Complex ontology matching

Cássia Trojahn

IRIT & Université de Toulouse 2 Jean Jaurès, Toulouse, France
cassia.trojahn@irit.fr

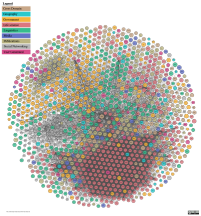
Séminaire résidentiel INRAE Semantic Linked Data, 11 au 14 octobre 2021

Context

Semantic web — Data exposed with annotations in a way that it can be used by machines

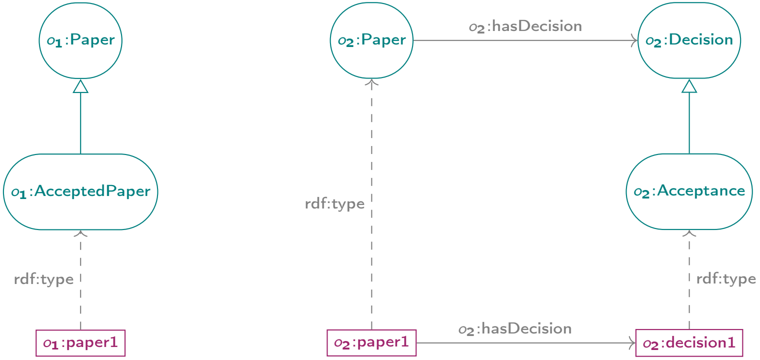
Ontology — Vocabulary describing a domain of interest and a formal specification of the meaning of its terms

Linked open data — Data as instances of ontologies, linked across knowledge bases

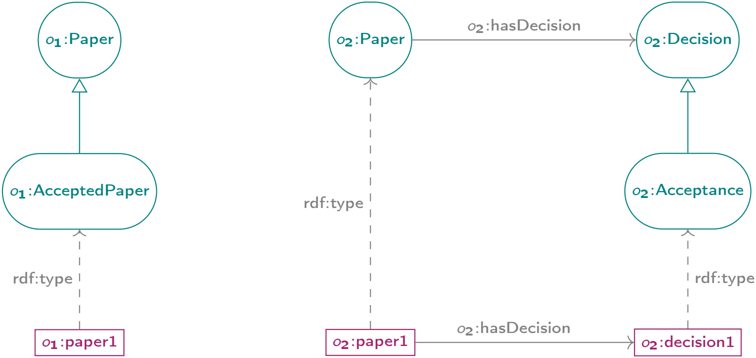


Ontology heterogeneity Ontology differences in terms of the terminology, coverage, granularity modelling strategies, or still level of generality

Ontology matching Task of generating a set of correspondences between different ontologies

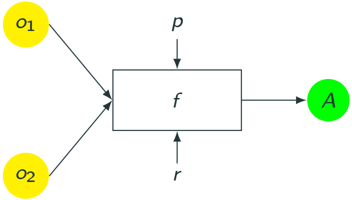


o_1:Paper $\equiv$ o_2:Paper



$$o_1\text{:Paper} \equiv o_2\text{:Paper}$$

$$o_1\text{:AcceptedPaper} \equiv o_2\text{:Paper} \sqcap \exists o_2\text{:hasDecision}.o_2\text{:Acceptance}$$



Adapted from [Euzenat and Shvaiko, 2013]

A is a set of **correspondences** $\{c_1, ..., c_n\}$, where c_i is a tuple (e_1, e_2, r)
e_1 and e_2 are the members of the correspondence:

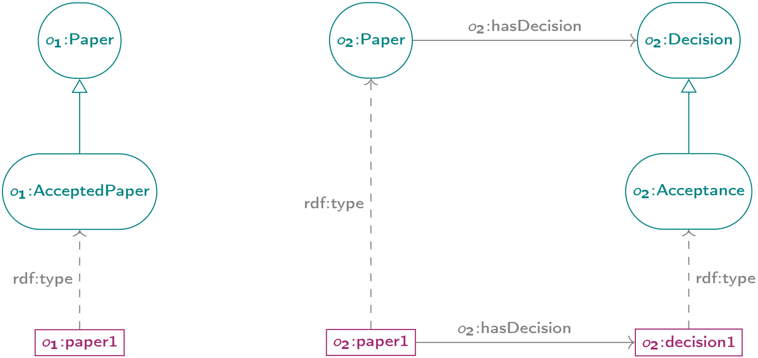
- **simple** correspondence (s:s): e_1 and e_2 are simple expressions $(o_1\texttt{:Paper}, o_2\texttt{:Paper}, \equiv)$
- **complex** correspondence (s:c, c:s, c:c): e_1 or/and e_2 is a complex expression $(o_1\texttt{:AcceptedPaper}, \exists\ o_2\texttt{:Paper} \sqcap o_2\texttt{:hasDecision}.o_2\texttt{:Acceptance}, \equiv)$
- r is a relation, e.g., $(\equiv, \sqsupseteq, \sqsubseteq, \perp)$

Need for complex correspondences

Simple correspondences are not expressive enough
to overcome the different kinds of ontology heterogeneity

Alignments between real-world ontologies contain many
relations uncovered by current systems

Need for more expressiveness in diverse domains and applications



$$o_1\text{:Paper} \equiv o_2\text{:Paper}$$

$$o_1\text{:AcceptedPaper} \equiv o_2\text{:Paper} \sqcap \exists o_2\text{:hasDecision}.o_2\text{:Acceptance}$$

- Higher search space for generating complex correspondences
- User needs are neglected in most matching approaches

- Reduce the matching space taking into account user's knowledge needs
  → **Competency Questions for Alignment**

Same as competency questions for ontology authoring [Suárez-Figueroa et al., 2012], but to be answered over two or more ontologies.

