime des données à caractère personnel. Le but de cet article est de donner un aperçu des problèmes pour la vie privée informationnelle qui pourrait résulter de l'utilisation des entrepôts de données. Après une vue d'ensemble du concept d'entrepôt de données et d'un résumé de la législation européenne et belge applicable, l'attention est portée sur le danger de recueillir de grands nombres de données, même anonymes, pour la sécurité et la vie privée, sur le défi de définir des objectifs spécifiques et légitimes de traitement, comme l'exige la loi dite vie privée, lorsqu' on utilise un outil si polyvalent et de souligner à ce propos l'importance du test de proportionnalité. Enfin, on évoque les soucis causés par l'emploi de copies des données au lieu des originaux.

• • •

Datawarehouses are particular types of databases designed in order to support data mining activities in organizations. Data coming from multiple databases are copied to the datawarehouse so that queries can be performed without disturbing the performance or the stability of the production systems. Datawarehouses are therefore likely to gather very large amount of information, including personal data, and suitable for several different uses. In particular, the huge capacities of processing present in theses datawarehouses will permit what is called data mining. It means the possibility to match a large number of datas of different types and without logical link and to define statistical correlations between them. So, a consumer having such kind of car, shopping mostly the evening and purchasing at least five liters of beer per week will normally be interested by a travel to Miami with a 80 % rate and might be suspected in the same proportion of fiscal fraud.

The existence of datawarehouses causes concerns on the security and rightful use of personal data. This article intends to give an insight in the problems for informational privacy that might arise from the use of datawarehouses. After an overview of the concept of datawarehouse and a summary of the European and Belgian legislation applicable, the focus is brought on the danger of gathering large amounts of data, even anonymous, for both security and privacy ; the challenge of defining

specific and rightful aims of treatment, as required by Privacy legislation, notably the importance of the proportionality test, when using such a polyvalent tool. Finally the paper underlines the concerns raised by the use of copies of data instead of originals.

INTRODUCTION

En 1974, le gouvernement français envisagea l'élaboration d'un « Système automatisé pour les fichiers administratifs et le répertoire des individus » (le système S.A.F.A.R.I.), qui prévoyait l'interconnexion des fichiers publics de l'État. Il ne s'agissait que d'une préfiguration de notre registre national. Le projet provoqua un tollé médiatique et une indignation du public tels que non seulement le gouvernement fut contraint de l'abandonner, mais que sur ses cendres naquirent la loi du 6 janvier 1978 « Informatique, fichiers et libertés » et sa gardienne, la C.N.I.L.[1]

Trente ans après, des projets d'envergure bien plus inquiétante pour nos libertés renaissent dans l'ombre par la création d'entrepôts de données, utilisées par le secteur privé autant que le secteur public, et permettant la fouille de données dans un but d'analyse décisionnelle. Bien que la législation sur la vie privée encadre aujourd'hui de telles pratiques, ce nouvel outil que sont les *data warehouses* présente des caractéristiques qui n'étaient pas envisagées par les législateurs, telles que la taille importante de ces entrepôts, leurs applications multiples et leur utilisation de copies des données originales, qui entraîne une absence de liens avec les systèmes de création (et de mise à jour) des données.

Cet article a pour but d'examiner l'application de la législation belge sur les données à caractère personnel à ces amas d'informations et à leur utilisation, afin de cerner les risques potentiels auxquels ils exposent la vie privée des personnes concernées et de proposer des solutions pour s'en prémunir. Il convient pour cela de définir l'entrepôt de données au travers de son historique, ses caractéristiques techniques et ses applications (I), puis de lui appliquer les dispositions protégeant les données à caractère personnel (II). Il sera ainsi possible d'envisager les inquiétudes que cette nouvelle forme de base de données peut susciter pour la vie privée des personnes concernées (III).

I. ENTREPÔTS DE DONNÉES : ORIGINES, CARACTÉRISTIQUES ET APPLICATIONS

A. Origines – l'ère de l'informatique décisionnelle et du *data mining*

L'entrepôt de données est l'instrument privilégié de l'informatique décisionnelle, qui a pour but de permettre aux responsables de la stratégie d'une organisation d'obtenir les informations nécessaires à la prise de décision. Dans un contexte d'internationalisation des échanges et de multiplications des bases de données opérationnelles, tout décideur doit pouvoir bénéficier d'un accès aux données de son organisation qui soit à la fois global et limité : global en ce qu'il n'est pas entravé par les séparations entre les sources de ces données, et limité à ce qui est pertinent pour les besoins de la prise de décision.

[1] Article du journal *Le Monde* du 21 mars 1974 intitulé : « S.A.F.A.R.I. ou la chasse aux Français » et « La C.N.I.L. en bref », brochure d'information sur la Commission nationale de l'informatique et des libertés françaises, http ://www.cnil.fr/fileadmin/documents/La_CNIL/publications/CNIL_EN_BREF-VFVD.pdf.

Or, les outils de l'informatique traditionnelle (ou informatique de production) destinés à traiter des opérations individuelles ne sont pas adaptés à une prise de décision efficace en ce qu'ils ne permettent pas de compiler et d'historiser les données dans le temps. Ces applications classiques permettent de stocker, modifier et restituer les données des services opérationnels mais sont rarement structurées ou codifiées de façon homogène. Pour avoir une vision synthétique et historique de l'organisation, il convient donc de filtrer, homogénéiser et croiser ces données.

À la suite de ce constat, de nouveaux outils se sont développés dans les années 1990 dans le but de rechercher des informations nouvelles ou cachées à partir de données : ce sont les logiciels de *fouille (ou forage) de données* ou *data mining*. Suivant les besoins, ces logiciels permettront d'obtenir des analyses sur le passé ou des analyses prédictives. Ils serviront par exemple à étudier les fichiers d'une compagnie d'assurance contenant les renseignements sur les assurés et l'énumération de leurs sinistres dans le but de déterminer les caractéristiques de l'assuré à risque qui, après validation, permettront de classer les nouveaux clients.

