# Extraction de clés de liage de données (résumé étendu)

Jérôme Euzenat

# Extraction de clés de liage de données (résumé étendu)

Jérôme Euzenat

INRIA & Univ. Grenoble Alpes
F-38000 Grenoble, France
Jerome.Euzenat@inria.fr
http://exmo.inria.fr/~euzenat/

**Résumé.** De grandes quantités de données sont publiées sur le web des données. Les lier consiste à identifier les mêmes ressources dans deux jeux de données permettant l'exploitation conjointe des données publiées. Mais l'extraction de liens n'est pas une tâche facile. Nous avons développé une approche qui extrait des clés de liage (link keys). Les clés de liage étendent la notion de clé de l'algèbre relationnelle à plusieurs sources de données. Elles sont fondées sur des ensembles de couples de propriétés identifiant les objets lorsqu'ils ont les mêmes valeurs, ou des valeurs communes, pour ces propriétés. On présentera une manière d'extraire automatiquement les clés de liage candidates à partir de données. Cette opération peut être exprimée dans l'analyse formelle de concepts. La qualité des clés candidates peut-être évaluée en fonction de la disponibilité (cas supervisé) ou non (cas non supervisé) d'un échantillon de liens. La pertinence et de la robustesse de telles clés seront illustrées sur un exemple réel.

## 1 Web des données

Le web des données est l'utilisation des technologies du web sémantique pour publier des données sur le web (Heath et Bizer, 2011). Les données sont publiées sous forme de graphe dans le langage RDF et le vocabulaire de ce graphe décrit dans une ontologie. De grandes quantités de données sont publiées de la sorte.

## 2 Lier les données sur le web

L'intérêt de RDF est de pouvoir lier les données provenant de différentes sources de manière à pouvoir les exploiter conjointement. Cela est souvent obtenu en connectant les ressources représentant la même entité à l'aide du prédicat `owl:sameAs`. Au regard des importantes quantités de données publiées sur le web, il est utile de pouvoir les lier automatiquement. Le *liage de données* (data interlinking) consiste à identifier les mêmes ressources dans deux jeux de données. Cette tâche est apparentée à des tâches connues de reconnaissance d'entités ou de déduplication.

Nous distinguons deux approches du liage de données :

— Une approche fondée sur la similarité dans laquelle plus deux ressources sont similaires, plus elles sont susceptibles d'être liées ;
— Une approche fondée sur les clés dans laquelle une clé détermine l'identité d'une ressource : deux ressources répondant à la même clé doivent être liées.

On se concentre ici sur la deuxième approche.

# 3 Clés de liage

Nous avons développé une approche qui extrait des *clés de liage* (link keys). Une clé de liage est une expression de type

$$\langle \{\langle p_i, p'_i \rangle\}_{i \in I} \{\langle q_j, q'_j \rangle\}_{j \in J} \text{ linkkey } \langle c, c' \rangle \rangle$$

où $c$ et $c'$ sont des classes d'objets (définies dans un langage d'ontologies), $p_i$, $p'_i$, $q_j$, et $q'_j$ sont des prédicats au sens de RDF. Dans chaque couple, le premier élément se trouve dans une source de données et le second dans l'autre source de données. Une telle expression, s'interprète comme : deux ressources appartenant respectivement aux classes $c$ et $c'$, ayant une valeur commune pour les propriétés $p_i$ et $p'_i$ et ayant les mêmes valeurs pour les propriétés $q_j$ et $q'_j$ représentent le même objet.

Les clés de liage étendent la notion de clé de l'algèbre relationnelle de plusieurs manières : (a) elles considèrent deux sources de données, (b) les propriétés ne sont plus fonctionnelles, d'où l'introduction de deux ensembles de couples de propriétés, et (c) l'appartenance à une classe et l'égalité entre deux propriétés s'interprète en fonction de la sémantique de l'ontologie.

# 4 Extraire des clés candidates

Nous avons développé une technique d'extraction de clés de liage pour une restriction des clés à un couple de classes et à la seule intersection des propriétés (Atencia et al., 2014a). Même dans ce cas, le nombre des clés de liage potentielles entre deux classes est exponentielle en fonction de la taille du produit des propriétés. Nous nous appuyons donc sur la notion de clés de liage candidates qui (a) engendrent au moins un lien et (b) sont maximales pour au moins un lien (ou sont l'intersection de telles candidates). Un algorithme pour l'extraction de telles clés a été développé tirant parti de l'indexation des propriétés et de leurs valeurs.

