
INSIGHT
Data Science Laboratory
Federal University of Ceará

Fragment Extraction from *RDF* Schemas using *Keywords*

INSIGHT

DEPARTAMENTO DE COMPUTAÇÃO

UNIVERSIDADE FEDERAL DO CEARÁ

HI!

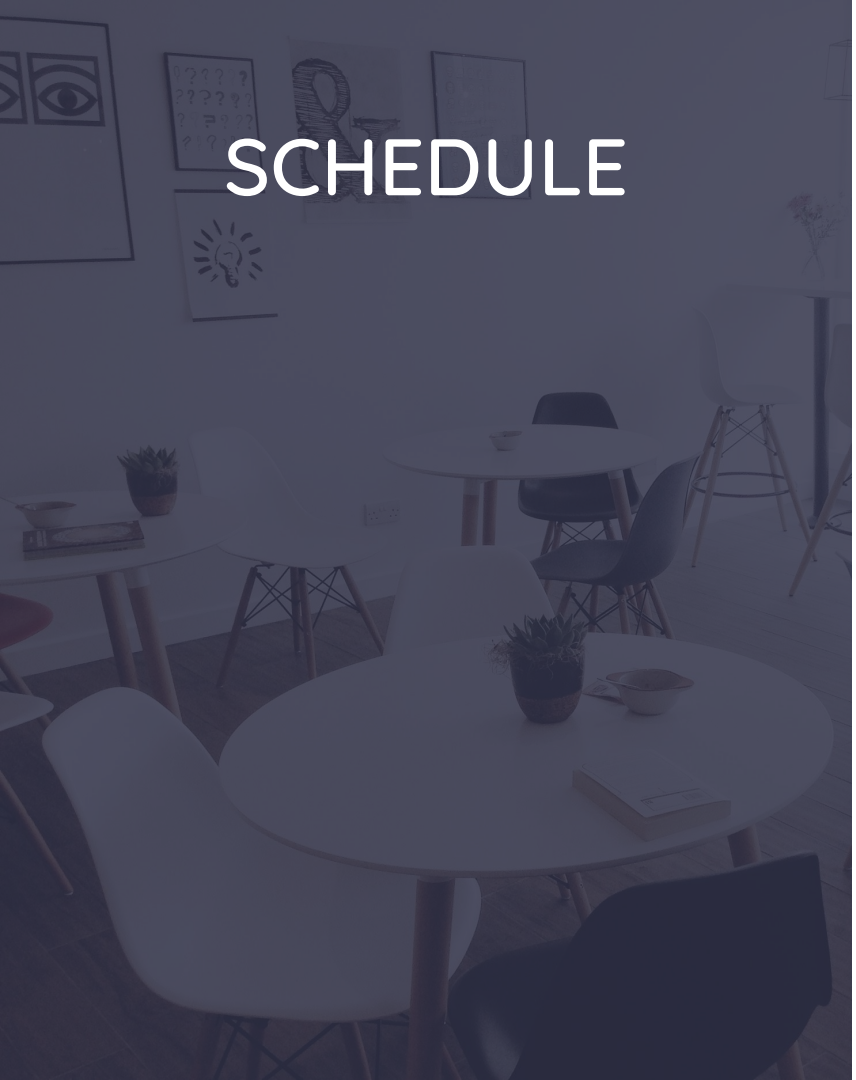
I'm **Lucas Peres!**

I'm a **researcher** at *Insight Data Science Lab*!

Graduated and **Master student** in Computer Science at Federal University of Ceará, working with *Information Retrieval* over ontology.

