Peuplement du jeu de données conférence

Thomas Cantié, Elodie Thiéblin, Cassia Trojahn dos Santos

Peuplement du jeu de données conférence

Thomas Cantié, Elodie Thiéblin, Cassia Trojahn

IRIT Institut de Recherche en Informatique de Toulouse, Toulouse, France
thomas.cantie.contact@gmail.com, elodie.thieblin@irit.fr, cassia.trojahn@irit.fr

Résumé : **Mots-clés** : Peuplement d'ontologie, Alignement d'ontologie, Évaluation d'alignements.

1 Introduction et travaux liés

Le jeu de données conférence proposé par Šváb Zamazal *et al.* (2005) est souvent employé pour de l'évaluation d'alignement d'ontologies, comme le montrent Zamazal & Svátek (2017). Ses alignements de référence utilisés dans l'OAEI (Ontology Alignment Evaluation Initiative), campagne annuelle d'évaluation de systèmes d'alignement, ont été déclinés sous plusieurs versions comme une version consensuelle (Cheatham & Hitzler, 2014) ou une version avec des alignements complexes (Thiéblin *et al.*, 2018a). Des approches d'alignement comme celles de Walshe *et al.* (2016) se basent sur des instances communes aux ontologies. Pour qu'elles puissent être évaluées et comparées aux autres approches d'alignement sur le jeu de données Conférence, il faut que celui-ci soit peuplé. D'autre part, dans la tâche OA4QA, Solimando *et al.* (2014) avaient succinctement peuplé quelques ontologies de Conférence pour évaluer des alignements sur de la réécriture de requêtes. Ce peuplement était limité à la portée des requêtes et était de fait difficilement réutilisable. Nous décrivons ici la méthodologie suivie pour peupler cinq ontologies du jeu de données Conférence et les jeux de données peuplés qui en résultent.

2 Méthodologie pour peupler le jeu de données

La méthodologie est basée sur des *questions de compétence pour alignement (CQAs)* qui sont définies comme des questions de compétence pouvant être couvertes par plusieurs ontologies (Thiéblin *et al.*, 2018b). Les CQAs définissent donc les besoins en connaissance devant être couverts (au mieux) par plusieurs ontologies et l'alignement entre elles. L'utilisation des CQAs dans le processus de peuplement assure qu'il soit homogène.

1. Création d'un ensemble de CQAs sur un scénario applicatif pour orienter l'interprétation des ontologies par les experts. Exemple de CQA : *Quels articles sont acceptés ?*
2. Création manuelle d'un format pivot (e.g., schéma JSON) pour couvrir les CQAs.
3. Pour chaque ontologie du jeu de données, créer des requêtes SPARQL INSERT pour traduire le format pivot. Chaque ontologie ne couvre pas forcément toutes les CQAs.
4. Instancier le format pivot avec des données réelles ou automatiquement générées.
5. Peupler les ontologies avec les instanciations du format pivot en utilisant les requêtes SPARQL INSERT de l'étape 3.
6. Appliquer un raisonneur aux ontologies peuplées. Si elles ne sont pas consistantes, changer son interprétation des ontologies et reprendre les étapes 3 à 5.

3 Jeu de données Conférence peuplé

La méthodologie a été suivie pour peupler cinq ontologies du jeu de données Conférence : *cmt, conference, confOf, edas, ekaw*. 152 CQAs ont été créées par un expert en suivant le

scénario réel d'organisation de Conférence sur l'édition de ESWC 2018 (Extended Semantic Web Conference) et étendues par exploration des ontologies. Le format pivot a été d'abord instancié avec les données du site Web d'ESWC 2018. Avec l'analyse de ces données, un script d'instanciation automatique du format pivot a été développé reprenant des statistiques telles que la proportion de membres du comité de programme auteur d'articles, etc. Le jeu de données, les instanciations du format pivot et le processus de peuplement sont disponibles [1].

En plus du peuplement avec les données d'ESWC, 6 jeux de données ont été générés pour proposer des ontologies partageant plus ou moins d'instances communes. Dans les jeux de données artificiels, chaque ontologie a été peuplée avec les données de 5 instanciations du format pivot (une instanciation du format pivot contient les informations pour l'organisation d'une conférence). Dans le jeu de données 0% toutes les ontologies ont été peuplées avec 5 instanciations du format pivot différentes. Dans le jeu de données 20%, les ontologies ont été peuplées avec 1 instanciation identique et 4 différentes. Les jeux de données 40%, 60%, 80% et 100% suivent la même logique. Le pourcentage qui sert de nom aux jeux de données est le pourcentage d'instanciations communes du format pivot utilisé dans le peuplement des ontologies. Comme la taille de chaque instanciation peut différer, le pourcentage d'instances communes entre deux ontologies varie. Par exemple, dans le jeu 20%, les instances *Articles scientifiques* communes aux ontologies représentent entre 7% des instances d'*Articles scientifiques* de l'ontologie *ekaw* et 11% des instances d'*Articles scientifiques* de l'ontologie *cmt*.

TABLE *1 – Nombre d'entités peuplées sur nombre d'entités total par ontologie. Nombre de CQAs couvertes par chaque ontologie.*

	cmt	conference	confOf	edas	ekaw
Classes	26 / 30	51 / 60	29 / 39	42 / 104	57 / 74
Obj. prop.	43 / 49	37 / 46	10 / 13	17 / 30	26 / 33
Data prop.	7 / 10	13 / 18	10 / 23	11 / 20	0 / 0
CQAs couvertes	46	90	67	60	84

Conclusion

Cinq ontologies du jeu de données Conférence ont été peuplées à l'aide de questions de compétence pour alignement. Sept jeux de données différents, ayant plus ou moins d'instances communes, résultent du peuplement. Ces jeux de données peuplés, couplés aux alignements de référence existants, peuvent servir à l'évaluation de systèmes d'alignements.

Références

CHEATHAM M. & HITZLER P. (2014). Conference v2. 0 : An uncertain version of the OAEI Conference benchmark. In *International Semantic Web Conference*, p. 33–48.

SOLIMANDO A., JIMÉNEZ-RUIZ E. & PINKEL C. (2014). Evaluating ontology alignment systems in query answering tasks. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, p. 301–304.

THIÉBLIN E., CHEATHAM M., TROJAHN C., ZAMAZAL O. & ZHOU L. (2018a). The first version of the oaei complex alignment benchmark. In *ISWC Posters and Demos*.

THIÉBLIN E., HAEMMERLÉ O. & TROJAHN C. (2018b). Complex matching based on competency questions for alignment : a first sketch. In *OM 2018 - 13th ISWC workshop on ontology matching*.

WALSHE B., BRENNAN R. & O'SULLIVAN D. (2016). Bayes-recce : A bayesian model for detecting restriction class correspondences in linked open data knowledge bases. *International Journal on Semantic Web and Information Systems (IJSWIS)*, **12**(2), 25–52.

ZAMAZAL O. & SVÁTEK V. (2017). The Ten-Year OntoFarm and its Fertilization within the OntoSphere. *Web Semantics : Science, Services and Agents on the World Wide Web*, **43**, 46–53.

ŠVÁB ZAMAZAL O., SVÁTEK V., BERKA P., RAK D. & TOMÁŠEK P. (2005). Ontofarm : Towards an experimental collection of parallel ontologies. *Poster Track of ISWC*, **2005**.

1. `https://framagit.org/IRIT_UT2J/conference-dataset-population`