L'activité de ces logiciels de *data mining* est susceptible de se baser sur l'ensemble des données conservées au sein de l'organisation, ce qui représente un volume d'informations considérable. S'il est possible grâce à ces outils d'extraire des connaissances à partir de données puisées directement dans les programmes informatiques de production, cette opération est largement facilitée par la mise en place d'un entrepôt de données ou *data warehouse*, qui rassemble les données puisées dans les bases de production ou en provenance de l'extérieur[2].

B. *Data warehouse* vs. *Database*

Le *data warehouse*, spécifiquement destiné à supporter le *data mining*, offre une organisation logique des données, conçue pour autoriser des recherches complexes. Il présente une architecture « en étoile » qui le différencie de la banque de données classique : le modèle utilisé est souvent multidimensionnel, permettant des mesures suivant plusieurs axes d'analyse simultanés (par exemple, la mesure du nombre de ventes par produit, par mois, par région et par profession des clients, quatre axes d'analyse). De plus, les données sont préparées avant d'être introduites dans le *data warehouse*, ce qui en facilite l'analyse : elles sont filtrées et validées pour maintenir la cohérence de l'ensemble, organisées autour des sujets majeurs et des métiers de l'organisation (« orientées sujet »), synchronisées, « historisées » pour permettre des analyses comparatives au niveau temporel et éventuellement agrégées selon les axes ou dimensions d'analyses prévus, nettoyées, normalisées, consolidées et parfois codées.

Une autre caractéristique du *data warehouse* vient du caractère non volatile, c'est-à-dire stable et non modifiable, des données qu'il contient. Si les systèmes de production se caractérisent par un besoin constant de modifier et d'interroger les données, les utilisateurs des systèmes d'information de décision n'ont pas ce besoin. Par contre, les interrogations, auxquelles ils soumettront le système, peuvent nécessiter des temps de calcul importants qui ne doivent pas interrompre l'activité des serveurs opérationnels. Le *data warehouse* présente donc l'avantage de regrouper de

[2] L'invention du *data warehouse* est généralement attribuée à W. H. Inmon à la suite de la publication en 1992 de son livre intitulé *Building the data warehouse* (QED Information Sciences, Inc., Wellesley, USA). Inmon y définit le *data warehouse* comme « une collection de données orientées sujet, intégrées, non volatiles et historisées, organisées pour le support d'un processus d'aide à la décision ».

grands jeux de données dans une structure qui permette la recherche d'informations.

Enfin, étant donné la variété des besoins des utilisateurs, le *data warehouse* peut être subdivisé en plusieurs sous-ensembles alimentés par l'entrepôt et consacrés à un métier particulier de l'entreprise (marketing, risque, contrôle de gestion…) : ce sont les *comptoirs/magasins de données* ou *data marts*.

C. Caractéristiques

Les caractéristiques d'un *data warehouse* peuvent donc se résumer à ceci :

1. il reprend des données collectées pour diverses finalités ;
2. il les compile et les prépare pour une réutilisation dans un but de *data mining ;*
3. il ne modifie pas les données mais les analyse pour créer de l'information nouvelle ;
4. ses finalités sont généralement d'ordre statistique mais peuvent être très variées et entraîner notamment des décisions relatives à des personnes[3] ;
5. son utilisation peut évoluer suivant les besoins.

D. Applications

Depuis l'avènement du *data warehousing*, ses utilisations se sont multipliées, que ce soit auprès des organismes privés ou des autorités publiques[4].

Les entreprises d'assurance ou les organismes bancaires voient leurs recherches statistiques facilitées. Pour toutes les entreprises commerciales, le *data mining* permet notamment d'optimiser les opérations de marketing en accroissant significativement les taux de réponse. Les centres d'enseignement s'en servent pour fournir aux professeurs des statistiques comparant sur une base standardisée les résultats à certains tests sur une certaine période, et aux élèves des informations concernant notamment les résultats nécessaires pour l'admission à certains types d'enseignements[5].

Au niveau des institutions publiques belges, les Services publics fédéraux Finances et Sécurité sociale ont constitué chacun un *data warehouse* reprenant une sélection de l'ensemble des informations qu'ils possèdent au sujet des citoyens. Ces *data warehouses* remplissent un double rôle : d'une part, un objectif purement statistique d'aide à la décision permettant notamment d'évaluer l'impact d'une législation ; d'autre part, un objectif de « gestion des risques ». Dans ce second cas, le but est de permettre aux services d'inspection d'effectuer des analyses sur des données codées provenant de diverses institutions qui peuvent se

[3] Pour des informations plus poussées sur le fonctionnement technique du *data warehouse* et son rôle dans la prise de décision, les sources ne manquent pas sur Internet : R. Gilleron, M. Tommasi, « Découverte de connaissances à partir de données », notes de cours, Université Charles-de-Gaulle – Lille 3, 2000, disponibles sur http://www.grappa.univ-lille3.fr/polys/fouille/index.html ; M. Santel-Leborgne, « Entrepôt de données », site réalisé dans le cadre des exposés-systèmes du cours de D. Revuz, maître de conférences à l'Université de Marne-La-Vallée, http://etudiant.univ-mlv.fr/~msantel/ ; le site « décisionnel.com » sur les outils de prise de décision, http://www.decisionnel.net/datawarehouse/dwh.htm.

[4] Signe de l'engouement récent pour ce type d'outil, le « pôle bases de données décisionnelles » de l'Université de Lyon 2 organise depuis 2005 des « Journées francophones sur les entrepôts de données et l'analyse en ligne » (http://bdd.univ-lyon2.fr/eda).