Same as competency questions for ontology authoring [Suárez-Figueroa et al., 2012], but to be answered over two or more ontologies.

Can be a NL question or SPARQL queries.

- "What are the accepted papers?"
- `SELECT ?x WHERE {?x a` o_1`:AcceptedPaper.}`
- `SELECT ?x WHERE {?x` o_2`:hasDecision ?y.  ?y a` o_2`:Acceptance.}`

Same as competency questions for ontology authoring [Suárez-Figueroa et al., 2012], but to be answered over two or more ontologies.

Can be a NL question or SPARQL queries.

- "What are the accepted papers?"
- `SELECT ?x WHERE {?x a` o_1`:AcceptedPaper.}`
- `SELECT ?x WHERE {?x` o_2`:hasDecision ?y.  ?y a` o_2`:Acceptance.}`

Unary set of instances Which are the accepted papers?
$\rightarrow$ {paper1, paper2, ...}

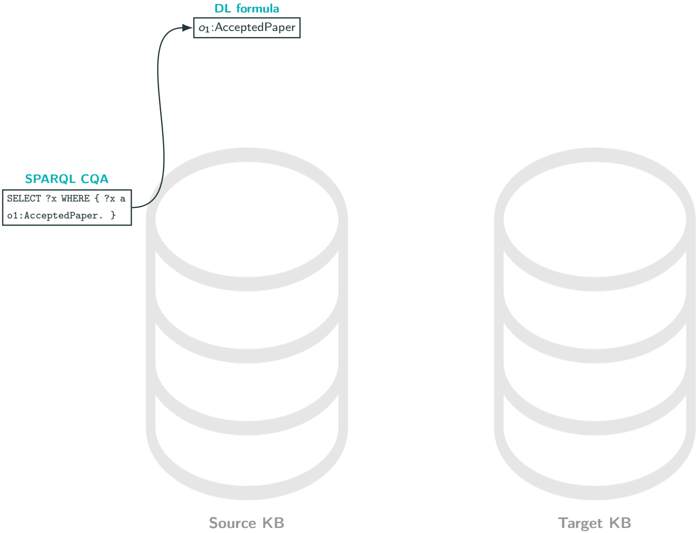
Binary set of pairs of instances Who is the author of which paper?
$\rightarrow$ {(paper1, person1), (paper2, person2), ...}

- Takes as input a set of CQAs in the form of SPARQL SELECT queries over o_1
- Requires o_1 and o_2 to have an Abox with at least one common instance for each CQA
  - answer (instances) to each input query are matched with those of a knowledge base described by o_2
  - matching is performed by finding the surroundings of the o_2 instances which are lexically similar to the CQA

SPARQL CQA

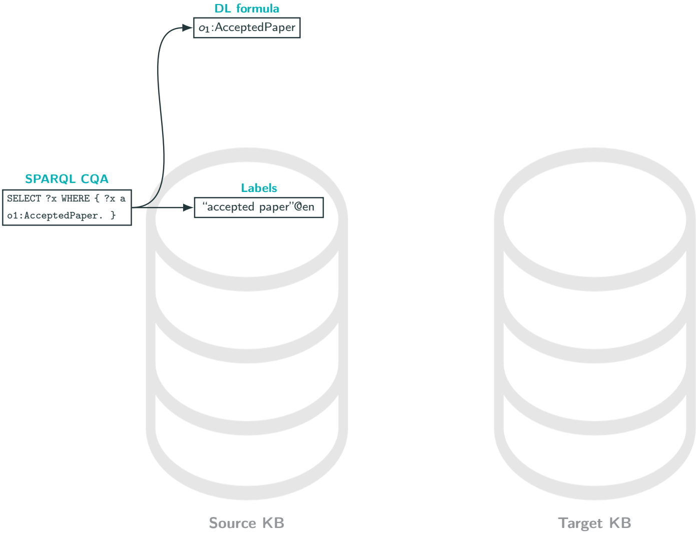
```
SELECT ?x WHERE { ?x a
o1:AcceptedPaper. }
```


