ELKER ANR Project
Enhancing Link Keys: Extraction and Reasoning

Manuel Atencia et Jérôme David

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG

10 juillet 2018

Info générale

- Date de début : 01/10/2017.
- Date de fin : 30/09/2021 (48 mois).
- Programme ANR : Société de l'information et de la communication (DS07) 2017
- Référence du projet : ANR-17-CE23-0007-01
- Site web : `https://project.inria.fr/elker/`

Consortium

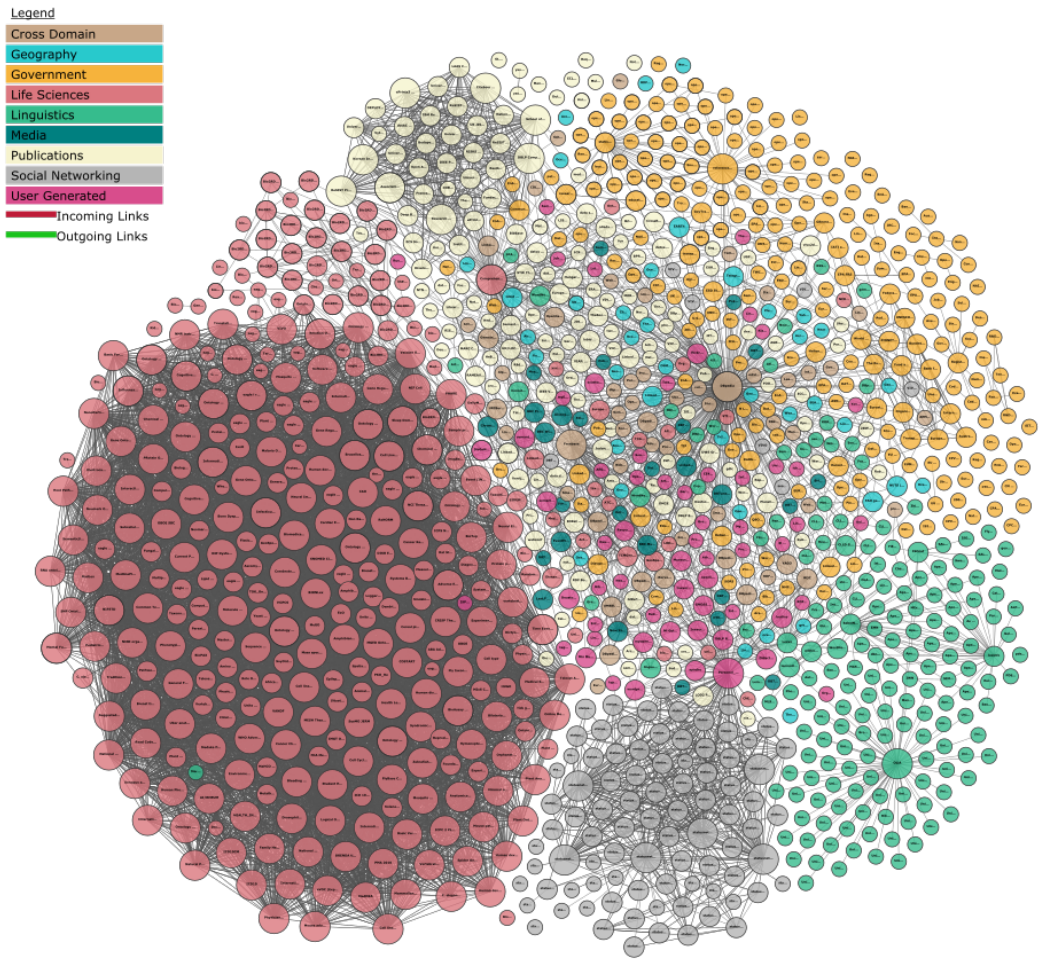
- **LIG**: Manuel Atencia, Jérôme David, Jérôme Euzenat et Marie-Christine Rousset
- **LIASD**: Mario Cataldi, Myriam Lamolle et Chan Le Duc
- **Inria (Nancy)**: Miguel Couceiro, Adrien Coulet, Amedeo Napoli et Chedy Raïssi