[5] Avantages décrits par W. Flaherty, « Data Warehousing and Decision Making », *Virginia Society for Technology in Education Journal*, 2003, vol. 17, n° 2, pp. 28-31, http://www.vste.org/publications/journal/attach/vj_1702/vj_1702_all.pdf. Pour des exemples, voy. le *data warehouse* de la Washington State University (http://infotech.wsu.edu/datawarehouse/), celui de l'Université de Pennsylvanie (http://www.upenn.edu/computing/da/dw/) ou celui de l'Université du Québec à Montréal (http://www.bri.uqam.ca/intra/entrepot.htm).

situer à l'extérieur des services publics fédéraux, et ce en vue d'élaborer des profils types de « personnes à risque » et de les appliquer afin de détecter des fraudes potentielles ; pour les personnes soupçonnées de fraude, les données sont réidentifiées pour les services d'inspection, afin qu'ils puissent effectuer des contrôles ciblés[6].

L'existence de bases de données combinées à une telle échelle et pour des buts aussi variés n'est pas sans provoquer des inquiétudes pour la confidentialité des données à caractère personnel. D'une part, le fait que les données appartiennent déjà à l'organisation peut faire craindre que le responsable de traitement ne respecte pas une partie des obligations auxquelles il est soumis en vertu de la loi du 8 décembre 1992 sur la protection de la vie privée à l'égard des traitements de données à caractère personnel[7], alors qu'il s'agira dans de nombreux cas d'un nouveau traitement. D'autre part, l'on peut se demander si les dispositions de la loi, même scrupuleusement appliquées, protègent adéquatement des données contre les risques spécifiques liés à leur agrégation à grande échelle.

II. LA CONFORMITÉ DES ENTREPÔTS DE DONNÉES À LA LÉGISLATION BELGE SUR LA VIE PRIVÉE

A. La détermination des finalités

Le premier pas dans l'application de la protection de la vie privée aux données à caractère personnel rassemblées dans le *data warehouse* est la détermination des finalités de celui-ci. En effet, de cette définition va dépendre la légalité de l'opération. Le *data warehouse* est constitué non pas de données collectées dans ce but spécifique, mais d'informations puisées dans les banques de données opérationnelles et réutilisées. Par conséquent, soit les finalités de cette réutilisation sont compatibles avec les buts premiers de la collecte, soit il faut réenvisager l'application de la loi sur la protection de la vie privée en ce qui concerne le *data warehouse*.

La détermination de finalités précises semble pourtant particulièrement ardue en ce qui concerne le *data warehouse* qui est conçu dans le but de remplir diverses fonctions très variées. Le fait que les données appartiennent déjà à l'organisation et que leur collecte ait éventuellement fait l'objet d'une déclaration peut amener les responsables de traitement à considérer qu'il leur est loisible de réutiliser les données comme bon leur semble. Rien n'est plus faux : pour vérifier la compatibilité des finalités d'origine avec leur réutilisation dans le *data warehouse*, il est nécessaire de décrire suffisamment précisément quel sera le rôle de ces deux traitements et de s'en tenir par la suite aux utilisations prévues par le premier.

Comme il a été exposé plus haut, dans certains cas, le *data warehouse* remplit une fonction essentiellement statistique. L'article 4, § 1er, 2°, de la loi sur la protection de la vie privée attache à ce type de finalité un certain nombre de conséquences particulières, en prévoyant que

6 Pour le *data warehouse* Sécurité sociale, voy. la délibération du comité sectoriel de la sécurité sociale n° 05/001 du 18 janvier 2005 relative à la création et à la gestion de la banque de données Oasis en vue de la lutte contre la fraude sociale (délibération n° 01/06 du 6 mars 2001, extension de l'autorisation ; pour le *data warehouse* Finances, voy. l'avis de la commission de protection de la vie privée n° 01/07 du 17 janvier 2007 sur l'avant-projet de loi relatif à certains traitements de données à caractère personnel par le Service public fédéral Finances.

7 *M.B.*,18 mars 1993.

« [Les données à caractère personnel doivent être] collectées pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement de manière incompatible avec ces finalités, compte tenu de tous les facteurs pertinents, notamment des prévisions raisonnables de l'intéressé et des dispositions légales et réglementaires applicables. Un traitement ultérieur à des fins historiques, statistiques ou scientifiques n'est pas réputé incompatible lorsqu'il est effectué conformément aux conditions fixées par le Roi, après avis de la Commission de la protection de la vie privée ».

Il convient donc de déterminer ce qu'on entend par « finalités statistiques » avant d'exposer les obligations légales auxquelles sont soumis les *data warehouses* suivant qu'ils entrent ou non dans cette catégorie.

B. La définition des « finalités statistiques »

L'article 4, § 1er, 4°, de la loi de 1992 emploie les termes « fins historiques, statistiques ou scientifiques » sans les définir. L'arrêté royal du 13 février 2001[8], qui porte exécution de cet article, n'apporte lui non plus aucune définition. Cependant, le rapport au Roi qui l'accompagne indique qu'à défaut de pouvoir se référer à une définition absente de la directive 95/46/CE comme le souhaitait la Commission de protection de la vie privée[9], il convient d'interpréter ces notions à la lumière du sens qui leur est donné dans la recommandation n° R(97)18 du comité des ministres du Conseil de l'Europe aux États membres concernant la protection des données à caractère personnel collectées et traitées à des fins statistiques.

Cette recommandation définit l'expression « à des fins statistiques » comme se référant à « toutes opérations de collecte et de traitement de données à caractère personnel nécessaires aux enquêtes statistiques ou à la production de résultats statistiques », c'est-à-dire réalisées « en vue de caractériser un phénomène collectif dans une population considérée ». La définition précise encore que « de telles opérations excluent toute utilisation de l'information obtenue pour des décisions ou des mesures relatives à une personne déterminée ».