Input: CQA, Source KB and Target KB

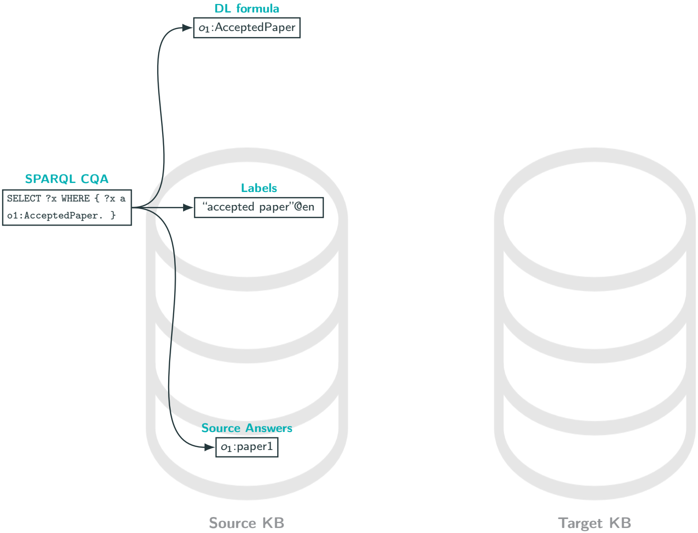


1 Extract formula

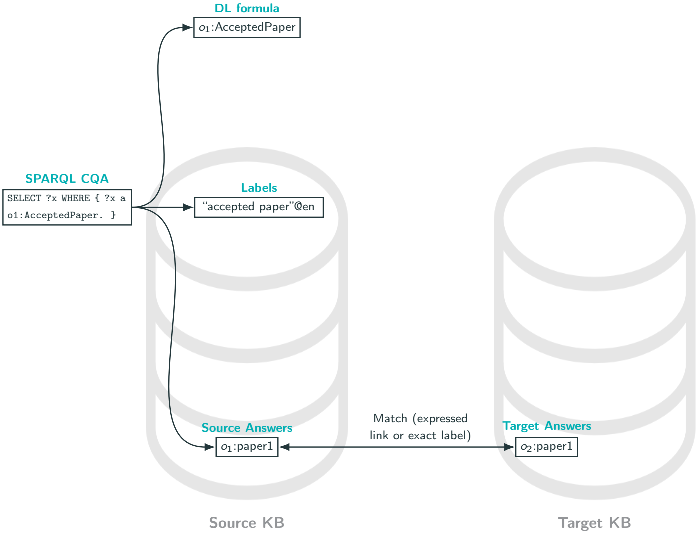
Complex alignment generation based on CQAs



