Approche guidée pour l'anonymisation de bases de données

Feten BenFredj[1], Nadira Lammari[1], Isabelle Comyn-Wattiau[2]

1. CEDRIC-CNAM, 2 Rue Conté, 75003 Paris, France
2. ESSEC Business School, 1 Av. Bernard Hirsch, 95021 Cergy, France
fetenbf@yahoo.fr, ilham-nadira.lammari@cnam.fr, wattiau@essec.edu

RESUME. L'anonymisation des données personnelles requiert l'utilisation d'algorithmes complexes 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 le propriétaire des données dans son processus d'anonymisation. Le guidage peut être informatif ou suggestif. Il permet de choisir l'algorithme le plus pertinent en fonction des caractéristiques des données mais aussi de l'usage ultérieur des données anonymisées. Le guidage a aussi pour but de définir les bons paramètres à appliquer à l'algorithme retenu. Dans cet article, nous nous focalisons sur les algorithmes de généralisation de micro-données. Les connaissances liées à l'anonymisation tant théoriques qu'expérimentales sont stockées dans une ontologie.

ABSTRACT. Personal data anonymization requires complex algorithms 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 may be informative or suggestive. It helps the data owner in choosing the most relevant algorithm given the data characteristics and the future usage of anonymized data. The guidance process also helps in defining the best input values for the algorithms. In this paper, we focus on generalization algorithms for micro-data. The knowledge about anonymization is composed of both theoretical aspects and experimental results. It is managed thanks to an ontology.

MOTS-CLES : 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.

1. Introduction

Le partage des données au-delà des frontières même de l'organisation s'est accentué, par exemple, par l'engagement des pays sur la voie de l'ouverture des données publiques, plus connue sous le nom d' «open data». Cette situation soulève 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 éditeurs de données sont toujours à la recherche d'une solution qui réponde au mieux à la confidentialité et à l'utilité de leurs données. Leur solution émerge après des prises de décision à différentes phases du déroulement de leur tâche. En effet, ils sont amenés entre autres à sélectionner un algorithme d'anonymisation, à opter pour un paramétrage adéquat de cet algorithme et à juger de la qualité du rendu après application du procédé ainsi paramétré. Ils sont donc engagés dans un processus de décision qui s'appuie sur leur connaissance du domaine.

Les outils existants, de par leur opacité et leur manque de guidage dans le choix et le paramétrage des algorithmes, ne simplifient pas cette activité pour un professionnel ayant une faible expertise dans le domaine. D'un point de vue académique, nous avons aussi constaté l'absence d'approches guidées pour l'anonymisation bien que la littérature abonde d'articles de recherche sur les algorithmes d'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. 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, que nous proposons dans cet article et que nous nommons MAGGO (Méthodologie pour une Anonymisation par Généralisation Guidée par une ontologie), sert de guide pour un professionnel dans sa prise de décision lors d'une anonymisation par généralisation de micro-données. Cela n'enlève rien à la généricité de notre approche. En effet, elle peut être instanciée pour une autre technique.

Après un rapide état de l'art (section 2), nous décrivons l'approche générale (section 3), puis ses étapes détaillées (section 4). En section 5, nous illustrons l'approche sur un exemple. Enfin, nous concluons en section 6 et présentons quelques axes de recherche future.

2 Etat de l'art

Plusieurs techniques d'anonymisation existent avec des degrés de fiabilité et des contextes d'applicabilité variables. Ces contextes sont caractérisés, entre autres, par l'usage souhaité des données (par exemple pour tester un logiciel ou encore pour publier les données à des fins d'analyse) et par le type de données à anonymiser (micro ou macro données tabulaires, données spatio-temporelles, graphes, images,

