
Normandie Université


UNIVERSITÉ
LE HAVRE
NORMANDIE


INSA
INSTITUT NATIONAL
DES SCIENCES
APPLIQUÉES
ROUEN NORMANDIE


UNIVERSITÉ
DE ROUEN
NORMANDIE

Knowledge Graph-based System for Technical Document Retrieval

A deductive reasoning-focused exploration

Matthias Sesboüé

September 5, 2024

Table of content

RESPONDING THESIS

Knowledge Graph-based System for Technical Document Retrieval

A deductive reasoning-focused exploration

- Research objective: Leveraging domain knowledge to enhance Information Retrieval in a technical context.
- CIFRE contract (financed by ANRT) between the Litis lab and the company Traceparts
- Began on March 15th 2021.

From a keyword-based search to a concept-based one.

Traceparts

One of the world's leading CAD-content platforms for Engineering, Industrial Equipment and Machine Design. The CAD-content platform traceparts.com provides access to over 1.8 thousand supplier-certified product catalogues with 2D drawings, 3D CAD models and product datasheets.

- Technical content aimed at an engineering audience from multiple industries
- Content available in 25 languages
- Users can search using :
  - A full text search
  - A list of catalogues
  - Different classifications

traceparts
Product Content Everywhere™

Information Retrieval

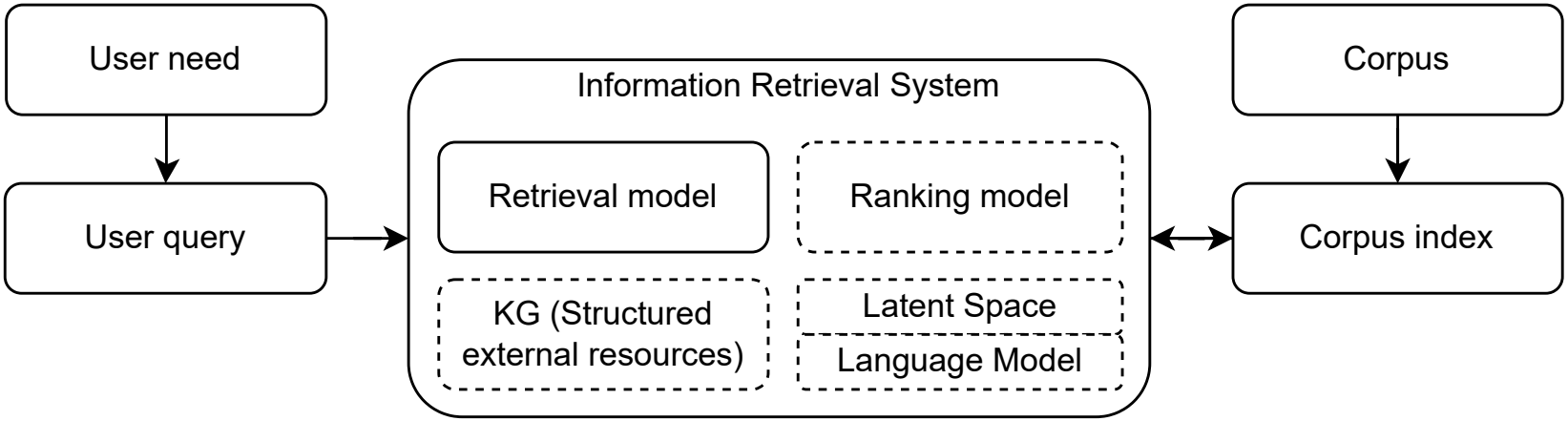


Figure: Information Retrieval System overview

Traditional approaches leverage statistics about the text corpus. Recent methods implement deep learning models and combines multiple approaches.

BM25

BM25 (and its many variants) is:

- based on the Term Frequencies and Inverse Document Frequencies (TF-IDF)
- still widely used in practice
- computes many statistics offline

Traceparts search system is largely based on a BM25 implementation.

Knowledge Graph and ontology

Knowledge Graph (Hogan et. al. 2021):

> *a knowledge graph as a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities. The graph of data (aka. data graph) conforms to a graph-based data model, which may be a directed edge-labelled graph, a property graph, etc. By knowledge, we refer to something that is known.*

Ontology (Hogan et. al. 2021):

> *In the context of computing, an ontology is then a concrete, formal representation of what terms mean within the scope in which they are used (e.g., a given domain).*

