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

Traceparts

- Founded in 1990 with activities worldwide
- One of the world's leading CAD-content platforms for Engineering, Industrial Equipment and Machine Design: traceparts.com
- Digital marketing services for more than 800 customers of all sizes and from all industries.

Traceparts CAD-content plateform

The plateform 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

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*.

Information Retrieval

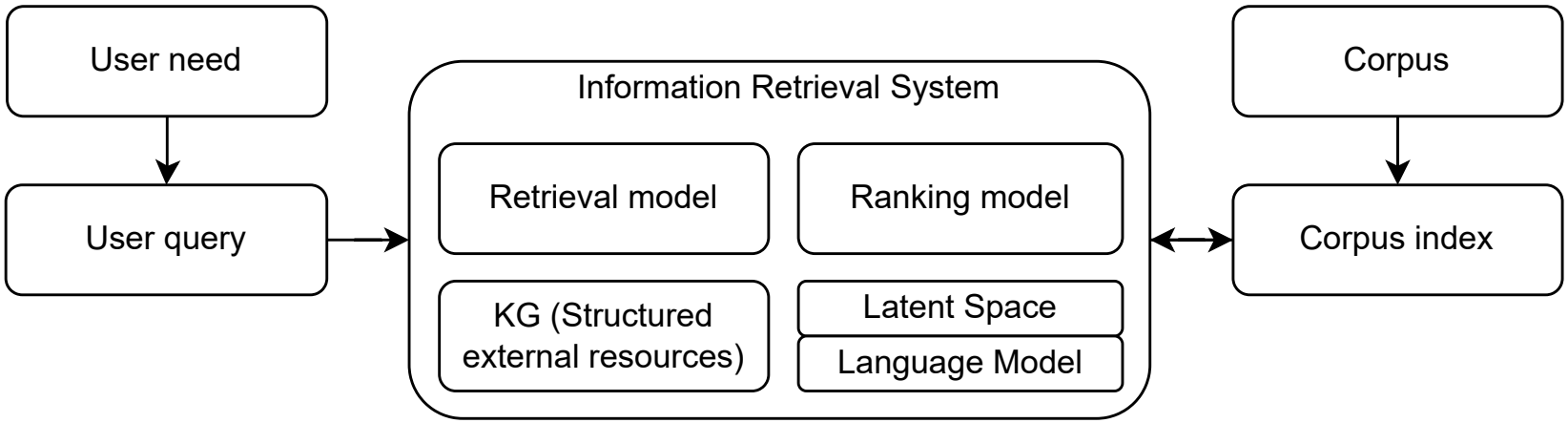


Figure: Information Retrieval System overview

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

BM25

- The more structured the content is the better.
- Systems combines multiple methods balanced with parameters.
- Deep Learning methods requires relatively long texts.

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.

Traceparts search system

A text-based search engine.

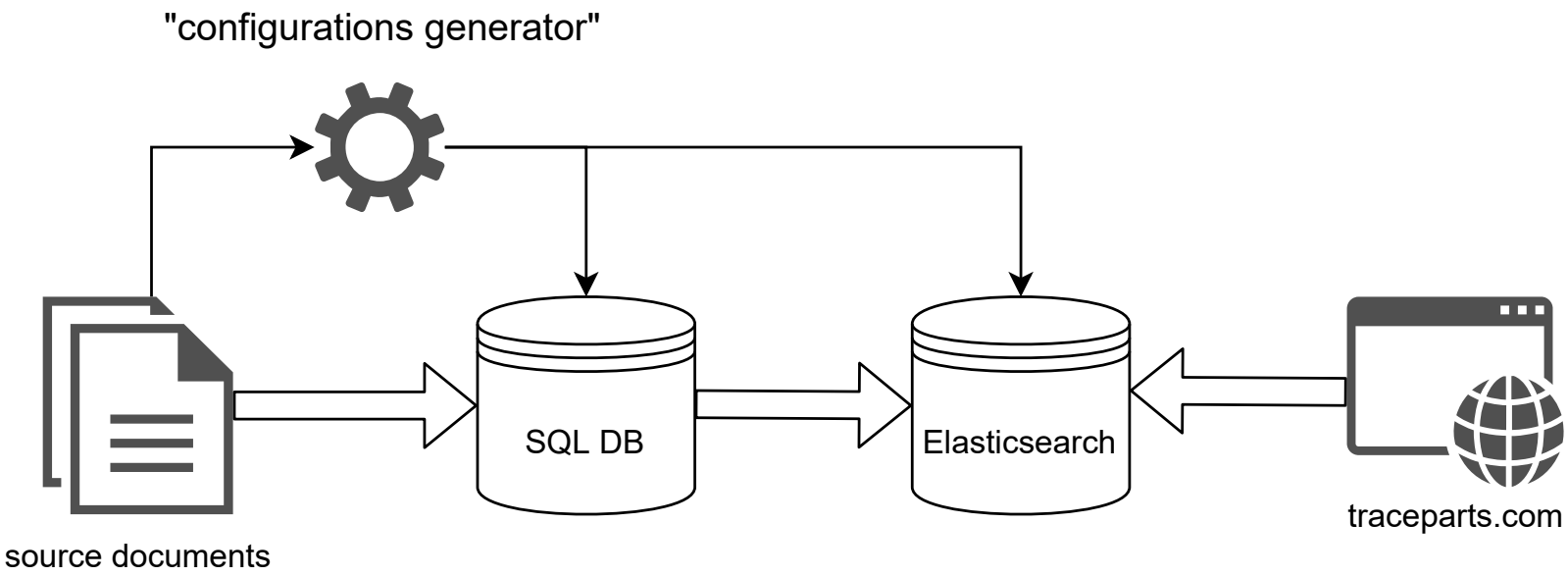


Figure: Traceparts current system

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

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

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.
- Technical content is implicitly structured by domain knowledge.
- This knowledge must be explicitly machine readable.
- KGs are machine readable structured knowledge artifacts.