- Khadija Jradeh, thésarde au LIG & LIASD
- Nacira Abbas, thésarde au LIG & Inria
- 1 postdoc au LIASD
- 1 postdoc à l'Inria

En bref

- ELKER attaque le probléme du liage de données dans le contexte du Web des données
- Il propose d'utiliser les **clés de liage** (*link keys*) pour résoudre ce probléme
- Il étudie les fondements et les algorithmes des clés de liage de deux façons complémentaires : l'**extraction** automatique de clés de liage et le **raisonnement** avec clés de liage pour l'inférence de liens.

Web des données



C'est quoi le Web des données ?

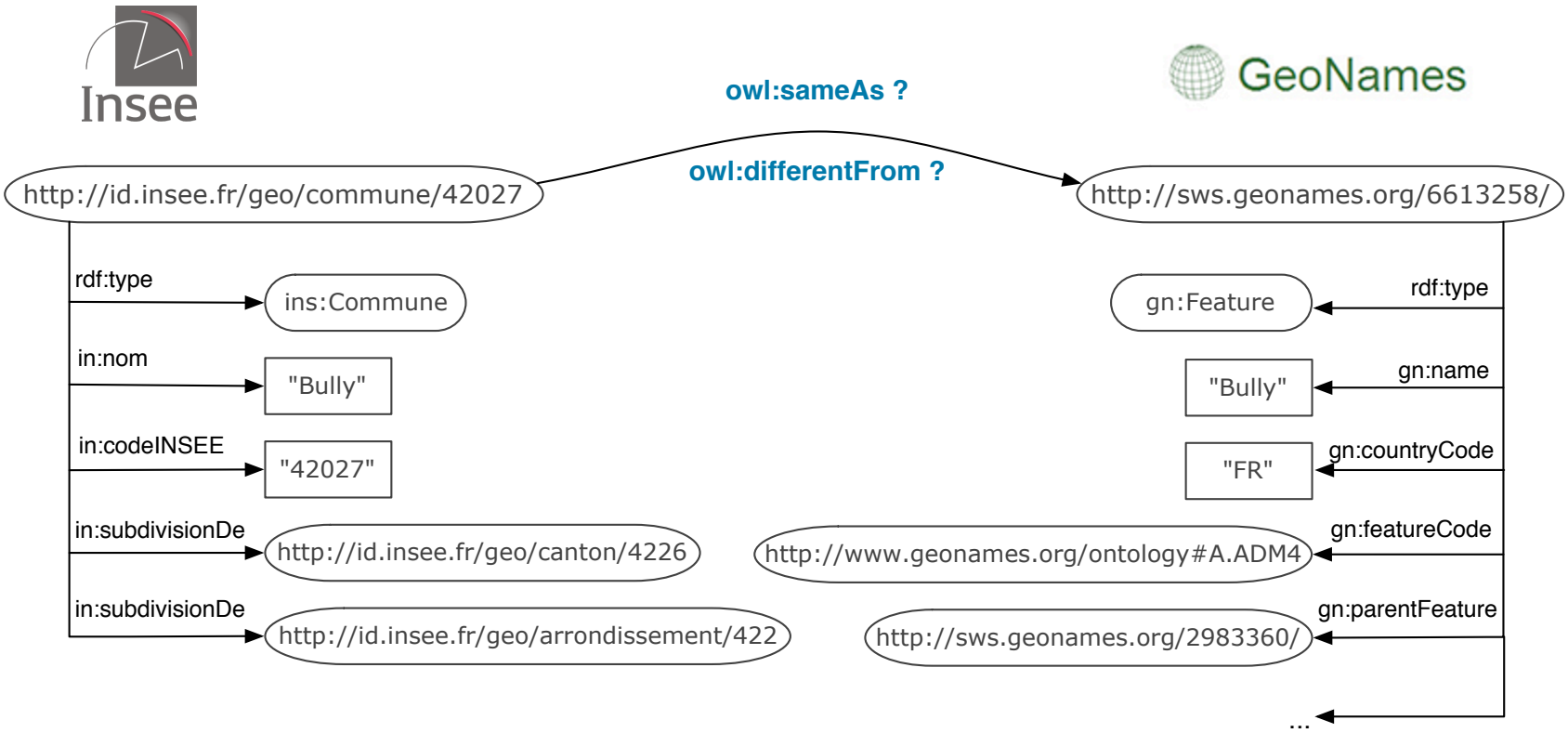
- Données structurées exprimées en utilisant les technologies du Web sémantique (RDF, OWL, etc.)
- Publiées sur le Web (URIs déréfénçables, points d'accès SPARQL, etc.), et
- **Liées** : ressources provenant de différents jeux de données et représentant la même entité du monde réel devraient être liées par `owl:sameAs`