You can find me at **@lucaspg96***

*Basically at all social networks

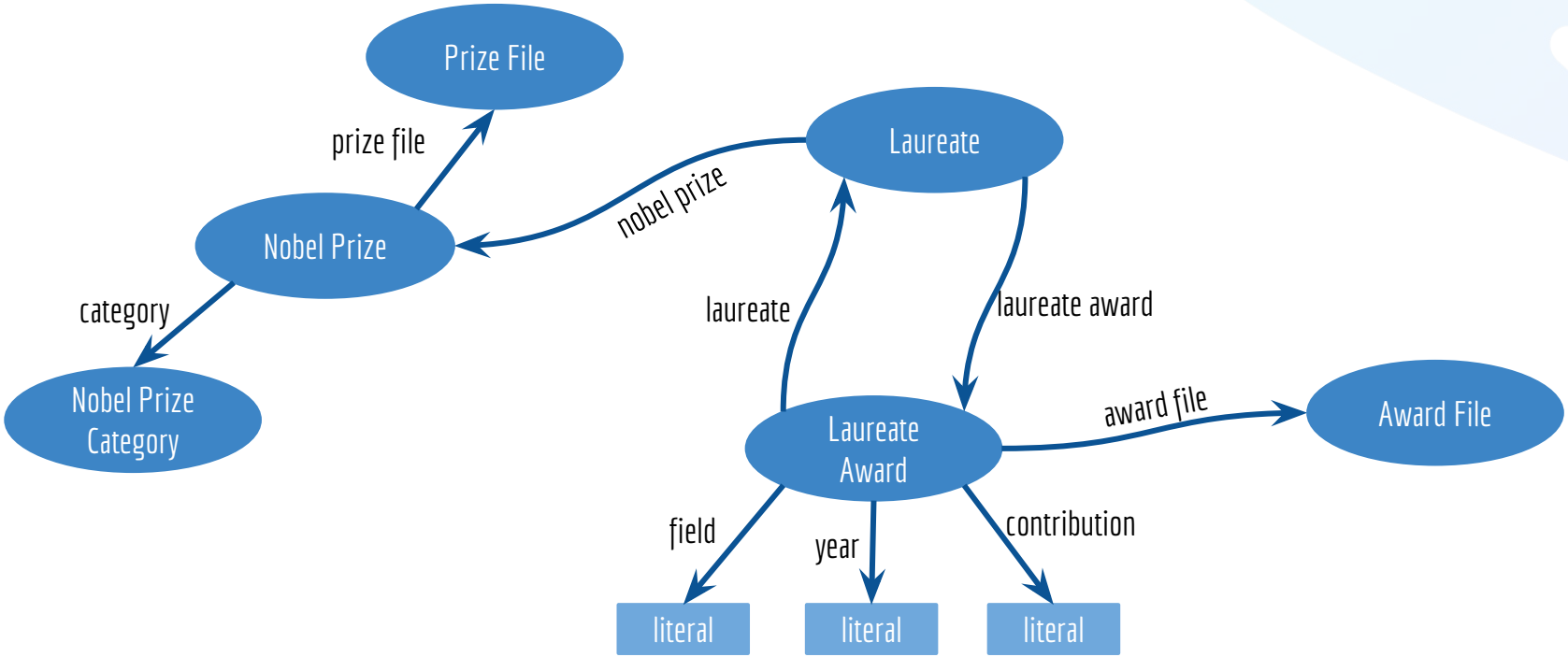


SCHEDULE

1. Introduction
2. Workflow
3. Fragment Extraction
4. Experiments
5. Conclusions and Future Works

1. Introduction

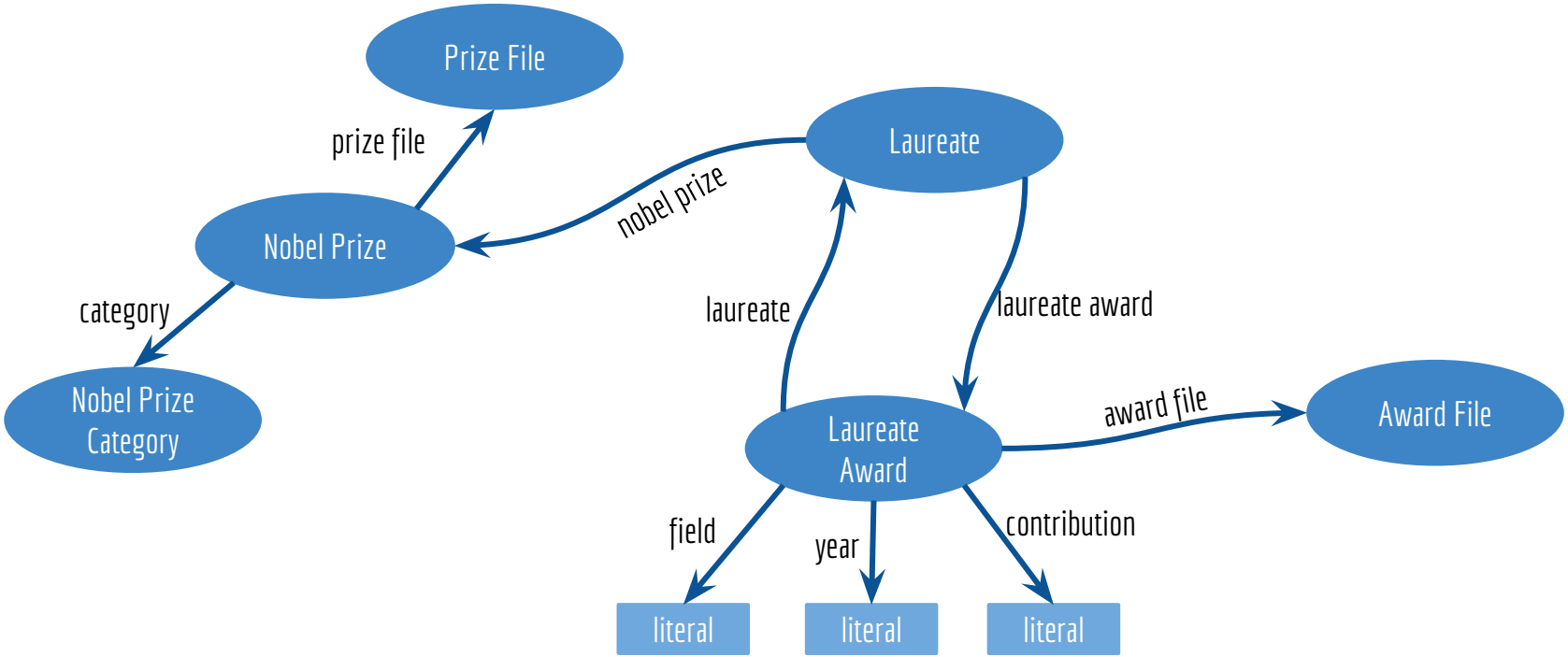
ONTOLOGIES


Prize File
prize file
Laureate
nobel prize
Nobel Prize
category
laureate
laureate award
Nobel Prize Category
Laureate Award
award file
Award File
field
year
contribution
literal
literal
literal

