Anonymisation de données par généralisation

Méthode avec guidage

Feten Ben Fredj[1], **Nadira Lammari**[1], **Isabelle Comyn-Wattiau**[2]

1. CEDRIC-CNAM, 2 rue Conté, 75003 Paris, France
fetenbf@yahoo.fr
ilham-nadira.lammari@cnam.fr,

2. ESSEC Business School, 1 av. Bernard Hirsch, 95021 Cergy, France
wattiau@essec.edu

RÉSUMÉ. De nombreux algorithmes ont été proposés pour l'anonymisation des données personnelles, permettant de minimiser le risque de ré-identification tout en préservant l'utilité des données. Dans cet article, nous décrivons une approche fondée sur les modèles qui guide l'éditeur de données dans son processus d'anonymisation. Le guidage, informatif ou suggestif, permet non seulement de choisir l'algorithme le plus pertinent, mais aussi de paramétrer cet algorithme en tenant compte des caractéristiques des données et du contexte. Dans cet article, nous nous intéressons aux algorithmes de généralisation de micro-données. Un processus de rétro-ingénierie des outils existants a permis d'extraire certaines connaissances. Nous les stockons avec toutes les connaissances liées à l'anonymisation, tant théoriques qu'expérimentales, dans une ontologie.

ABSTRACT. Many algorithms allow data owners to anonymize personal data, aiming at avoiding disclosure risk without losing data utility. In this paper, we describe a model-driven approach guiding the data owner during the anonymization process. The guidance, informative or suggestive, helps the data owner not only in choosing the most relevant algorithm but also in defining the best input values for the algorithm, given the characteristics of data and the context. In this paper, we focus on generalization algorithms for micro-data. We conducted a reverse engineering process in order to extract some knowledge from existing anonymization tools. The knowledge about anonymization, both theoretical and experimental, is managed thanks to an ontology.

MOTS-CLÉS : guidage, sécurité, ontologie, méthodologie, respect de la vie privée, anonymisation, approche guidée par les modèles.

KEYWORDS: guidance, security, ontology, methodology, privacy, anonymization, model-driven approach.

DOI:10.3166/ISI.23.1.63-87

1. Introduction

Avant le 25 mai 2018, les organisations soumises aux réglementations européennes doivent urgemment mettre en place un processus de vérification de leur niveau de protection des données personnelles (Règlement général européen sur la protection des données – RGPD). C'est la question du risque de divulgation de données sensibles, et plus particulièrement, le risque de violation de la vie privée *via* l'utilisation de données personnelles. La norme ISO/TS 25237:2008 définit l'anonymisation comme « un processus qui supprime l'association entre l'ensemble de données identifiant et le sujet des données ». C'est un processus complexe, notamment parce qu'il tente de satisfaire deux objectifs contradictoires que sont : l'utilité des données (c'est-à-dire leur qualité) et leur sécurité (c'est-à-dire leur confidentialité). Par conséquent, les détenteurs de données doivent mettre en œuvre un processus de protection qui réponde au mieux à la confidentialité et à l'utilité de leurs données. Ce processus inclut des prises de décision à différentes phases. En effet, ils sont amenés entre autres à sélectionner un algorithme d'anonymisation, à paramétrer cet algorithme et à évaluer la qualité du rendu après application du procédé. Ces décisions s'appuient sur leur connaissance du domaine.

Les outils existants ne sont pas suffisants en raison de leur opacité et de leur manque de guidage dans le choix et le paramétrage des algorithmes. Dans la littérature scientifique qui abonde d'articles de recherche proposant des algorithmes d'anonymisation, nous avons aussi constaté l'absence d'approches guidées pour l'anonymisation.

Ces constats ont motivé notre démarche de création d'une ontologie de domaine pour l'anonymisation de micro-données[1] ainsi que d'une approche guidée s'appuyant sur cette ontologie (BenFredj *et al.*, 2017). L'ontologie produite (BenFredj *et al.*, 2015), que nous avons nommée OPAM, permet de capitaliser les connaissances du domaine. Cependant, elle ne stocke qu'une portion d'expertise du domaine. En effet, OPAM, n'a été, pour l'instant, instanciée que par les connaissances récoltées sur la technique de généralisation de micro-données. Par conséquent, l'approche MAGGO (Méthodologie pour une Anonymisation par Généralisation Guidée par une ontologie) que nous décrivons dans cet article, sert de guide pour un professionnel dans sa prise de décision lors d'une anonymisation par généralisation de micro-données. Nous avons construit OPAM à l'aide d'une méthode incrémentale décrite dans (BenFredj *et al.*, 2015). La connaissance sur les techniques de généralisation constitue le résultat du premier incrément.

Après un état de l'art (section 2), nous proposons un processus d'anonymisation (section 3), puis nous décrivons l'approche générale qui s'appuie sur ce processus (section 4), puis ses étapes détaillées (section 5). En section 6, nous illustrons l'approche sur un exemple. En section 7, nous concluons et présentons des axes de recherche future.

1. Données atomiques décrivant des individus (Hand, 1992).

2. Etat de l'art

Les tables relationnelles peuvent contenir aussi bien des macro-données que des micro-données. Une macro-donnée, telle que le mentionne (Ciriani *et al.*, 2007), est une donnée agrégée décrivant un ensemble d'individus. Une micro-donnée est une information de base caractérisant un individu vis-à-vis d'un attribut (par exemple, le prénom d'une personne). Elle peut contribuer à l'identification d'un individu ou à sa quasi-identification. Elle peut aussi être sensible ou non sensible (Ciriani *et al.*, 2007). Ainsi, dans une table décrivant des individus, on peut trouver quatre groupes distincts d'attributs : l'identifiant explicite[2], le quasi-identifiant[3], les attributs sensibles[4] ainsi que les attributs non sensibles.

Les macro-données ont pendant longtemps constitué les publications traditionnelles des organismes nationaux de statistique. Ainsi, la recherche sur la protection de ce type de données est la plus ancienne et la plus établie (Dalenius, 1977). Elle a été essentiellement menée par des statisticiens travaillant sur le contrôle de divulgation statistique, connu sous le terme de SDC (*Statistical Disclosure Control*) et/ou sous celui de SDL (*Statistical Disclosure Limitation*). En revanche, la protection des micro-données est plus récente. La recherche liée à la définition et la mise en œuvre des techniques d'anonymisation pour ce type de données bénéficie non seulement de l'apport de la communauté de statisticiens mais aussi de celle des informaticiens s'intéressant à la préservation de la vie privée à des fins de fouille de données connue sous l'acronyme anglais PPDM (*Privacy Preserving Data Mining*) ou à des fins de publication connue sous l'acronyme PPDP (*Privacy Preserving Data Publishing*). Le PPDM est un domaine de recherche visant à étendre les techniques traditionnelles de fouille de données de façon à pouvoir manipuler des données où l'information sensible a été masquée (Aïmeur, 2009 ; Vaghashia et Ganatra, 2015 ; Vassilios *et al.*, 2004). Le PPDP étudie comment immuniser les données contre les attaques de la vie privée (Amita *et al.*, 2014 ; (Kiran et Kavya 2012 ; Fung *et al.*, 2010). Plusieurs publications proposent des revues de littérature de ces trois domaines (SDC/SDL, PPDM, PPDP), par exemple

2. Un identifiant est un attribut ou un ensemble d'attributs qui désigne directement un individu (par exemple, un numéro de sécurité sociale, un prénom, un nom). Ce n'est pas nécessairement un identifiant au sens de la modélisation conceptuelle, puisqu'un prénom et/ou un nom peuvent être partagés par plusieurs individus. Toutefois, au sein d'un jeu de données, ce type d'information nominative peut facilement conduire à une ré-identification.

3. Un quasi-identifiant (QI) est un ensemble d'attributs dont la sélectivité est telle qu'ils présentent un risque de ré-identification, Par exemple {sexe, code postal, date de naissance} forme un quasi-identifiant connu dans de nombreux ensembles de données. Ils sont suffisamment discriminants pour permettre de retrouver une seule personne dans une base de données.

4. Un attribut sensible représente les données que les individus ne veulent généralement pas publier, comme des informations médicales ou des salaires.

(Hussien *et al.*, 2013). Pour des raisons d'espace, nous nous concentrons, dans cet état de l'art, sur les techniques les plus fréquentes dédiées aux micro-données.

Les techniques d'anonymisation de micro-données ont des degrés de fiabilité et des contextes d'applicabilité variables. Sous le vocable de contexte, on intègre l'usage souhaité des données (par exemple tester un logiciel ou encore publier des données à des fins d'analyse) et le type de données à anonymiser (micro ou macro données tabulaires, données spatiotemporelles, graphes, images, textes, etc.). Le degré de fiabilité est en lien direct avec le risque de ré-identification des données anonymes. En effet, face à l'évolution des technologies de l'information qui rendent possible le lien entre des données de différentes sources, il est quasiment impossible d'effectuer une anonymisation qui garantirait un risque de ré-identification nul.