In our work, we consider an ontology a particular component of a Knowledge Graph

Knowledge Acquisition

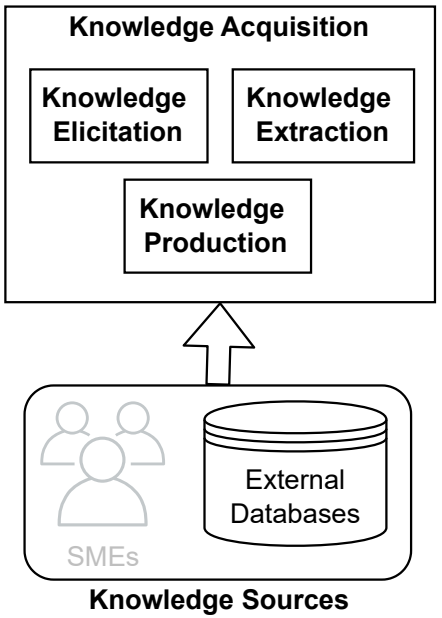


Figure: KGBS: Knowledge acquisition

Knowledge modeling

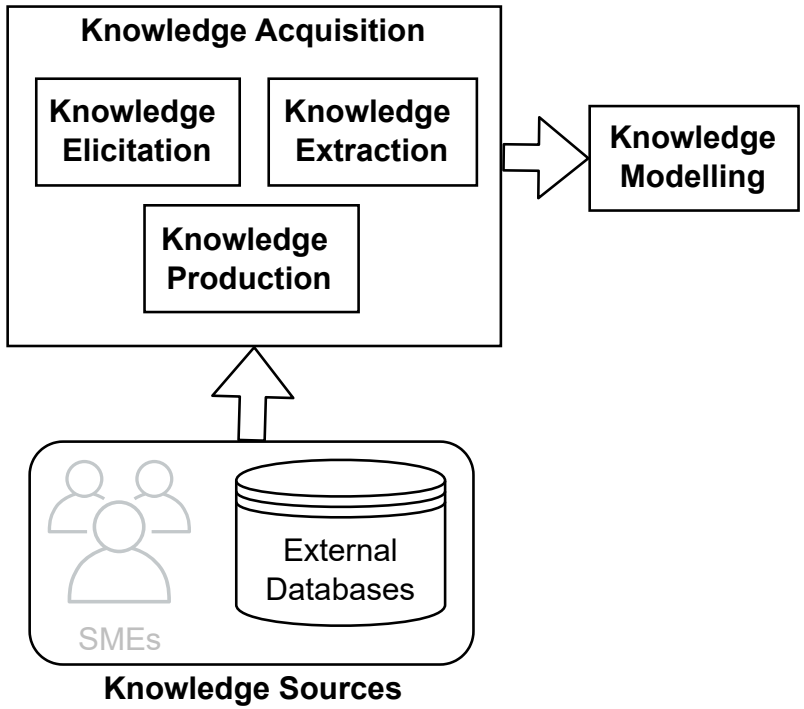


Figure: KGBS: Knowledge modeling

Knowledge Graph

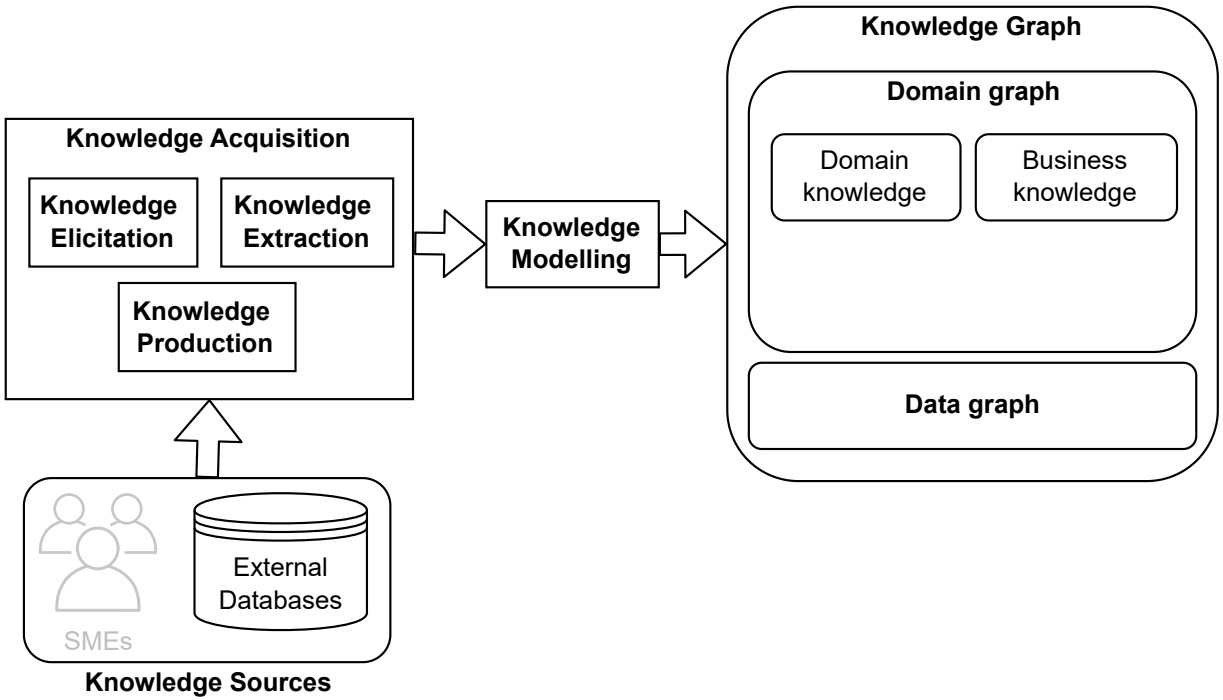


Figure: KGBS: Knowledge Graph

