OLAF : Ontology Learning Applied Framework

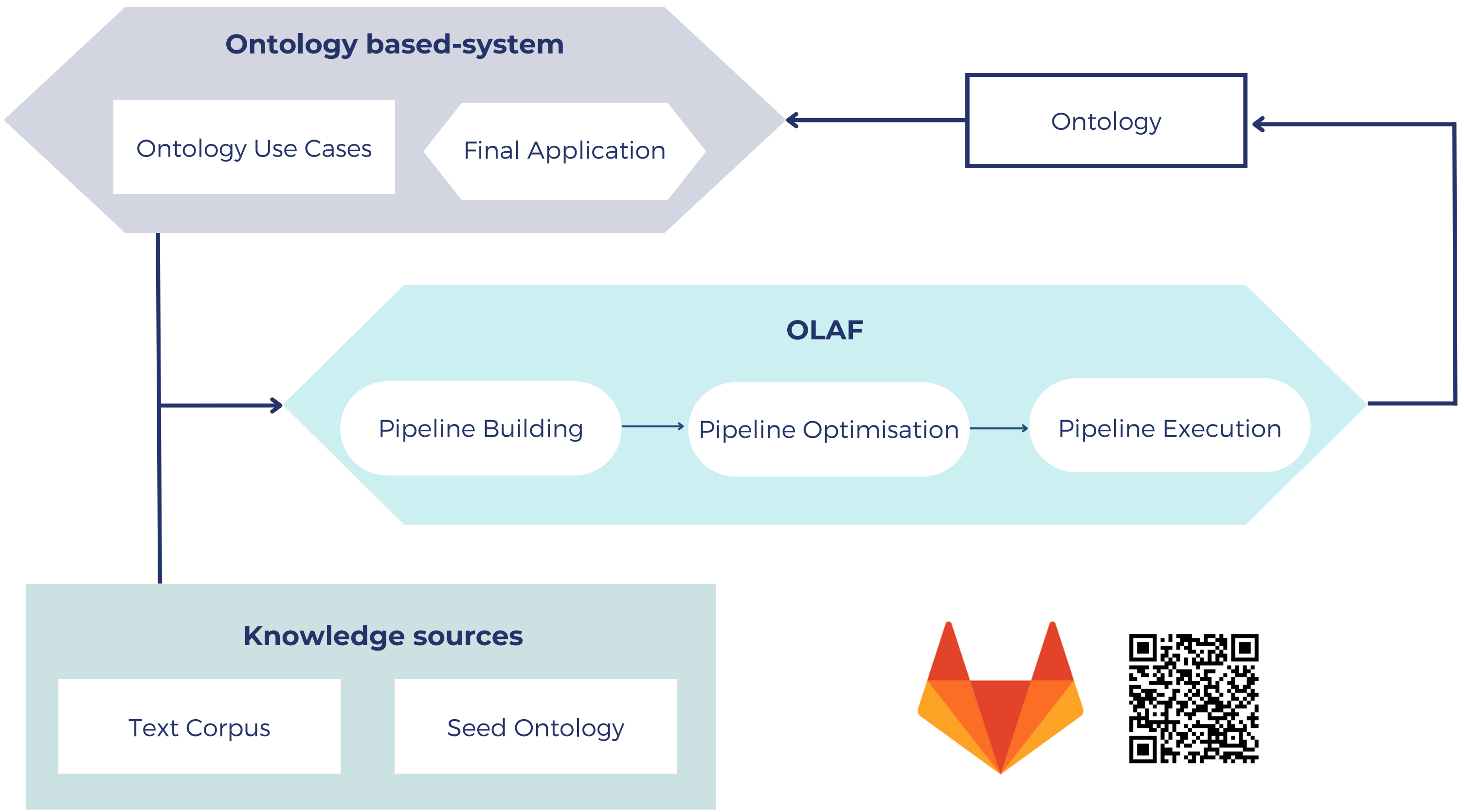
Marion SCHAEFFER (marion.schaeffer@insa-rouen.fr) - Matthias SESBOUE (matthias.sesboue@insa-rouen.fr)
Jean-Philippe KOTOWICZ - Nicolas DELESTRE - Cecilia ZANNI-MERK

Since the beginning of the century, research on ontology learning has gained popularity. Automatically **extracting and structuring knowledge** relevant to a domain of interest from unstructured data is a major scientific challenge. We propose a new approach with a **modular ontology learning framework** considering tasks from data pre-processing to axiom extraction. Whereas previous contributions considered ontology learning systems as tools to help the domain expert, we developed the proposed framework with **full automation** in mind. An implementation as an **open-source and collaborative python library** is available at https://gitlab.insa-rouen.fr/msesboue/ontology-learning.