RETRIEVING INFORMATION

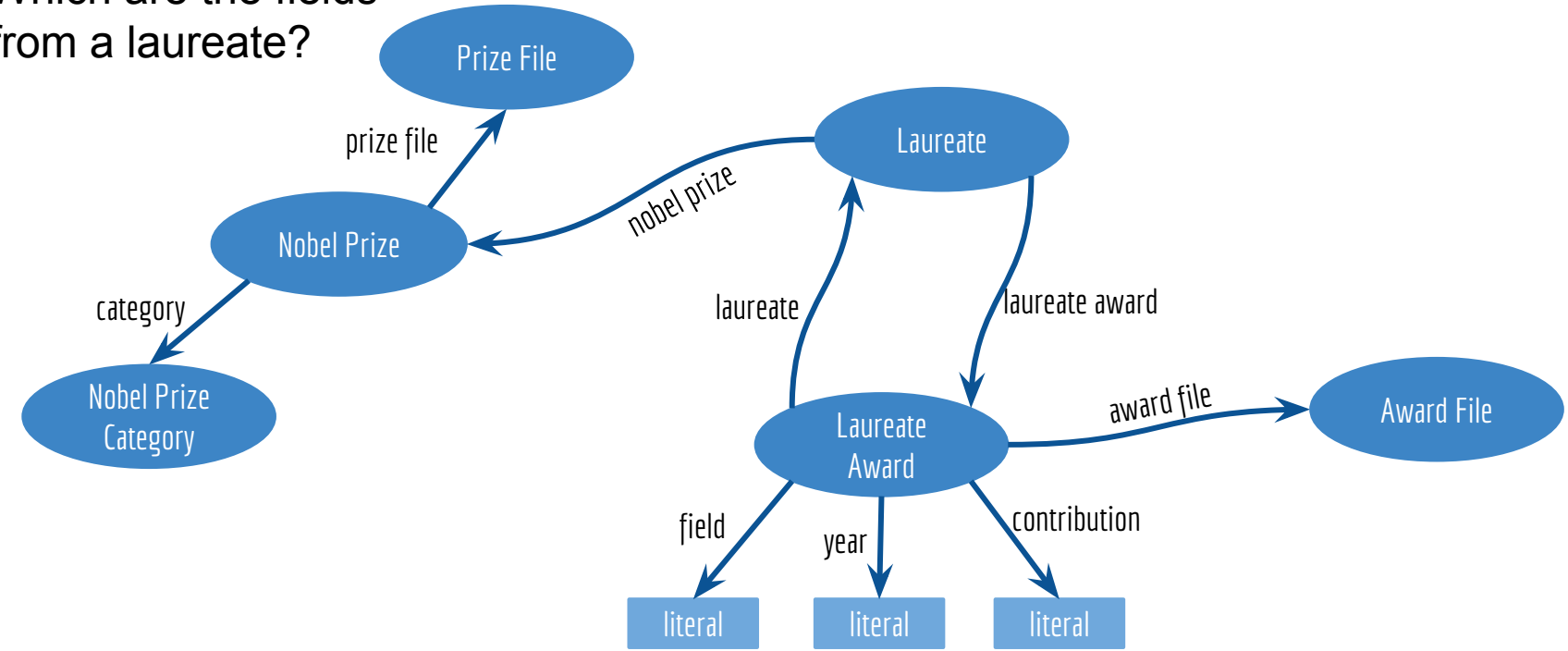
- Build queries over the ontology
- Retrieve relation of an element
- Search for keywords
  - get triples with certain words
- Exploratory search
  - explore over the ontology

RETRIEVING INFORMATION



RETRIEVING INFORMATION

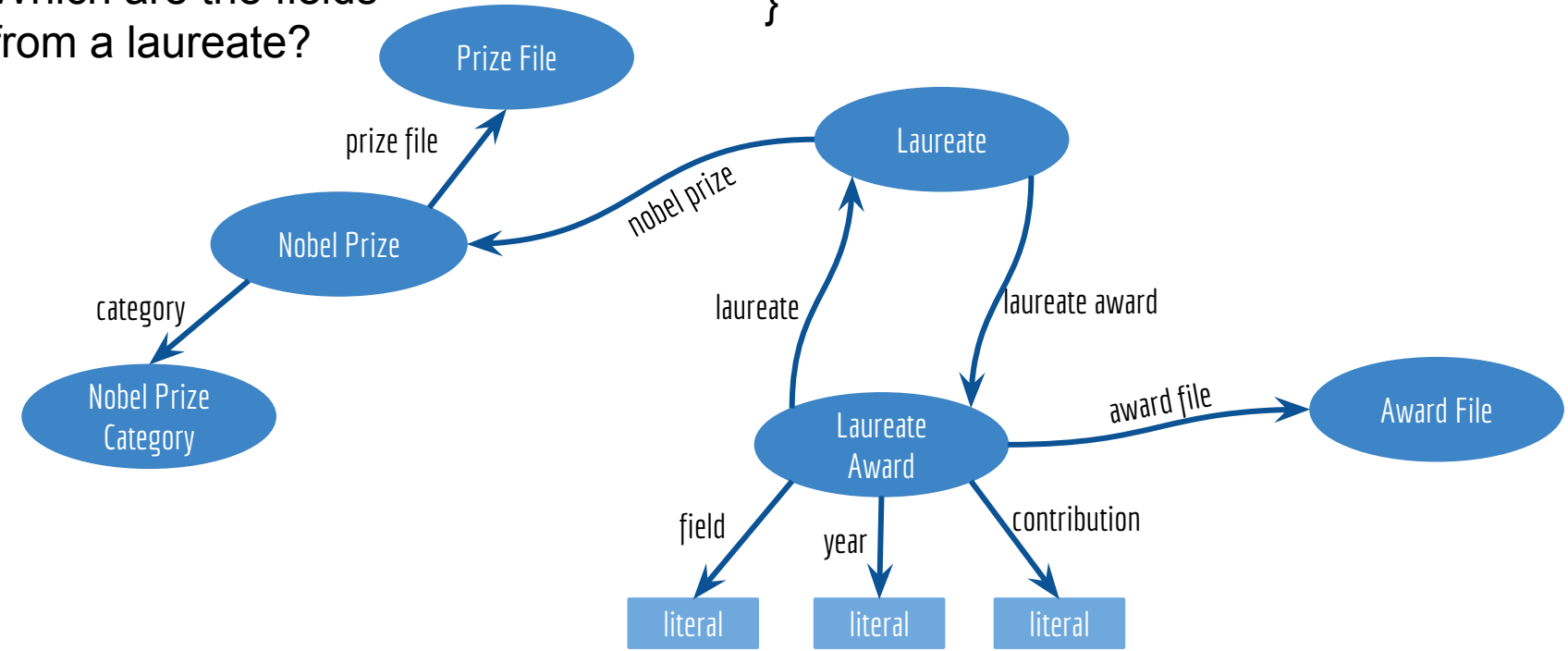
Which are the fields from a laureate?