Knowledge consumption

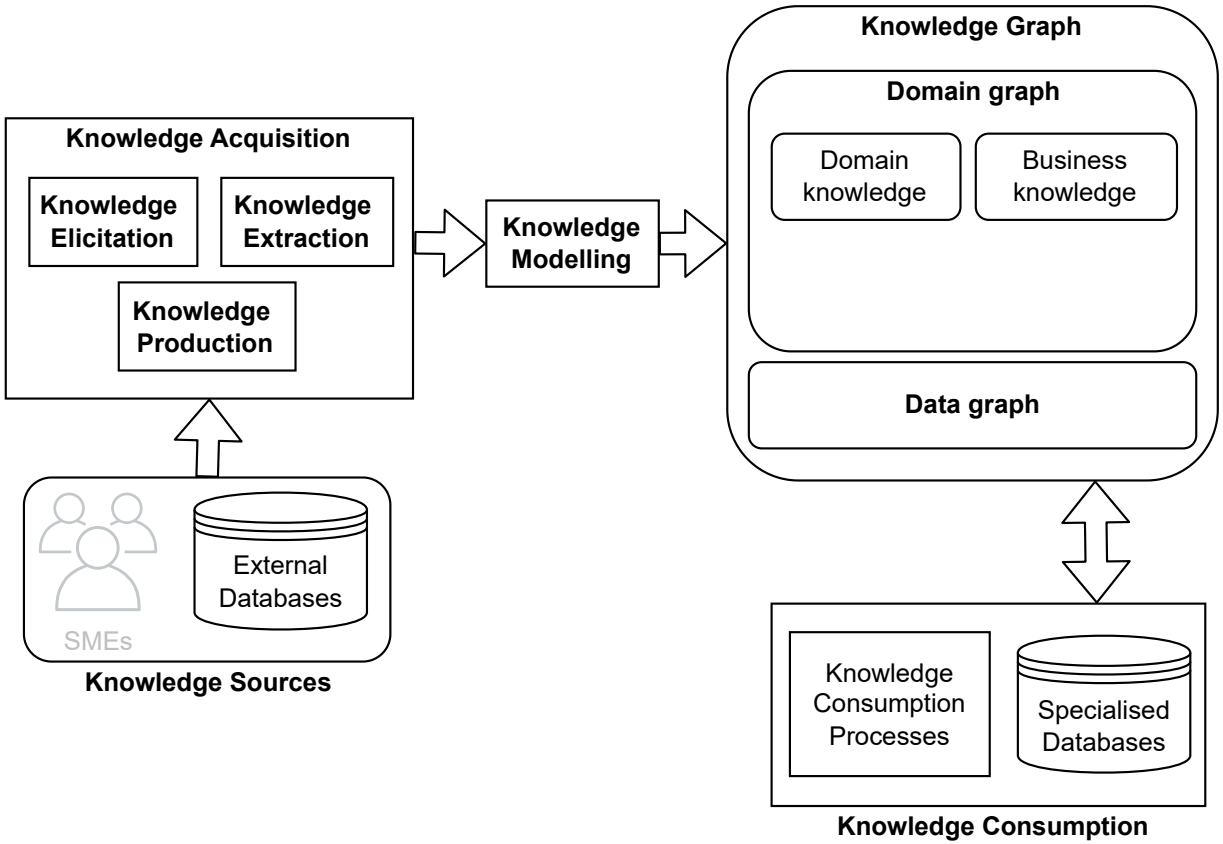


Figure: KGBS: knowledge consumption

Contribution: Knowledge Graph-Based System architecture

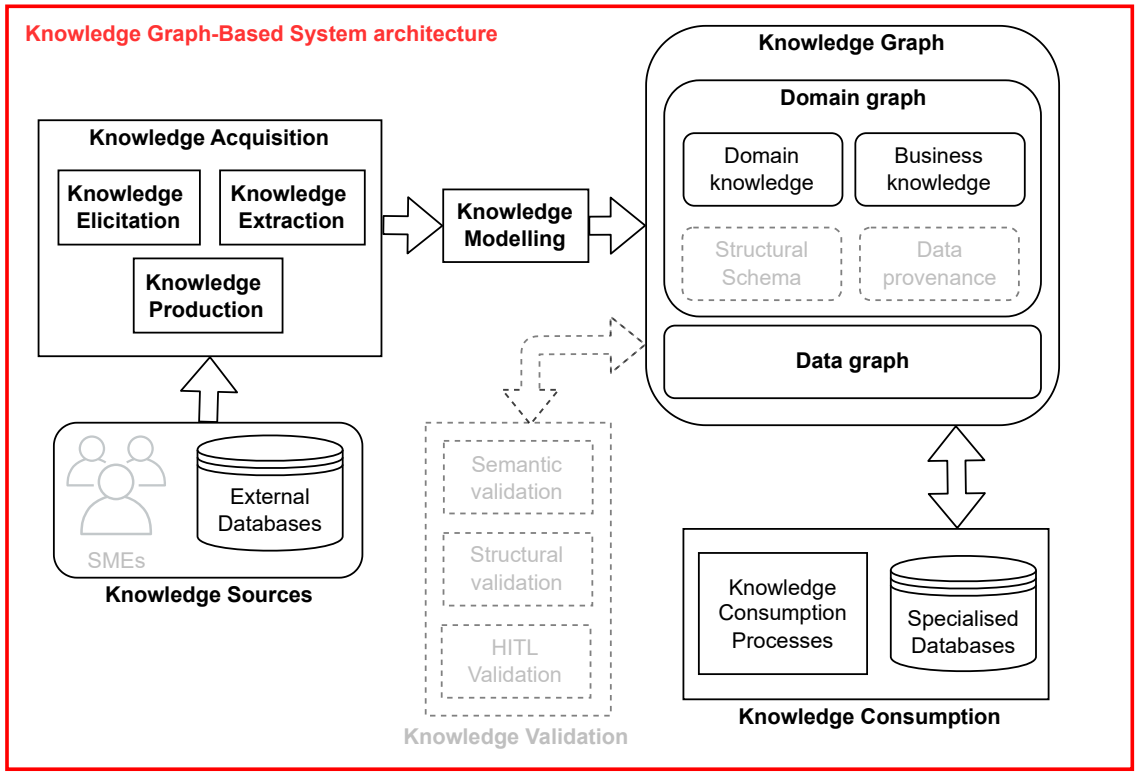


Figure: KGBS architecture

Knowledge Graph vs Ontology

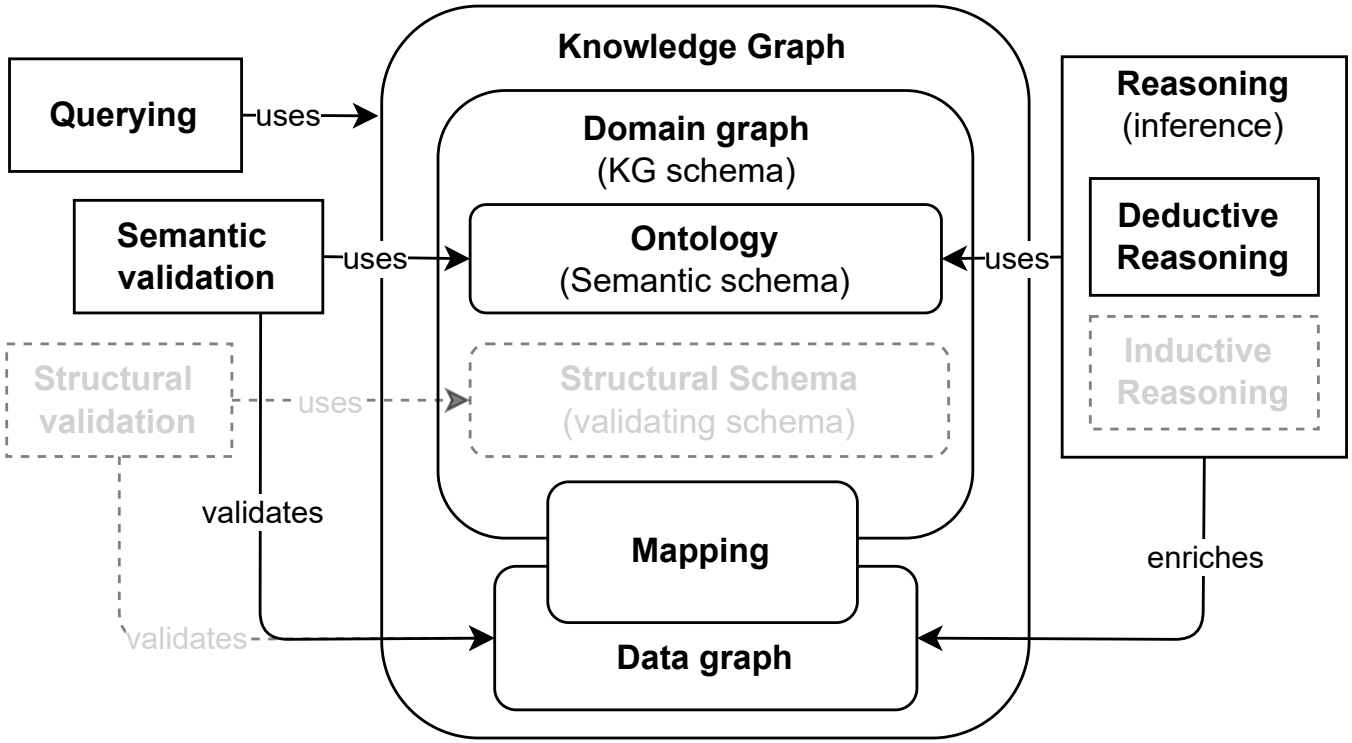


Figure: Knowledge Graph definition

Contribution: Ontology Learning Applied Framework (OLAF)