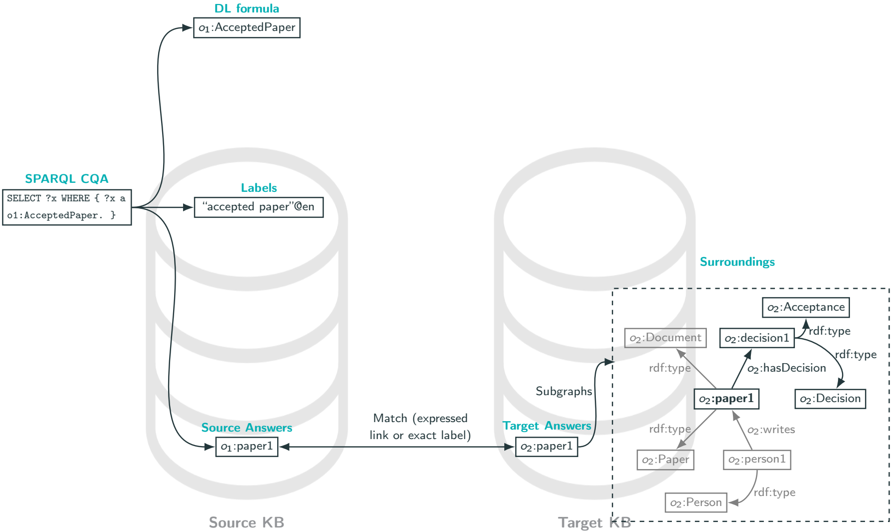
2 Extract CQA labels



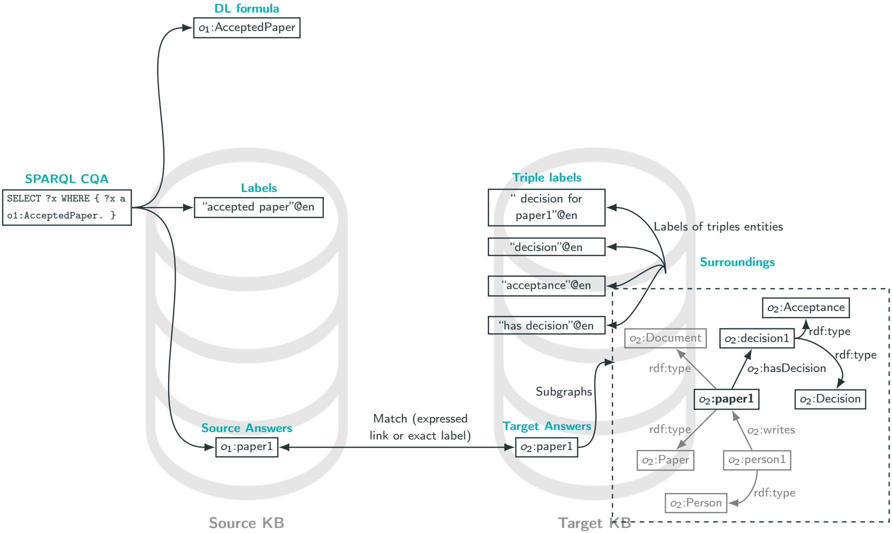
3 Retrieve answers



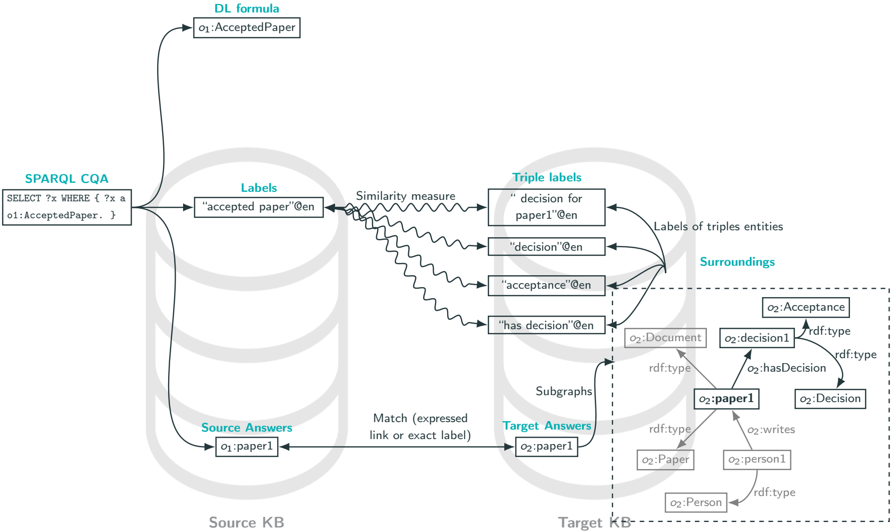
4 Match answers with target instances



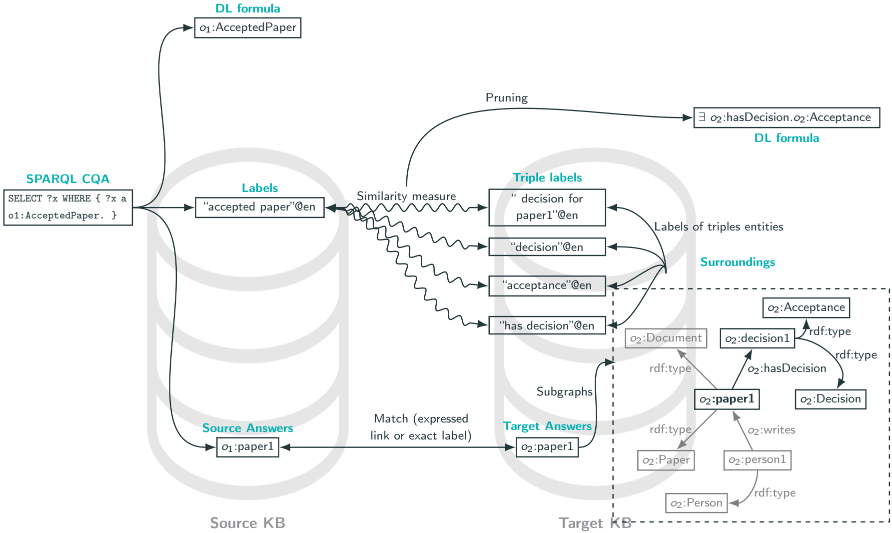
5 Get target subgraphs



6 For each triple, get entity labels

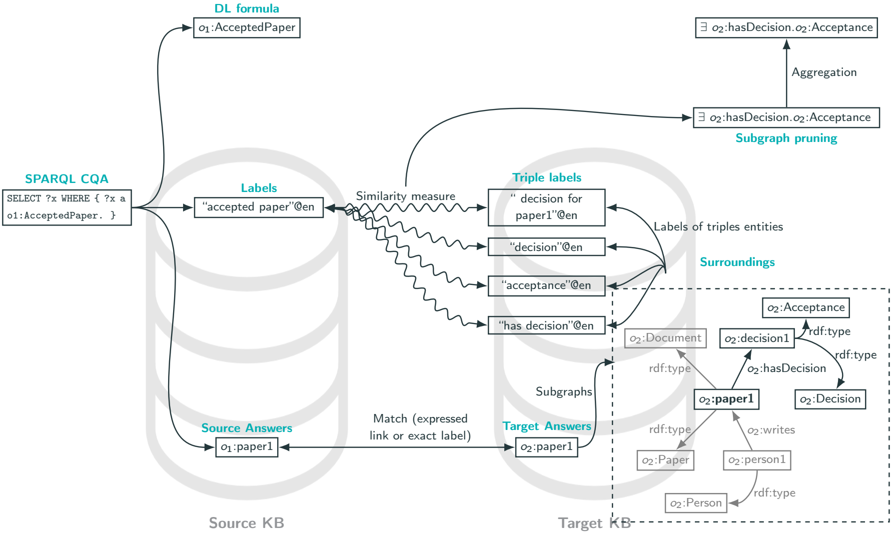


7 Compare the triple entities labels with the CQA labels



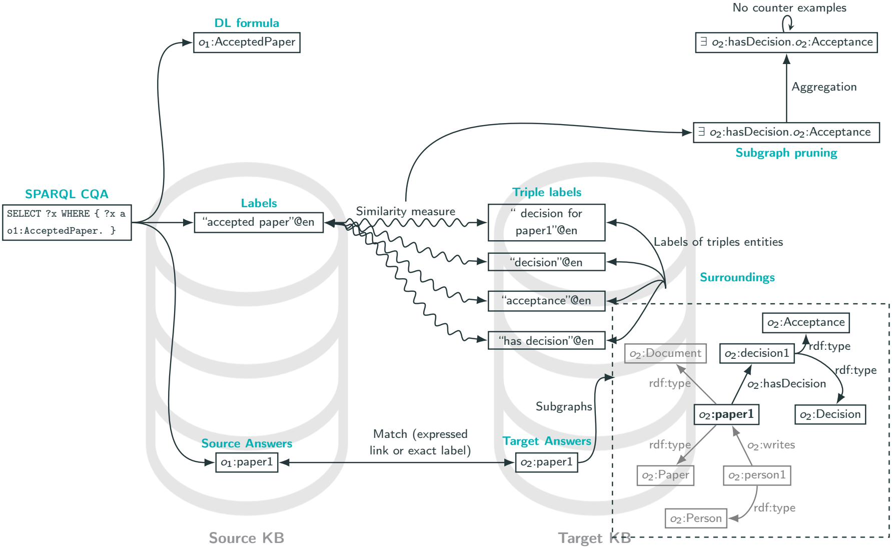
8 Prune the subgraph, transform it into a DL formula