[1] Données atomiques décrivant des individus (Hand, 1992)

textes, etc.). Le degré de fiabilité est en lien direct avec le risque de ré-identification des données anonymes. Cependant, 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 données de la deuxième catégorie sont dénaturées, c'est-à-dire inexactes, ce qui n'empêche pas leur usage à des fins de test ou de statistique par exemple. La technique de 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. La technique de généralisation consiste à remplacer des valeurs par des valeurs plus générales (Samarati, 2001) : les données sont vraies, mais moins précises. La généralisation est appliquée à un ensemble d'attributs formant un quasi-identifiant (QI). Elle nécessite la définition d'une hiérarchie pour chaque attribut composant le QI. L'âge peut être généralisé à l'aide d'intervalles de valeurs de plus en plus grands vers la racine de la hiérarchie. Généraliser consiste à remplacer une valeur par son ancêtre direct dans la hiérarchie de généralisation, 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. 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 variable aléatoire. Chacune de ces techniques a donné lieu à un ou plusieurs algorithmes. Par exemple, la généralisation peut s'appliquer de différentes façons et des dizaines d'algorithmes ont été proposés dans cette catégorie.

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). Certaines sont certes orientées usage mais restent non accessibles à des éditeurs de données avec de faibles compétence dans le domaine. De plus, les algorithmes associés aux techniques 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[2] (Poulis *et al*., 2014 ; Xiao *et al*., 2009 ; Dai *et al.*, 2009). Le plus souvent, ils sont opaques. Même s'ils proposent plusieurs techniques, ils mettent en œuvre, en général, un seul algorithme par technique sans mentionner lequel. La plupart de ces outils ne fournissent pas de guidage dans le choix de la technique et de l'algorithme. Ils n'offrent pas d'aide au paramétrage des algorithmes proposés. Le guidage est réduit à l'application de métriques sur les données anonymisées qui permettent à l'éditeur de données d'évaluer notamment le risque résiduel et la dégradation due à l'anonymisation.

L'état de l'art comprend aussi de nombreuses métriques permettant d'évaluer la qualité des données anonymisées, en termes de perte d'information ou de précision, ou le risque de ré-identification (Fung *et al*., 2010).

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 et dans son paramétrage.

Dans la suite, nous présentons l'approche MAGGO en détaillant ses étapes principales.

3. 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 a) repérer les données identifiantes[3], quasi-identifiantes (QI)[4] et sensibles[5], b) proposer des techniques appropriées avec une orchestration adéquate. Il lui faut aussi, pour chaque technique, identifier

[2] PARAT est un exemple en ligne (http://www.privacyanalytics.ca/software/parat/).

[3] 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.

[4] 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

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

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 fera en évaluant plusieurs algorithmes avec plusieurs combinaisons possibles de paramètres.

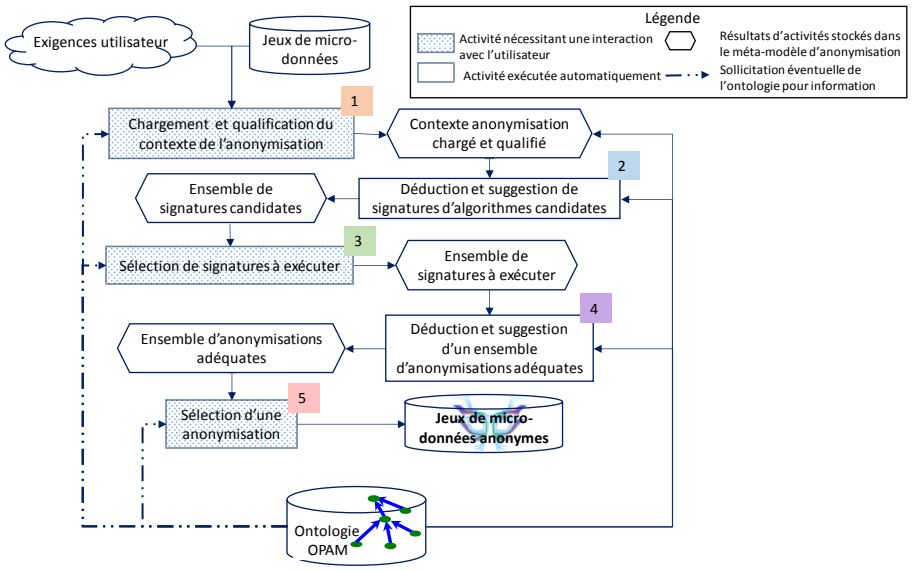

Figure 1. Les étapes de MAGGO

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. Ainsi, à chacune des étapes, MAGGO fait intervenir des connaissances expertes en vue d'un guidage suggestif ou informatif. Nous reprenons ces concepts de (Silver, 2006) : le