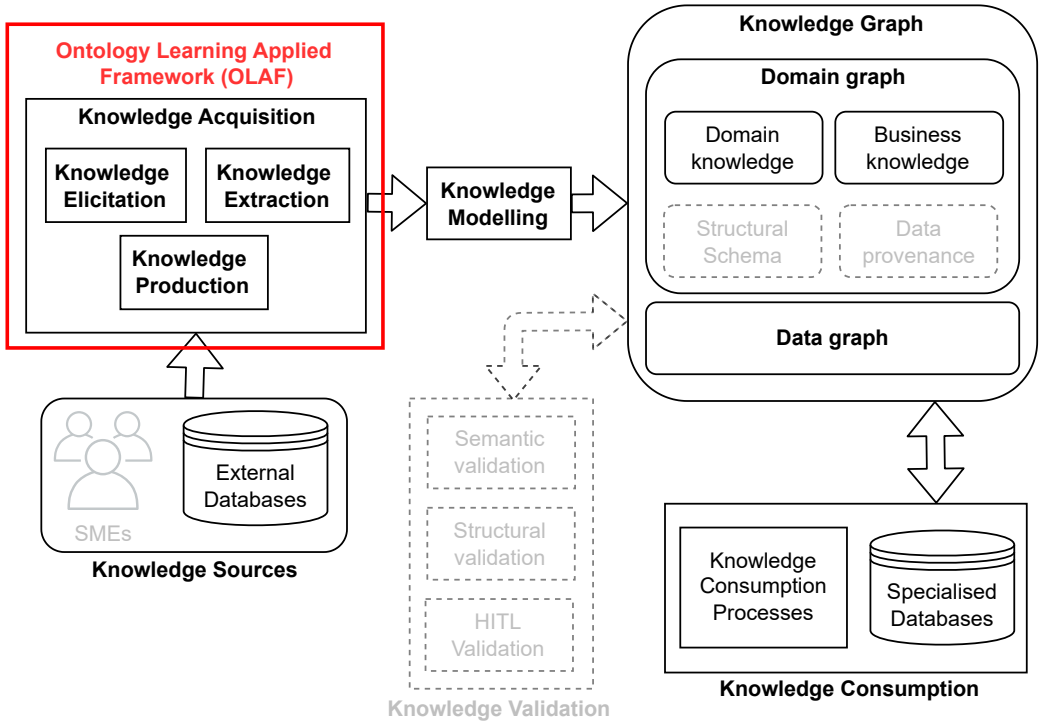


Figure: KGBS architecture: OLAF

Contribution: Information Retrieval ontology

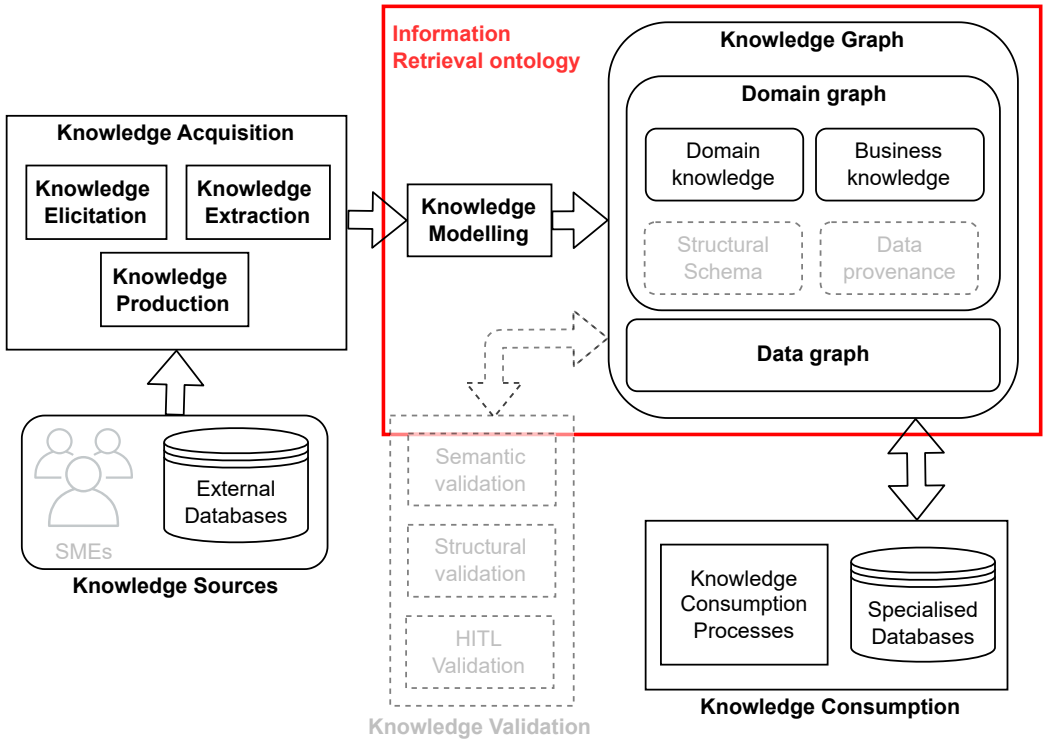


Figure: KGBS architecture: IR ontology

Contribution: KG-based IR system

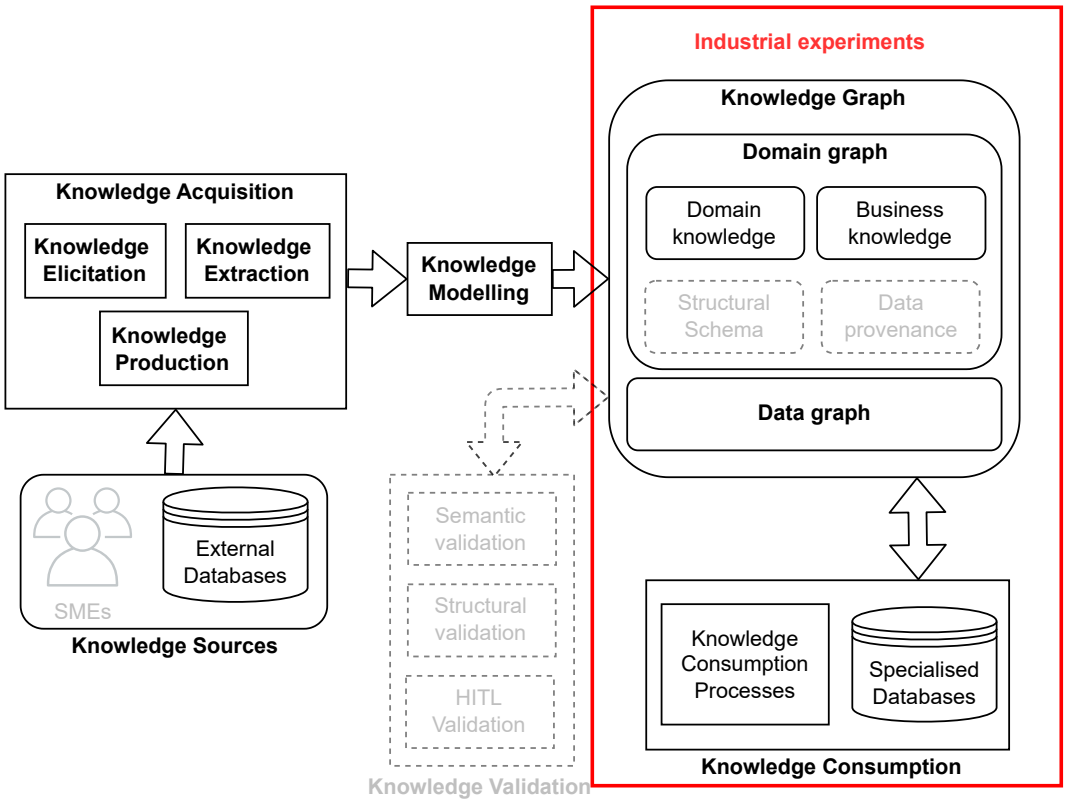


Figure: KGBS architecture: IR ontology

In this presentation

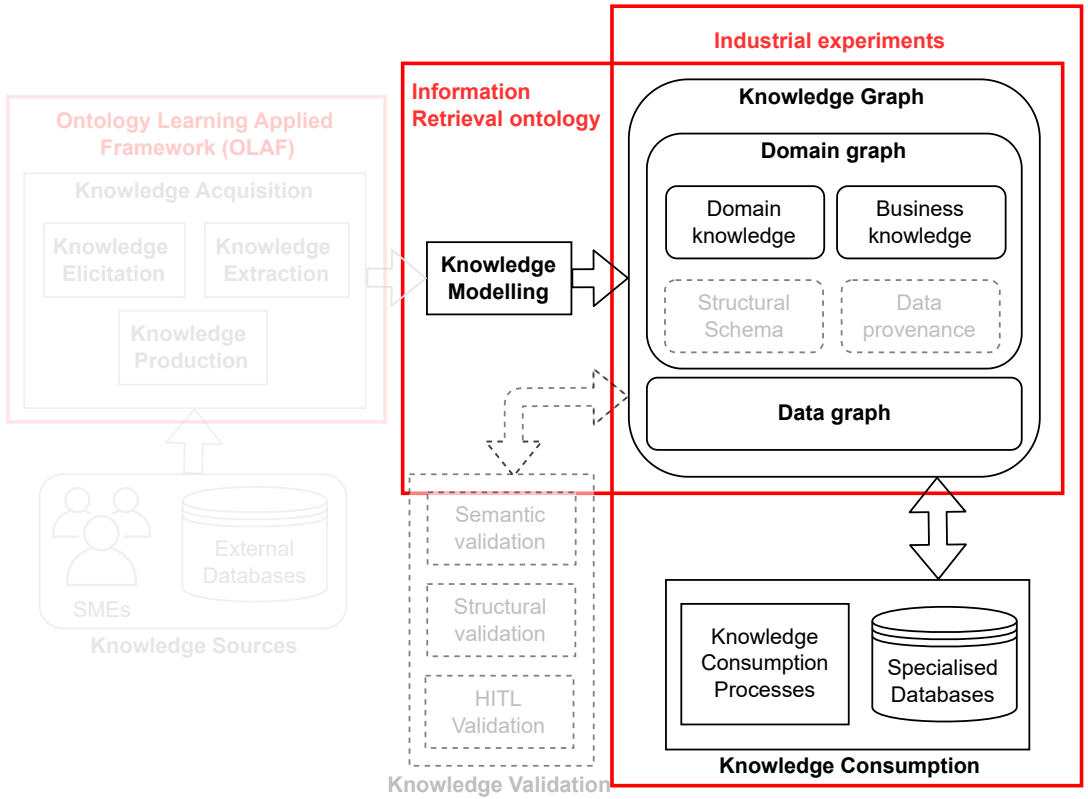


Figure: This presentation KGBS architecture components focus

Corpus

- Over 1.1 million Document Families
- Over 127.8 millions individual documents
- 25 languages
- Documents' texts contain average 50 characters and 7 words
- Over 210 thousand tags, amongst which:
  - Over 2.5 thousand suppliers and manufacturers
  - Over 1.9 thousand catalogues
  - Over 208 thousand categories

Some text content examples are:

- *DIN 912*
- *The P01 to P08 pumps are designed to pump lubricating fluids (oil, diesel oil, etc.). Their flow rate is from 1 to 24 L / min; maximum working pressure 10 bar.*

User searches

User text searches:

- are composed of domain-specific keywords, notations, identifiers, and acronyms.
- contain on average 13 characters separated into 2 words.
- can come in any languages

Some common search examples are:

motor, *din 912*, and *ball valve*.

Traceparts search system challenges

Traceparts search challenges come from:

- Short multilingual texts
- Technical texts with many synonyms, acronyms, homonyms, and notations
- A large and heterogeneous corpus
- Multiple engineering domains coverage
- High recall but low precision

Traceparts search system

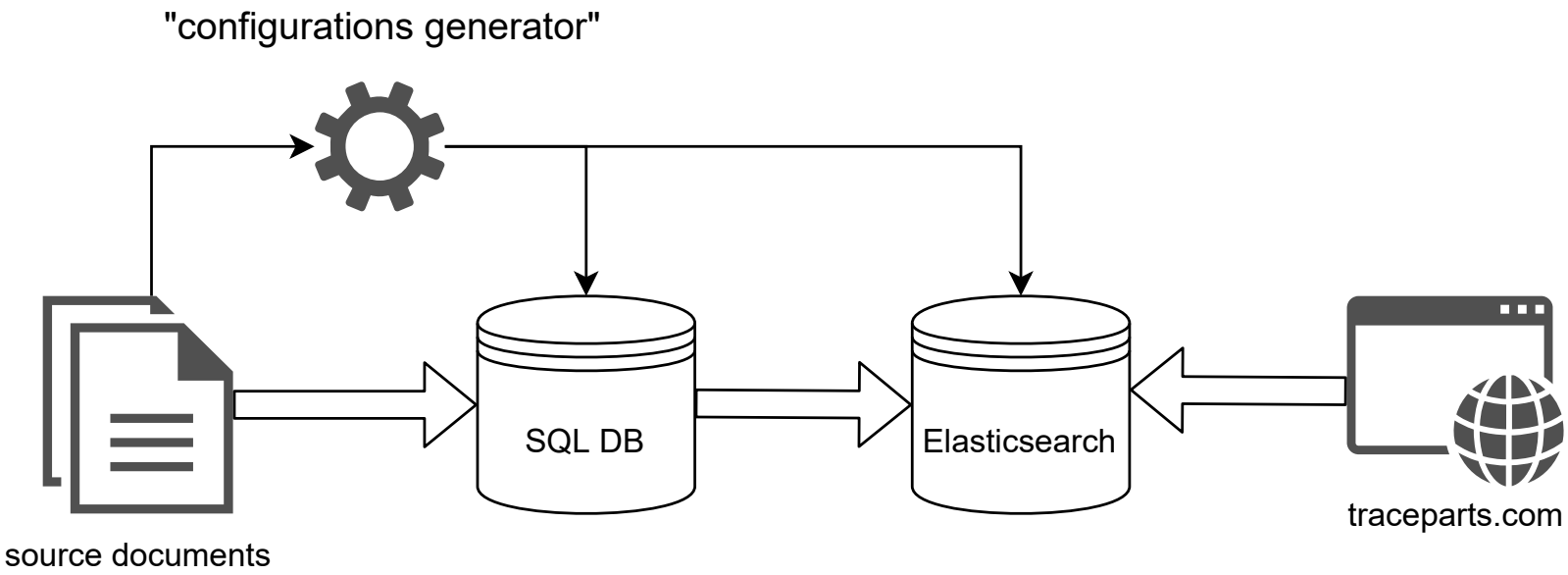


Figure: Traceparts current system

Parts configurations are generated with their text content to be searchable.