Les techniques d'anonymisation peuvent être classées en deux catégories : les techniques perturbatrices et les techniques non perturbatrices (Patel et Gupta, 2013). La première catégorie représente les procédures dans lesquelles les données résultantes ne sont pas dénaturées, c'est-à-dire que les données sont vraies mais qu'elles peuvent manquer de détails, alors que les techniques de la deuxième catégorie peuvent dénaturer les données, c'est-à-dire les rendre inexactes, ce qui n'empêche pas leur usage à des fins de test ou de statistique par exemple. Par exemple, la suppression consiste à retirer des données de la table pour éviter leur divulgation. C'est une technique non perturbatrice. La technique de recodage global (« global recoding ») s'applique à toutes les valeurs d'un attribut afin d'uniformiser au plus les enregistrements et donc de diminuer le risque de ré-identification. Ainsi, on peut remplacer l'âge d'un individu par un intervalle. Le « data swapping » (Fienberg et McIntyre, 2004) consiste à permuter les valeurs d'un même attribut entre des paires d'enregistrements. La micro-agrégation (Defays et Nanopoulos, 1993) répartit les données originales en groupes homogènes. Par la suite, les valeurs originales sont remplacées par la moyenne ou la médiane du groupe auquel elles appartiennent. La technique de bruit aléatoire (« random noise ») (Brand, 2002) s'applique à un seul attribut à la fois. Elle fonctionne en ajoutant ou en multipliant chaque valeur de l'attribut à anonymiser par une valeur aléatoire. La technique de généralisation consiste à remplacer des valeurs par d'autres plus générales (Samarati, 2001) : les données sont vraies, mais moins précises, comme l'année que l'on substitue à la date de naissance.

Chacune de ces techniques a donné lieu à un ou plusieurs algorithmes. Ainsi, il existe une grande variété de techniques d'anonymisation et encore plus d'algorithmes qui les mettent en œuvre. Des comparaisons de techniques sont proposées (Ilavarasi *et al.*, 2013 ; Fung *et al.*, 2010). Certains états de l'art sont certes orientés usage mais restent non accessibles à des éditeurs de données avec de faibles compétences dans le domaine de l'anonymisation. De plus, les algorithmes ne sont accessibles qu'à travers les publications de recherche. Leur spécification se rapproche du code de programmation. Ils sont, le plus souvent, partiellement illustrés à l'aide d'exemples. Leurs principes fondamentaux sont décrits

textuellement. Par conséquent, ils ne sont compréhensibles que par des informaticiens ou des professionnels ayant des compétences en programmation.

Il existe aussi des logiciels d'anonymisation. La plupart de ces outils ne fournissent pas de guidage dans le choix de la technique et/ou de l'algorithme proposé ni d'aide à leur paramétrage. Le guidage se limite à fournir en parallèle de l'anonymisation des valeurs de métriques appliquées aux données anonymisées afin d'évaluer notamment le risque résiduel et la dégradation due à l'anonymisation.

La littérature comprend ainsi de nombreuses métriques permettant d'évaluer la qualité des données anonymisées, en termes de perte d'information ou de précision, ou de risque de ré-identification (Fung *et al.*, 2010).

Enfin, il existe d'autres travaux connexes qui pourraient utilement être exploités pour compléter l'automatisation de notre approche. Par exemple, la technique de généralisation suppose de connaître le ou les attributs formant les quasi-identifiants, concept qui sera défini ci-après. Dans cet article, nous supposons qu'ils sont connus. Il existe des travaux de recherche visant la détection automatique de ces attributs. Citons (Agrawal *et al.*, 2014) qui combine une analyse statique et une analyse dynamique des programmes qui accèdent aux données sensibles. (Motwani et Xu, 2007) proposent des métriques (*distinct ratio* et *separation ratio*) et des algorithmes utilisant ces ratios pour rechercher les clés et les quasi-identifiants. (Akoka *et al.*, 2014) utilisent des techniques d'analyse syntaxique et sémantique pour déduire de façon semi-automatique la sensibilité des attributs.

2.1. La technique de généralisation de micro-données et ses algorithmes

La généralisation de micro-données est la technique la plus explorée dans le cadre du PPDM et PPDP. Elle s'applique à un ensemble d'attributs formant un quasi-identifiant (QI). Elle met en œuvre *a minima* le k-anonymat[5]. Elle nécessite la définition d'une hiérarchie pour chaque attribut composant le QI. Elle consiste à remplacer une valeur par son ancêtre direct dans la hiérarchie de généralisation, et cela à chaque étape de la généralisation. Ainsi, on peut appliquer une seule étape de généralisation à l'attribut Ville et deux étapes de généralisation à l'attribut Age. La généralisation de micro-données est mise en œuvre *via* plusieurs algorithmes dont les plus connus sont : µ-Argus (Burton *et al.*, 1997), Datafly (Sweeney 1997), l'algorithme de Samarati (Samarati 2001), Incognito (LeFevre *et al.*, 2005), « Bottom up generalization » (Wang *et al.*, 2004), « Top down specialization » (TDS) (Fung *et al.*, 2005), « Median Mondrian » (LeFevre *et al.*, 2006), « InfoGain Mondrian » et « LSD Mondrian » (LeFevre *et al.*, 2008).

5. C'est le premier modèle de protection de la vie privée proposé dans la littérature (Samarati et Sweeney, 1998). Lorsqu'il est mis en œuvre, il offre l'assurance que chaque n-uplet de valeurs du quasi-identifiant apparaît au moins k fois dans la table à publier.

µ-Argus propose une exécution itérative du processus d'anonymisation. A chaque itération, (a) l'utilisateur choisit l'attribut à généraliser, (b) µ-Argus remplace chaque valeur de cet attribut par la valeur de son parent dans la hiérarchie de généralisation correspondante, (c) vérifie la satisfaction du k-anonymat et rend compte à l'utilisateur qui peut choisir entre la poursuite du processus ou encore la suppression locale de données (remplacement de valeurs par la valeur nulle).

Datafly, contrairement à µ-Argus, exécute automatiquement des suppressions globales (c'est-à-dire des suppressions de tuples). Afin de minimiser la perte d'information qui pourrait être engendrée par des généralisations excessives, il utilise la métrique DA (*Distinct Attribute*) qui consiste à choisir de généraliser, parmi les attributs du QI, celui qui a le plus de valeurs distinctes. Il poursuit la généralisation tant que le nombre de tuples ne satisfaisant pas le k-anonymat est au-dessus du seuil de suppressions autorisées par l'utilisateur. Puis, s'il reste des tuples ne satisfaisant pas le k-anonymat, Datafly procède à leur suppression.

L'algorithme de Samarati est fondé sur un treillis qui représente les combinaisons possibles des niveaux de généralisation de tous les attributs du quasi-identifiant. Chaque nœud, dans le treillis, définit un niveau de généralisation pour chaque attribut du QI, c'est-à-dire une généralisation possible de la table originale. Dans le treillis, les nœuds les plus bas définissent une généralisation minimale pour réduire la perte d'information. Les nœuds les plus hauts garantissent une meilleure protection, au prix d'une plus forte généralisation. Ainsi, la généralisation optimale est un compromis qui peut être atteint au niveau de certains nœuds. Afin de trouver ces nœuds « optimaux », l'algorithme agit par itération dichotomique en considérant les nœuds du niveau h/2, h étant la hauteur de la partie inexplorée du treillis. Le processus s'arrête lorsqu'au moins un nœud du niveau h/2 satisfait le k-anonymat. L'algorithme mémorise alors ce nœud ainsi que tous les nœuds de ce niveau permettant de satisfaire le k-anonymat (avec ou sans suppression globale). Il se poursuit par exploration de la moitié inférieure du treillis, dans le cas contraire. Il s'arrête lorsque h est égal à 0 et restitue à l'utilisateur les derniers nœuds mémorisés.

Incognito est également fondé sur un treillis. Cependant, le treillis est construit de manière itérative et incrémentale afin d'atteindre une plus grande efficacité. En d'autres termes, à la première itération, Incognito construit tous les treillis liés à un attribut du QI. Chaque treillis correspond à une hiérarchie de généralisation. Ces différents treillis sont nettoyés en supprimant tous les nœuds qui ne conduisent pas à une généralisation k-anonyme. A chaque itération, Incognito fusionne les treillis résultant de l'étape précédente et procède à leurs nettoyages. Ce processus se poursuit de façon itérative jusqu'à ce que le treillis regroupant tous les attributs du QI soit construit et nettoyé.

L'algorithme « Bottom up generalization » est destiné à préserver les données pour un type spécifique de traitement statistique qu'est la classification. Il parcourt les hiérarchies de généralisation des feuilles vers la racine. Il exécute, successivement, des généralisations qu'il considère comme « bonnes » dans le sens

où elles préservent au mieux la qualité de la classification tout en fournissant le k-anonymat souhaité. Chaque généralisation est sélectionnée parmi un ensemble de généralisations candidates. Elle est considérée comme bonne si elle renvoie le meilleur score calculé par application de la métrique de compromis IL/AG(G) agrégeant une mesure de la perte d'information liée à la généralisation G et une autre calculant le gain d'anonymat engendré par G.

L'algorithme « Top down specialization » a le même objectif que le précédent mais parcourt la hiérarchie de la racine vers les feuilles. Il utilise par conséquent la métrique de compromis IG/AL(S) pour sélectionner la spécialisation S qui préserve au mieux le k-anonymat tout en générant le moins de perte d'information.