STATE OF THE ART

System	Overview	Pros and cons
Text2Onto, 2005, [1]	It is the reference in the field as it defines a representation-agnostic structure with modular steps and takes into account uncertainty. The system is implemented as a GATE module.	Ontologies can be exported in various formats. GATE system adds great visualisations. But it is not maintained since 2011.
OntoGain, 2010, [2]	It focuses on multiword terms to construct a "lexicalised ontology" by adapting an agglomerative clustering and an FCA method. It implements 4 steps: text preprocessing, concept extraction (C/NC-value), taxonomy construction, and non-taxonomic relation acquisition (rule-based and probabilistic).	It considers only multiword terms and relies on WordNet and POS tags. It does not distinguish between terms and concepts and implements different adaptable approaches.
OntoLearn (Reloaded), 2013, [3]	It focuses on "lexicalised ontologies" and uses seed knowledge. It implements 5 steps: terminology extraction, hypernym graph construction, domain filtering of hypernyms, hypernym graph pruning and edge recovery.	It relies on WordNet and POS tags and does not distinguish between terms and concepts. It implements different adaptable approaches.

OLAF IN A PRACTICAL CONTEXT

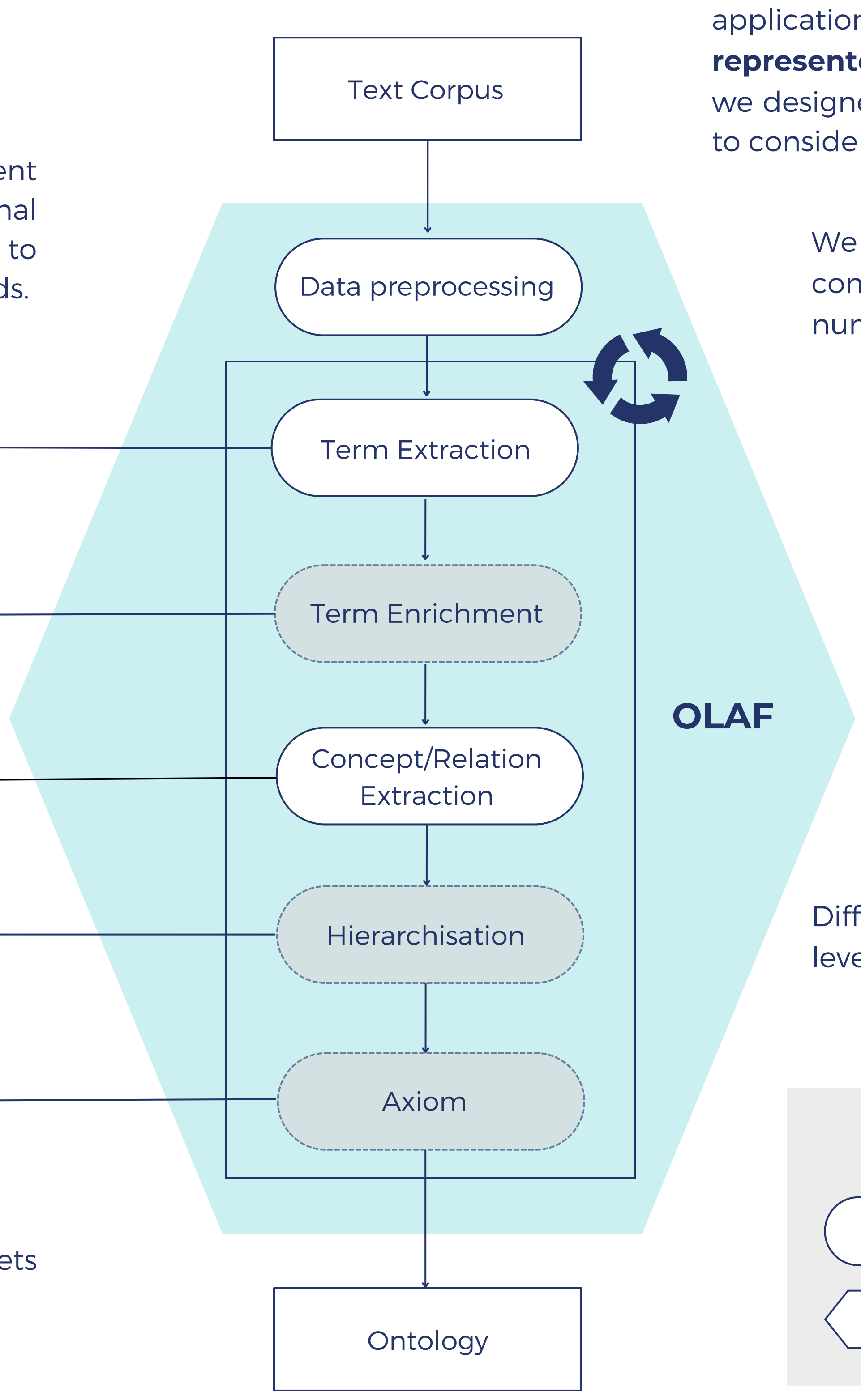


Ontology based-system
Ontology Use Cases
Final Application
Ontology
OLAF
Pipeline Building
Pipeline Optimisation
Pipeline Execution
Knowledge sources
Text Corpus
Seed Ontology

Most ontology learning systems do not consider the targeted ontology-based system. Though an ideal ontology should model a domain in an application-independent manner, in practice, **concepts and relations represented largely depend on one or more business use cases**. As we designed our framework with industry application in mind, we need to consider it within its **real-world usage context**.

ONTOLOGY LEARNING FRAMEWORK ARCHITECTURE

Our framework provides several algorithms for the different stages of the pipeline. The algorithms are taken from external libraries or directly implemented in the framework. The goal is to have as many methods as possible to cover the maximum needs.



Text Corpus
Data preprocessing
Term Extraction
Term Enrichment
Concept/Relation Extraction
Hierarchisation
Axiom
Ontology
OLAF

C-value-based filtering
Linguistic-based filtering
TF-IDF value-based filtering

Embedding-based similar term extraction
ConceptNet synonym extraction
WordNet synonym extraction

ConceptNet-based extraction
Grouping terms based on synonyms
Term cooccurrences-based extraction
Similarity-based extraction
Formal concept Analysis

Term subsumption algorithm
Hierarchical clustering

Rule-based axiom extraction