Complex alignment generation based on CQAs



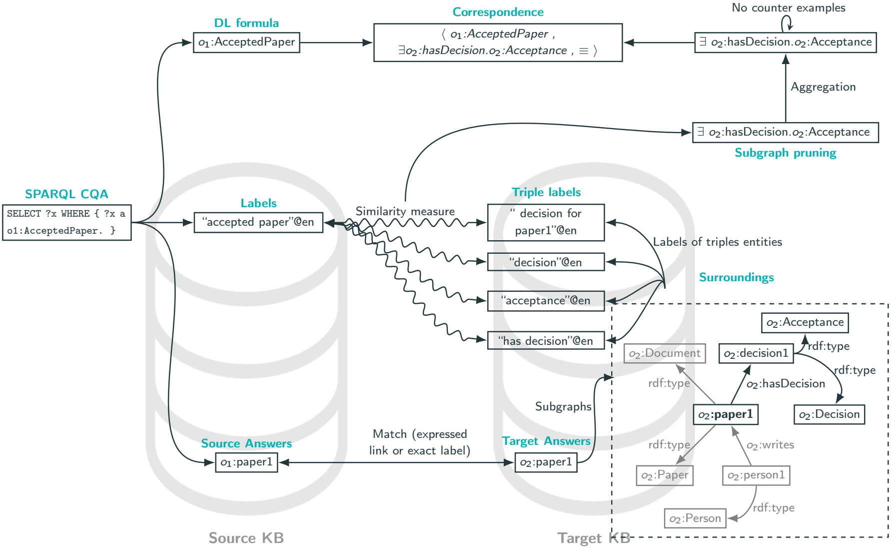
8 Aggregate the formula

Complex alignment generation based on CQAs



9 Look for counter-examples and compute the confidence value

Complex alignment generation based on CQAs



(10) Filter the formulae + (11) Generate correspondence

Comparison of instance sets I_{ref} and I_{ev} and different scoring functions

$$classical(I_{ref}, I_{ev}) = \begin{cases} 1 & \text{if } I_{ev} \equiv I_{ref} \\ 0 & \text{otherwise} \end{cases}$$

$$precision\ oriented(I_{ref}, I_{ev}) = \begin{cases} 1 & \text{if } I_{ev} \sqsubseteq I_{ref} \\ 0.5 & \text{if } I_{ev} \sqsupseteq I_{ref} \\ 0 & \text{otherwise} \end{cases}$$

$$query\ \ Fmeasure(I_{ref}, I_{ev}) = 2 \times \frac{QR \times QP}{QR + QP} \qquad QP = \frac{|I_{ev} \cap I_{ref}|}{|I_{ev}|} \qquad QR = \frac{|I_{ev} \cap I_{ref}|}{|I_{ref}|}$$

Others: *recall-oriented*, *overlap*, *non-disjoint*

CQA coverage

- Measures the overall coverage of the alignment with respect to the knowledge needs

$$coverage(A_{eval}, cqa_{pairs}, KB_s, KB_t, f) = \underset{\langle cqa_s, cqa_t \rangle \in cqa_{pairs}}{\text{average}} f(I_{cqa_t}^{KB_t}, I_{bestq_t}^{KB_t}) \quad (1)$$

Intrinsic precision

- Balancing strategy consists in calculating the intrinsic alignment precision based on common instances

$$precision(A_{eval}, KB_s, KB_t, f) = \underset{\langle e_s, e_t \rangle \in A_{eval}}{\text{average}} f(I_{e_s}^{KB_s}, I_{e_t}^{KB_t}) \quad (2)$$

Matcher implemented in Java under GNU LGPL v2.1

https://framagit.org/IRIT_UT2J/ComplexAlignmentGenerator

Evaluation system implemented in Java under GNU LGPL v2.1

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

2 evaluation datasets

- OAEI dataset about conference organisation
- 4 knowledge bases about plant taxonomy (species classification)

4 ontologies which describe the classification of species:

- AgronomicTaxon [Roussey et al., 2013]
- AgroVoc [Caracciolo et al., 2012]
- DBpedia [Auer et al., 2007]
- TaxRef-LD [Michel et al., 2017]

Version	AgronomicTaxon	AgroVoc	DBpedia	TaxRef-LD
Taxa (original)	32	8,077	306,833	570,531
Plant taxa (reduced)	32	4,563	58,257	47,058

6 CQAs from AgronomicTaxon competency questions.
Uneven population: manual evaluation

Tested	Nb ans.	Lev. thr.	Inst. matching	Co.-ex.	CQAs
v1	1	0.4	owl:sameAs then labels		✓
v10	10	0.4	owl:sameAs then labels		✓

	v1	v10
runtime	28h	32h
nb. corr.	134	328
Precision	0.3-1	0.3-1
CQA Coverage	0.3-0.7	0.5-0.8

Worst values
Best values

Because of the uneven population, more support instances entail a better CQA Coverage

- Works with only **1** common instance
- Depends on the quality of the instance matches
- Depends on the evenness of the instances
- Extremely long runtime

Short-term perspectives

Investigate linguistic similarities (lemmatisation, disambiguation, synset distance)

Improve instance matching step

Long-term perspectives

Community-driven ontology matching (each user's CQAs grows the alignment between ontologies)
Also comes with visualisation, validation and edition of correspondences Mixing the approach and instance matching techniques based on complex alignments