Pour satisfaire le k-anonymat, l'algorithme « Median Mondrian » représente les enregistrements dans un espace multidimensionnel où chaque dimension correspond à un attribut du QI. La position initiale d'un enregistrement dans l'espace multidimensionnel tient compte de la valeur des attributs du QI. Le découpage de l'espace en zones permet la constitution de groupes d'individus. Plus précisément, à chaque itération, l'algorithme choisit une dimension (c'est-à-dire un attribut du QI) et vérifie la possibilité de diviser un groupe en deux sous-groupes (en divisant la zone selon la valeur médiane de cette dimension). La division est possible si, dans chaque groupe résultant, il existe au moins k individus (satisfaction du k-anonymat). Chaque groupe trop petit pour être divisé est marqué. Lorsque tous les groupes sont marqués pour une dimension, l'algorithme passe à une autre dimension (un autre attribut). Il arrête son processus de marquage lorsque toutes les dimensions ont été explorées. Ensuite, il applique les généralisations proposées, en remplaçant les différentes valeurs d'une même zone par la valeur de leur premier parent commun. On parle de processus de recodage pour désigner cette dernière étape.

L'algorithme « InfoGain Mondrian » (respectivement « LSD Mondrian ») étend « Median Mondrian » pour des données à usage de classification (respectivement régression). Pour ce faire, ils combinent le recodage multidimensionnel de Median Mondrian avec des heuristiques de partitionnement orientées vers la classification ou vers la régression. Intuitivement, il s'agit dans Infogain, de choisir, à chaque itération, le partitionnement qui minimise l'entropie pondérée de l'ensemble des partitions résultantes tout en préservant la contrainte d'anonymat. L'utilisation de cette métrique favorise l'obtention de partitions homogènes. LSD Mondrian, quant à lui, s'inspire de l'algorithme CART de construction d'un arbre de régression (Breiman *et al.*, 1984). A chaque itération, il choisit la division qui minimise la somme pondérée des erreurs quadratiques moyennes pour les partitions résultantes.

2.2. Evaluation de la qualité d'une anonymisation par généralisation

Pour évaluer la qualité des données obtenues suite à une anonymisation par généralisation, plusieurs métriques (Fung *et al*., 2010) sont proposées dans la

littérature[6]. Parmi ces métriques, on peut citer la complétude qui, comme son nom l'indique, mesure la complétude de la table anonyme qui est directement fonction du taux de suppressions effectuées. Elle est utile, bien entendu, dans le cas où l'on utilise un algorithme qui procède à des suppressions de données. D'autres métriques telle que la métrique DM (*Discernability Metric*) (Bayardo et Agrawal, 2005) renseignent l'utilisateur sur la qualité des données résultant du degré de différenciation des individus. En effet, le k-anonymat obligeant k individus à partager le même QI, plus la taille de la classe d'équivalence est grande, plus l'utilité des données est amoindrie. Pour mesurer l'effet négatif d'une anonymisation générant des classes d'équivalence de taille dépassant de beaucoup k, LeFevre *et al.* (2006) proposent la métrique CAVG. Pour mesurer la perte en précision dans l'anonymisation par généralisation (Sweeney, 2002) a introduit la métrique PREC. Cette dernière exploite le postulat selon lequel, plus la hiérarchie associée à un attribut est profonde, plus cette donnée est généralisable conduisant à plus de perte de la précision. GenIloss (Generalized Information Loss) (Iyengar, 2002) est aussi une métrique permettant d'évaluer la qualité des données anonymes. Elle est fondée sur l'hypothèse selon laquelle si un attribut du QI prend ses valeurs dans une grande plage de valeurs, alors, après anonymisation, les données sont moins précises que dans le cas où l'intervalle est moins large.

2.3. Les outils pour l'anonymisation de micro-données par généralisation

Dans cet article, nous nous limitons aux outils qui implémentent la technique de généralisation. Nous avons recensé les plus cités dans la littérature : CAT *(*Xiao *et al*. 2009), TIAMAT (Dai *et al*., 2009), SECRETA (Poulis *et al*., 2014), ARX[7], PARAT[8]. L'outil CAT (*Cornell Anonymization Toolkit*) met en œuvre l'algorithme de Samarati[9]. Dans TIAMAT (*Tool for Interactive Analysis of Microdata Anonymization Techniques*), deux algorithmes de généralisation du QI sont implémentés : le Median Mondrian et l'algorithme k-Member (Dai *et al*. 2009). SECRETA (*System for Evaluation and Comparing RElational and Transaction Anonymization algorithms*) est aussi un prototype. Il intègre entre autres quatre algorithmes de généralisation de micro-données : Incognito, Cluster, Top-down and Full subtree bottom-up. ARX utilise un unique algorithme baptisé 'flash' qui permet de construire un treillis de généralisation similaire à Incognito ou Samarati. Contrairement aux outils précédemment décrits, l'outil PARAT (Privacy Analytics Risk Assessment Tool) est commercialisé. Dans tous les outils commercialisés, les algorithmes mis en oeuvre ne sont pas dévoilés. Mener une anonymisation est un processus dont il faut définir les objectifs et les étapes. C'est à partir de

6. Il n'existe pas de mesures standard qui soient largement acceptées par la communauté de chercheurs (Kiran et Kavya, 2012).
7. http://arx.deidentifier.org/development/algorithms/
8. https://www.nahdo.org/sites/nahdo.org/files/conference_sessions/PrivacyAnalytics.pdf
9. Selon (Xiao *et al*., 2009), CAT utilise l'algorithme Incognito.

l'observation de ces outils que nous avons conçu ce processus. Ce dernier est décrit dans la section suivante.

3. Définition du processus d'anonymisation

L'objectif de notre méthode est d'assister l'éditeur de données dans son processus d'anonymisation. Pour ce faire, il a à sa disposition plusieurs outils tels ceux décrits plus haut. L'utilisation de ces derniers requiert une compréhension de leur fonctionnement et, notamment, du processus que ces outils soutiennent.

3.1. Les étapes du processus d'anonymisation

En étudiant ces outils, nous avons reconstruit, par une approche de rétro-conception, les étapes de ce processus (figure 1). La première étape est l'importation dans l'environnement de l'outil de la base de données originale. La seconde est la détermination du quasi-identifiant et des attributs sensibles. Ensuite, le seuil de risque est défini. Puis un algorithme est choisi avec le paramétrage adéquat. Plusieurs paramétrages peuvent être testés. Nous appelons signature un paramétrage possible d'un algorithme. Comme le montre la figure 1, on peut généralement tester de façon itérative différents algorithmes et différentes signatures de ces algorithmes. Suit l'évaluation du résultat en termes de risque mais aussi de qualité. Si l'évaluation n'est pas satisfaisante, d'autres algorithmes et/ou d'autres signatures peuvent être testés. Enfin, la base de données anonymisée est exportée.

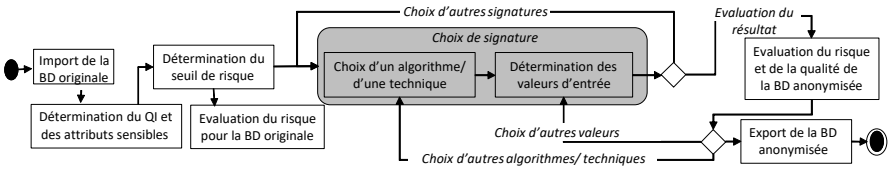

Figure 1. Les étapes du processus d'anonymisation

3.2. Le guidage pendant le processus d'anonymisation

L'objectif de notre approche est de guider l'éditeur de données pour mener à bien son anonymisation. En matière d'aide, chaque outil d'anonymisation propose un tutoriel décrivant les connaissances du domaine et parfois certains conseils pour mener à bien son processus d'anonymisation. Cependant, ces descriptions ne sont pas, la plupart du temps, explicites ou faciles à comprendre. L'outil PARAT est le seul outil à offrir une interface interactive permettant à l'utilisateur de suivre les étapes du processus d'anonymisation. Nous avons pu constater qu'aucun de ces outils n'offre un guidage dans le choix d'un algorithme ou d'une technique ni pour leur paramétrage. L'outil TIAMAT est le seul à offrir un guidage pour la

détermination des constituants d'un QI. L'outil PARAT est le seul qui contient un guidage dans la définition du seuil de risque toléré. Il propose un taux selon un contexte fourni par l'utilisateur. Enfin, tous les outils offrent un guidage informatif pour l'évaluation des signatures. Notre comparaison des outils nous a permis de dégager leurs limites. En effet, même si certains outils offrent un guidage au moment de l'évaluation de la base de données anonymisées, il n'est pas compréhensible par des éditeurs de données avec de faibles compétences dans le domaine de l'anonymisation. Notre approche vise à fournir un guidage à chaque étape de l'anonymisation.

3.3. Le choix de l'algorithme

Il n'y a pas d'outil, à notre connaissance, qui implémente tous les algorithmes de généralisation décrits dans l'état de l'art. Toutefois, dans (Benfredj *et al.*, 2014), nous avons analysé tous les éléments qui permettent de comparer ces algorithmes. Ils peuvent concerner aussi bien les prérequis à leur exécution que leurs entrées, leurs processus et leurs sorties. Dans (BenFredj *et al.*, 2016), nous avons proposé un arbre de décision pour guider le choix d'un algorithme en fonction de tous ces critères.