Experiments objective

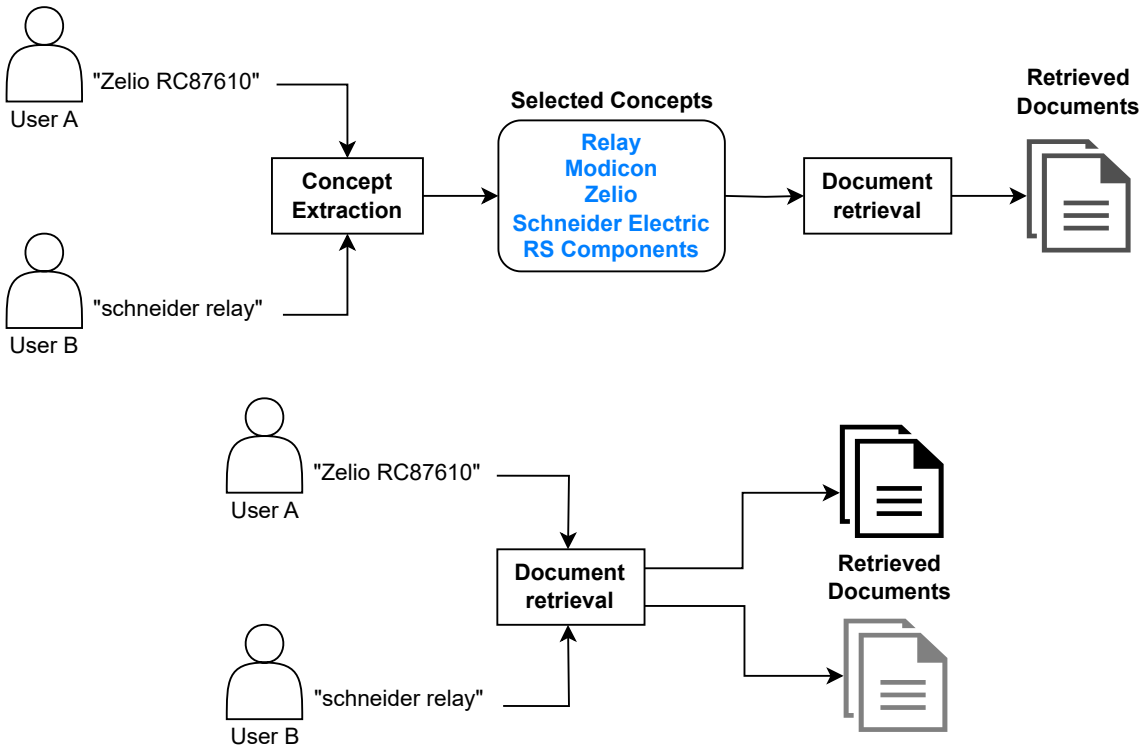


Figure: Text-based vs concept-based search.

Evaluation metrics

- Mean Average Precision at k (MAP@k):
  - A sliding (or growing) precision window, averaged over a set of query examples.
  - Ranges from 0 to 1 (1 is the best value).
  - Gives information about the amount and positions of positive results in the k first ones.
- Binary Mean at k (BM@k):
  - Binary average over a set of query examples.
  - Ranges from 0 to 1 (1 is the best value).
  - Provides information about the amount of queries with a positive result in the k first ones.
  - Does not give any detail on the positive result position.

Experiments protocol

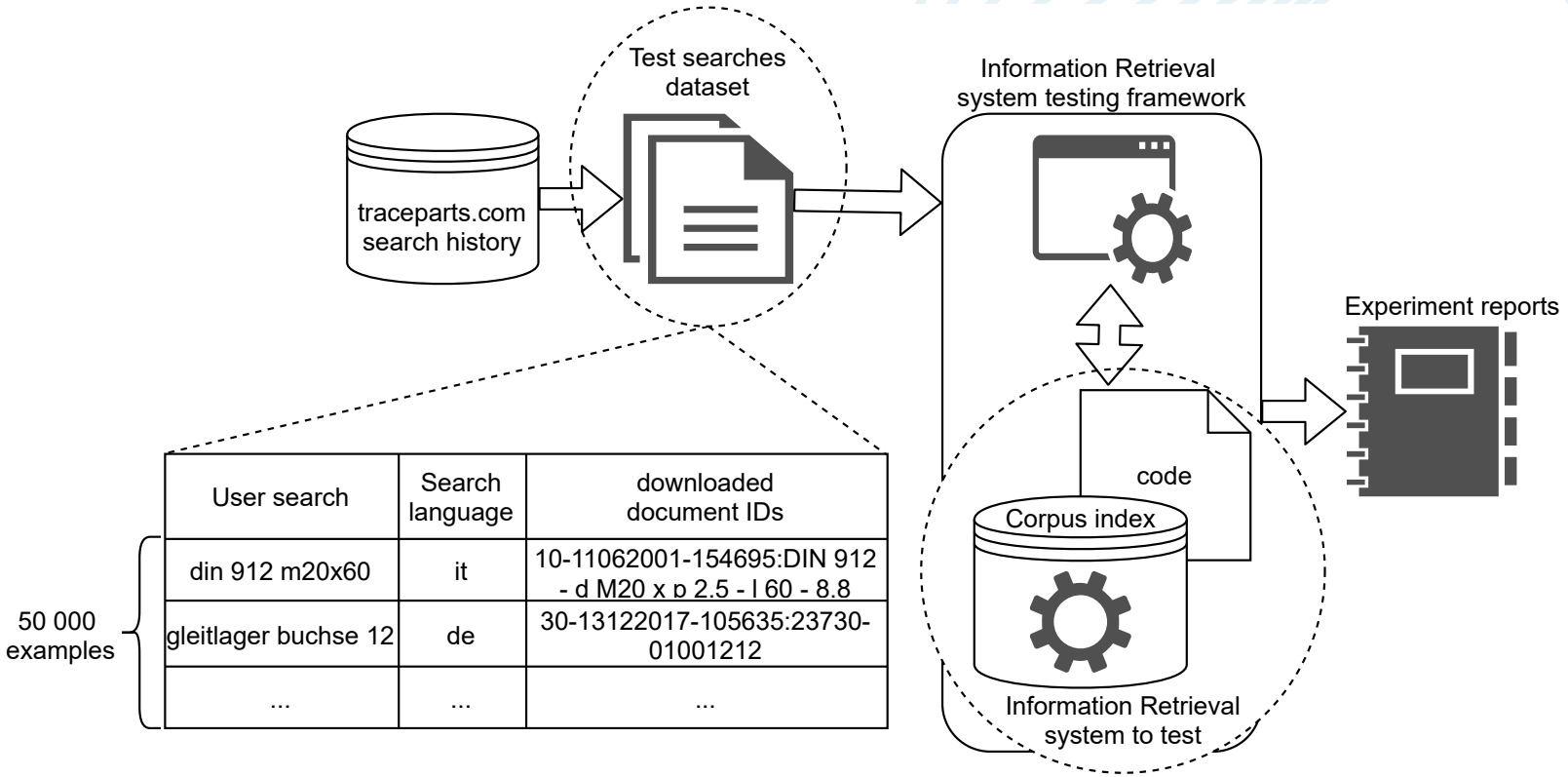


Figure: Experiments Protocol.

Experiments

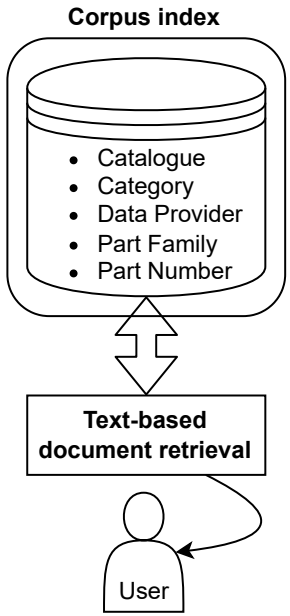
6 distinct systems built iteratively:

- *Text-based system (baseline)*
- **Concept-based system**
- Knowledge Graph-based system
- Text-based system with implicit knowledge
- Concept-based system with implicit knowledge
- **Knowledge Graph-based system with implicit knowledge**

Systems implementations:

- User search concept matching problem as an information retrieval task.
- Leverage user search history as implicit knowledge.
- Query concept enrichment as a graph traversal task.