Nombreux exemples : DBpedia, YAGO, BBC Music, MusicBrainz, GeoNames, data.gov.uk, Insee, Ina, BNF, etc.

Liage de données

- Problème : étant donné deux jeux de données RDF, trouver des paires de ressources qui identifient la même entité
- Essentiel pour élargir et améliorer le Web de données
- Défis :
  - Passage à l'échelle
  - Hétérogénéité des descriptions de données
  - Qualité des données

Liage de données : Un exemple



Approches existantes

- Méthodes numériques basées sur l'agrégation de valeurs de similarité entre les valeurs de propriétés pertinentes
  - Silk (Volz et al. ISWC'09)
  - LIMES (Ngomo and Auer IJCAI'11)
- Méthodes symboliques basées sur les clés
  - L2R (Saïs et al. AAAI'07)
  - SAKey (Symeonidou et al. ISWC'14)
  - **Link keys** (Atencia et al. ECAI'14)
  - ProbFR (Al-Bakri et al. ECAI'16)
  - KeyRanker (Farah et al. K-CAP'17)

Clés de liage

- Les clés de liage généralisent les clés des bases de données relationnelles aux jeux de données RDF différents
- Par conséquent, elles peuvent êtres utilisées pour découvrir des liens d'égalité
- Elles diffèrent des clés des bases de données relationnelles dans divers aspects :
  - Les propriétées RDF peuvent ne pas être fonctionnelles
  - Elles s'appliquent à deux sources de données qui peuvent dépendre d'ontologies (et donc interprétées logiquement)

Clés de liage

Exemple : Une clé de liage

$$\{\langle \text{auteur}, \text{creator}\rangle, \langle \text{titre}, \text{title}\rangle\}\, \textit{linkkey}\, \langle \text{Livre}, \text{Book}\rangle$$

Intuitivement : si une instance de la classe *Livre* a les mêmes valeurs pour les propriétés *auteur* et *titre* qu'une instance de la classe *Book* a pour les propriétés *creator* et *title*, alors elles désignent la même entité.

$\Rightarrow$ pourrait être utilisée pour identifier les mêmes livres dans deux sources de données bibliographiques, l'une française et l'autre anglaise.

Clés de liage

Exemple : Une clé de liage

$$\{\langle \text{auteur}, \text{creator}\rangle, \langle \text{titre}, \text{title}\rangle\} \mathit{linkkey} \langle \text{Livre}, \text{Book}\rangle$$

Intuitivement : si une instance de la classe *Livre* a les mêmes valeurs pour les propriétés *auteur* et *titre* qu'une instance de la classe *Book* a pour les propriétés *creator* et *title*, alors elles désignent la même entité.

$\Rightarrow$ pourrait être utilisée pour identifier les mêmes livres dans deux sources de données bibliographiques, l'une française et l'autre anglaise.