Enfin, tous les algorithmes de généralisation sauvegardent de façon implicite la cohérence, l'exactitude et la véracité des données. Cependant, ils affectent la précision des données. Pour éviter que le processus de transformation des données par généralisation ne dégrade trop la précision de celles-ci et ne remette en cause leur utilité, la plupart des algorithmes de généralisation utilisent, au cours de leur processus, une métrique permettant d'orienter le codage des données.

3.4. La connaissance expérimentale sur les algorithmes

Plusieurs articles décrivent des résultats d'expérimentation donnant lieu à des évaluations d'algorithmes en termes de qualité et de temps d'exécution (Ayala-Rivera *et al.*, 2014 ; Fung *et al.*, 2005 ; LeFevre *et al.*, 2008). Ces deux critères varient la plupart du temps selon un certain nombre de paramètres en entrée tels que la valeur de k, la taille du QI, la distribution de la base de données originale, etc. En ce qui concerne l'évaluation de la qualité, les auteurs la mesurent selon différents critères tels que la complétude ou la précision, en utilisant différentes métriques pour un même critère. Notons que, certains articles, au vu des expérimentations et de la complexité de certains algorithmes, fournissent quelques recommandations quant à l'utilisation de ces algorithmes (Ayala-Rivera *et al.*, 2014 ; Fung *et al.*, 2010). Ainsi, la comparaison des algorithmes étant une tâche très difficile en soi du fait de la variété des métriques, il peut s'avérer intéressant de conserver les évaluations d'algorithmes, fournies dans les publications, comme base d'exemples pour un apprentissage supervisé sur les algorithmes d'anonymisation. Dans le cadre d'une anonymisation par généralisation de micro-données, l'ensemble des paramètres est

composé de l'algorithme servant à l'anonymisation, du critère d'évaluation, de la métrique d'évaluation associée au critère, de la valeur de k, du nombre d'attributs constituant le QI ainsi que de la distribution des données.

Enfin, à notre connaissance, à l'exception de notre ontologie OPAM (BenFredj *et al.*, 2015), il n'existe pas de base de connaissance dans laquelle le professionnel chargé de la désidentification des données pourrait rechercher les connaissances le guidant vers une anonymisation utile et préservant au mieux la vie privée. Il n'existe pas non plus de méthode qui puisse concrétiser le processus d'anonymisation de données tout en offrant des aides à la décision.

Ainsi, dans cet article, nous définissons une approche d'aide à la décision permettant, à l'aide de l'ontologie OPAM, de guider l'éditeur de données dans le choix d'un algorithme de généralisation de micro-données et dans son paramétrage. Dans la suite, nous présentons l'approche MAGGO en détaillant ses étapes principales.

4. Présentation générale de l'approche

L'anonymisation de données est une des mesures de sécurité qui peuvent être préconisées dans le cadre de la protection de la vie privée. Dès lors que cette mesure est décidée, le responsable de l'anonymisation doit concevoir et exécuter un processus de brouillage. Pour cela, il doit 1) repérer les données identifiantes, quasi identifiantes (QI) et sensibles, 2) proposer des techniques appropriées avec une orchestration adéquate. Il lui faut aussi, pour chaque technique, identifier l'algorithme à appliquer, trouver un paramétrage reflétant ses besoins et évaluer la qualité des données anonymisées en termes d'utilité et de sécurité en se conformant au cahier des charges de l'anonymisation. Ce processus comprend plusieurs points de décisions-clés dont la qualité affecte le résultat final. Il exige du responsable de l'anonymisation une grande maîtrise du domaine. Offrir une aide sur la totalité du processus exigerait des efforts considérables, compte tenu de la variété des données susceptibles d'être brouillées (micro-données, données liées, données géographiques, etc.) et de la diversité des techniques existantes et des algorithmes de mise en œuvre de ces techniques. Dans cet article, nous nous focalisons sur une partie du processus d'anonymisation, c'est-à-dire une technique (la généralisation) et un type de donnée (les micro-données contenues dans une table relationnelle). En effet, nous proposons une approche guidée permettant, compte tenu d'un contexte d'anonymisation (défini dans un cahier des charges) de choisir l'algorithme de généralisation de micro-données le meilleur – au regard des exigences du cahier des charges – et de l'exécuter. Le meilleur algorithme est celui qui offre le meilleur compromis entre les exigences contradictoires de sécurité et d'utilité. Plus précisément, la recherche du compromis se fait en évaluant plusieurs algorithmes avec plusieurs combinaisons possibles de paramètres.

Pour aider l'utilisateur dans la spécification du contexte, dans la sélection de signatures et de solutions d'anonymisation, MAGGO met à disposition de l'utilisateur des connaissances nécessaires pour le rendre apte à décider (figure 2). Ainsi, à chacune des étapes, MAGGO fait intervenir des connaissances expertes en vue d'un guidage suggestif ou informatif. Le guidage suggestif guide l'utilisateur dans ses choix alors que le guidage informatif lui fournit des informations qui peuvent éclairer son choix (Silver, 2006). Dans notre cadre, le guidage suggestif aide l'éditeur de données dans la sélection de l'algorithme et de la signature tandis que le guidage informatif lui fournit des informations pour éclairer son choix.

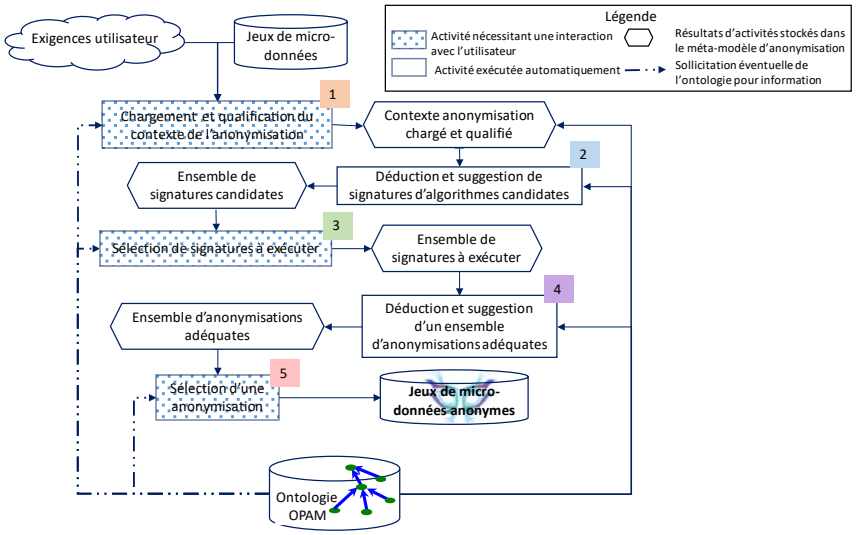

Figure 2. Les étapes de MAGGO

Tableau 1. Type de guidage pour chaque étape

Etape	Activité	Guidage
1	Chargement et qualification du contexte de l'anonymisation	informatif
2	Déduction et suggestion d'un ensemble de signatures candidates	**suggestif**
3	Sélection de signatures à exécuter	informatif
4	Déduction et suggestion d'un ensemble d'anonymisations adéquates	**suggestif**
5	Sélection d'une anonymisation	informatif

Le tableau 1 récapitule les types de guidage offerts dans MAGGO selon l'étape. Ces connaissances sont rendues disponibles *via* OPAM. Le guidage de MAGGO est incrémental dans le sens où il est introduit à différents points de décisions clés tout au long du processus.

La notion de méta-modèle joue un rôle central dans notre approche. En effet, alors que l'ontologie met à disposition de l'approche les connaissances nécessaires à

l'anonymisation, le méta-modèle (figure 3) réunit les abstractions conceptuelles des artefacts cibles et sources de notre approche.

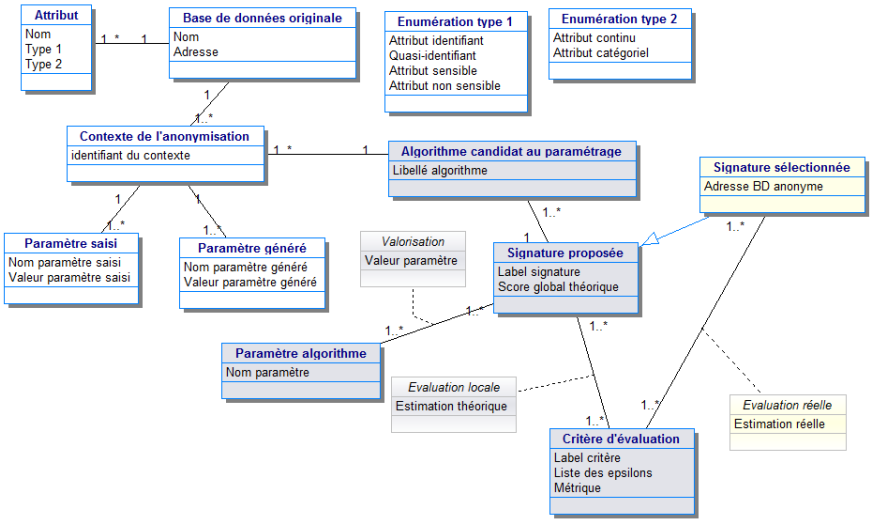

Figure 3. Le méta-modèle du processus d'anonymisation

La base de données originale est caractérisée par un ensemble d'attributs dont on définit le type 1 (identifiant, quasi identifiant, sensible, non sensible) et le type 2 (continu ou catégoriel). Son contexte[10] est défini à partir de paramètres saisis (par l'utilisateur) ou générés (calculés automatiquement ou déduits de paramètres saisis). En fonction du contexte, on déduit des algorithmes candidats. Pour ces algorithmes, on déduit des signatures de paramètres à évaluer. L'évaluation peut être théorique, c'est-à-dire déduite d'évaluations comparables contenues dans l'ontologie, ou réelle, c'est-à-dire déduite de l'exécution de l'algorithme sur le jeu de données. Ainsi, la première étape de MAGGO permet d'instancier le méta-modèle en décrivant et en qualifiant le contexte de l'anonymisation. Un enrichissement du modèle, par des données complémentaires issues de chacune des différentes étapes, est effectué.

