

# Tab2Onto: Unsupervised Semantification with Knowledge Graph Embeddings

Hamada M. Zahera, Stefan Heindorf, Stefan Balke, Jonas Haupt,  
Martin Voigt, Carolin Walter, Fabian Witter, Axel-Cyrille Ngonga Ngomo

## Highlights

- ❖ Tabular data lacks semantic information
- ❖ Learning Ontology by **embedding-based clustering** and **human-in-the-loop**.
- ❖ Ontology organizes data into **hierarchical classes**; simplifies data integration and automatic reasoning

## Introduction

- Data is collected from various sources to improve business analytics
- In most cases, this data comes in tabular format e.g. CSV, Spreadsheet
- In tabular data, the lack of semantic information makes it difficult for machines to understand their meaning
- **Semantification** is the process of converting data into a representation with unique semantics, e.g., an RDF knowledge graph

## Research Questions?

- Which KG embeddings yield the best clusters of entities in the embedding space? (→ Table 1)
- Which clustering approach yields the best clusters of entities? (→ Table 1)
- How well does our pipeline work for the semantification of tabular data? (→ Table 2)

## Tab2Onto Approach

- Construct a KG from tabular data
- Represent entities & relations using **knowledge graph embeddings**
- Clustering** similar entities together
- Labeling clusters by annotating **sampled entities** using **LabENT<sup>1</sup>**
- Ontology generation

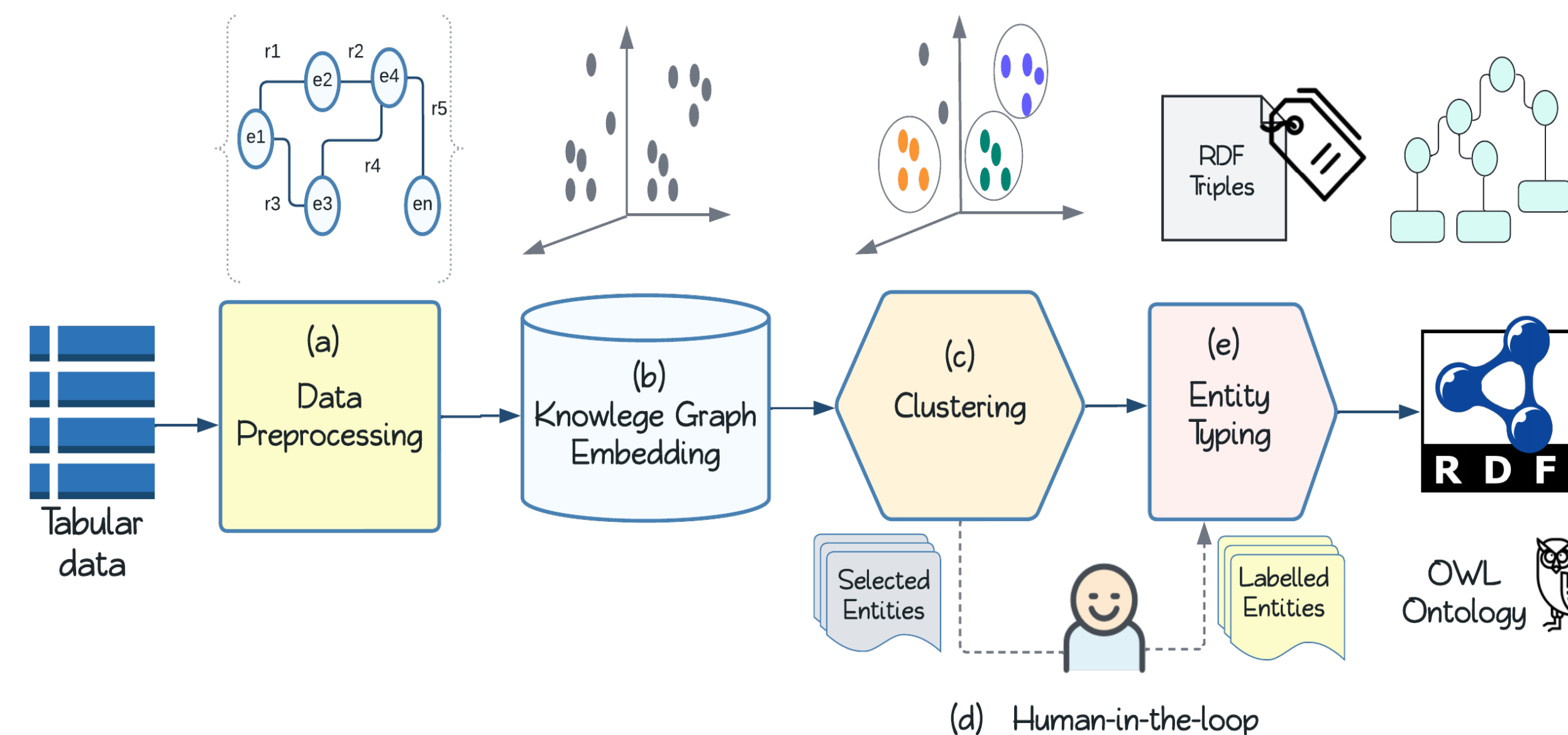


Fig.1 Tab2Onto Pipeline for Semantification

Table 1: Clustering for type prediction on FB15k-237. Best results in bold.

Algorithm	TRANSE		DISTMULT		ROTATE		QMULT	
	Acc.	F <sub>1</sub>	Acc.	F <sub>1</sub>	Acc.	F <sub>1</sub>	Acc.	F <sub>1</sub>
K-Means	0.784	0.751	0.771	0.741	0.282	0.200	<b>0.785</b>	<b>0.803</b>
Agglomerative	0.779	0.746	0.781	0.749	0.284	0.201	0.744	0.775
HDBSCAN	0.678	0.624	0.475	0.362	0.276	0.119	0.276	0.119

<sup>1</sup><https://github.com/dice-group/LabENT>

## Use Case:

### Lymphography Semantification

- **Input:** CSV file of lymphography data
- **Output:** OWL ontology → a taxonomy of OWL classes: Normal, Fibrosis, Metastases, Malign-Lymph

Table 2: Tab2Onto semantification of *Lymphography* with QMULT embeddings and K-Means clustering.

Approach	Acc.	F <sub>1</sub>
Tab2Onto ( <i>unsupervised</i> )	<b>0.666</b>	<b>0.728</b>
Random ( <i>unsupervised</i> )	0.533	0.485
Logistic regression ( <i>supervised</i> )	0.833	0.818

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xml:base="http://daikiri-semantification.de/onto.owl"
  xmlns="http://daikiri-semantification.de/onto.owl#">

  <owl:Ontology rdf:about="http://daikiri-semantification.de/onto.owl"/>

  <owl:Class rdf:about="#Normal">
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
  </owl:Class>

  <owl:Class rdf:about="#Fibrosis">
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
  </owl:Class>
```

Fig.2 OWL Ontology of Lymphography