En d'autres termes, la distinction entre traitements statistiques ou non statistiques se base sur le résultat que l'on vise à obtenir : s'il concerne un groupe de personnes (la « population considérée »), le traitement est statistique ; mais il ne l'est pas s'il se réfère à une personne particulière. L'exposé des motifs de la recommandation précise que « dans tous les cas (…) ces résultats [statistiques] ne disent rien de spécifique d'aucune des personnes dont l'information a été utilisée ». Le fonctionnement de l'arrêté royal du 13 février 2001 suit d'ailleurs cette logique : puisque le traitement statistique ne caractérise que des phénomènes collectifs, il doit être en principe réalisé sur des données rendues anonymes. Cependant, la qualification en traitement statistique et l'application des règles d'anonymisation qui en découlent n'excluent pas que des données soient recoupées et permettent l'identification des personnes, ou que le traitement ne soit l'objet d'un détournement de finalité[10].

Le *data warehouse* n'étant qu'un outil pouvant être utilisé dans des buts très variés, il n'est pas possible de classer définitivement les finalités qu'il remplit dans l'une ou l'autre catégorie (traitements statistiques, traitements non statis-

[8] 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, *M.B.*, 13 mars 2001.

[9] Avis n° 25/99 du 23 juillet 1999 sur le projet d'arrêté royal 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.

[10] C. de Terwangne et S. Louveaux, « Protection de la vie privée face au traitement de données à caractère personnel : le nouvel arrêté royal », *J.T.*, 2001, p. 457.

tiques). Cependant, l'outil a été conçu pour certains types de recherches, celles aboutissant à la prise de décision, et pour certaines catégories d'utilisateurs, à savoir les organisations publiques ou privées. S'il remplit le rôle pour lequel il a été créé, il réalise donc des tâches qui peuvent être en principe classifiées.

Cette classification a été faite dans l'exposé des motifs de la recommandation R (97)18 (point 13) : « La finalité d'aide à la planification et à la décision s'adresse à des responsables qui sont appelés à prendre deux types de décisions : des décisions générales – lois, barèmes, campagnes de vaccination, organisation des transports, conception de modèles, mise en production, etc. – et des décisions individuelles – admission ou radiation, imposition, allocation, traitement, etc. Or les données à caractère personnel qui sont collectées et traitées à des fins statistiques ne doivent servir qu'au premier type de décisions »[11].

Dans les décisions du premier type, on retrouve pour l'organisme privé les choix stratégiques, commerciaux, budgétaires, de gestion des ressources humaines, de marketing, etc., dans la limite où ils ne concernent pas une ou plusieurs personnes déterminées mais décident d'une orientation générale. Pour l'organisme public il s'agira notamment d'évaluer l'impact d'une législation, d'étudier une population dans un but sociologique, de cibler des groupes à atteindre ou informer, de servir d'appui à une décision réglementaire, etc. Les décisions du second type recouvrent les décisions individuelles : engager ou licencier un employé, accorder une assurance ou un avantage social, proposer une offre à un client particulier, décider d'un contrôle sur un employé, un contribuable, un bénéficiaire social, etc.

Remarquons au passage qu'il n'est pas exclu de classer dans la finalité statistique, scientifique ou historique des opérations qui ont un but final de type commercial, comme l'élaboration de profils de consommateurs à des fins de marketing ciblé. Le *data warehouse* est un outil particulièrement performant pour ce genre de recherche.

Il est important de noter que les deux types de finalités pourront coexister. C'est le cas notamment dans les *data warehouses* des Services publics fédéraux Finances et Sécurité sociale. Comme énoncé plus haut, ils remplissent à côté d'un rôle purement statistique un rôle de « gestion des risques », lequel peut se diviser en deux opérations : d'une part, l'élaboration de profils de risques, permettant sur base d'études statistiques de décrire les caractéristiques d'une population présentant un taux élevé de fraude et, d'autre part, l'application de ces profils aux données recueillies concernant les citoyens pour cibler les contrôles vers les personnes jugées « à risque » en fonction des profils. Ces deux opérations sont à l'évidence intimement mêlées ; pourtant, en vertu des critères déterminés plus haut, l'élaboration des profils appartient à la catégorie « opérations statistiques », tandis que la détermination des personnes à risques ne peut être qualifiée de la même façon.

[11] À l'appui de cette interprétation, on retrouve également le considérant n° 28 de la directive 95/46/CE du 24 octobre 1995 du Parlement européen et du Conseil relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (*J.O.C.E.* n° L 281 du 23 novembre 1995, pp. 0031-0050) : « Considérant que le traitement ultérieur de données à caractère personnel à des fins historiques, statistiques ou scientifiques n'est pas considéré en général comme incompatible avec les finalités pour lesquelles les données ont été auparavant collectées, dans la mesure où les États membres prévoient des garanties appropriées ; que *ces garanties doivent notamment empêcher l'utilisation des données à l'appui de mesures ou de décisions prises à l'encontre d'une personne* ».

C. Application des dispositions légales aux entrepôts de données

1. Principes généraux

La réutilisation des données détenues par une entreprise ou une organisation dans le cadre d'un *data warehouse* constitue un nouveau traitement auquel s'appliquent les dispositions de la loi sur la protection de la vie privée. Rappelons brièvement les conséquences que cela implique.

La loi sur la protection de la vie privée stipule en son article 4 que des données à caractère personnel doivent toujours être « collectées pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement de manière incompatible avec ces finalités, compte tenu de tous les facteurs pertinents, notamment des prévisions raisonnables de l'intéressé et des dispositions légales et réglementaires applicables ». Selon l'interprétation qui est faite de cette disposition dans l'exposé des motifs[12], la loi laisse donc la possibilité au responsable de traitement de réutiliser les données qu'il détient soit pour réaliser une finalité qui soit compatible avec le but de la collecte, soit en modifiant fondamentalement la finalité, mais, dans ce cas, il s'agira d'un nouveau traitement au sens de la loi.