Revenons sur ELKER...

- ELKER attaque le probléme du liage de données dans le contexte du Web des données
- Il propose d'utiliser les **clés de liage** pour résoudre ce probléme
- Il étudie les fondements et les algorithmes des clés de liage de deux façons complémentaires : l'**extraction** automatique de clés de liage et le **raisonnement** avec clés de liage pour l'inférence de liens.

Extraction de clés de liage

- deux phases interdépendantes: :
  - identifier les **clés de liage candidates** (ensemble maximal de paires de propriétés qui, s'il était utilisé comme une clé de liage, générerait au moins un lien)
  - sélectionner les meilleures clés de liage candidates selon certaines mesures de qualité
- nous prévoyons d'utiliser les techniques de l'**Analyse Formelle de Concepts** (FCA)
  - formalisme permettant l'analyse et la classification de données
  - bien adapté aux problèmes de découverte de connaissances
  - des années d'expérience dans le développement de modèles et d'algorithmes

Extraction de clés de liage - Tâches spécifiques

- Adapter FCA pour l'extraction de clés de liage, en considérant
  - le cas où les valeurs des propriétés ne sont pas égales mais similaires (dépendances fonctionnelles floues)
  - différents types de clés de liage (*eq-link keys* et *in-link keys*)
- Extensions utilisant RCA (Analyse Relationnelle de Concepts)
  - RCA permet de gérer les relations entre les objets explicitement
  - nécessaire pour extraire conjointement des clés de liage à partir de classes dépendantes (par ex. `Livre` dépend de `Personne` comme valeur de la propriété `auteur`)
- Mesures de qualité
  - précision et rappel, couverture et discriminabilité
  - critères globaux et locaux en présence de classes dépendantes
  - utilisées pour l'optimisation
- Responsables : LIG et Inria

Raisonnement avec clés de liage

- Les données RDF, avec les ontologies RDFS ou OWL sont des théories logiques
- Les clés de liage peuvent être exprimées sous la forme d'axiomes logiques
- Il est possible de raisonner avec les clés de liage de manières diffrentes :
  - inférer des clés de liage à partir de déclarations OWL
  - inférer des clés de liage à partir d'autres clés de liage
  - inférer des déclarations OWL (liens `owl:sameAs`) à partir de clés de liage

Raisonnement avec clés de liage

Exemple : Raisonnement avec clés de liage

Connaissance d'un expert du domaine :

$$\text{in:Commune} \equiv \text{gn:Municipality}$$

$$\text{gn:Municipality} =_{def} \text{gn:Feature} \sqcap \exists \text{gn:countryCode}.\{\text{FR}\} \sqcap \exists \text{gn:featureCode}.\{\text{A.ADM4}\}$$

Outil de découverte de clés :

$$(\{\text{gn:name}, \text{gn:parentFeature}, \text{gn:featureCode}, \text{gn:countryCode}\} \text{ key } \text{gn:Feature})$$

Outil de mise en correspondance d'ontologies :

$$\text{in:nom} \equiv \text{gn:name}$$

$$\text{in:subdivisionDe} \sqsubseteq \text{gn:parentFeature}$$

Raisonnement avec clés de liage

Exemple : Raisonnement avec clés de liage (suite)

De ces connaissances, il peut être inféré :

$$(\{\langle \text{nom}, \text{name}\rangle\}\{\langle \text{subdivisionDe}, \text{parentFeature}\rangle\} \text{ linkkey } \langle \text{Commune}, \text{Municipality}\rangle)$$

qui peut être exprimée comme sous la forme d'une règle logique :

$$\text{nom}(?x, ?n), \text{name}(?y, ?n), \text{subdivisionDe}(?x, ?u), \text{parentFeature}(?y, ?v),$$
$$\text{sameAs}(?u, ?v), \text{Commune}(?x), \text{Municipality}(?y) \rightarrow \text{sameAs}(?x, ?y)$$