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

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

Manuscript overview

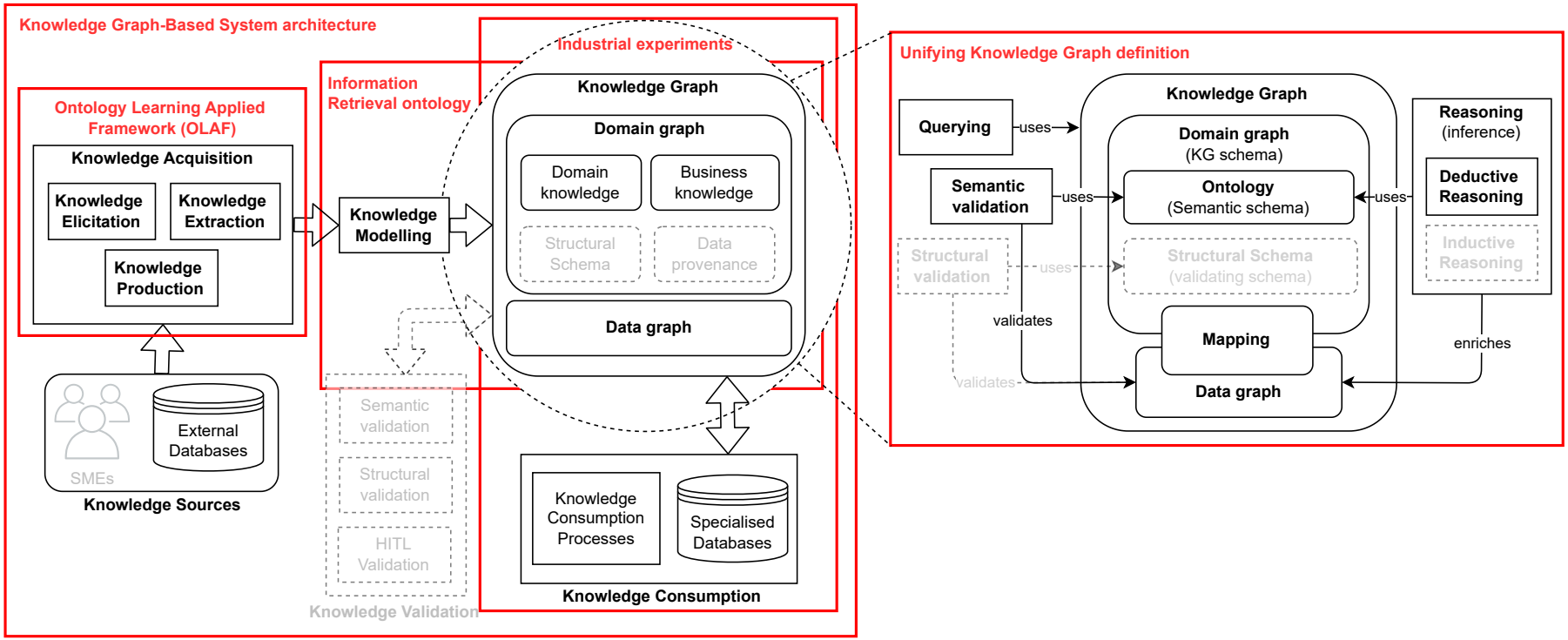


Figure: Manuscript overview

Knowledge Graph vs Ontology