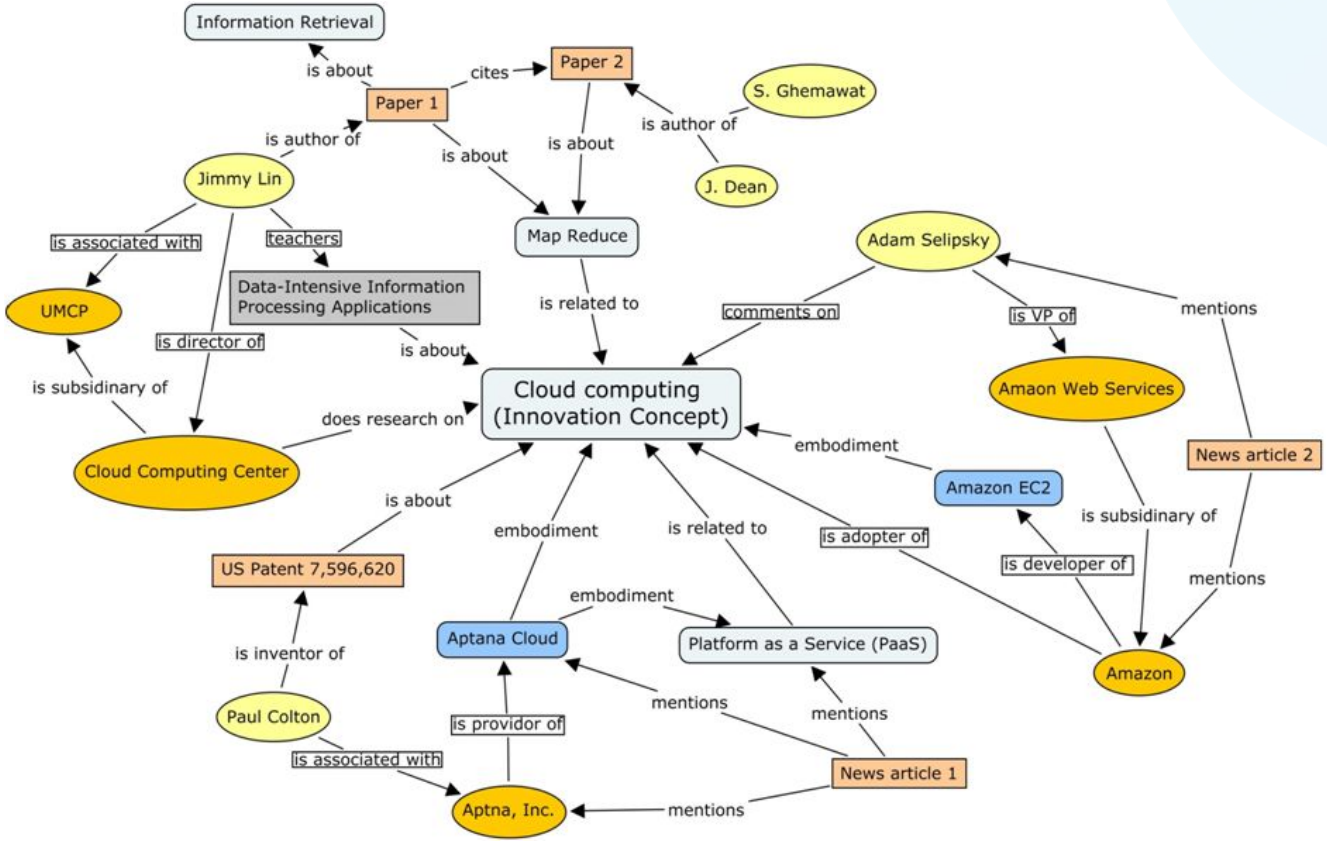
RETRIEVING INFORMATION

```
select ?laureate ?field
where{
    ?laureate a :Laureate;
    ?laureate :laureate_award ?award
    ?award :field ?field;
}
```

Which are the fields from a laureate?



RETRIEVING INFORMATION



RETRIEVING INFORMATION

STATE OF THE ART

- **(ELBASSUONI, 2011)**
  - Retrieve ontology *subgraphs* from a set of keywords
- **(MUSETTI, 2012)**
  - Retrieve an entity from keywords and presents its more relevant relations
- **(OUKSILI, 2017)**
  - Retrieve ontology *subgraphs* from a set of keywords using graph patterns

STATE OF THE ART

- **(ELBASSUONI, 2011)**
  - Retrieve ontology *subgraphs* from a set of keywords
- **(MUSETTI, 2012)**
  - Retrieve an entity from keywords and presents its more relevant relations
- **(OUKSILI, 2017)**
  - Retrieve ontology *subgraphs* from a set of keywords using graph patterns
- **Common Task**: retrieve **fragments** from an **ontology**

How can we retrieve a ***fragment*** *from the* ***ontology*** *using* ***keywords****?*

2. Workflow

OUR APPROACH

- Work only with the schema
  - the data may not fit in memory
  - no previous knowledge about the domain

OUR APPROACH

- Work only with the schema
  - the data may not fit in memory
  - no previous knowledge about the domain

- Use solution to graph problems
  - Since the ontology can be represented as a graph, we can map the fragment extraction to a graph problem