# 5 Clés candidates et analyse formelle de concepts

La détermination de clés dans les bases de données se fait à partir de relations bien formées (supposées sans duplicats). Elle peut alors être ramenée à l'extraction de dépendances fonctionnelles. L'extraction de dépendances fonctionnelles peut être exprimée dans l'analyse formelle de concepts (Ganter et Wille, 1999). L'extraction de clés candidates (avec duplicats) ne nécessite pas d'extraire les dépendances fonctionnelles, mais peut réutiliser une partie de cette approche.

On a montré que les clés de liage candidates sont les concepts d'un encodage du problème dans l'analyse formelle de concepts, tout d'abord sur une réduction du problème dans le modèle relationnel (Atencia et al., 2014b), puis sur le problème décrit ci-dessus.

# 6 Sélectionner de bonnes clés de liage

Les clés candidates n'ont aucune garantie d'engendrer un grand nombre de liens ou de couvrir l'ensemble des individus. Il est donc nécessaire de sélectionner les plus prometteuses d'entre elles. Pour ce faire, nous avons développé deux mesures permettant d'estimer leur qualité (Atencia et al., 2014a) :

— en fonction de la disponibilité d'un échantillon de liens (cas supervisé), par une approximation de la précision et du rappel sur cet échantillon,
— ou en son absence (cas non supervisé), par une évaluation de la discriminabilité et de la couverture des clés sur les données qui peuvent aussi être considérées comme une approximation de la précision et du rappel.

La pertinence des clés extraites a pu être vérifiée sur un exemple réel. Nous avons aussi déterminé la robustesse de ses mesures en évaluant la qualité du résultat obtenus par les clés sélectionnées en présence de dégradation des jeux de données initiaux.

# 7 Travaux futurs

Nous développons ce travail dans différentes dimensions, tout en préservant le parallèle entre la partie algorithmique et l'analyse formelle de concepts :

— en considérant les clés de liage avec égalité de propriétés,
— en développant des méthodes alternatives de restriction fondées sur des classes et propriétés construites,
— en développant des méthodes de co-extraction de plusieurs clés de liage simultanément, et l'analyse relationnelle de concepts devrait s'avérer utile (Hacene et al., 2013),
— en prenant en compte la circularité dans l'approche précédente.

Enfin, nous travaillons aussi sur le raisonnement à l'aide de clés de liage qui peuvent être considérées comme des assertions de logiques de descriptions susceptibles d'être inférées et de permettre d'inférer d'autres assertions.

# 8 Remerciements

Cet exposé est le résultat d'un travail conjoint avec Manuel Atencia, Jérôme David et Amedeo Napoli. Ce travail a été en partie subventionné par l'ANR sur le projet blanc international Lindicle (12-IS02-0002).

# Biographie

Jérôme Euzenat est directeur de recherches à l'INRIA. Il dirige l'équipe Exmo du Laboratoire d'Informatique de Grenoble qui se consacre particulièrement à l'interopérabilité sémantique. Ces dernières années il a fortement contribué au développement de l'alignement d'ontologies : il est coauteur de l'ouvrage de référence sur le sujet. Avec ses collègues, il prolonge ce travail vers le liage de données RDF.

# Références

Atencia, M., J. David, et J. Euzenat (2014a). Data interlinking through robust linkkey extraction. In *Proc. 21st european conference on artificial intelligence (ECAI), Praha (CZ)*, pp. 15–20.

Atencia, M., J. David, et J. Euzenat (2014b). What can fca do for database linkkey extraction ? In *Proc. 3rd ECAI workshop on What can FCA do for Artificial Intelligence ? (FCA4AI), Praha (CZ)*, pp. 85–92.

Ganter, B. et R. Wille (1999). *Formal concept analysis : mathematical foundations*. Springer.

Hacene, M., M. Huchard, A. Napoli, et P. Valtchev (2013). Relational concept analysis : mining concept lattices from multi-relational data. *Annals of Mathematics and Artificial Intelligence 67*(1), 81–108.

Heath, T. et C. Bizer (2011). *Linked Data : Evolving the Web into a Global Data Space*. Morgan & Claypool.

# Summary

Large data sets are published on the web of data. In order to exploit them jointly, it is necessary to identify resources from different data sources representing the same entity and link them. We have developed a data interlinking approach which uses link keys. Link keys extend the relational notion of keys to several data sources. They are based on sets of property pairs identifying objects. We present an algorithm for link key extraction which first extract a small number of candidate link keys. These candidate link keys have also been characterised in formal concept analysis. The quality of such candidates is evaluated depending of the availability of sample links (supervised case) or in absence of such links (non supervised case). A real-world data set is used to illustrate the relevance and robustness of extracted keys.