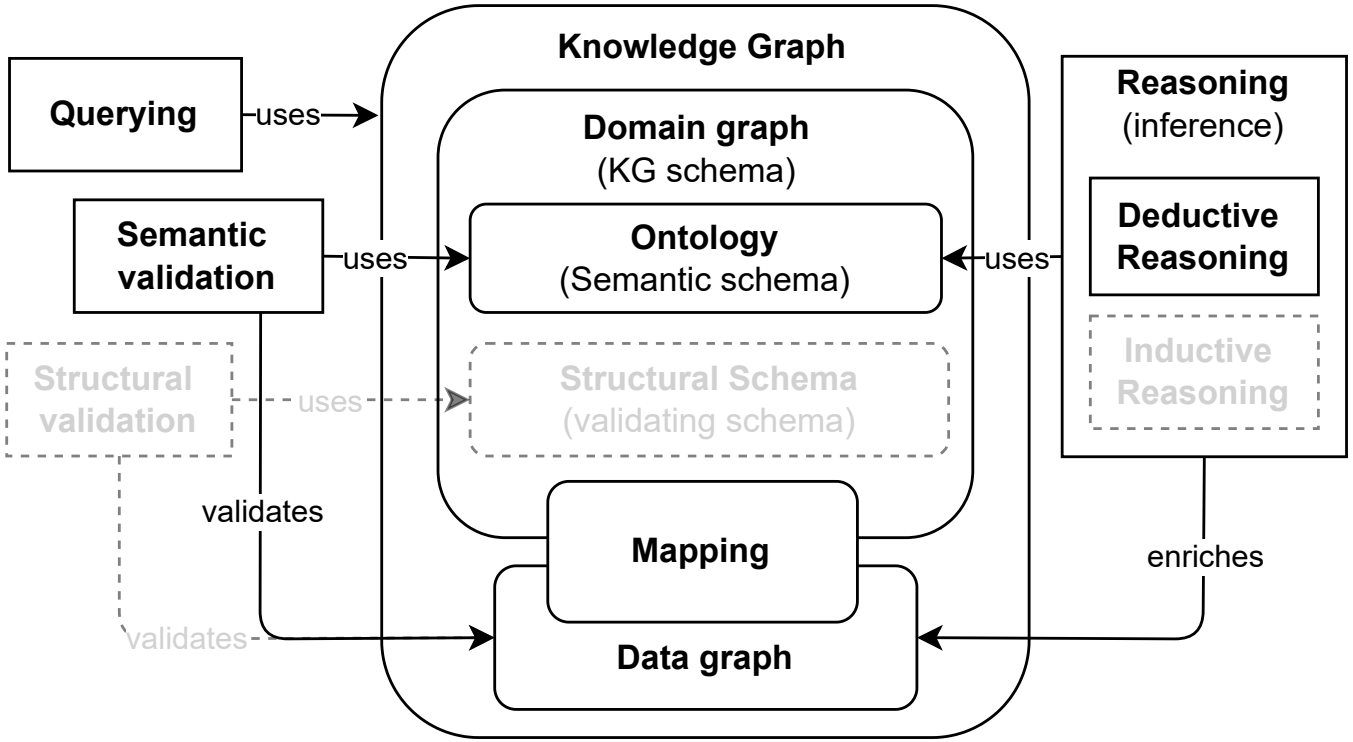


Figure: Knowledge Graph definition

Experiments objective

From a text-based to a concept-based search.

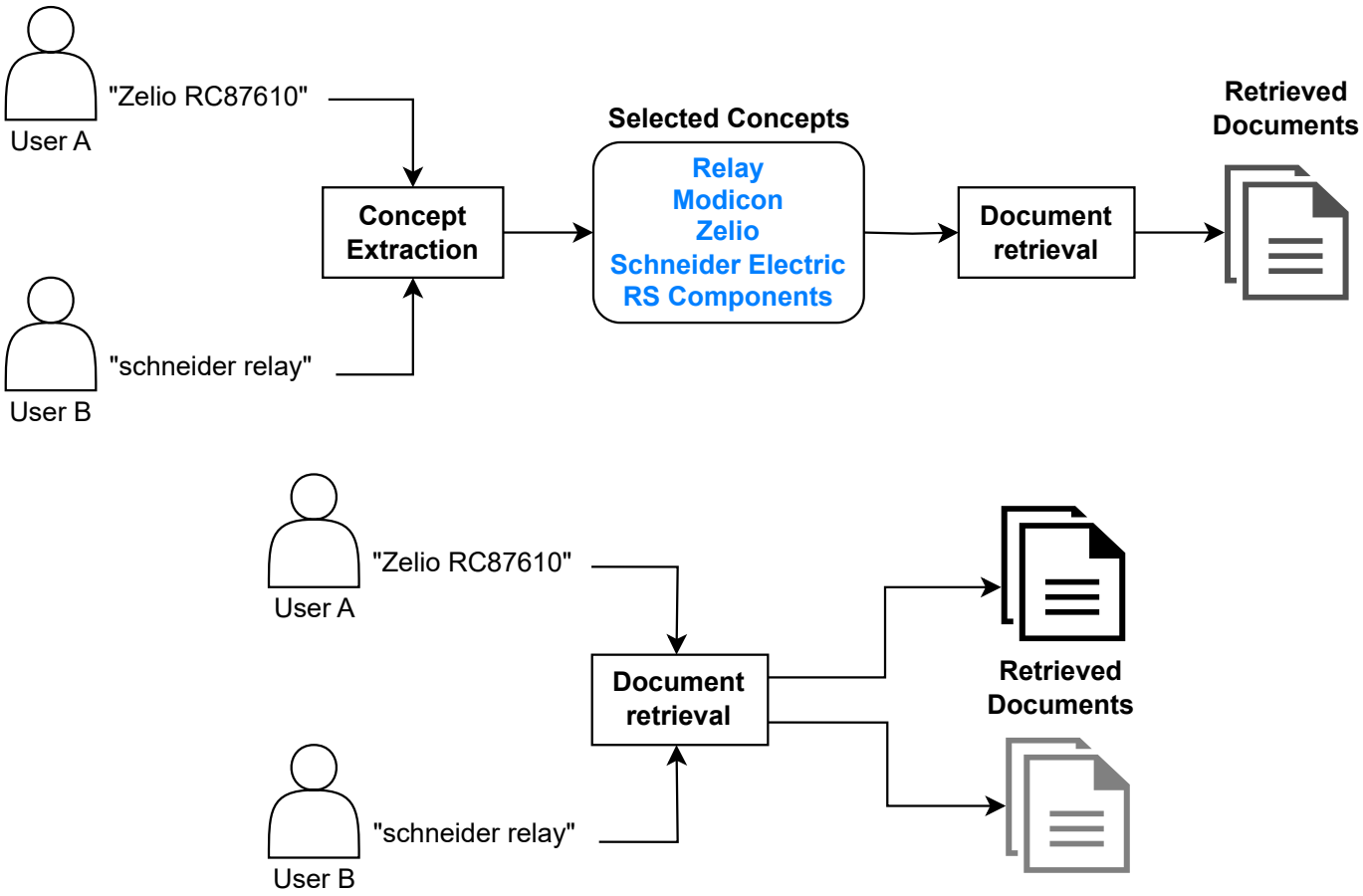


Figure: Text-based vs concept-based search.