L'enrichissement du méta-modèle est réalisé à l'aide de deux techniques statistiques : l'analyse hiérarchique multicritère ou *Analytical Hierarchy Process* (AHP) (Saaty et Sodenkamp, 2008) et la régression. La première est utilisée aux étapes 2 et 4 pour évaluer les résultats soumis à l'utilisateur. La régression est utilisée sous la forme d'apprentissage supervisé afin de prédire la valeur d'une métrique d'utilité ou d'une métrique de sécurité, compte tenu de données expérimentales disponibles dans OPAM.

10. Le tableau 2 fourni plus loin liste les paramètres de chaque catégorie. Par exemple, les attributs quasi-identifiants sont déductibles de l'examen du jeu de données.

Avant de présenter les différentes étapes de MAGGO, pour faciliter la compréhension de l'approche, nous rappelons à la figure 4 les éléments principaux du méta-modèle d'OPAM (BenFredj *et al.*, 2015). Les classes sur fond blanc dans la partie haut gauche de la figure sont celles qui permettent la représentation de la connaissance « théorique » relative aux techniques et algorithmes d'anonymisation. Les classes sur fond gris permettent de décrire les concepts sous-jacents au processus d'anonymisation. Enfin, les classes sur fond blanc dans la partie bas droit de la figure sont le support de la représentation de la connaissance empirique extraite des articles de recherche ou accumulée grâce à nos tests des outils.

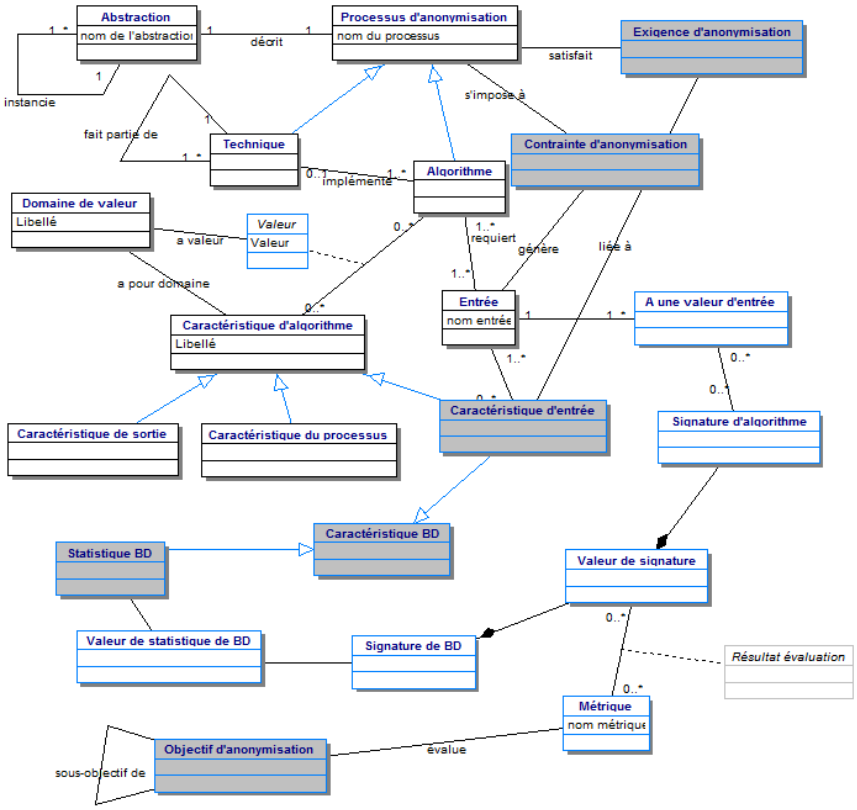

Figure 4. Méta-modèle sous-jacent de l'ontologie OPAM (extrait)

5. Description des étapes de la méthode MAGGO

Dans cette section, nous décrivons chacune des cinq étapes de la démarche MAGGO schématisée à la figure 2.

5.1. Etape 1 - Chargement et qualification du contexte de l'anonymisation

Une anonymisation vise la prévention contre des attaques potentielles portant atteinte à la vie privée. Sa mise en œuvre nécessite la sélection d'une ou plusieurs techniques qui mettent en œuvre le modèle de protection censé contrer ces attaques. Ainsi se pose le problème de choix d'algorithmes pour mettre en œuvre l'anonymisation qui répond aux attentes de son initiateur. Ces attentes constituent les deux catégories d'exigences que doit satisfaire l'anonymisation. La première catégorie rassemble les exigences indépendantes de la technique par exemple l'usage prévu des données anonymes (publication, test, classification, etc.), le seuil de risque de ré-identification toléré, le taux de suppression à ne pas dépasser ainsi que la qualité minimale exigée. Cette dernière peut être exprimée par l'importance relative accordée aux critères de qualité que doivent vérifier les données anonymes. La deuxième catégorie regroupe des exigences dépendantes de la technique choisie (ici la généralisation) et influent sur le choix d'un algorithme implémentant cette technique. Dans le cas de la généralisation, le type de généralisation souhaité peut constituer une exigence spécifique. A titre d'exemple, une anonymisation peut être demandée pour un besoin de classification des données, tout en exigeant de ne pas accepter un taux de suppression de plus de 5 % (qui réduit l'échantillon et éventuellement le déforme) ni un résultat qui engendre un risque de ré-identification de plus de 10 %. Le demandeur peut aussi préciser qu'il accorde plus d'importance à la sécurité qu'à la complétude des données anonymes (dans ce cas, il exprime une préférence pour la suppression qui permet d'effacer des « outliers », présentant un risque élevé de ré-identification). Quand bien même on dispose de ces informations, elles ne suffisent pas pour sélectionner des algorithmes adéquats. En effet, comme on a pu le montrer dans (BenFredj *et al*., 2014), le choix des algorithmes repose sur des méta-données qui, si elles ne peuvent pas être déduites automatiquement, doivent être fournies avec les données. Par exemple, la qualification des attributs en identifiant, quasi identifiant, sensible ou non sensible, catégoriel ou continu peut être automatisée. Certaines méta-données, par exemple, la liste des attributs formant le quasi-identifiant, sont nécessaires quelle que soit la technique. D'autres, comme la distribution des données (dense ou non) sont spécifiques à une technique. En résumé, dans un souci de généricité, le contexte d'une anonymisation sollicitée par un utilisateur pour ses données est construit en deux temps (figure 5). Dans un premier temps, MAGGO récupère de l'ontologie les types d'exigences utilisateur à renseigner ainsi que le type de méta-données à connaître pour le type d'anonymisation sollicitée.

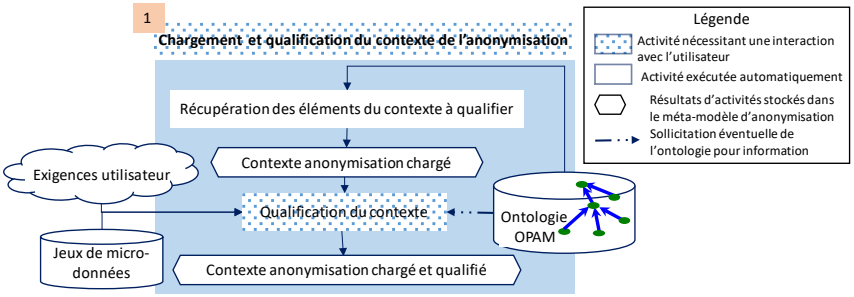

Figure 5. Chargement et qualification du contexte de l'anonymisation

Tableau 2. Paramètres de contexte de la généralisation de micro-données

Paramètres fournis par l'utilisateur
Seuil de risque toléré
Taux de suppression autorisé
Besoin d'usage
Jeu de données original
Propriétés de qualité attendues
Importance relative des propriétés de qualité

Paramètres pouvant être déduits automatiquement
Attributs identifiants
Attributs quasi-identifiants (QI)
Attributs sensibles
Nature des attributs (catégoriel ou continu)
Type de généralisation attendu
Distribution des données
Taille du jeu de données
k, taille minimale des classes d'équivalence de QI
MaxSup, nb maximum de tuples pouvant être supprimés

A titre d'exemple, dans le cas d'une anonymisation par généralisation, notre approche MAGGO, après interrogation de l'ontologie OPAM, construira le contexte d'anonymisation par généralisation. Ce contexte est constitué des paramètres de contexte décrits dans le tableau 2. Ces paramètres de contexte, intégrés dans le méta-modèle d'anonymisation, seront renseignés, dans la seconde phase de l'étape 1. La plupart des paramètres sont déductibles de l'analyse des jeux de données. Deux paramètres spécifiques à la généralisation, MaxSup et k, sont calculés. MaxSup définit le nombre maximum de lignes qui pourront être supprimées pendant l'anonymisation. L'attribut k fait référence au k-anonymat (Fung *et al.*, 2010), modèle de protection de la vie privée ciblé par la technique de généralisation. Il correspond à la taille minimale des classes d'équivalence de quasi-identifiants anonymes pouvant être générés par généralisation. Par exemple, si le sexe et le code postal forment un quasi-identifiant et que k vaut 10, le jeu de données anonymisées ne pourra pas comprendre moins de 10 lignes pour le même sexe et le même code postal. Si nécessaire, soit le code postal sera généralisé au numéro du département soit les lignes correspondantes seront supprimées.