guidage suggestif guide l'utilisateur dans ses choix alors que le guidage informatif lui fournit des informations qui peuvent éclairer son choix. Dans notre cadre, le guidage suggestif aide l'éditeur de données dans la sélection de l'algorithme approprié tandis que le guidage informatif lui fournit des informations pour éclairer son choix sur un algorithme ou sur une technique.

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 (Fig. 2) réunit les abstractions conceptuelles des artefacts cibles et sources de notre approche.

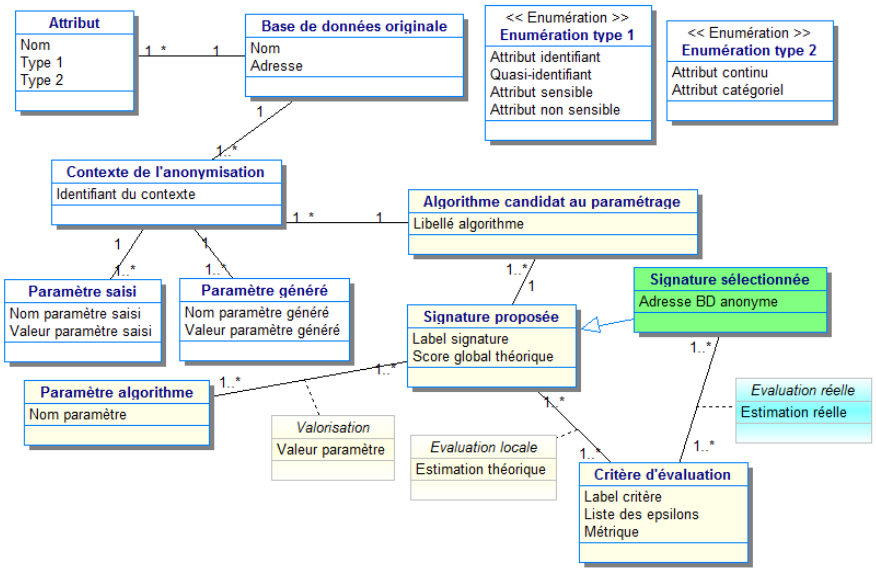

Figure 2. Le méta-modèle du processus d'anonymisation

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[6] est défini à partir de paramètres saisis (par l'utilisateur) ou générés (calculés automatiquement ou déduits des 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 des évaluations comparables contenues dans l'ontologie, ou réelle c'est-à-dire déduite d'une exécution de l'algorithme sur le jeu de données. Chaque couleur dans la figure correspond à l'étape de MAGGO au cours de laquelle les éléments correspondants sont sollicités.

Ainsi, l'exécution de la première étape de MAGGO permet d'instancier notre méta-modèle à l'aide des données relatives au cahier des charges. Cette instanciation correspond à une description du contexte de l'anonymisation ainsi qu'à sa qualification. Un enrichissement du modèle, par des données complémentaires issues de chacune des différentes étapes, est effectué.

