DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **13 décembre 2019**

Nom de famille et prénom de l'auteur : **DELANAUX Vincent**

Titre de la thèse : « Intégration de données liées respectueuse de la confidentialité ».

Résumé

La confidentialité des données personnelles est un souci majeur et un problème peu étudié pour la publication de données dans le Web des données ouvertes (ou LOD cloud, pour *Linked Open Data* cloud). Ce nuage formé par le LOD est un réseau d'ensembles de données interconnectés et accessibles publiquement sous la forme de graphes de données modélisés dans le format RDF, et interrogés via des requêtes écrites dans le langage SPARQL. Ce cadre très standardisé est très utilisé de nos jours par des organismes publics et des entreprises. Mais certains acteurs notamment du secteur privé sont toujours réticents à la publication de leurs données, découragés par des soucis potentiels de confidentialité.

Pour pallier cela, nous présentons et développons un cadre formel déclaratif pour la publication de données liées respectant la confidentialité, dans lequel les contraintes de confidentialité et d'utilité des données sont spécifiées sous forme de politiques (des ensembles de requêtes SPARQL). Cette approche est indépendante des données et du graphe considéré, et consiste en l'analyse statique d'une politique de confidentialité et d'une politique d'utilité pour déterminer des séquences d'opérations d'anonymisation à appliquer à n'importe quel graphe RDF pour satisfaire les politiques fournies. Nous démontrons la sûreté de nos algorithmes et leur efficacité en termes de performance via une étude expérimentale.

Un autre aspect à prendre en compte est qu'un nouveau graphe publié dans le nuage LOD est évidemment exposé à des failles de confidentialité car il peut être relié à des données déjà publiées dans d'autres données liées. Dans le second volet de cette thèse, nous nous concentrons donc sur le problème de construction d'anonymisations *sûres* d'un graphe RDF garantissant que relier le graphe anonymisé à un graphe externe quelconque ne causera pas de brèche de confidentialité. En prenant un ensemble de requêtes de confidentialité en entrée, nous étudions le problème de sûreté indépendamment des données du graphe, et la construction d'une séquence d'opérations d'anonymisation permettant d'assurer cette sûreté. Nous détaillons des conditions suffisantes sous lesquelles une instance d'anonymisation est sûre pour une certaine politique de confidentialité fournie. Par ailleurs, nous montrons que nos algorithmes sont robustes même en présence de liens de type *sameAs* (liens d'égalité entre entités en RDF), qu'ils soient explicites ou inférés par de la connaissance externe.

Enfin, nous évaluons l'impact de cette contribution assurant la sûreté de données en la testant sur divers graphes. Nous étudions notamment la performance de cette solution

et la perte d'utilité causée par nos algorithmes sur des données RDF réelles comme synthétiques. Nous étudions d'abord les diverses mesures d'utilité existantes et nous en choisissons afin de comparer le graphe original et son pendant anonymisé. Nous définissons également une méthode pour générer de nouvelles politiques de confidentialité à partir d'une politique de référence, via des modifications incrémentales. Nous étudions le comportement de notre contribution sur 4 graphes judicieusement choisis et nous montrons que notre approche est efficace avec un temps très faible même sur de gros graphes (plusieurs millions de triplets). Cette approche est graduelle : le plus spécifique est la politique de confidentialité, le plus faible est son impact sur les données. Pour conclure, nous montrons via différentes métriques structurelles (adaptées aux graphes) que nos algorithmes ne sont que peu destructeurs, et cela même quand les politiques de confidentialité couvrent une grosse partie du graphe.

Via une méthode simple et efficace pour assurer la confidentialité et l'utilité de graphes RDF dans des cas d'utilisation plausibles, cette nouvelle approche suggère de nombreuses extensions possibles et, sur le long terme, plus de travail sur la publication de données liées respectueuse de la confidentialité.