Ainsi, dans MAGGO, MaxSup est calculé à partir de la taille du jeu de données et du taux de suppression autorisé par l'utilisateur en appliquant la formule suivante : *MaxSup = Taille du jeu de micro-données * taux de suppression autorisé*

Pour calculer k, nous utilisons la formule suivante de l'outil PARAT :

k = 100/ taux de risque de ré-identification

Cette formule exprime le fait que le taux de risque de ré-identification est inversement proportionnel à k. En d'autres termes, plus k est petit, plus le risque de ré-identification est grand.

Une fois le contexte d'anonymisation renseigné, MAGGO suggère à l'utilisateur, dans sa seconde étape, sous forme de signatures, un ensemble potentiel d'algorithmes paramétrés susceptibles de satisfaire ses exigences.

5.2. Etape 2 - Suggestion de signatures d'algorithmes candidats

Le jeu de données brouillé, renvoyé par application d'une technique d'anonymisation, dépend fortement de la signature de l'algorithme exécuté sur le jeu de données original. La construction, l'évaluation et la proposition à l'utilisateur, de signatures d'algorithmes se rapprochant le plus de ses exigences de qualité, est l'objet de l'étape 2 (figure 6). La première phase de cette étape consiste à construire des signatures pertinentes. Dans un premier temps, on extrait les algorithmes applicables au contexte de l'anonymisation et on les dote de valeurs de paramètres conformes aux contraintes spécifiées dans le contexte. La seconde phase a pour objectif de proposer à l'utilisateur, parmi les signatures pertinentes, celles offrant le meilleur score en termes de concordance avec les exigences de qualité. Les paragraphes qui suivent détaillent chacune de ces phases.

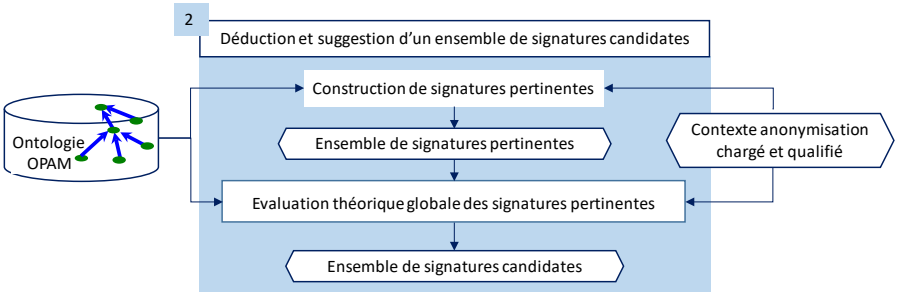

Figure 6. Déduction et suggestion de signatures candidates

5.2.1. Construction des signatures pertinentes

Le ciblage des algorithmes applicables au contexte exploite certains paramètres de ce contexte. A titre d'exemple, pour une anonymisation par généralisation, si l'utilisateur n'a pas d'exigence sur le type de généralisation à obtenir alors, de ce point de vue, tous les algorithmes de généralisation sont candidats au paramétrage.

En revanche, si son souhait est d'obtenir des généralisations multidimensionnelles[11], alors cet ensemble se restreint aux algorithmes fournissant ce type de généralisation tels que le « Median Mondrian ».

Pour effectuer ce filtrage d'algorithmes, l'ontologie OPAM est exploitée car elle dispose des connaissances permettant de confronter les exigences des algorithmes aux exigences de l'anonymisation. Ces connaissances sont représentées dans le schéma d'OPAM représenté à la figure 4.

Les algorithmes sélectionnés permettent bien sûr d'instancier le méta-modèle de l'anonymisation (les classes en gris du méta-modèle de la figure 3). Cette instanciation contient aussi, pour chaque algorithme sélectionné, l'ensemble des combinaisons possibles de valeurs de paramètres pouvant lui être affectées. Chaque algorithme sélectionné, couplé avec chaque combinaison de valeurs de paramètres possible, constitue une signature pertinente.

Il s'agit d'octroyer au paramètre de l'algorithme, la valeur de contexte générée suite à la prise en compte de la contrainte d'anonymisation imposée par l'utilisateur. A titre d'exemple, dans le cas d'une anonymisation par généralisation, l'utilisateur exprime un taux de risque de ré-identification et un taux de suppression tolérés (paramètres saisis). Ces contraintes génèrent dans le contexte de l'anonymisation une valeur pour k et MaxSup (paramètres générés). Ces deux valeurs, combinées avec chaque algorithme retenu, constituent autant de signatures candidates.

5.2.2. Evaluation théorique des signatures pertinentes

Cette phase vise à fournir à l'utilisateur les signatures se rapprochant le plus de ses exigences de qualité et de sécurité. C'est un processus de décision multicritère pour lequel nous appliquons la méthode AHP. Cette dernière, sur la base de comparaisons par paires, détermine le score global de chacune des signatures afin de retenir les mieux classées. On peut ainsi décider de fournir à l'utilisateur les trois signatures pertinentes ayant le score le plus élevé.

La hiérarchie fournie à AHP a pour premier niveau l'objectif de cette étape. Le niveau intermédiaire correspond à la hiérarchie des exigences emmagasinée dans OPAM. Son dernier niveau, c'est-à-dire les feuilles de l'arbre, regroupe les signatures pertinentes à évaluer. La figure 7 contient, à titre d'exemple, la hiérarchie construite par cette phase pour une anonymisation à des fins de classification.

Une fois la hiérarchie construite, les jugements sur l'importance relative des éléments de cette hiérarchie sont déterminés. Les jugements entre les éléments du niveau intermédiaire de la hiérarchie (les critères) sont ceux émis par l'utilisateur et spécifiés dans le contexte de l'anonymisation. Les jugements sur l'importance

11. Une généralisation multidimensionnelle est telle que, dans la table résultat, les données ne sont pas nécessairement au même niveau de généralité. Ainsi, on peut imaginer qu'une tranche d'âge pourra être plus ou moins large selon les individus.

relative des signatures sont, quant à eux, déterminés de façon automatique après une évaluation de chaque signature selon un critère donné. Cette évaluation approximative, que l'on nomme « évaluation théorique locale », est déduite des expérimentations faites par les experts en anonymisation et qui sont emmagasinées dans OPAM (classes sur fond blanc dans le bas droit de la figure 4). L'importance relative de chaque signature est aussi déterminée automatiquement. Elle est fondée sur leur évaluation locale et sur une échelle de comparaison disponible dans MAGGO.

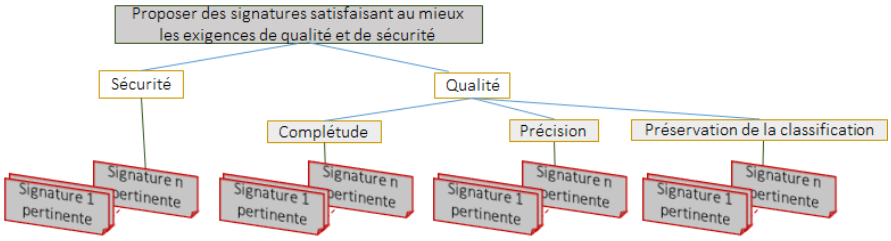

Figure 7. Hiérarchie multicritère pour l'anonymisation

Les paragraphes qui suivent décrivent respectivement les processus d'évaluation locale et globale (le score) d'une signature.

5.2.2.1. Evaluation théorique locale des signatures pertinentes

Plusieurs évaluations théoriques d'algorithmes d'anonymisation de micro-données sont disponibles dans la littérature. Chacune fournit la qualité d'un jeu de données anonyme vis-à-vis d'un critère (sécurité, précision, complétude, etc.) compte tenu d'une signature d'algorithme et des caractéristiques spécifiques du jeu de données originales. Le critère en question est mesuré à l'aide d'une métrique. Dans le cas où il n'y a pas d'évaluation théorique, pour la signature et les caractéristiques du jeu de données spécifiées dans le contexte d'anonymisation, MAGGO applique une technique d'apprentissage supervisée afin de prédire la qualité de cette signature vis-à-vis d'un critère. La régression se prête bien à notre problématique en raison du type des variables explicatives et de la variable cible. Le modèle retenu est l'arbre de régression en raison de la petite taille de la base d'expérimentations disponibles (Loh, 2011). La variable à expliquer est le critère de qualité à mesurer. Les variables explicatives sont les différents éléments de contexte influençant la variable cible. Le jeu de données d'entrainement est extrait de l'ontologie OPAM. Ainsi, à titre d'exemple, pour une anonymisation par généralisation à des fins de classification, il nous faut quatre jeux de données : un par critère constituant une feuille du niveau intermédiaire de la hiérarchie AHP (sécurité, complétude, précision, préservation de la classification) décrite à la figure 7. Tous les jeux de données contiennent les mêmes informations : une valeur pour k, une valeur pour le nombre d'attributs constituant le QI et une valeur pour caractériser la distribution du jeu de données original. En revanche, ces jeux

d'exemples se distinguent par la sortie qui correspond à la mesure du critère cible. Après évaluation de chaque signature, le méta-modèle est enrichi de ces nouvelles estimations.