Le principe de compatibilité traduit l'exigence de transparence des traitements : chaque personne doit pouvoir prévoir l'utilisation des données la concernant. La jurisprudence a condamné sur la base de ce principe l'utilisation par des établissements bancaires de données personnelles produites lors de transcriptions financières dans un but de marketing ou de contrôle des agents[13].

En ce qui concerne le *data warehouse*, étant donné la variété des applications pour lesquelles il pourra être utilisé, il est fort probable que, dans de nombreux cas, l'emploi des données qui le constituent dépasse les prévisions raisonnables des intéressés. Ce sera de toute façon le cas si les données sont collectées dans un but non commercial et sont réutilisées à des fins de marketing sans information de la personne concernée. Dans d'autres hypothèses, la jurisprudence précitée semble indiquer une tendance vers une limitation des cas où la réutilisation des données pour usage interne sera considérée comme prévisible par l'intéressé[14].

Si le traitement des données au sein du *data warehouse* ne peut être considéré comme compatible avec la finalité pour laquelle elles ont été collectées, cette réutilisation impliquera la naissance d'un nouveau traitement au sens de la loi. Dans ce cas, le responsable devra respecter à nouveau l'intégralité de ses dispositions applicables lors de l'établissement d'un traitement, à savoir :

- déclarer le traitement auprès de la Commission de protection de la vie privée (article 17) ;
- se trouver dans l'une des situations autorisant le traitement (article 5) ; même si le but de la collecte initiale rencontrait cette condition, une réutilisation des données implique une vérification de la légitimité de l'objectif poursuivi ;

[12] Voy. à ce propos l'analyse de T. Léonard et Y. Poullet, « La protection des données à caractère personnel en pleine (r)évolution », *J.T.*, 1999, p. 377, n° 30.

[13] Comm. Anvers, 7 juillet 1994, *D.C.C.R.*, 1994-1995, p. 77, note T. Léonard ; Anvers, 3 mai 1999, *A.J.T.*, 1999-2000, p. 437 ; Comm. Bruxelles, 23 octobre 2003, *Computerrecht*, 2004, p. 7, note D. de Bot ; Bruxelles, 15 février 2005, *Forum Financier / Droit bancaire et financier*, 2005/IV, p. 273, note D. Daubies et M. Mairlot.

[14] Voy. à ce propos la note de D. De Bot sous Comm. Bruxelles, 23 octobre 2003, ci-dessus.

– informer la personne concernée de l'utilisation nouvelle qui sera faite de ses données, et pour cela, déterminer avec précision les finalités du *data warehouse* (article 4) ;
– ne pas traiter les données dites sensibles (articles 7 et 8) ; par exemple, l'entreprise qui traite des données sensibles concernant les membres de son personnel afin d'exécuter ses obligations et ses droits spécifiques en matière de droit du travail ne sera plus autorisée à les réutiliser dans un autre cadre.

2. *Les entrepôts de données à finalité statistique*

La loi sur la protection de la vie privée accorde un régime de faveur aux traitements de données qui sont destinés à effectuer des opérations statistiques, historiques ou scientifiques : d'une part, leur traitement est réputé compatible avec les finalités pour lesquelles les données ont été collectées et, d'autre part, les données peuvent être conservées au-delà de la durée nécessaire à la réalisation des finalités d'origine. En pratique, cette déclaration de compatibilité des finalités implique qu'il ne faudra pas, comme lors de la création d'un nouveau traitement, justifier d'une base légale pour la collecte, informer les personnes collectées de ces nouvelles opérations ou effectuer une nouvelle déclaration à la Commission de protection de la vie privée.

Ces deux exceptions ne seront cependant accordées que si le traitement ultérieur respecte les conditions énumérées dans l'arrêté royal du 13 février 2001[15]. Cet arrêté énonce le principe selon lequel les recherches statistiques ou scientifiques devront s'effectuer autant que possible sur la base de données anonymes, c'est-à-dire des « données qui ne peuvent être mises en relation avec une personne identifiée ou identifiable »[16]. Ces données n'étant plus des données à caractère personnel, l'exception s'applique sans autre condition.

S'il n'est pas possible d'atteindre le ou les buts recherchés avec des données anonymes, l'arrêté royal impose, autant que possible, l'usage de données codées, c'est-à-dire des données « qui ne peuvent être mises en relation avec une personne identifiée ou identifiable que par l'intermédiaire d'un code »[17]. Ce sera par exemple le cas lorsqu'il sera nécessaire de mettre en relation les données collectées à propos d'une même personne. Une série d'obligations sont alors prévues pour tenter de garantir un codage efficace des données en fonction des hypothèses : traitement ultérieur par le responsable du traitement initial, par un tiers ou par un sous-traitant.

Dans l'hypothèse d'un *data warehouse* à vocation décisionnelle, le cas de figure le plus fréquent sera le cas d'un traitement ultérieur par le responsable du traitement initial. Le responsable du traitement peut alors soit confier le codage à un organisme indépendant, soit coder lui-même les données mais en s'assurant que les mesures techniques et organisationnelles appropriées soient prises pour éviter le décodage. Il doit en outre justifier de la nécessité de l'usage de données codées lors de la déclaration de traitement auprès de la Commission de protection de la vie privée.

Certains *data warehouses* peuvent cependant être étoffés par l'obtention de données provenant de l'extérieur, comme des données achetées à des organisations privées spécialisées dans le marketing, ou des données obtenues de la part d'autorités publiques pour

[15] Pour une analyse détaillée des exigences de l'arrêté royal, voy. C. de Terwangne et S. Louveaux, « Protection de la vie privée face au traitement de données à caractère personnel : le nouvel arrêté royal », *J.T.*, 2001, p. 457.