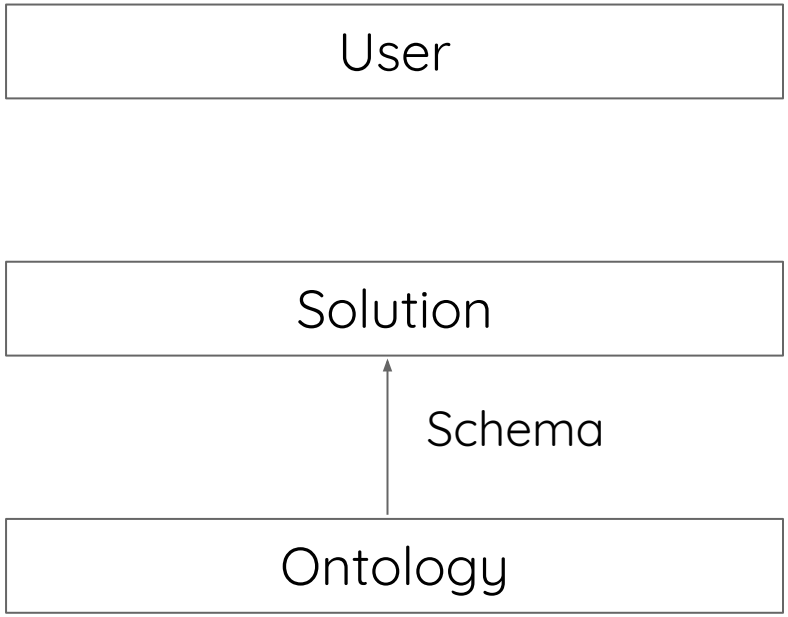
OUR APPROACH

User

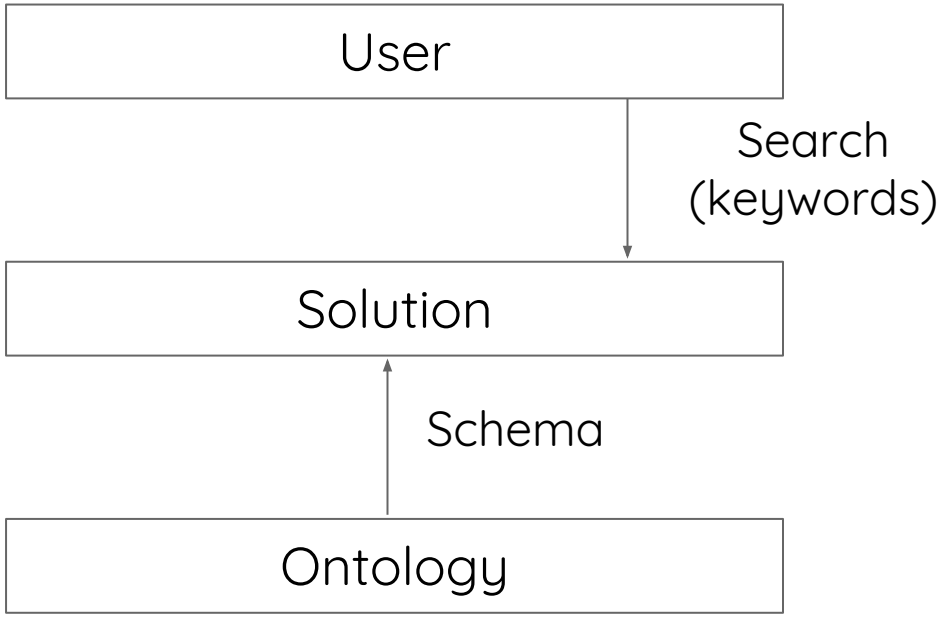
Solution

Ontology

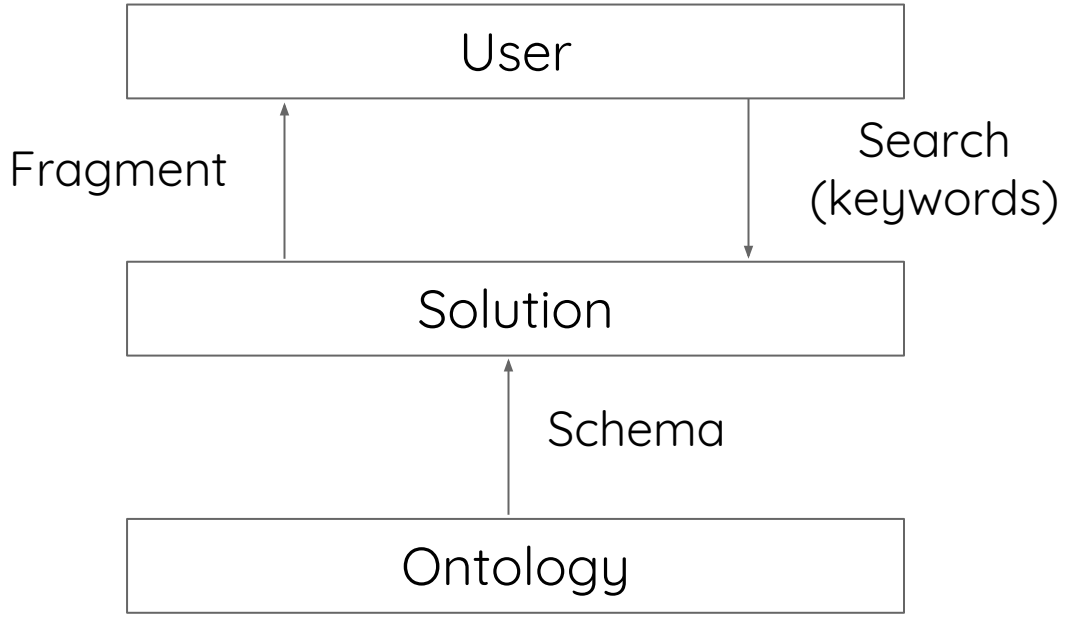
OUR APPROACH



OUR APPROACH



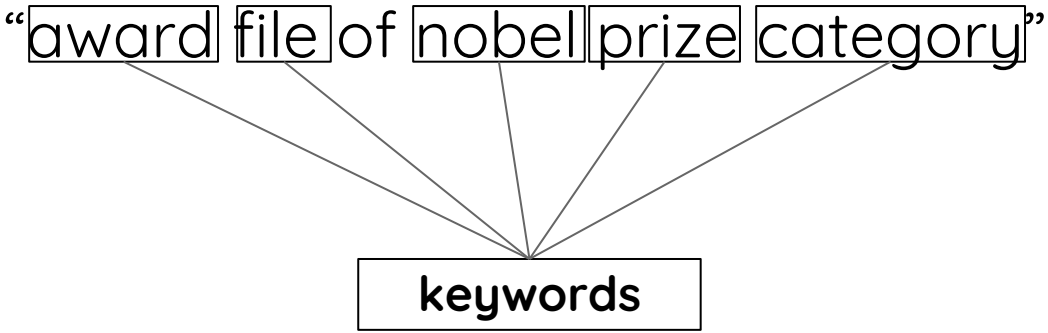
OUR APPROACH



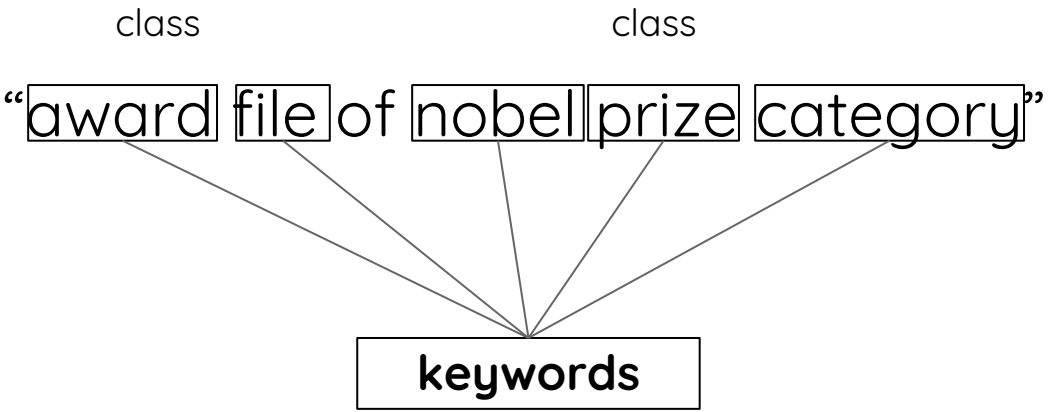
SEARCH

“award file of nobel prize category”