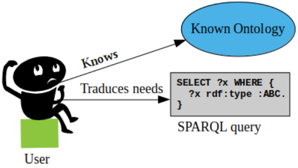
Plan

An approach to cross information based on SPARQL query rewriting

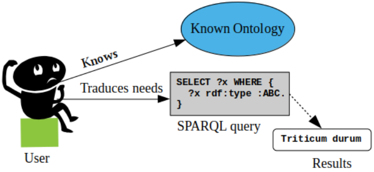
SPARQL

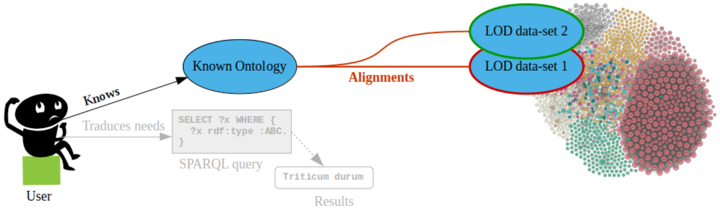
- Used for querying LOD data-sets
- Query from the ontology terms

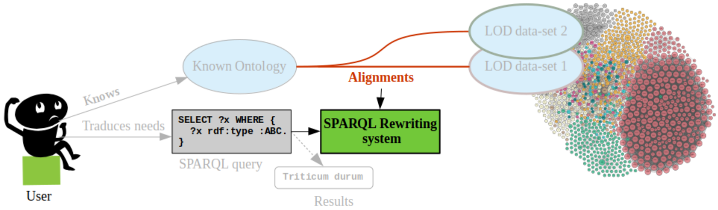
Query LOD datasets - Context

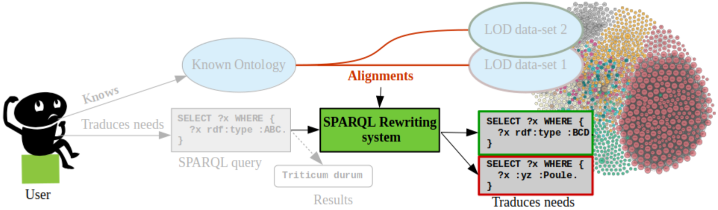


Query LOD datasets - Context

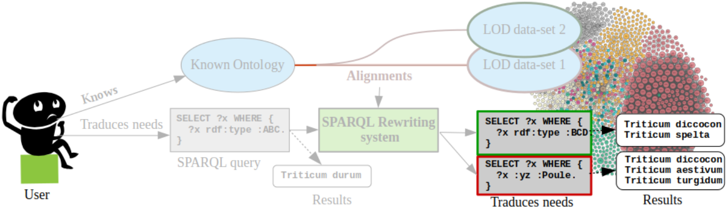



Knows
Known Ontology
Alignments
LOD data-set 2
LOD data-set 1
Traduces needs
SELECT ?x WHERE {
?x rdf:type :ABC.
}
SPARQL query
Triticum durum
Results
User


LOD data-set 2
LOD data-set 1
Known Ontology
Alignments
Knows
Traduces needs
SELECT ?x WHERE {
?x rdf:type :ABC.
}
SPARQL query
SPARQL Rewriting system
Triticum durum
Results
User


LOD data-set 2
LOD data-set 1
Known Ontology
Knows
Alignments
Traduces needs
SELECT ?x WHERE {
?x rdf:type :ABC.
}
SPARQL Rewriting system
SELECT ?x WHERE {
?x rdf:type :BCD.
}
SELECT ?x WHERE {
?x :yz :Poule.
}
SPARQL query
Triticum durum
Results
User
Traduces needs

Query LOD datasets - Context



Original query (AgronomicTaxon)

```
SELECT DISTINCT ?specy WHERE {
    ?taxon agro:prefScientificName ?label.
    ?taxon agro:hasLowerRank ?specy.
    ?specy rdf:type agro:Taxon.
FILTER (regex(?label, "^triticum$","i")).}
```

Rewritten query (Agrovoc)

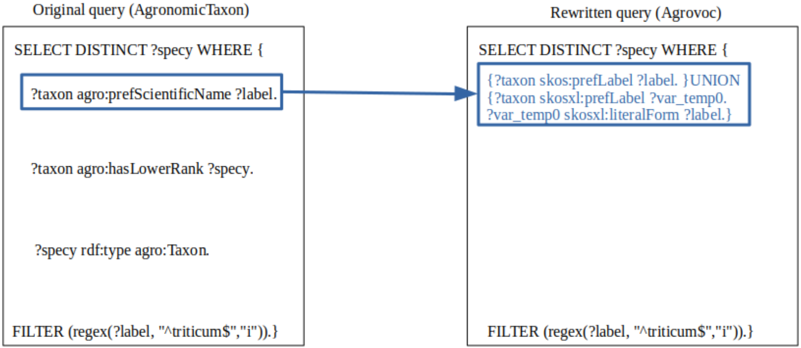
SPARQL Rewriting system [Thiéblin et al., 2016]

Original query (AgronomicTaxon)

```
SELECT DISTINCT ?specy WHERE {
  ?taxon agro:prefScientificName ?label.
  ?taxon agro:hasLowerRank ?specy.
  ?specy rdf:type agro:Taxon.
FILTER (regex(?label, "^triticum$","i")).}
```