[16] Article 1er, 5°, de l'arrêté royal.

[17] Article 1er, 3°.

les *data warehouses* des S.P.F. Sécurité sociale et Finances. Les données devront alors être codées par le tiers fournisseur ou par une entreprise intermédiaire avant leur réception.

Si la réutilisation des données sous forme codée ne permet pas d'atteindre les buts historiques, statistiques ou scientifiques recherchés, le traitement de données à caractère personnel non codées est admis sous une série de conditions plus strictes que celles auxquelles est soumise une collecte primaire dans le même but. Cette rigueur est destinée à compenser la perte de maîtrise des données par la personne concernée et l'avantage du responsable de traitement qui n'a plus à collecter et contrôler les données. Le régime de l'arrêté royal impose une information stricte de la personne concernée et son autorisation expresse à la réutilisation des données. Deux exemptions sont prévues pour les cas d'utilisation de données publiques ou lorsque cette procédure se révèle impossible ou requiert des efforts disproportionnés.

3. Les entrepôts de données à finalités statistiques et non statistiques

Quelle que soit l'hypothèse, le responsable de traitement devra veiller à respecter les droits de la personne concernée, notamment à ne prendre de décisions produisant des effets juridiques à propos d'une personne sur le seul fondement d'un traitement de *data mining* que dans les cas énumérés à l'article 12*bis*, alinéa 2, et à l'informer dans ce cas de la logique qui sous-tend le traitement (article 10, § 1er, *c*). Les données devront toujours être pertinentes, c'est-à-dire limitées à ce qui est strictement nécessaire pour la réalisation du but recherché (d'où, encore une fois, la nécessité de déterminer précisément les finalités du traitement). Elles devront également être mises à jour régulièrement.

III. LES PROBLÈMES PARTICULIERS POSÉS PAR LES ENTREPÔTS DE DONNÉES EN MATIÈRE DE VIE PRIVÉE

Après avoir analysé les implications de la loi sur la protection de la vie privée aux traitements de données via un *data warehouse*, il convient de s'interroger sur la pertinence des mécanismes de protection mis en place. Le phénomène en expansion du *data mining* et des *data warehouses* pose un certain nombre de questions nouvelles en terme de protection des données. Ces questions sont liées notamment aux risques liés à l'agrégation d'un grand nombre de données (même anonymes), aux difficultés de définition des finalités, à la sortie des données de leur contexte original et à la sécurité de ces entrepôts.

A. L'agrégation de données anonymes

Comme constaté plus haut, la protection des données ne s'étend qu'aux données dites personnelles ; il suffit donc, pour pouvoir traiter librement des données au sein d'un *data warehouse*, de les anonymiser de façon irréversible (sans que le risque de dévoilement ne puisse être toujours rigoureusement nul)[18] avant de les agréger. À partir du moment où ces données ne sont plus identifiables directement ou indirectement, la législation dans son ensemble ne s'applique plus. D'autre part, une grande partie des obligations liées à la réutilisation de données au sein d'un *data warehouse* seront levées si les données, à défaut de pouvoir être rendues anonymes, sont codées dans les règles avant d'être traitées dans un but statistique. Le législateur estime ainsi que l'anonymisation ou le codage dans un but statistique représente une protection suffisante de la confidentialité des données.

[18] *Cfr* exposé des motifs de l'arrêté royal du 13 février 2001.

Cependant, l'agrégation de données au sein de vastes banques ajoute une dimension au problème qui incite à remettre en cause la pertinence de cette présomption. Que se passe-t-il lorsque les données anonymes en provenance de plusieurs sources sont recoupées et analysées de façon à parvenir à l'identification d'individus avec un très haut taux de probabilité ? « Des statistiques pourtant anonymes peuvent, si elles présentent un niveau trop fin d'agrégation, soulever des difficultés au regard des principes issus de la législation de protection des données à caractère personnel »[19]. Si la probabilité d'identification n'est pas de 100 %, et si les données étaient à la base anonymes ou codées, l'on peut pourtant s'interroger sur l'application formelle de la loi.

Ceci prend encore davantage d'importance quand les données disponibles auprès du maître de traitement sont rassemblées, au sein du *data warehouse*, avec des données en provenance de l'extérieur. Prenons l'exemple d'une société de vente qui posséderait légalement les noms et les coordonnées de cibles potentielles pour des opérations de marketing. En recoupant ces informations avec des statistiques anonymes concernant, par exemple, le niveau de revenu des ménages en fonction de leur localisation sur le territoire, le responsable marketing crée une nouvelle donnée qui lui permettra de mieux cibler sa campagne publicitaire ; c'est là une application typique du *data mining*. Cependant, du point de vue de la personne concernée, il découvre une nouvelle information confidentielle, avec la crainte que les obligations légales ne soient pas correctement respectées sous le prétexte qu'il ne s'agit pas d'une donnée certaine, seulement d'une présomption.

Bien entendu, l'application de statistiques ne donne pas naissance à des certitudes ; mais un *data mining* efficace qui se baserait sur un nombre de sources suffisamment important pourrait, en théorie, engendrer de nouvelles informations personnelles avec un taux de probabilité suffisamment élevé pour être pris en compte lors de décisions[20]. Il se crée alors un schéma de profilage[21] où la personne concernée se voit attribuer, à son insu, des caractéristiques personnelles supplémentaires d'autant plus dangereuses qu'elles sont incontrôlées et qu'elles se basent sur des probabilités. Qu'arrive-t-il à la personne à qui est attachée une caractéristique erronée ? Dans certains cas, l'inconvénient sera mineur, comme l'envoi de publicités mal ciblées ; mais d'autres situations pourront aboutir à des décisions injustes, comme le refus d'une assurance ou des contrôles fiscaux répétés parce que le contribuable se trouve par erreur dans une catégorie « à risques ».