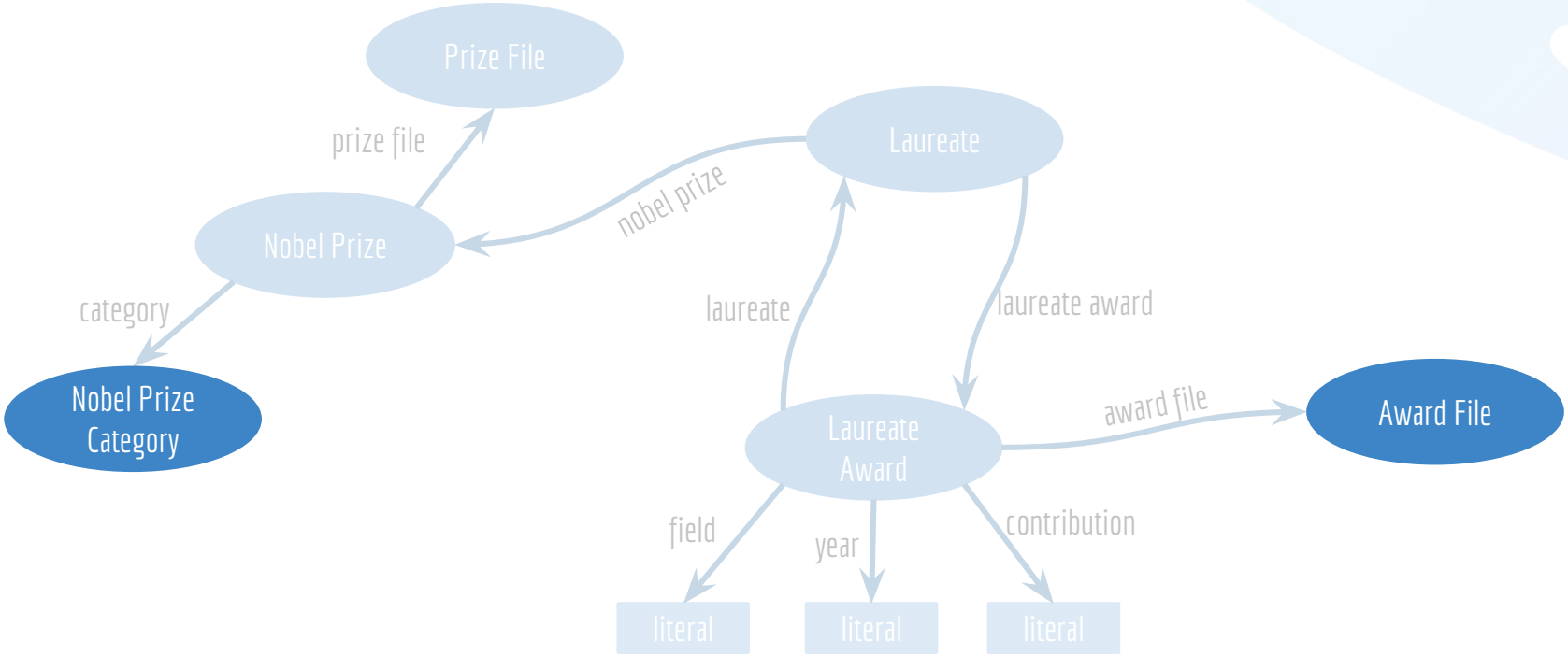
SEARCH



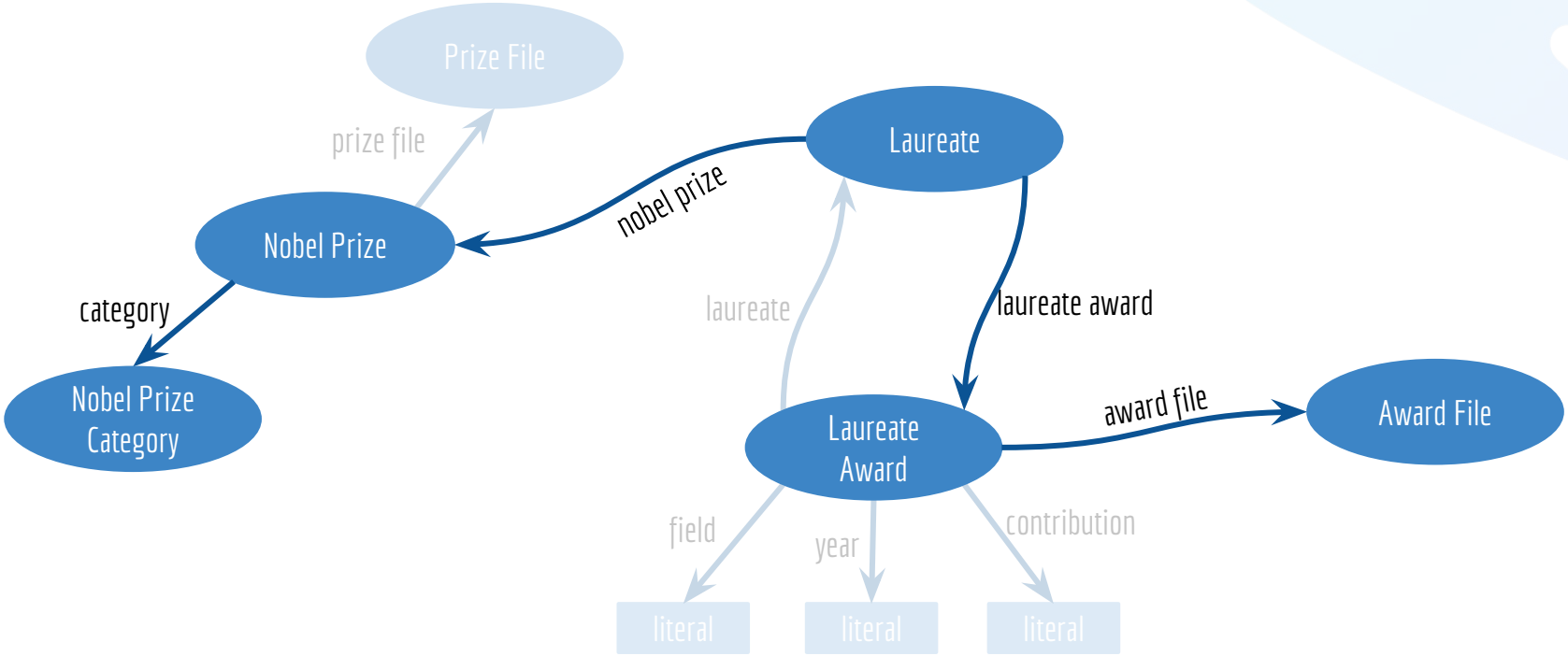
IDENTIFY ELEMENTS



FRAGMENT GENERATION

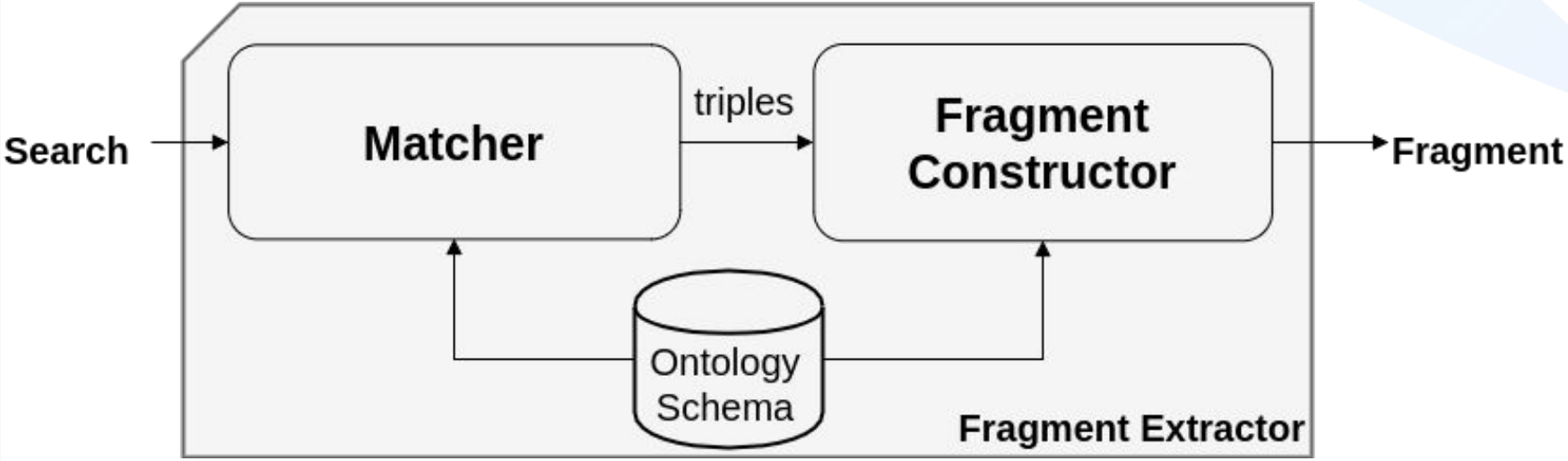


FRAGMENT GENERATION