5.2.2.2. Mesure de l'importance relative des signatures

Une fois les évaluations locales des différentes signatures effectuées, MAGGO procède à des comparaisons par paires de signatures afin de déduire l'importance relative des signatures vis-à-vis de chaque critère. Pour ce faire, nous nous sommes inspirés de l'échelle sémantique de (Saaty et Sodenkamp, 2008) afin de permettre une comparaison automatique des signatures, qui résulte en une matrice de valeurs à livrer à AHP. Si l'on considère deux couples E(Ci,Sj) et E(Ci,Sj') où E(Ci, Sj) (resp. E(Ci,Sj')) représente l'évaluation locale de la signature Sj (resp. Sj') pour le critère Ci, nous construisons la table d'échelle sémantique d'AHP comme suit (tableau 3). Cette table sert pour la comparaison deux par deux des signatures. Pour chaque critère de qualité, on a $\varepsilon 1 < \varepsilon 2 < \varepsilon 3 < \varepsilon 4 < \varepsilon 5$.

Tableau 3. Comparaison des signatures par une échelle sémantique

Intensité	Signification	Interprétation formelle de la signification
(Sj, Sj', 1)	Sj et Sj' sont d'égale qualité vis-à-vis du critère Ci	$E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_1$
(Sj, Sj', 2)	Sj est d'une qualité légèrement meilleure que celle de Sj' vis-à-vis du critère Ci	$\varepsilon_1 < E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_2$
(Sj, Sj', 3)	Sj est d'une qualité meilleure que celle de Sj' vis-à-vis du critère Ci	$\varepsilon_2 < E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_3$
(Sj, Sj', 4)	Sj est d'une qualité nettement meilleure que celle de Sj' vis-à-vis du critère Ci	$\varepsilon_3 < E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_4$
(Sj, Sj', 5)	Sj est d'une qualité très nettement meilleure que celle de Sj' vis-à-vis du critère Ci	$\varepsilon_4 < E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_5$

5.3. Etapes 3,4 et 5 de MAGGO

Une fois la comparaison par paires effectuée, AHP calcule le score global de chaque signature pertinente ; ce qui permet de classer ces signatures et de proposer à l'utilisateur, dans l'étape 3 de MAGGO, les signatures qui ont le meilleur score. L'utilisateur a la possibilité de choisir une ou plusieurs signatures à faire exécuter sur son jeu de micro-données. L'exécution de ces signatures fait l'objet de l'étape 4 de MAGGO. Dans cette étape, un jeu de données anonyme est livré pour chaque signature pertinente de score le plus élevé, choisie par l'utilisateur. Pour guider l'utilisateur dans son choix de jeux de données anonymes, différentes évaluations cette fois-ci réelles, c'est-à-dire sur le jeu de données fourni, sont effectuées. Chaque évaluation permet à l'utilisateur de positionner le jeu anonyme vis-à-vis d'une exigence de qualité attendue.

6. Exemple d'illustration

Pour illustrer notre approche, on suppose que le contexte est caractérisé comme suit. Le risque maximum toléré est 10 %. De même, on admet que l'on ne peut supprimer plus de 20 % des données. De plus, la table à anonymiser est de grande taille (1 000 tuples). La distribution des données est dense. Le quasi-identifiant comprend trois attributs. L'usage prévu des données anonymisées est la classification. L'utilisateur accorde autant d'importance à l'utilité des données qu'au respect de la vie privée.

Etape 1 – chargement et qualification du contexte de l'anonymisation

Au cours de cette première étape, l'éditeur de données doit entrer son contexte. Certains éléments (taille, distribution, nombre d'attributs du QI) peuvent être calculés automatiquement après chargement de la table.

Etape 2 – Sélection d'algorithmes et signatures pertinentes

Les paramètres k et MaxSup peuvent être calculés en fonction du taux de risque et du taux de suppression. Ici k vaut donc 1/0,1 = 10 et MaxSup = 1000*20 % = 200. Plus précisément 10 est la valeur minimale de k et 200 la valeur maximale de MaxSup. On peut aussi tester des signatures où k = 12 et MaxSup = 150 par exemple.

L'algorithme de Samarati (2001) ne peut pas être appliqué à une table de cette taille, car il est de complexité exponentielle (Ciriani *et al*., 2007). Cette information fait partie des connaissances contenues dans l'ontologie. Supposons donc que seuls les algorithmes Datafly, Median Mondrian et TDS remplissent les contraintes.

Les deux phases précédentes de l'étape 2 ont généré deux valeurs de k (10 et 12), deux valeurs de MaxSup (200 et 150) et trois algorithmes (Datafly, Median Mondrian et TDS). Seul Datafly effectue des suppressions. On va donc tester Datafly avec les quatre combinaisons possibles pour k et MaxSup. Pour Median Mondrian et TDS, on va tester les deux valeurs de k. Par conséquent, les signatures générées sont récapitulées dans les quatre premières colonnes du tableau 4. Elles sont évaluées selon les critères feuilles de la hiérarchie des buts (figure 7). Les évaluations liées aux deux critères 'sécurité' et 'complétude' ont été déduites à partir respectivement des valeurs de k et MaxSup dans la mesure où l'on a :

$$\text{Sécurité} + 1/k = 1 \text{ et } \text{Complétude} + \text{MaxSup}/\text{Taille} = 1$$

Les critères « Précision » (métrique de discernabilité DM (Fung *et al.*, 2010)) et « Préservation de la classification » ont été déduites en appliquant une technique de régression sur les données expérimentales issues d'OPAM (tableau 4).

Le passage des évaluations individuelles des signatures à des comparaisons deux à deux est nécessaire afin de pouvoir appliquer la méthode AHP. Par exemple, pour le critère classification, les signatures 5 et 8 sont évaluées respectivement à 0,65 et 0,71, ce qui représente une différence de 6 %. On suppose que l'échelle utilisée

induit ainsi une intensité de 3. Les huit signatures sont ainsi comparées deux à deux pour chacun des critères. Après application d'AHP (grâce à l'outil Weka JAHP), on agrège les quatre critères pour chaque signature. On aboutit à un score final fourni en dernière colonne du tableau 4. Ce score permet à l'utilisateur de choisir d'exécuter les signatures qui donnent le meilleur compromis entre les quatre critères, compromis qui résulte de l'application d'AHP à chaque paire de signatures, par exemple les quatre dernières.

Tableau 4. Evaluation des signatures

Signature	Algorithme	k	Maxsup	Sécurité	Complétude	Précision métrique DM	Usage Classification	Score final
Sig 1	Datafly	10	150	0,9	0,85	50000	0,54	0,1
Sig 2	Datafly	10	150	0,9	0,85	50000	0,54	0,05
Sig 3	Datafly	12	200	0,92	0,8	60000	0,61	0,04
Sig 4	Datafly	12	200	0,92	0,8	60000	0,61	0,05
Sig 5	Mondrian	10	0	0,9	1	15000	0,65	0,27
Sig 6	Mondrian	12	0	0,92	1	20000	0,63	0,18
Sig 7	TDS	10	0	0,9	1	35000	0,79	0,19
Sig 8	TDS	12	0	0,92	1	40000	0,71	0,12

L'exemple ci-dessus a permis d'illustrer le fonctionnement de notre approche. La valeur ajoutée de cette approche et notamment du guidage fourni à l'éditeur de données permettent d'améliorer l'efficacité et l'efficience de l'anonymisation puisque l'éditeur obtient plus rapidement un jeu de données publiables. Elle permet aussi l'apprentissage de l'anonymisation par l'éditeur. Ces trois critères ainsi que la satisfaction de l'éditeur sont ceux qui feront l'objet d'une mesure lors de l'expérimentation de l'approche, en combinant l'évaluation du temps passé par l'éditeur, de la qualité de l'anonymisation obtenue, de son degré de satisfaction et de sa compréhension de l'approche. Les premières expérimentations ont permis de montrer une amélioration de toutes ces mesures, grâce à l'approche, mais le nombre de tests n'est pas suffisant pour assurer la validité des conclusions.