Experiments protocol

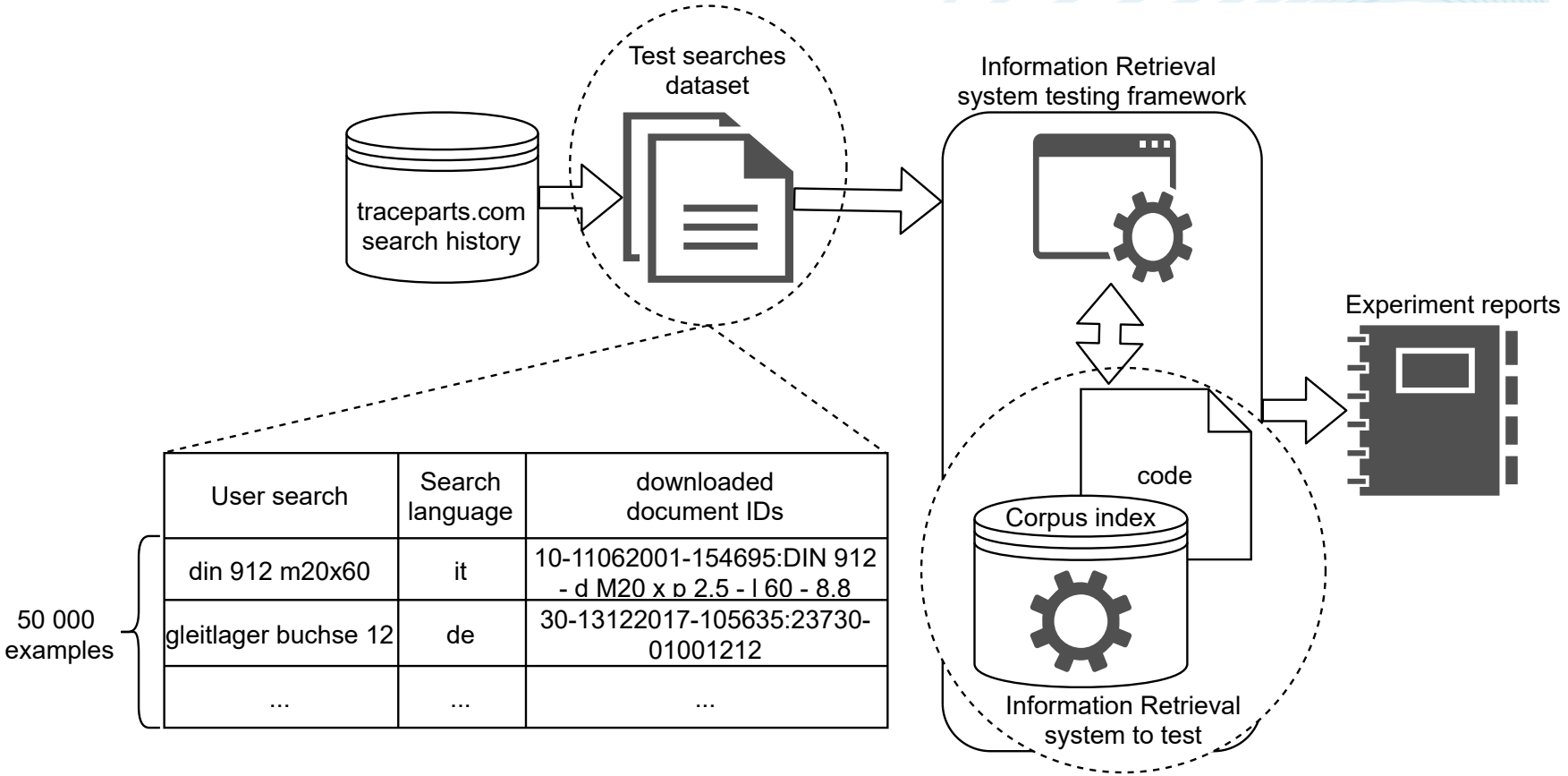


Figure: Experiments Protocol.

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.

Information Retrieval systems

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

Implementation:

- 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.