Text-based system (baseline)

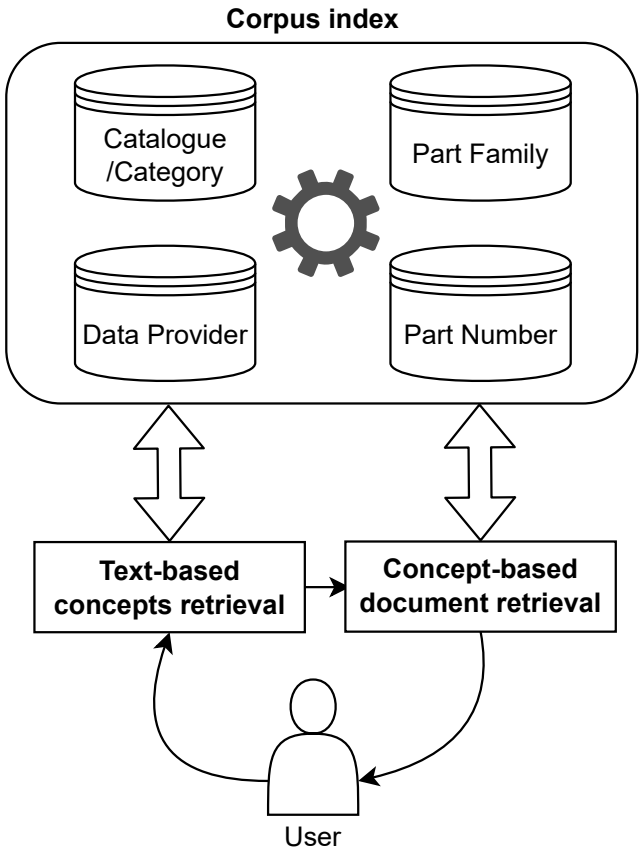


Text-based system (baseline) results

	Text-based system (baseline)	
@k ↓	**MAP@k**	
@5	0.061	0.114
@25	0.064	0.148
@50	0.064	0.157
@100	0.064	0.161
@350	0.064	0.164

Table: Text-based system (baseline) results for different k values.

Concept-based system

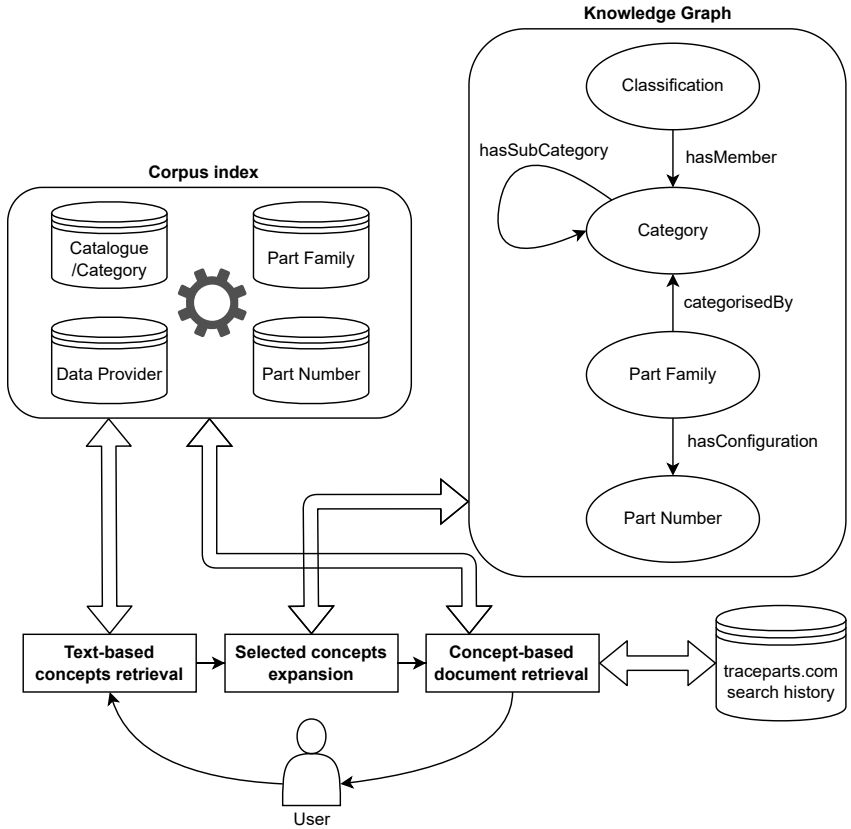


Concept-based system results

	Text-based system (baseline)		Concept-based system	
@k ↓	**MAP@k**	**BM@k**	**MAP@k**	**BM@k**
@5	0.061	0.114	0.152	0.243
@25	0.064	0.148	0.159	0.334
@50	0.064	0.157	0.160	0.371
@100	0.064	0.161	0.161	0.403
@350	0.064	0.164	0.161	0.429

Table: Text and concept-based systems results for different k values.

Knowledge Graph-based system with implicit knowledge



Knowledge Graph-based system with implicit knowledge results

	Text-based system (baseline)		Concept-based system		KG-based system with search history	
@k ↓	**MAP@k**	**BM@k**	**MAP@k**	**BM@k**	**MAP@k**	**BM@k**
@5	0.061	0.114	0.152	0.243	0.115	**0.291**
@25	0.064	0.148	0.159	0.334	0.122	**0.471**
@50	0.064	0.157	0.160	0.371	0.123	**0.552**
@100	0.064	0.161	0.161	0.403	0.123	**0.624**
@350	0.064	0.164	0.161	0.429	0.124	**0.715**

Table: Text, concept, and KG-based systems results for different k values.

Quantitative results

	No results	Less than 400 results (non empty)
Text-based system (baseline)	64.48%	35.44%
Concept-based system	11.43%	**88.36%**
KG-based system with search history	**8.10%**	51.59%

Table: Comparing all search systems results set corpus.

Online OWL reasoning-based approach

An approach focusing on OWL.

- An Information Retrieval ontology.
- Push knowledge closer to the data.
- Model domain knowledge as linked sets of taxonomies.

Knowledge modelling

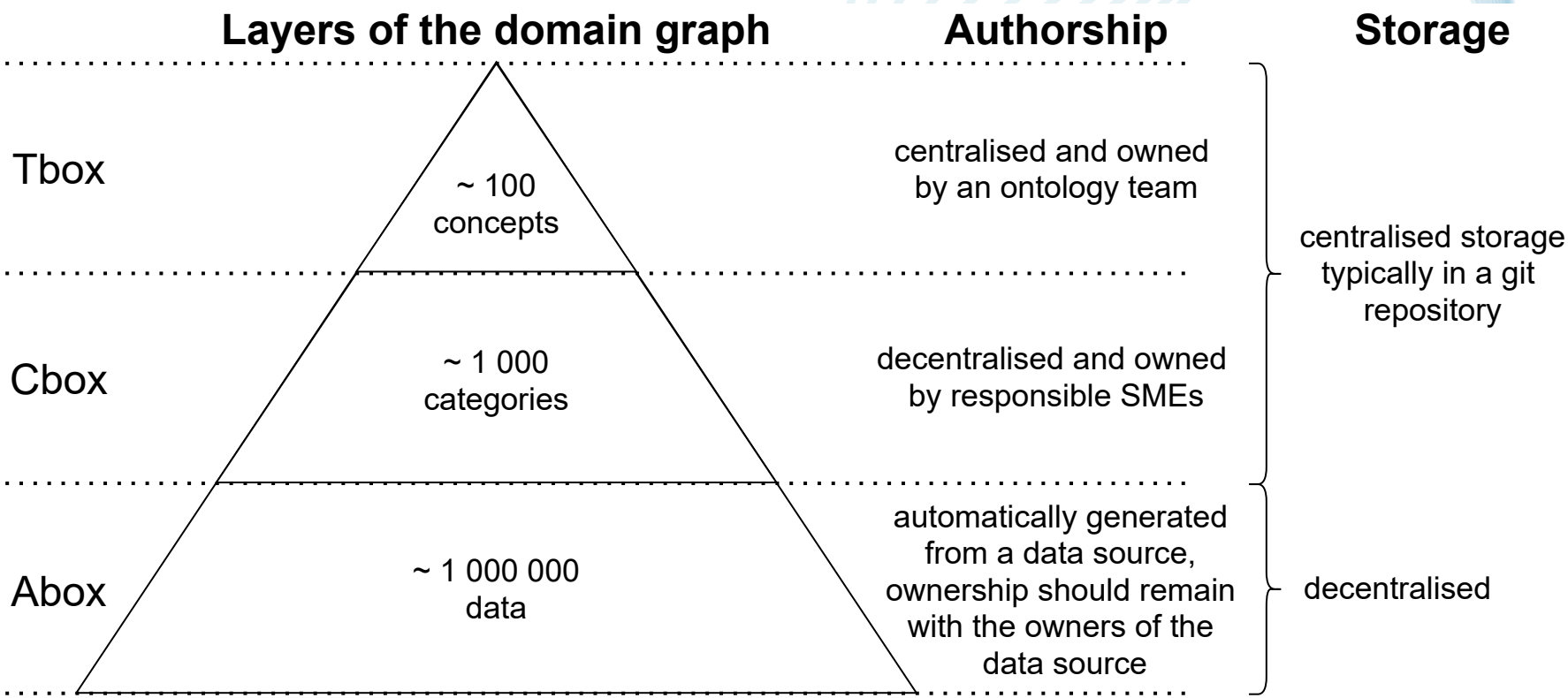


Figure: "C-box" knowledge modelling approach

Information Retrieval Ontology

Competency questions:

- CQ1 What are the categories in the user search?
- CQ2 What are the documents relevant to a search?
- CQ3 What categories are enabled to refine the search?