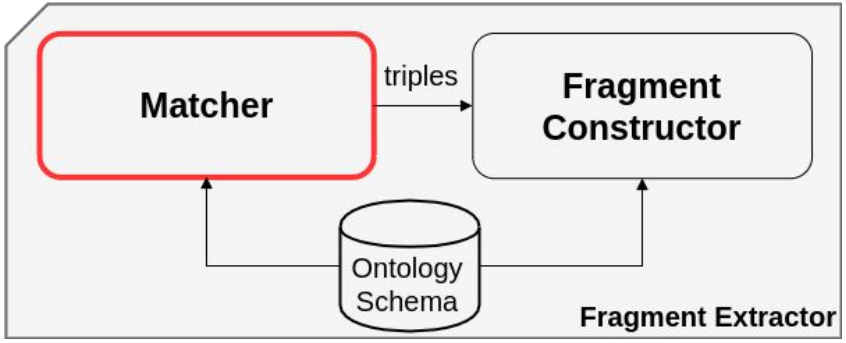


3. Fragment Extraction

OVERVIEW

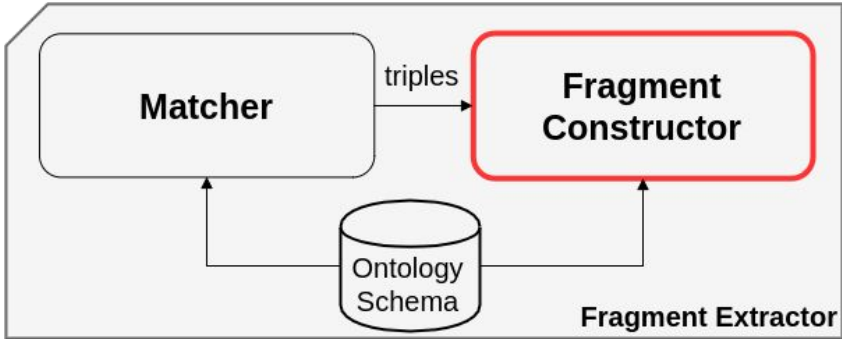

Search
Matcher
triples
Fragment Constructor
Fragment
Ontology Schema
Fragment Extractor