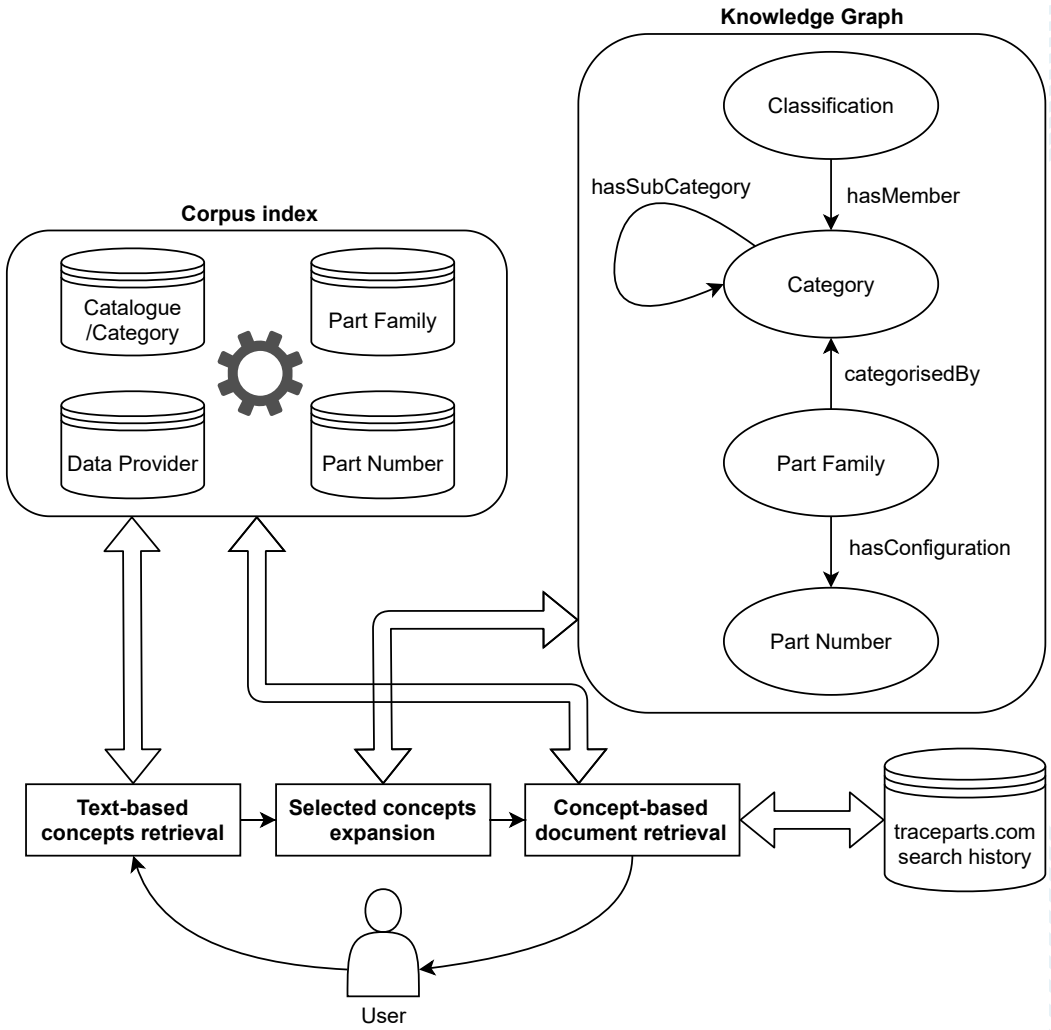


Figure: Knowledge Graph-based system with implicit knowledge.

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: Comparing text, concept, and KG-based systems for different k values.

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

A second approach focusing on OWL.