7 classes:

- *Candidate Document* subclass of *Document*
- *Selected Category* and *Enabled Category* subclasses of *Category*
- *Search Context* subclass of *Search*

6 Object properties:

- *categorises* inverse of *categorised By*
- *has Search Category* subproperty of *enables Category*
- *has Direct Subcategory* subproperty of *has Subcategory*

Pizza ontology

Pizza ontology:

- Well-knowledge ontology built to introduce RDF/RDFS/OWL with examples (and even SHACL)
- Simple ontology with class hierarchies of:
  - Pizzas (has topping, has base)
  - Pizza bases
  - Pizza Toppings
- We use the Pizza ontology for demonstration in interest of time constraints

Pizza ontology Knowledge Graph

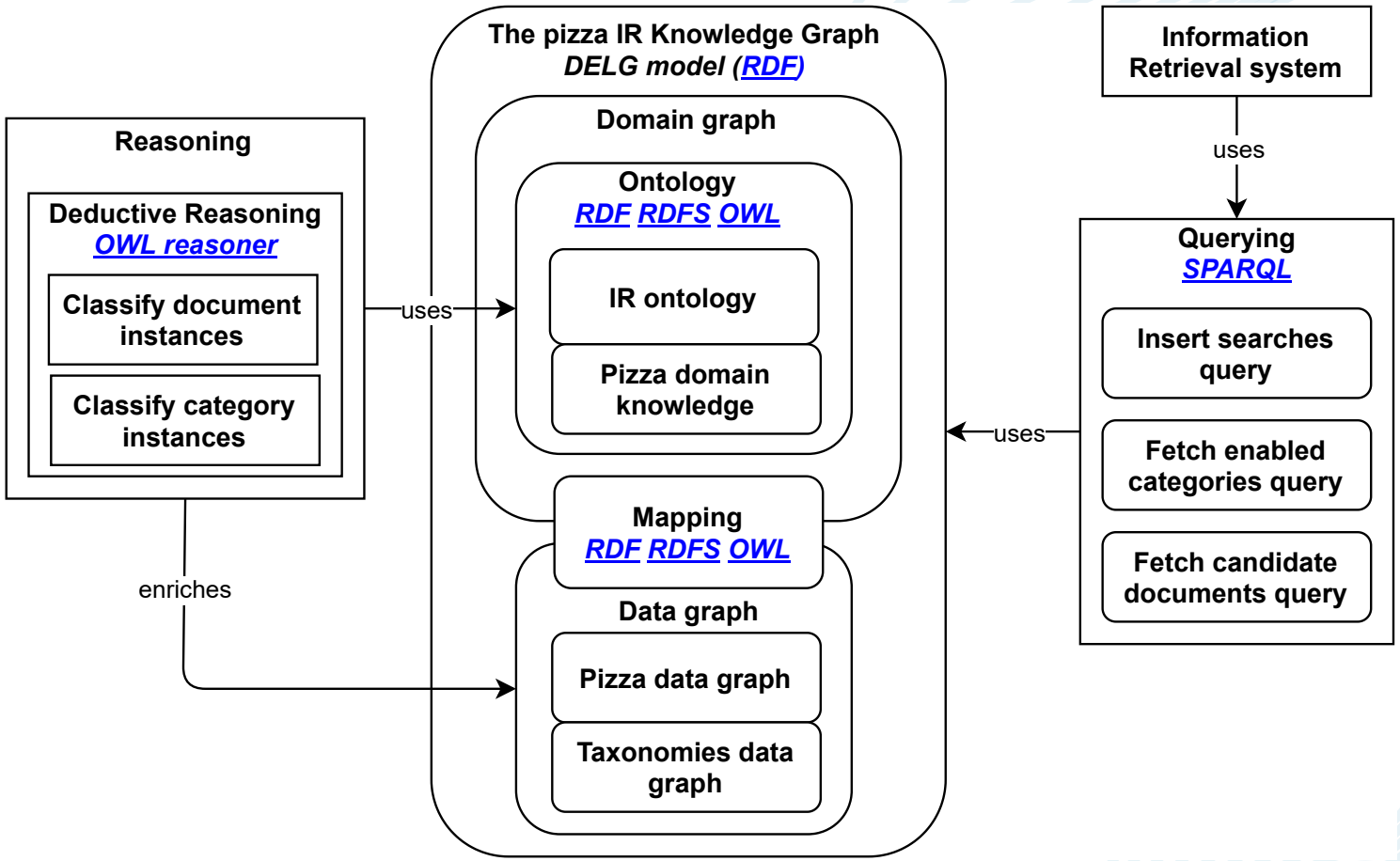


Figure: Pizza ontology Knowledge Graph

OWL reasoning-based Information Retrieval

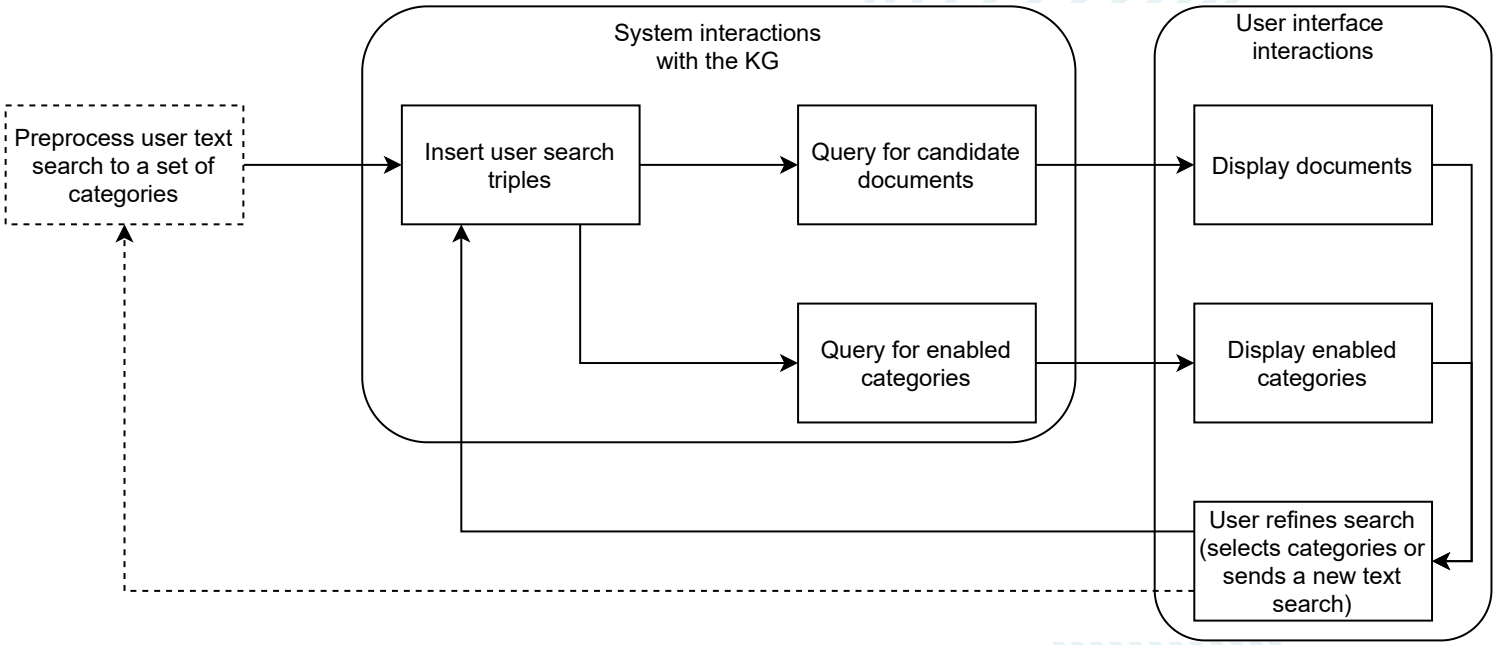


Figure: OWL reasoning-based Information Retrieval process.