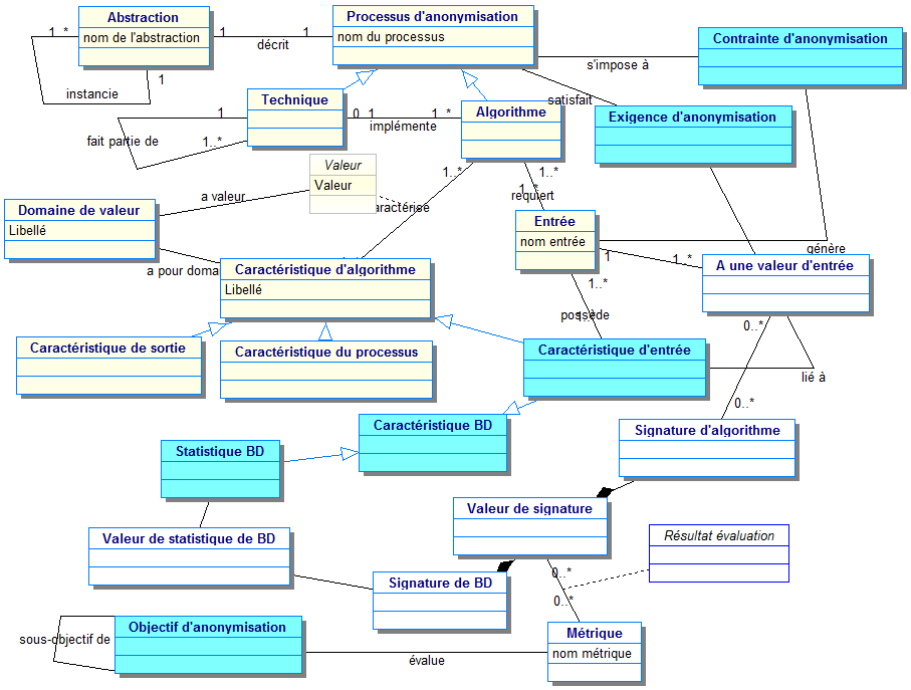

Figure 3. Méta-modèle sous-jacent de l'ontologie OPAM (extrait)

De plus, MAGGO exploite pour l'enrichissement du méta-modèle deux techniques statistiques : la technique d'aide à la décision multicritère 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

[6] 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.

régression est utilisée sous la forme d'apprentissage supervisé afin de prédire la valeur d'un critère d'utilité ou d'un critère de sécurité, compte tenu de données expérimentales disponibles dans OPAM.

Enfin, l'approche s'appuie sur l'ontologie OPAM (BenFredj *et al.*, 2015). Enfin, avant de présenter les différentes étapes de MAGGO, pour faciliter la compréhension de l'approche, nous rappelons à la figure 3 les éléments principaux du méta-modèle d'OPAM. Les classes sur fond jaune sont celles qui permettent la représentation de la connaissance « théorique » relative aux techniques et algorithmes d'anonymisation. Les classes sur fond bleu permettent de décrire les concepts que nous avons définis pour décrire le processus d'anonymisation. Enfin, les classes sur fond blanc sont le support de la représentation de la connaissance empirique que les expérimentations décrites dans les articles de recherche ou accumulées au cours de nos tests des outils ont permis de constituer.

4. Description des étapes de la méthode MAGGO

Dans cette section, nous décrivons chacune des cinq étapes de la démarche schématisée à la figure 1.

4.1. 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 l'ensemble des exigences que doit satisfaire l'anonymisation. A ce titre, on peut considérer deux catégories d'exigences pour l'anonymisation de micro-données. 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 par généralisation pourrait 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 pourrait aussi préciser qu'il accorderait 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 constater dans notre état de l'art sur l'anonymisation par généralisation (BenFredj *et al.*, 2014), le choix des algorithmes repose sur des données descriptives de la base qui, si elles ne peuvent pas être déduites automatiquement, doivent être fournies par le demandeur. A cet effet, on peut citer la qualification des attributs (identifiant/quasi-identifiant/sensible/non sensible, catégoriel/continu). De plus, certaines données descriptives (par exemple, la liste des attributs formant le quasi-identifiant) sont nécessaires quelle que soit la technique. D'autres (par exemple, la distribution des données) sont spécifiques à une technique. En résumé, dans un souci de généricité, le contexte d'une anonymisation sollicitée par un utilisateur pour ses micro-données est construit en deux temps (Fig. 4). Dans un premier temps, MAGGO construit le contexte à qualifier en récupérant de l'ontologie ses paramètres, c'est-à-dire les types d'exigences utilisateur à renseigner ainsi que les types de données descriptives à connaître pour le type d'anonymisation sollicitée.