À ce propos, le groupe de travail de l'« article 29 » sur la protection des données a adopté le 20 juin 2007 un avis analysant en profondeur

19 C. de Terwangne et S. Louveaux, « Protection de la vie privée face au traitement de données à caractère personnel : application en Belgique de la directive européenne », *Actualités du droit des technologies de l'information et de la communication,* Liège, formation permanente C.U.P., 2001, p. 12. Voy. aussi L. Van Wel et L. Royakkers, « Ethical issues in web data mining », *Ethics and Information Technology*, Kluwer, 2004, 6 : 131 : « Even non-identifiable data can become identifiable when merged ».

20 S. Sumathi et S.N. Sivanandam (« Major and Privacy Issues in Data Mining and Knowledge Discovery », *Studies in Computational Intelligence (SCI)*, 29, 277 (2006)) présentent l'exemple de la mise en corrélation de données bancaires et téléphoniques afin de déterminer si les clients d'une banque possèdent un fax à domicile et quel en est l'impact sur leur potentiel à s'engager dans un prêt.

21 Un problème connexe est celui de la perte d'individualité, la stigmatisation des personnes en fonction du groupe auquel elles appartiennent. Voy. à ce sujet L. van Wel et L. Royakkers, « Ethical issues in web data mining », *Ethics and Information Technology*, Kluwer, 2004, 6 : 129-140.

le concept de données à caractère personnel[22]. Le groupe de travail y énumère notamment les conditions dans lesquelles une donnée peut être considérée comme concernant une personne physique, en distinguant un élément de *contenu*, de *finalité* ou de *résultat*. Il en résulte qu'une donnée concerne une personne soit lorsqu'elle a directement trait à une personne, soit lorsque son traitement a pour finalité d'évaluer, de traiter d'une certaine manière ou d'influer sur le statut ou le comportement d'une personne, ou encore lorsque l'utilisation de cette donnée aura pour résultat un impact, même mineur, sur certains des droits et intérêts d'une personne. L'avis précise également qu'une donnée *a priori* anonyme peut devenir indirectement identifiable (et donc personnelle) via la combinaison d'un ou plusieurs éléments spécifiques liés à l'identité physique, physiologique, psychique, économique, sociale ou culturelle. Autrement dit, « une combinaison de détails à un niveau catégoriel (tranche d'âge, origine régionale, etc.) peut également s'avérer assez concluante dans certaines circonstances, notamment si l'on a accès à des informations supplémentaires »[23].

Ces précisions règlent le sort des données recueillies et créées au sein d'un *data warehouse* : il s'agit de données à caractère personnel soumises à la législation appropriée, même si elles ne sont *a priori* pas identifiantes. En effet, elles sont personnelles dans la mesure où elles peuvent servir de base à une décision relative à une personne ou avoir un impact sur la situation d'une personne (élément de résultat), et elles le sont même si l'identification à une personne particulière ne résulte que d'une combinaison de détails qui ne soient pas totalement concluants (identification indirecte).

B. La définition des finalités

Un second problème se pose pour l'application de la loi aux entrepôts de données : celui de la définition de la finalité du traitement. On l'a vu, un grand nombre de dispositions de la loi dépend des finalités déclarées. Pourtant, la principale justification du rassemblement des données dans un même entrepôt est la possibilité de les analyser de multiples façons et, partant, de les utiliser dans des buts variés. Le même outil pourra servir à orienter une campagne publicitaire, évaluer l'efficacité des travailleurs, mesurer la rentabilité d'un produit, décider de l'opportunité d'une action, etc. Il est même soutenu que, poussé à son niveau de performance le plus élevé, le *data mining* ne permet pas de prévoir le résultat d'une analyse avant de l'entamer : une recherche ne doit pas nécessairement s'appuyer sur une hypothèse de départ mais découvrir « à l'aveugle » de nouvelles corrélations entre les données[24].

De là naît la tentation, pour le maître de traitement, de définir largement la finalité de rassemblement des données : les termes « finalité interne », *data mining* ou toute dénomination se référant à la mission globale d'une organisation, ne peuvent suffire. La loi précise bien que les finalités doivent être « déterminées, explicites et légitimes » et, comme nous l'avons vu, l'ensemble des dispositions relatives à l'information de la personne concernée ou à la limitation du traitement perdent tout leur sens sans une définition précise des buts initiaux. La personne concernée doit être capable de prévoir, au moins largement, l'utilisation qui sera faite de ses données. L'imprévisibilité des corrélations entre données n'empêche pas que la recherche est toujours entamée dans un but particulier : cibler une campagne de marketing,

[22] Avis 4/2007 sur le concept de données à caractère personnel, document n° 01248/07/FR WP 136.

[23] *Idem*, p. 15.

[24] A. Cavoukian, « Data Mining : Staking a Claim on Your Privacy », *Publication of the Information and Privacy Commissionner*, Ontario, janvier 1998, pp. 12-13.

traquer la fraude, etc. Ce but doit être connu de la personne concernée.

C. Le danger des données sorties de leur contexte

Comme il a été souligné plus haut, l'un des avantages essentiels du *data warehouse* dans le processus de *data mining* est de pouvoir travailler non pas sur des données telles que conservées dans les systèmes de production, mais sur des copies stables et homogénéisées. Le but est, entre autre, de ne pas gêner les utilisateurs en ralentissant le système lors d'analyses complexes. Cependant, l'utilisation du *data warehouse* risque de mettre la pertinence des données en cause de deux façons.