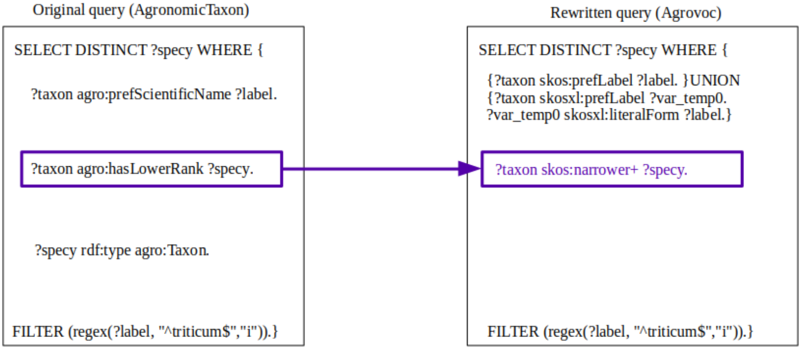
Rewritten query (Agrovoc)

```
SELECT DISTINCT ?specy WHERE {
  FILTER (regex(?label, "^triticum$","i")).}
```



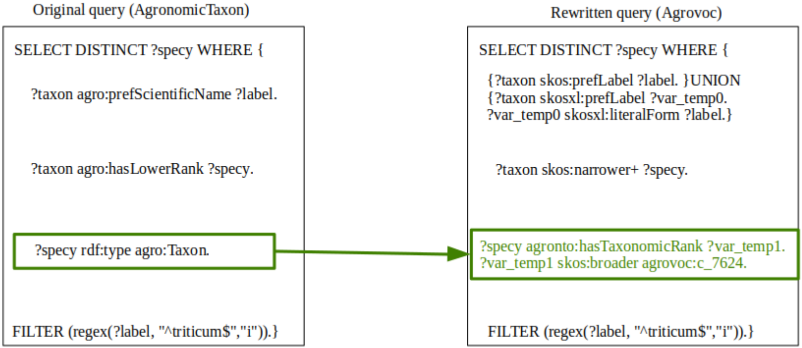
$\forall x,y$, agro:prefScientificName $\leq$ skos:prefLabel(x,y) $\vee$ ($\exists$ z, skosxl:prefLabel(x,z) $\wedge$ skosxl:literalForm(z,y))

SPARQL Rewriting system [Thiéblin et al., 2016]



∀x,y, agro:hasLowerRank(x,y) ≤ skos:narrower+(x,y)

SPARQL Rewriting system [Thiéblin et al., 2016]



$\forall x$, agro:Taxon(x) $\equiv \exists y$, agronto:hasTaxonomicRank(x,y) $\wedge$ skos:broader(y,agrovoc:c 7624)

- Known ontology AgronomicTaxon
- Users' needs AgronomicTaxon's design competency questions
  - 5 needs from agronomy experts
- LOD datasets DBpedia, Agrovoc
- Alignment (1:n) correspondences :
  - AgronomicTaxon-DBpedia: 29 correspondences
  - AgronomicTaxon-Agrovoc: 31 correspondences
  - Only 6 simple correspondences !

What is the kingdom of the Triticum taxon ?

- ✔ Query successfully rewritten
- ✔ Same information in all datasets : *Plantae*

What is the kingdom of the Triticum taxon ?

- ✔ Query successfully rewritten
- ✔ Same information in all datasets : *Plantae*

What are the common names of the Triticum taxon in French/English ?

- ✔ Query successfully rewritten
- ✔ Information present in *DBpedia*
- ✖ Information missing in *Agrovoc*

What are the different wheat species ?

✔ Query sucessfully rewritten

- Different classifications
  - Taxa missing in some datasets
  - Subspecies distinction in Agrovoc

⇒ Different points of view, complementarity of the sources

What are the different wheat species ?

✔ Query sucessfully rewritten

- Different classifications
  - Taxa missing in some datasets
  - Subspecies distinction in Agrovoc

⇒ Different points of view, complementarity of the sources

What is the rank of the taxon Triticum ?

✖ Fail in the query rewriting process

- Expected answers
  - in AgronomicTaxon: class *agro:GenusRank*
  - in DBpedia: property *dbo:genus*
  - in Agrovoc: concept *agronto:c_11125*

⇒ Different types of entities: what are the semantics behind such correspondences ?

- Use natural language to SPARQL systems to generate the original query
- Class-instance correspondences: how to model them
  - *Genus* is a class in an ontology but an instance in an other
- Towards an ontology alignment repository ?

Contributors

Fabien Amarger (PhD student)

Pascal Gillet (Master student)

Ollivier Haemmerlé (UT2/IRIT)

Nathalie Hernandez (UT2J/IRIT)

Camille Pradel (PhD student)

Catherine Roussey (IRSTEA)

Élodie Thiéblin (PhD student)

Thank you !

Questions ?

Bibliography i

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007).
DBpedia: A Nucleus for a Web of Open Data.
In Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., and Cudré-Mauroux, P., editors, *The Semantic Web: The 6th International Semantic Web Conference ISWC and the 2nd Asian Semantic Web Conference ASWC*, volume 4825 of *LNCS*, pages 722–735, Busan, Korea. Springer Berlin Heidelberg.

Caracciolo, C., Stellato, A., Rajbahndari, S., Morshed, A., Johannsen, G., Keizer, J., and Jaques, Y. (2012).
Thesaurus maintenance, alignment and publication as linked data: the AGROVOC use case.
International Journal of Metadata, Semantics and Ontologies, 7(1):65.

Euzenat, J. and Shvaiko, P. (2013).
Ontology Matching, Second edition.
Springer Berlin Heidelberg, Berlin, Heidelberg.

Michel, F., Gargominy, O., Tercerie, S., and Faron-Zucker, C. (2017).
A Model to Represent Nomenclatural and Taxonomic Information as Linked Data.Application to the French Taxonomic Register, TAXREF.
In Algergawy, A., Karam, N., Klan, F., and Jonquet, C., editors, *Proceedings of the 2nd International Workshop on Semantics for Biodiversity (S4BioDiv 2017) co-located with 16th International Semantic Web Conference (ISWC 2017)*, volume 1933, Vienna, Austria. CEUR-WS.org.