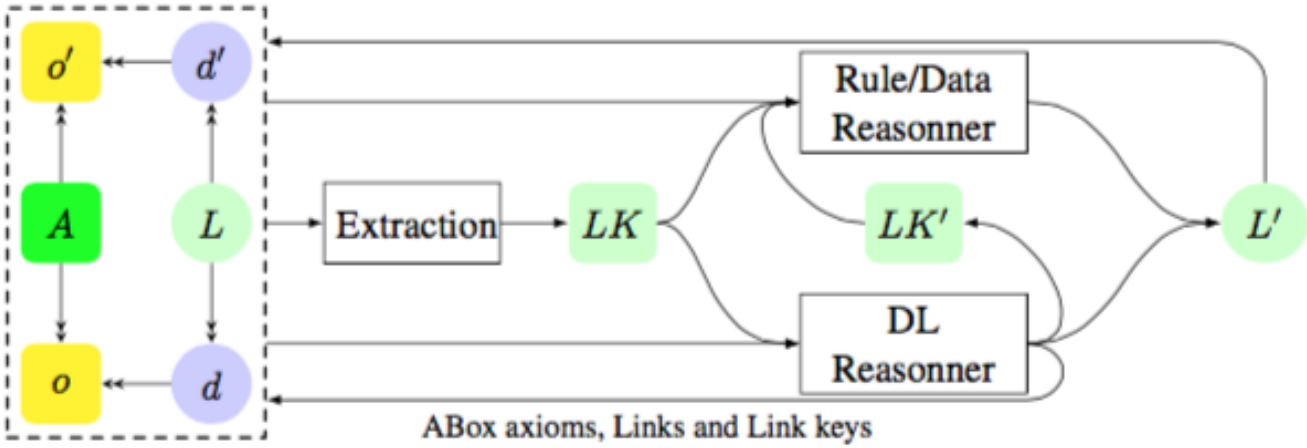
Raisonnement avec clés de liage

- Le raisonnement avec les clés de liage est adressé de deux manières :
  (1) **inférence de clés de liage** : inférer des clés de liage additionnelles à partir d'autres clés de liage et des connaissances
  (2) **inférence de liens** : combiner des clés de liage, des connaissances et données pour inférer des liens d'égalité.
- Pour (1) nous prévoyons d'étendre les **méthodes des tableaux** pour les logiques de description (DLs) en ligne avec les travaux sur le raisonnement avec les ontologies OWL
- Pour (2), et afin d'assurer le passage à l'échelle, nous utiliserons les **méthodes basées sur les règles** de type **Datalog** (quand la traduction est possible)

Raisonnement avec clés de liage - Tâches spécifiques

- Développer des méthodes de tableaux et de tableaux compressés pour raisonner avec des clés de liage dans le contexte des logiques de description
- Raisonnement distribué avec tableaux optimisés quand les connaissances sont distribuées sur des sources différentes
- Extensions de notre travail précédent sur les méthodes basées sur les règles pour le liage de données
  - Traduction des clés de liages en règles Datalog
  - Configuration automatique du seuil d'acceptation des liens
  - Version probabiliste de notre algorithme "import-by-query"
- Extraction de clés de liage utilisant le raisonnement
- Responsables : LIG et LIASD

Extraction et raisonnement



Implementation et évaluation

- Les méthodes d'extraction et de raisonnement avec les clés de liage seront implémentées et évaluées
- L'implémentation étendra le logiciel existant
  - Extraction de clés de liage : Alignment API, Linkex
  - Inférence de liens basée sur les règles : ProbFR
  - Raisonneurs OWL : DRAOn, Staré
- Evaluation
  - Correction et passage à l'échelle
  - Liage de données du monde réel : domaine des bibliothèques (BNF, BNE, BNB) et domaine de l'académie (HAL, DBLP) à l'aide de "gold standards" disponibles ou fabriqués à la main