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

Knowledge modelling

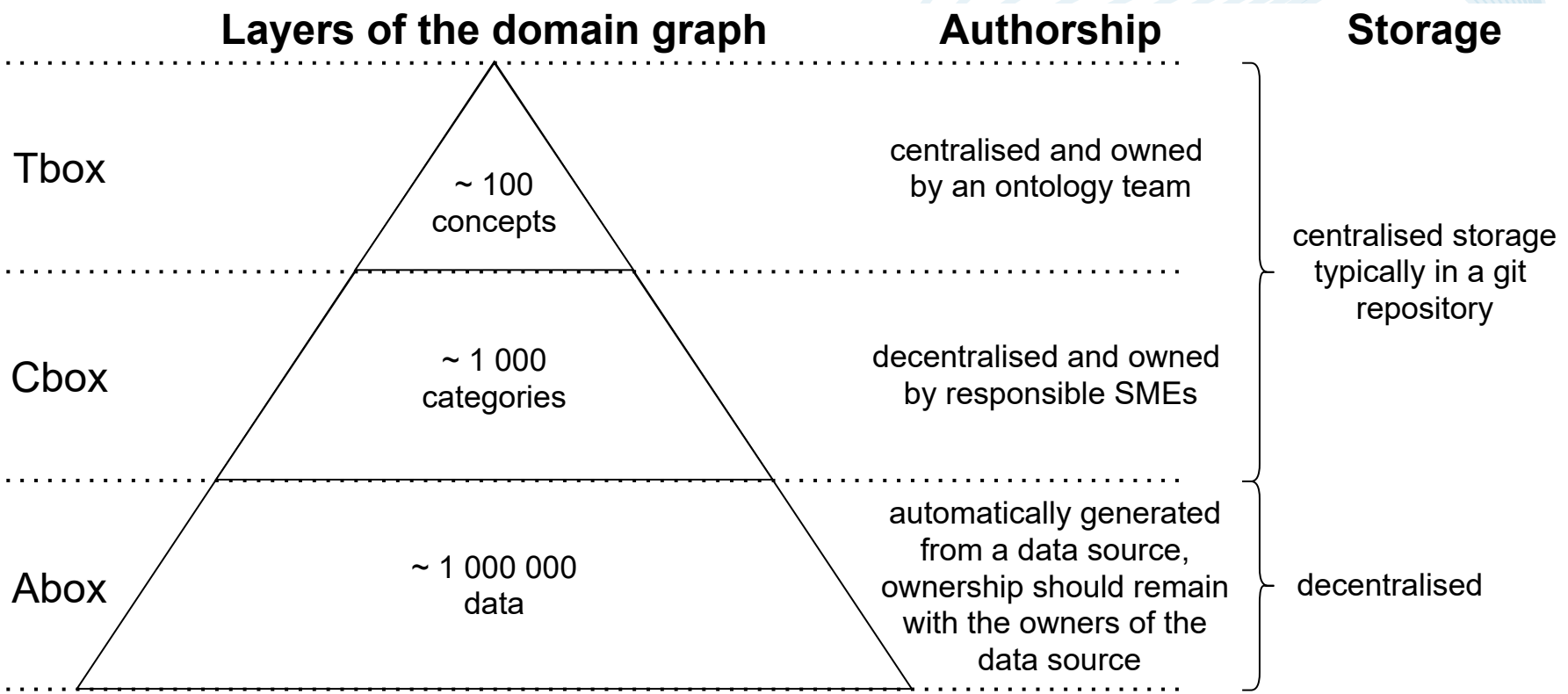


Figure: "C-box" knowledge modelling approach

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*

Knowledge Graph

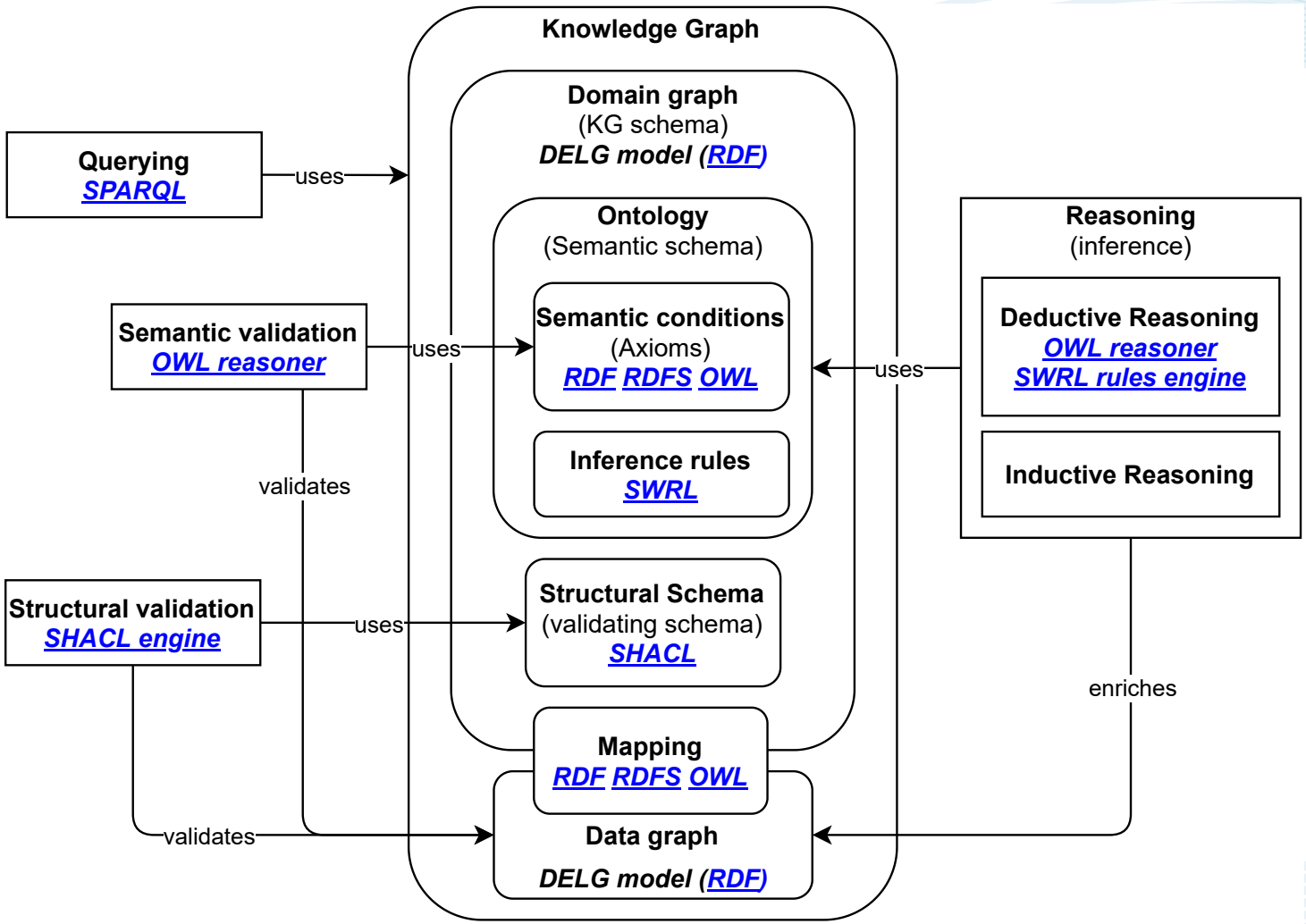