Roussey, C., Chanet, J., Cellier, V., and Amarger, F. (2013).
Agronomic taxon.
In Christophides, V. and Vodislav, D., editors, *Proceedings of the 2nd International Workshop on Open Data, WOD 2013, Paris, France, June 3, 2013*, pages 5:1–5:4. ACM.

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2012).
The neon methodology for ontology engineering.
In Suárez-Figueroa, M. C., Gómez-Pérez, A., Motta, E., and Gangemi, A., editors, *Ontology Engineering in a Networked World.*, pages 9–34. Springer.

Thiéblin, É., Amarger, F., Haemmerlé, O., Hernandez, N., and Trojahn, C. (2016).
Rewriting select sparql queries from 1: n complex correspondences.
In *11th Workshop on Ontology Matching*.

Zamazal, O. and Svátek, V. (2017).
The Ten-Year OntoFarm and its Fertilization within the Onto-Sphere.
Web Semantics: Science, Services and Agents on the World Wide Web, 43:46–53.

Šváb Zamazal, O., Svátek, V., Berka, P., Rak, D., and Tomášek, P. (2005).
Ontofarm: Towards an experimental collection of parallel ontologies.
Poster Track of ISWC, 2005.

Conference dataset

OAEI dataset proposed in [Šváb Zamazal et al., 2005] and used a lot since [Zamazal and Svátek, 2017]

Population of 5 ontologies (cmt, conference, confOf, edas, ekaw)

Population based on 152 CQAs: equivalent population for ontologies covering the CQA

100 CQAs are kept for the evaluation

Evaluation outline – Conference dataset

Evaluated variant	Nb ans.	Lev. thr.	Inst. match	Co.-ex.	CQAs
baseline	10	0.4	owl:sameAs		✓
Levenshtein	10	0.0–1.0	owl:sameAs		✓
Support answers	1-100	0.4	owl:sameAs		✓
query	10	0.4	owl:sameAs		
Counter-examples	10	0.4	owl:sameAs	✓	✓

- **Path max length** 3 properties
- **Similarity metric** $sim(L_s, L_t) = \sum_{l_s \in L_s} \sum_{l_t \in L_t} strSim(l_s, l_t)$

$$strSim(l_s, l_t) = \begin{cases} \sigma \text{ if } \sigma > \tau \text{, where } \sigma = 1 - \dfrac{levenshteinDist(l_s, l_t)}{\max(|l_s|, |l_t|)} \\ 0 \text{ otherwise} \end{cases}$$

- **Formula filtering threshold** confidence value > 0.6 or best formula

Impact of Levenshtein threshold

Evaluated variant	Nb ans.	Lev. thr.	Inst. match	Co.-ex.	CQAs
Levenshtein	10	0.0–1.0	owl:sameAs		✓

The higher the Levenshtein threshold, the more formulae are filtered out (not similar enough).

When Levenshtein threshold increases:

→ Stagnation of runtime

↘ Decrease of number of correspondences

↗ Increase of Precision

↘ Decrease of CQA Coverage

Impact of number of support answers

Evaluated variant	Nb ans.	Lev. thr.	Inst. match	Co.-ex.	CQAs
Support answers	1-10, 20, 100	0.4	owl:sameAs		✓

The higher the number of support answers, the more accidental correspondences appear.
Satisfying results with only 1 support answer.

When the number of support answers increases:

- $\nearrow$ Increase of runtime
- $\nearrow$ Increase of number of correspondences
- $\searrow$ Decrease of Precision
- $\rightarrow$ Stagnation of CQA Coverage

Impact of CQAs

Evaluated variant	Nb ans.	Lev. thr.	Inst. match	Co.-ex.	CQAs
baseline (CQAs)	10	0.4	owl:sameAs		✓
query	10	0.4	owl:sameAs		

Generated queries: instantiated classes, instantiated properties, attribute-value pairs.

	CQAs	queries
runtime	2h	2h
nb. corr.	1699	3098
Precision (query F-measure)	0.63	0.47
CQA Cov. (query F-measure)	0.76	0.64

Best values

Impact of counter-examples computing

Evaluated variant	Nb ans.	Lev. thr.	Inst. match	Co.-ex.	CQAs
no Counter-ex.	10	0.4	owl:sameAs		✓
Counter-ex.	10	0.4	owl:sameAs	✓	✓

Computing counter examples increases the Precision of the alignment.

	no Counter-ex.	Counter-ex.
runtime	2h	46h
nb. corr.	1699	1320
Precision (query F-measure)	0.63	0.74
CQA Cov. (query F-measure)	0.76	0.76

Worst values

Best values

Comparison with existing approaches/alignments

	baseline	Counter-ex.	Ritze 2010	AMLC	ra1	Onto merg.	Query rew.
corr. type[1]	(c:c)	(c:c)	(s:c)	(s:c)	(s:s)	(s:c)	(s:c)
runtime	2h	46h	1h	0h03			
nb. corr.	1699	1320	360	441	348	628	842
Precision[2]	0.3-1	0.4-1	0.8	0.4-0.6	0.6-1	0.4-1	0.4-1
CQA Cov.[3]	0.8	0.8	0.4	0.5	0.4	0.6	0.7

Worst values
Best values

[1] most complex correspondence form
[2] classical - not disjoint
[3] query Fmeasure