# Découverte d'expressions référentielles dans les graphes de connaissances

Armita Khajeh Nassiri *, Nathalie Pernelle *
Fatiha Saïs *

* LRI, Université Paris Saclay, CNRS UMR 8623, France
nom.prenom@lri.fr,

**Résumé.** Dans un graphe de connaissance, une expression référentielle est une formule logique qui permet d'identifier de façon unique une entité. De telles expressions peuvent être exploitées pour répondre à des requêtes, lier des données, annoter des ressources textuelles, ou encore anonymiser des données. Il peut potentiellement exister de nombreuses expressions logiques pour identifier de maniére unique une entité. Nous proposons une approche permettant de découvrir efficacement certaines expressions référentielles en nous concentrant sur celles qui ne peuvent être trouvées en instanciant des clés. Les premières expérimentations montrent que cette approche passe à l'échelle de jeux de données de plusieurs millions de triplets RDF et que ces expressions peuvent permettre de lier efficacement les instances de classes de différents jeux de données.

## 1 Introduction

Une *expression référentielle* (ER) est une description en langage naturel ou une formule logique permettant d'identifier de manière unique une entité. Par exemple, le 44ème président des Etats-Unis caractérise sans ambiguité Barack Obama. Les expressions référentielles trouvent des applications en désambiguisation, en anonymisation des données, en réponse à une requête ou encore en liage de données. Il peut potentiellement exister de nombreuses expressions référentielles pour une entité. La génération de telles expressions est une tâche bien étudiée en génération de langage naturel, (Dale, 1992). Différents algorithmes avec différents objectifs ont été proposés pour découvrir automatiquement les ER. Ces approches varient en fonction de l'expressivité des formules logiques qu'elles peuvent générer. Par exemple, dans (Dale, 1992; Krahmer et al., 2003), les ER créées sont des conjonctions d'atomes. (Ren et al., 2010) découvrent des ER plus complexes, représentées en logique de description et pouvant comporter des quantificateurs universels. Pour être efficace et réduire l'espace de recherche, les méthodes mettent l'accent sur la minimalité des descriptions qu'elles découvrent et/ou sur des préférences définis sur les propriétés utilisées (Galárraga et al., 2019). Cependant, la plupart de ces méthodes ne sont pas capables de s'adapter à de grands graphes de connaissances tels que YAGO ou DBpedia, graphes qui comportent des millions d'instances.
Dans ce travail, nous proposons un algorithme de découverte automatique de ER pour chacune des instance d'une classe d'un graphe de connaissances. Nous nous focalisons sur les ER qui ne peuvent pas résulter de l'instanciation d'une clé associée à la classe considérée. En effet, les