D'une part, l'actualité des informations n'est plus garantie, puisqu'elles ne sont plus liées au système de production ; une modification de la donnée dans l'ensemble original d'où elle provient ne se reflétera pas dans le *data warehouse*. Or l'article 4 de la loi sur la vie privée prévoit explicitement que « Les données à caractère personnel doivent être (…) exactes et, si nécessaire, mises à jour ; toutes les mesures raisonnables doivent être prises pour que les données inexactes ou incomplètes, au regard des finalités pour lesquelles elles sont obtenues ou pour lesquelles elles sont traitées ultérieurement, soient effacées ou rectifiées ». Il est donc essentiel que le *data warehouse* soit mis à jour suffisamment régulièrement que pour refléter au plus près l'actualité des données telles qu'elles figurent dans les fichiers actifs, et notamment prendre en compte les modifications apportées par la personne concernée sur base de son droit de rectification.

Le second problème qui se pose est également lié à cette disposition de l'article 4. Le *data warehouse* a typiquement pour vocation de rassembler en un seul ensemble les données issues des différents milieux de production. Or l'on peut s'interroger sur la pertinence de données ainsi sorties de leur contexte original et interprétées dans un tout autre but. Il convient également de prêter attention au processus de filtrage qui intervient généralement préalablement à l'entrée des données dans le *data warehouse* ; s'il a pour but d'uniformiser les données, il peut avoir pour effet une perte de qualité ou de précision préjudiciable aux résultats de l'analyse. Il est de l'intérêt du responsable de traitement autant que de la personne concernée de veiller à ce que les données analysées restent fiables et gardent leur sens initial, au vu des informations qui les accompagnaient à l'origine.

D. Problèmes de sécurité

Le *data warehouse* est l'outil des grands ensembles de données. Il est conçu pour rassembler en un seul lieu toutes les données d'une organisation. Cependant, un tel rassemblement augmente les risques pour la sécurité et la confidentialité des données, surtout dans les organisations de taille importante. La dispersion des données et la limitation des accès, qui caractérisent les systèmes de production multiples, permettent aussi de réduire la gravité des fuites, volontaires ou involontaires. À la suite de la création d'un *data warehouse*, de telles fuites peuvent potentiellement être d'une gravité beaucoup plus importante.

E. Proportionnalité

Au vu des problèmes énoncés ci-dessus, il convient de s'interroger sur les alternatives à la création d'un *data warehouse*. La loi prévoit en effet que les traitements de données doivent être proportionnés au but poursuivi, et la jurisprudence a rappelé que les effets d'un traitement sont disproportionnés par rapport à l'objectif poursuivi si cet objectif avait pu être atteint d'une manière moins dommageable

pour les intéressés[25]. Or des solutions techniques ont été élaborées en vue de limiter les risques pour la confidentialité des données tout en respectant les objectifs et l'efficacité du processus de *data mining*. L'on peut notamment envisager l'utilisation de données exclusivement codées, et dont le codage serait assuré par un service totalement indépendant du service chargé du *data mining*. Il serait également conseillé de diviser l'entrepôt en « comptoirs » de données indépendants, de façon à limiter la fouille de données aux ensembles strictement nécessaires à la recherche. Dans tous les cas, le rassemblement et le traitement de données doit se limiter à ce qui est strictement nécessaire pour réaliser le but légitime poursuivi.

CONCLUSION : LE *DATA WAREHOUSE*, UN OUTIL INDISPENSABLE ?

Le *data mining* et son support favori le *data warehouse* font depuis longtemps partie des pratiques habituelles du secteur privé autant que du secteur public. Pouvoir extraire de la connaissance nouvelle à partir de données déjà en possession de l'organisation et l'utiliser pour rentabiliser l'activité offre un atout non négligeable. L'outil est puissant et particulièrement flexible, ce qui garantit son succès croissant.

Cependant, l'opinion publique ignore souvent la taille et le potentiel de ces entrepôts où les données personnelles sont rassemblées et analysées. En fournissant des informations à un service particulier, le citoyen est loin de s'imaginer qu'elles pourront être recoupées avec d'autres informations qui lui ont été demandées dans un tout autre contexte. Il ne se doute pas que dans des entrepôts virtuels, des amas de données parfois gigantesques s'assemblent et que ceux qui les manipulent sont susceptibles d'apprendre sur lui des choses qu'il ignore.

La législation actuelle apporte des garde-fous qui répondent à certaines des inquiétudes liées à cette tendance. Pourtant, tous les risques ne trouvent pas leur réponse. À partir de ce constat, certains auteurs ont proposé des solutions techniques, allant jusqu'à démontrer que le *data mining* peut se faire sans rassemblement des données en un même lieu[26]. Dans tous les cas, il est important d'attirer l'attention des dirigeants d'organisations utilisant les techniques du *data mining* sur les dangers de celles-ci pour la confidentialité des données et la vie privée des personnes concernées. Il n'est pas impossible d'allier les avantages d'un *data warehouse* avec le respect des dispositions légales en matière de données à caractère personnel, à la condition que cet aspect soit envisagé préalablement à la mise en place d'un tel système. La question de la proportionnalité de tels traitements doit également être posée. Dans quelle mesure les résultats escomptés ne pourraient-ils pas être atteints par des moyens moins attentatoires à la vie privée ? En matière de lutte contre la fraude, finalité certes légitime, l'ampleur des conséquences sur des catégories de population du choix de telles méthodes doivent les faire réserver à des cas extrêmes et non se substituer aux méthodes traditionnelles d'investigation.

[25] C.A., arrêt n° 16/2005 du 19 janvier 2005.

[26] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin et M. Y. Zhu, « Tools for Privacy Preserving Distributed Data Mining », SIGKDD Explorations (Purdue University), Vol. 4, Issue 2, p. 1-7, disponible à l'adresse http://www.cs.purdue.edu/homes/clifton/DistDM/kddexp.pdf.