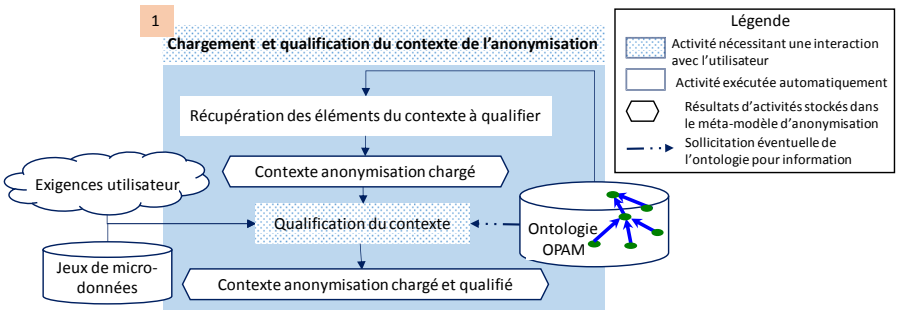

Figure 4. Chargement et qualification du contexte de l'anonymisation

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

Paramètres fournis par l'utilisateur	Paramètres pouvant être déduits automatiquement
Seuil de risque toléré	Attributs du QI
Taux de suppression autorisé	Attributs identifiants
Besoin d'usage	Attributs sensibles
Jeu de micro-données original	Nature de chaque attribut du QI : catégoriel ou continu
Propriétés de qualité attendues	Type de généralisation attendu
Importance relative des propriétés de	Distribution des données
	Taille du jeu de micro-données
	k : taille maximale des classes d'équivalence de QI
	MaxSup : le nombre maximal de tuples à supprimer

Certains de ces paramètres, rappelons-le, sont spécifiques à une technique. 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 « chargement et qualification du contexte ». 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 à ses exigences.

4.2. 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 cette étape de MAGGO (Fig. 5). 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.

4.2.1 Construction des signatures pertinentes

Le ciblage des algorithmes applicables au contexte exploite certains paramètres de 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[7], alors cet ensemble se restreint aux algorithmes fournissant ce type de généralisation tel 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 celles se trouvant dans le schéma d'OPAM représenté à la figure 3.

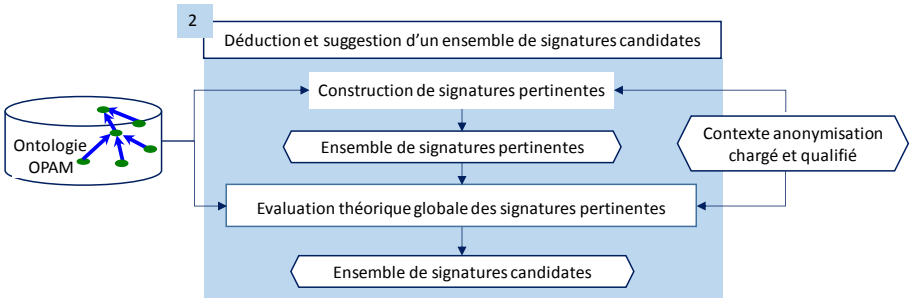

Figure 5. Déduction et suggestion de signatures candidates

Les algorithmes sélectionnés permettent bien sûr d'instancier le méta-modèle de l'anonymisation (les classes jaunes du méta-modèle de la figure 2). 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 deux 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.

4.2.2. 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 des 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.

[7] 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.

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 6 contient, à titre d'exemple, la hiérarchie construite par cette phase pour une anonymisation à des fins de classification.

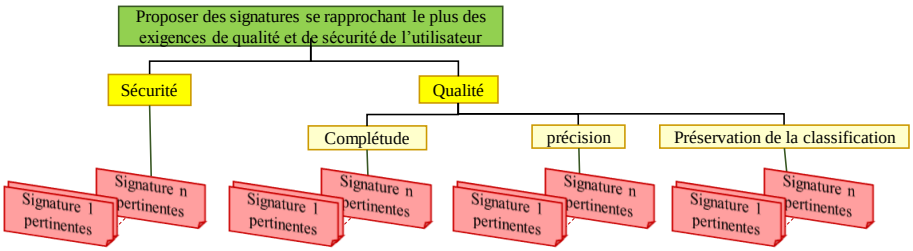

Figure 6. Hiérarchie multicritère pour l'anonymisation

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 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 de la figure 3). 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.