Figure: Knowledge Graph definition

Pizza ontology Knowledge Graph

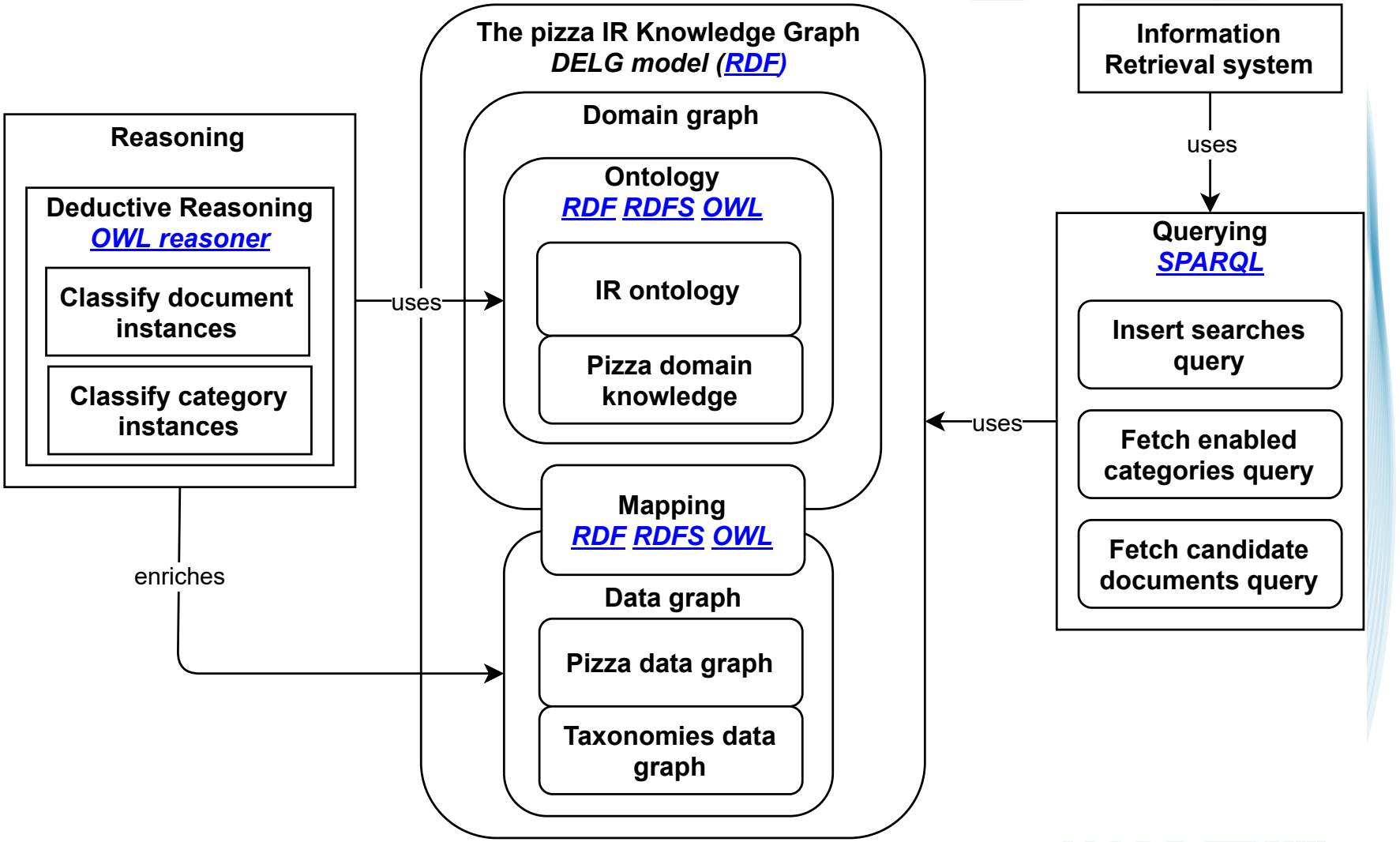


Figure: Pizza ontology Knowledge Graph

OWL reasoning-based Information Retrieval

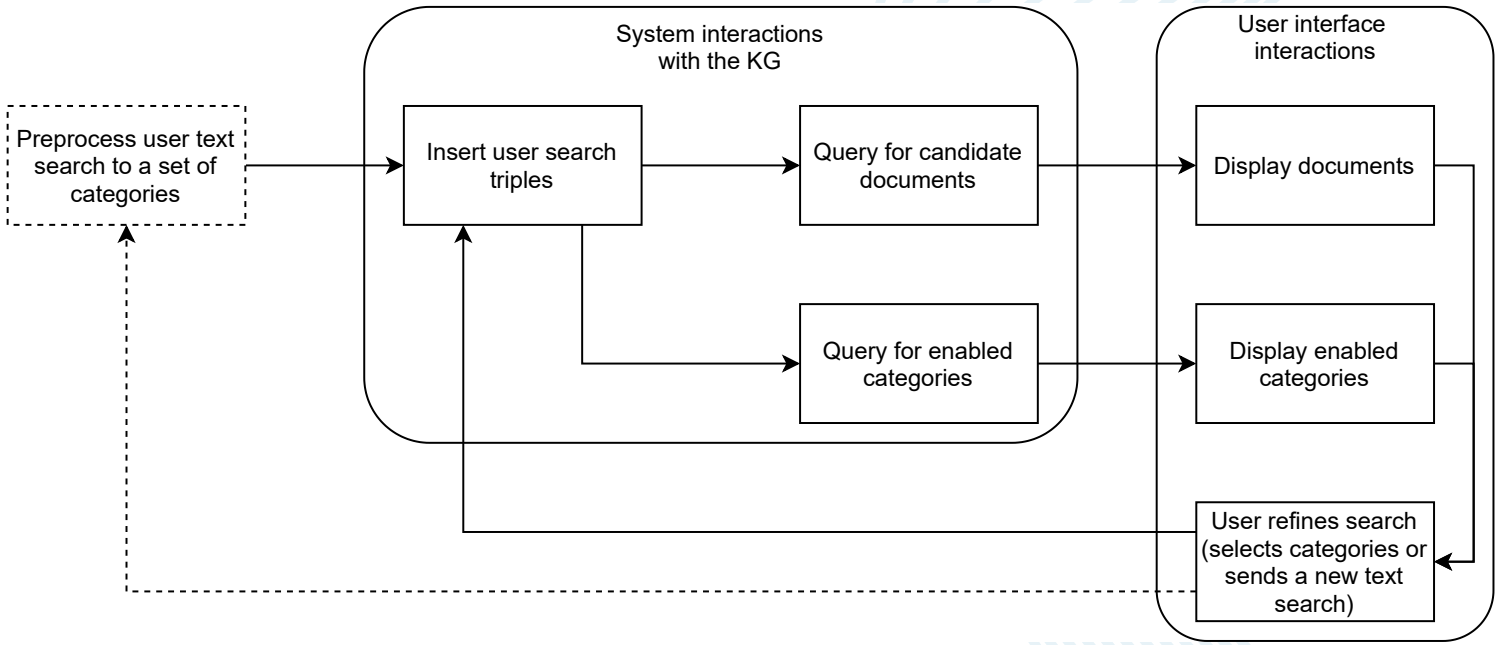


Figure: OWL reasoning-based Information Retrieval process.