7. Conclusion

Les éditeurs de données sont confrontés à deux difficultés majeures lors d'un processus d'anonymisation. La première concerne le choix de l'algorithme adéquat au contexte. La seconde est le paramétrage à opérer pour que l'algorithme génère des données sécurisées (difficiles à ré-identifier) mais encore utiles (dont la qualité reste conforme avec l'objectif). Notre approche MAGGO automatise ces deux tâches en utilisant une ontologie. Cette dernière peut aussi être consultée par l'éditeur de données afin de recueillir les connaissances nécessaires lui permettant

de décrire son contexte et de répondre de façon adéquate aux questions qui lui sont posées lors du déroulement du processus. La sécurisation des données par anonymisation, d'une part, et le maintien de la précision et de la complétude des données, d'autre part, sont des objectifs contradictoires. C'est pourquoi, le processus d'anonymisation vise un compromis entre ces deux objectifs, en fonction de l'usage des données. Notre approche est, pour le moment, limitée aux algorithmes fondés sur la technique de généralisation. Toutefois, nous nous sommes efforcées de la rendre la plus générique possible afin qu'elle puisse être appliquée à d'autres techniques d'anonymisation de micro-données. Enfin, pour rendre l'approche évolutive et son implémentation incrémentale, nous avons utilisé une conception dirigée par les modèles.

En termes de recherche future, nous envisageons trois axes : 1) la mise au point d'un outil support de l'approche, 2) la conduite d'une expérimentation à plus grande échelle incluant des utilisateurs pour mesurer l'utilité et l'utilisabilité de la méthode et de l'outil, 3) l'extension à d'autres techniques pour pouvoir choisir à la fois une technique d'anonymisation, un algorithme et une signature.

Bibliographie

Agrawal H., Cochinwala M., Horgan J.R. (2014). Automated Determination of Quasi-Identifiers Using Program Analysis, U.S. Patent N° 8661423B2, Date: Feb. 25.

Aïmeur E. (2009). Data Mining and Privacy. In *Encyclopedia of Data Warehousing and Mining*, Second Edition, p. 388-393. IGI Global.

Akoka J., Comyn-Wattiau I., Du Mouza C., Fadili H., Lammari N., Metais E., Cherfi S. S. S. (2014). A semantic approach for semi-automatic detection of sensitive data. *Information Resources Management Journal (IRMJ)*, vol. 27, n° 4, p. 23-44.

Amita S., Ranjan Baghel, Puneeta Panday, Praveen Saini (2014). A Survey on Techniques for Privacy Preserving Data Publishing (PPDP). *MIT International Journal of Computer Science and Information Technology*, vol. 4, n° 2, August, p. 60-64.

Ayala-Rivera V., McDonagh P., Cerqueus T., et Murphy L. (2014). A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners. *Trans. Data Privacy*, vol. 7, n° 3, p. 337-370.

Bayardo R. J. et Agrawal R. (2005). Data privacy through optimal k-anonymization. *ICDE*. p. 217-228.

BenFredj F., Lammari N., Comyn-Wattiau I (2014). Characterizing Generalization Algorithms-First Guidelines for Data Publishers. *International Conference on Knowledge Management and Information Sharing*, Rome, Italy.

BenFredj F., Lammari N., Comyn-Wattiau I. (2015) Building an Ontology to Capitalize and Share Knowledge on Anonymization Techniques. *European Conference on Knowledge Management, p. 122-131*. Kidmore End: Academic Conferences International Limited.

BenFredj F., Lammari N., Comyn-Wattiau I. (2016). L'anonymisation des données par généralisation - un arbre de décision. *Ingénierie et management des systèmes d'information*, Cepadues, ISBN 978.2.36493.573.0, p. 159-171.

BenFredj F., Lammari N., Comyn-Wattiau I. (2017). Approche guidée pour l'anonymisation de bases de données, *Actes de la conférence INFORSID*, Toulouse.

BenFredj F. (2017). Méthode et outil de brouillage des données sensibles. Thèse de doctorat, CNAM, Paris, juillet.

Breiman L., Friedman J., Stone C. J. et Olshen R.A. (1984). Classification and Regression Trees. Wadsworth Statistics/Probability.

Brand R, (2002). Microdata protection through noise addition. *Inference Control in Statistical Databases*, Domingo-Ferrer J. (ed.), LNCS vol. 2316, p. 97-116, Springer.

Burton R., Hundepool A. J., Willenborg L. CRJ, Nitz L. H., Kim K. E. (1997). Record Linkage. In Record Linkage Techniques-1997, Proceedings of an International Workshop and Exposition, March 20-21, Arlington, Va, 139. National Academies.

Ciriani V., De Capitani di Vimercati S., Foresti S., Samarati P. (2007). Microdata Protection. *Secure Data Management in Decentralized Systems 2007*, Advances in Information Security, p. 291-321, Springer.

Dai C., Ghinita G., Bertino E., Byun J., Li N. (2009). TIAMAT: a Tool for Interactive Analysis of Microdata Anonymization Techniques. *PVLDB*, vol. 2, n° 2, p. 1618-1621.

Dalenius T. (1977). *Towards a methodology for statistical disclosure control*. Statistisk Tidskrift.

Defays D., Nanopoulos P. (1993) Panels of enterprises and confidentiality: the small aggregates method, Paper read at the 92nd *Symposium on Design and Analysis of Longitudinal Surveys*, Ontorio, Canada, November.

Fienberg S.E, McIntyre J. (2004). Data swapping: Variations on a theme by dalenius and reiss. In *International Workshop on Privacy in Statistical Databases*, p. 14-29. Springer

Fung B., Wang K., Yu P. S. (2005). Top-down specialization for information and privacy preservation. *ICDE'05*, p. 205-216.

Fung, B. C. M., Ke Wang, Chen R., et Yu. P. S. (2010). Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys*, vol. 42, n° 4, p. 1-53.

Hand D.J., 1992. Microdata, macrodata, and metadata. *Computational Statistics*, In Dodge Y., Wittaker J. (Eds), Physica Verlag, Heidelberg, p. 325-340.

Hussien A. A., Hamza N., Hefny H. A. (2013). Attacks on Anonymization-Based Privacy-Preserving: A Survey for Data Mining and Data Publishing. *Journal of Information Security*, 4, p. 101-112

Ilavarasi B., Sathiyabhama A. K., Poorani S. (2013). A survey on privacy preserving data mining techniques. *Int. Journal of Computer Science and Business Informatics*, vol. 7, n° 1

Iyengar V. S. (2002). Transforming data to satisfy privacy constraints. *ACM SIGKDD'02*, p. 279-288.

Kiran P. et Kavya N. P. (2012). A Survey on Methods, Attacks and Metric for Privacy Preserving Data Publishing. *International Journal of Computer Applications*, vol. 53, n° 18.

LeFevre K., DeWitt D. J., Ramakrishnan R. (2005). Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD*, p. 49-60.

LeFevre K., DeWitt D.J., Ramakrishnan R. (2006). Mondrian multidimensional k-anonymity. *ICDE'06*. p. 25-25.

LeFevre K., DeWitt D. J., et Ramakrishnan R. (2008). Workload-Aware Anonymization Techniques for Large-Scale Datasets. *ACM Transactions on Database Systems*, vol. 33, n° 3, p. 1-47.

Loh W.-Y. (2011). Classification and regression trees. Wiley Interdisc. *Rew.: Data Mining and Knowledge Discovery*, vol. 1, n° 1, p. 14-23.

Motwani R., Xu Y. (2007). Efficient Algorithms for Masking and Finding Quasi-Identifiers, In : *Proceedings of the Conference on Very Large Data Bases (VLDB)*, p. 83-93.

Patel L., Gupta R. (2013) A Survey of Perturbation Technique for Privacy-Preserving of Data. *Int. Journal of Emerging Technology and Advanced Engineering*, vol. 3, n° 6.

Poulis G., Gkoulalas-Divanis A., Loukides G., Skiadopoulos S., Tryfonopoulos C. (2014). SECRETA: A System for Evaluating and Comparing RElational and Transaction Anonymization algorithms. *EDBT'14.*

Saaty T.L, Sodenkamp M.A. (2008). Making decisions in hierarchic and network systems. *IJADS*, vol. 1, n° 1. p. 24-79

Samarati P., Sweeney L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International. http://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf.

Samarati P. (2001). Protecting respondents' identities in microdata release, *IEEE Transactions on Knowledge and Data Engineering*, vol. 13, n° 6, p. 1010-1027.

Silver M. S. (2006). Decisional Guidance. Broadening the Scope. Human-Computer *Interaction in Management Information Systems*, Galleta D. et Zhang P. (Eds.). *International handbooks on information systems* vol. 6, p. 90-119. Armonk, NY: M.E. Sharp.

Sweeney L. (1997). Datafly: a System for Providing Anonymity in Medical Data. *Eleventh International Conference on Database Securty XI: Status and Prospects*, p. 356-381.

Sweeney L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, n° 05, p. 571-588.

Vaghashia H., Amit G. (2015). A survey: privacy preservation techniques in data mining. *International Journal of Computer Applications*, vol. 119, n° 4.

Vassilios V.S., Bertino E., Nai Fovino I., Parasiliti Provenza L., Saygin Y., et Theodoridis Y. (2004). State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, vol. 33, n° 1, p. 50-57.

Wang K., Yu P.S., Chakraborty S. (2004). Bottom-up generalization: A data mining solution to privacy protection. *In ICDM'04*, p. 249-256.

Xiao X., Wang G., Gehrke G. (2009). Interactive Anonymization of Sensitive Data. *SIGMOD'09*, June 29-July 2, Providence, Rhode Island, USA, p. 1051-1054.