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

4.2.2.1. 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, une technique d'apprentissage supervisée est mise en place 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. Un exemple est constitué d'une entrée et d'une sortie.

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 6. Tous les jeux de données contiennent les mêmes informations : une valeur pour « k », une valeur pour « nombre d'attributs constituant le QI », une valeur pour « distribution du jeu de micro-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.

4.2.2.2. Mesure de l'importance relative des signatures

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	$\varepsilon_4 < E_{Sj}^{Ci} - E_{Sj'}^{Ci} \leq \varepsilon_5$

Une fois les évaluations locales des différentes signatures effectuées, il s'agit de procéder 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 inspirons de l'échelle sémantique de (Saaty et Sodenkamp, 2008) afin de permettre une comparaison automatique par paires de signatures pour livrer à AHP la matrice de comparaison des signatures pertinentes. 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). Dans cette table servant de comparaison par paires de signatures $\varepsilon1 < \varepsilon2 < \varepsilon3 < \varepsilon4 < \varepsilon5$. Ces valeurs sont définies par l'approche pour chaque critère de qualité.

4.3. Etapes 3,4 et 5 de MAGGO

Une fois la comparaison par paires effectuée, AHP se charge de fournir 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. Ce dernier 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 signatures pertinentes, de score le plus élevé, choisies par l'utilisateur. Pour guider l'utilisateur dans son choix de jeu de données anonymes, différentes évaluations cette fois-ci réelles, sont effectuées. Chaque évaluation permet de positionner le jeu anonyme vis-à-vis d'une exigence de qualité attendue.

5. 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 (1000 tuples). La distribution des données est dense. Le quasi-identifiant comprend trois attributs. L'usage 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

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 10 et MaxSup=20*1000/100=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 trop gourmand en temps de réponse. 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, Media Mondrian et TDS). Seul Datafly effectue des suppressions. 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 (Fig. 6). Les évaluations liées aux deux critères 'sécurité' et 'complétude' ont été déduites à partir respectivement des valeurs de k et MaxSup. Celles liées aux critères 'Précision' (métrique de discernabilité DM (Fung *et al.,* 2010)) et 'Préservation de la classification' ont été déduites en appliquant la 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. On aboutit à un score final fourni en dernière colonne du tableau 4. Après application d'AHP, il agrège les quatre critères pour chaque signature. Ce score permet à l'utilisateur de choisir d'exécuter les signatures (par exemple les quatre dernières) qui donnent le meilleur compromis entre les quatre critères, compromis qui résulte de l'application d'AHP à chaque paire de signatures.

Tableau 4. Evaluation locale (individuelle) 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

6. Conclusion

Les propriétaires 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 de telle sorte qu'il délivre des données sécurisées (difficiles à ré-identifier) et 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 le propriétaire des 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 et le maintien de la précision et de la complétude des données sont 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

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: 122-131*. Kidmore End: Academic Conferences International Limited.

Brand, R, (2002) Microdata protection through noise addition. In Domingo-Ferrer J, editor, *Inference Control in Statistical Databases*, LNCS Vol. 2316, pp 97-116, Springer.

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

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.

Fung, B. C. M., Wang, K., Chen, R., Yu, P. S. (2010) Privacy preserving data publishing: a survey of recent developments. In ACM Computing Surveys (CSUR), Vol. 42(14).

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

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

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

Loh W-Y (2011) Classification and regression trees. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery 1(1): 14-23.

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

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

Saaty T.L, Sodenkamp M.A. (2008) Making decisions in hierarchic and network systems. IJADS 1(1): 24-79

Samarati, P. (2001) Protecting respondents' identities in microdata release, *IEEE Transactions on Knowledge and Data Engineering*, Vol 13, No. 6, pp 1010-1027.

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

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