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

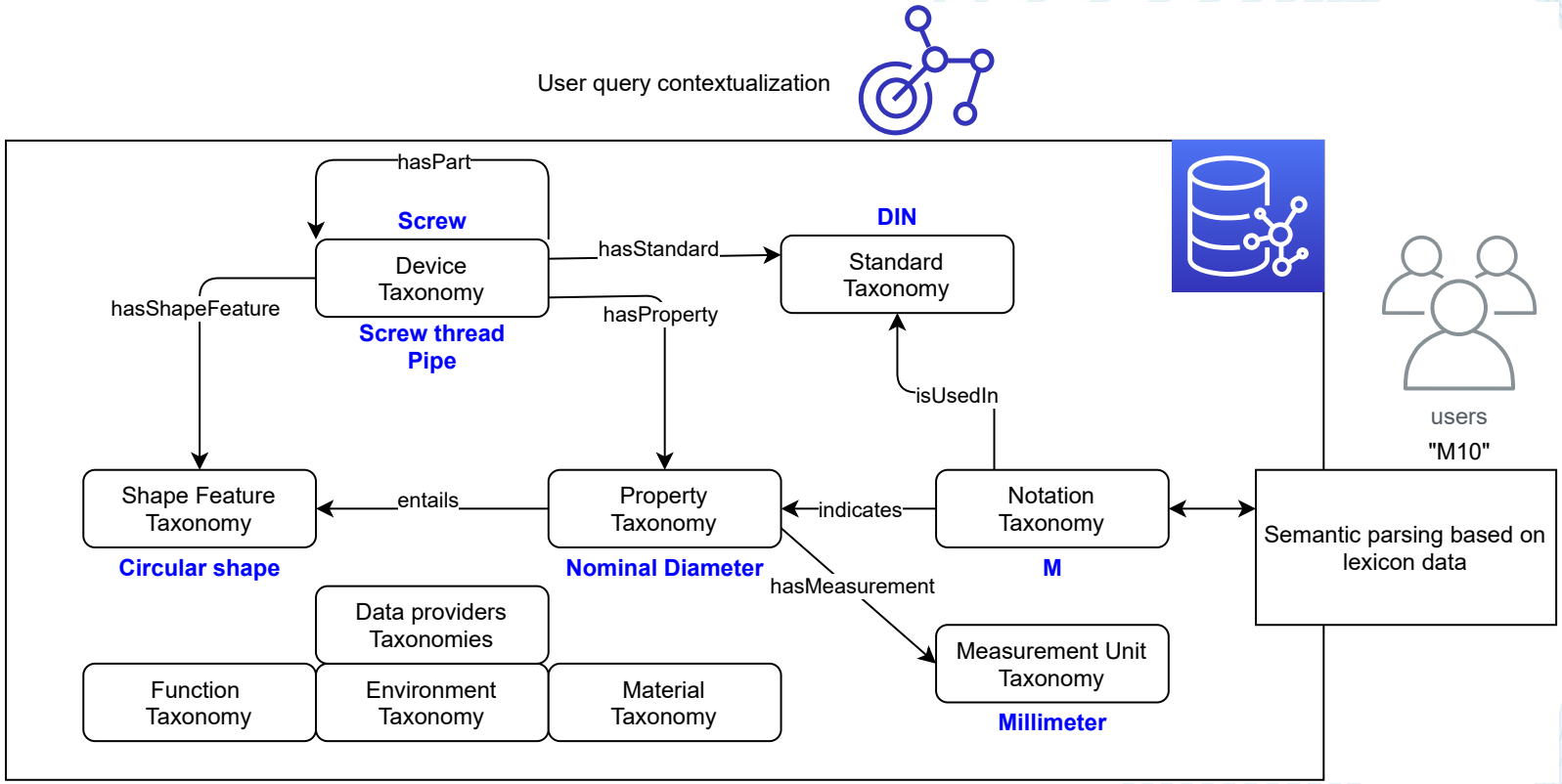
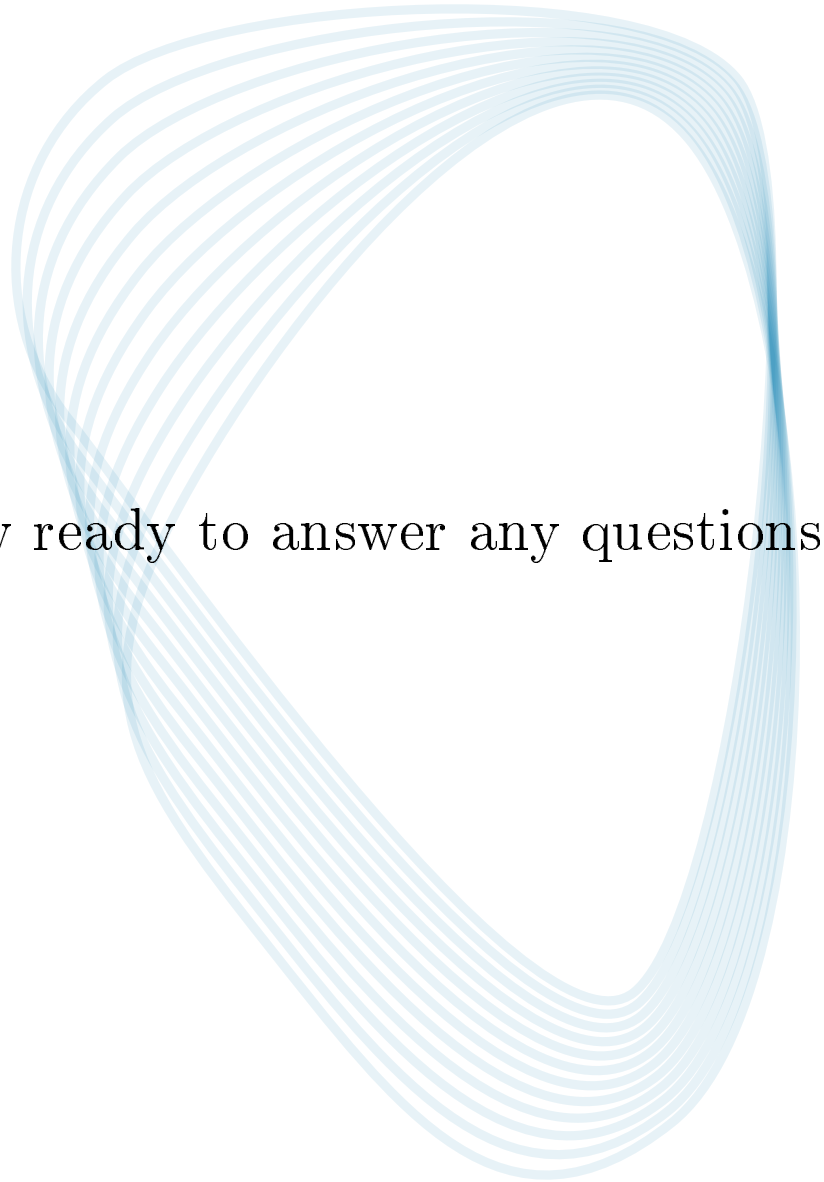


Figure: Extended semantic search example.

Thank you!



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