Inductive Logic Programming

We only work on **unstructured textual data**.
We apply the framework in two different use cases and datasets to validate our results :

- a search engine on Schneider Electric products
- a chatbot on Human Resources issues.

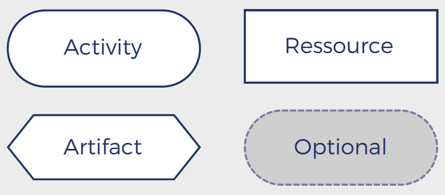
We choose **Python** as it eases access to the vast python community and its library ecosystem, particularly **NLP tools** and numerous **Machine Learning (ML) libraries**.

Our implementation is largely based on the **Python NLP library spaCy**. The text processing on spaCy helps us work with data in **many different languages** while staying flexible on the methods used. The only constraint is to end up with a list of **spaCy Doc objects**.

Our vision is to implement a **toolbox of methods** we can gather to build **pipelines**. These pipelines can be run, optimised and analysed to learn the best possible ontology.

Different **serialization techniques** can be used to export and leverage the learned ontology in an application system.

CAPTION



Activity
Ressource
Artifact
Optional

Algorithm implemented
Upcoming implementation
: Iterative process

We designed the proposed framework focusing on **automation** with very little, if any, human involvement in mind. Unlike most existing approaches, particular attention is brought to the **learned ontology final production use case**. We implement the framework as an open-source and open-access python library. We aim to **gather feedback and grow a community** to develop and test multiple algorithms. Various satellite tools could be developed to enhance the framework implementation. However, we should focus on developing **axiom extraction** and **automatic ontology evaluation**. One exciting research area might be the adaptation of the software industry's "DevOps" concepts to knowledge management. The latter field is known as "SemOps".

1. Cimiano P, Völker J. Text2Onto. Natural Language Processing and Information Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005.p. 227-238. ISBN: 978-3-540-32110-1
2. Drymonas E, Zervanou K, Petrakis EGM. Unsupervised Ontology Acquisition from Plain Texts: The OntoGain System. Natural Language Processing and Information Systems. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010. p. 277-87. ISBN: 978-3-642-13881-2
3. Paola Velardi, Stefano Faralli, Roberto Navigli; OntoLearn Reloaded: A Graph-Based Algorithm for Taxonomy Induction. Computational Linguistics 2013; 39 (3): 665–707. DOI: 10.1162/COLI_a_00146
4. Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood, Hafiza Mahnoor Abbasi, A survey of ontology learning techniques and applications, Database, Volume 2018, 2018, bay101, DOI: 10.1093/database/bay101