Contributions

A top-down approach from a system perspective down to solution implementations.

Contributions:

- A unifying definition of Knowledge Graph
- An architecture for Knowledge Graph-Based Systems
- A framework for Ontology Learning
- An OWL Information Retrieval ontology
- A study of a text-based compared to a Knowledge Graph-based Information Retrieval system

Conclusion

We have explored:

- A Knowledge Graph definition incorporating ontologies
- A Semantic Web-focused implementation of this Knowledge Graph definition
- An OWL Information Retrieval Ontology
- Two Knowledge Graph-based Information Retrieval System approaches:
  - A real-world use case moving from a text-based to a Knowledge Graph-based Information Retrieval System.
  - An online OWL reasoning-based Information Retrieval use case.

Future works

- Knowledge Graph-based Information Retrieval system:
  - Expand the Knowledge Graph
  - Expand the approach to other domains
- OWL reasoning-based Information Retrieval system:
  - Experiment with a real-world use case at scale
  - Explore distinguishing between searches with no matching documents and incoherent ones
- Implement an end-to-end Knowledge Graph-Based System architecture use case.

Perspectives: Knowledge Graph

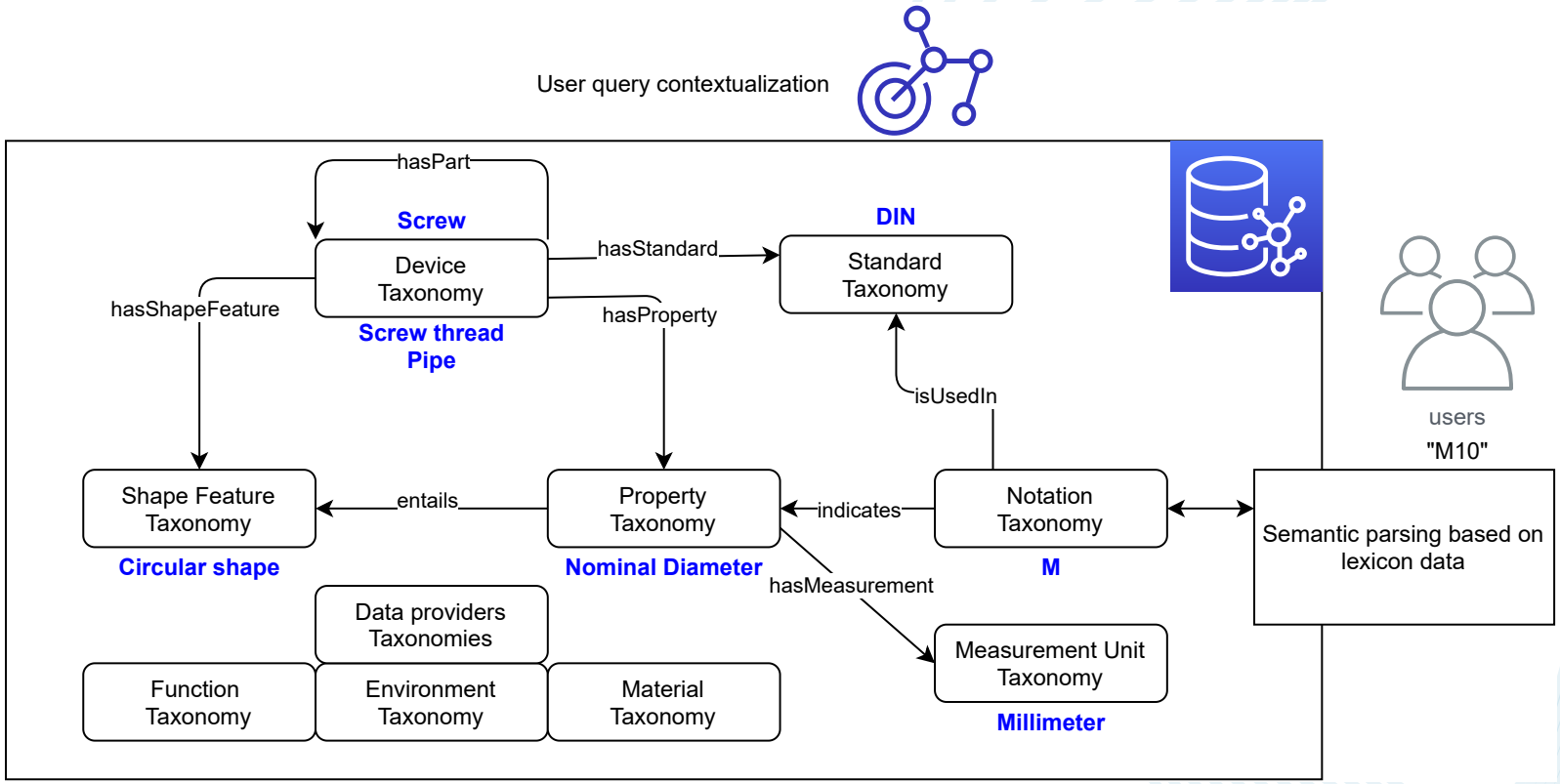


Figure: Extended semantic search example.

Scientific productions

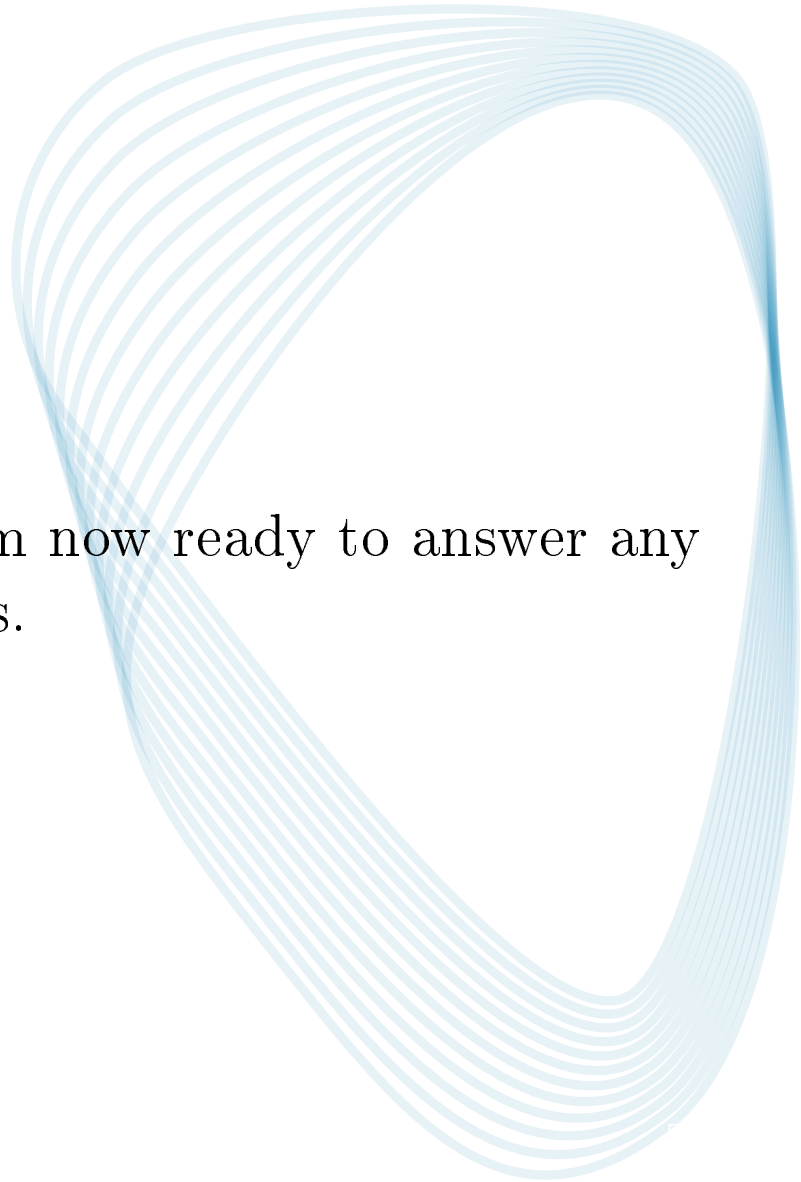
Peer-reviewed international conference papers:

- An operational architecture for knowledge graph-based systems. Proceedings of the 26th International Conference KES2022.
- (with Marion Schaeffer) Olaf: An ontology learning applied framework. Proceedings of the 27th International Conference KES2023.

Open-source software library (with Marion Schaeffer):

- Ontology Learning Applied Framework Python library implementation: https://wikit-ai.github.io/olaf/

Thank you!



Thank you for your attention. I am now ready to answer any questions.