COMPONENTS



- **Matcher**
  - Identify, from the keywords, the classes and properties known by the ontology
  - Generate triples representing the classes and properties identified

COMPONENTS



- **Matcher**
  - Identify, from the keywords, the classes and properties known by the ontology
  - Generate triples representing the classes and properties identified

- **Fragment Constructor**
  - Build a graph from triples, representing an initial fragment
  - Identify which properties are reachable from the fragment

MATCHER

- Algorithm ***Keyword Matcher***
  - Identify *keywords* inside the search
  - Iterate over the *keywords* in sequence
  - Retrieve the ontology terms (classes and properties) most similars:
    - *Jaro-Winkler*(WINKLER, 1999) similarity
    - Synonyms map (optional)
    - Some keywords can be composed (multiple tokens):
      - Algorithm ***Permuted Jaro-Winkler***
  - Return triples

MATCHER

- ***Permuted Jaro-Winkler***
  - Calculate the similarity among the tokens
  - Pair the most similars
  - Calculate the mean similarity among then

- Example: **nobel file prize** x **prize nobel file**
  - *Jaro-Winkler*: 72%
  - Permuted *Jaro-Winkler*: 100%

FRAGMENT CONSTRUCTOR

- Algorithm ***Fragment Constructor*** (SAHA, 2016)
  - Generates the graph from the triples retrieved by *Keyword Matcher*
  - If the graph is disconnected:
    - calculate the *Steiner Tree* from the graph over the ontology
    - apply *Dijkstra* shortest path algorithm to generate compressed edges between two classes
    - calculate the *Minimum Spanning Tree* with *Prim*'s algorithm
    - uncompress the edges
  - Update the graph with all classes properties
  - Return the graph as a fragment

4. Experiments

EXPERIMENTS

- We used two different ontologies:
  - Nobel Prizes (datahub.io)
  - LinkedBrainz

- *Gold Standard*
  - We proposed, from the ontologies, and the minimum fragments that must be returned
  - The accuracy is represented by the *gold standard* contingency inside the fragment

EXPERIMENTS

- To each search we ran 100 executions at the following scenarios:

SCENARIO	DESCRIPTION
C1	Search tokens randomly permuted
C2	Search tokens' characters randomly replaced
C3	Union of C1 and C2
C4	Search tokens randomly replaced by synonyms
C5	Union of C1 and C4

EXPERIMENTS

- Nobel Prize

SEARCH	C1	C2	C3	C4	C5
prize file from nobel prize laureate	93%	98.5%	88.5%	100%	82%
field of laureate award from laureate	85%	98%	66%	100%	65%
award file and prize file from laureate	61.75%	84.5%	52.25%	100%	56.5%

EXPERIMENTS

- Nobel Prize

SEARCH	C1	C2	C3	C4	C5
prize file from nobel prize laureate	93%	98.5%	88.5%	100%	82%
field of laureate award from laureate	85%	98%	66%	100%	65%
award file and prize file from laureate	61.75%	84.5%	52.25%	100%	56.5%

EXPERIMENTS

- LinkedBrainz

SEARCH	C1	C2	C3	C4	C5
musicArtist records	100%	85.5%	87%	100%	87%
track release	100%	86%	84%	100%	91%
musicalWork from musicArtists	100%	83%	85.5%	100%	87.5%

EXPERIMENTS

- LinkedBrainz

SEARCH	C1	C2	C3	C4	C5
musicArtist records	100%	85.5%	87%	100%	87%
track release	100%	86%	84%	100%	91%
musicalWork from musicArtists	100%	83%	85.5%	100%	87.5%

EXPERIMENTS

- LinkedBrainz

SEARCH	C1	C2	C3	C4	C5
musicArtist records	100%	85.5%	87%	100%	87%
track release	100%	86%	84%	100%	91%
musicalWork from musicArtists	100%	83%	85.5%	100%	87.5%

- Nobel Prize

SEARCH	C1	C2	C3	C4	C5
prize file from nobel prize laureate	93%	98.5%	88.5%	100%	82%
field of laureate award from laureate	85%	98%	66%	100%	65%
award file and prize file from laureate	61.75%	84.5%	52.25%	100%	56.5%

EXPERIMENTS

- LinkedBrainz

SEARCH	C1	C2	C3	C4	C5
musicArtist records	100%	85.5%	87%	100%	87%
track release	100%	86%	84%	100%	91%
musicalWork from musicArtists	100%	83%	85.5%	100%	87.5%

- Nobel Prize

SEARCH	C1	C2	C3	C4	C5
prize file from **nobel prize** laureate	93%	98.5%	88.5%	100%	82%
field of l**aureate award** from laureate	85%	98%	66%	100%	65%
award file and **prize file** from laureate	61.75%	84.5%	52.25%	100%	56.5%

5. Conclusions and Future Works

CONCLUSIONS

- Some tokens permutations can imply in a smaller accuracy
- Work only with the schema allowed use a generic approach
- Graph algorithms proved effective to retrieve the fragments

FUTURE WORKS

- Perform more experiments with larger ontologies
- Try different approaches to identify the terms inside *Matcher*
- Identify most important properties to use in fragment construction
- Use the fragment in some *Information Retrieval* scenarios

Thank you!

Questions?

You can find me at

- @lucaspg96
- lucasperes@lia.ufc.br

REFERENCES

- **Elbassuoni**, Shady, and Roi Blanco. "Keyword search over RDF graphs." Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.
- **Musetti**, Alberto, et al. "Aemoo: Exploratory search based on knowledge patterns over the semantic web." Semantic Web Challenge 136 (2012).
- **Ouksili**, Hanane, et al. "Using Patterns for Keyword Search in RDF Graphs." EDBT/ICDT Workshops. 2017.
- **Saha**, Diptikalyan, et al. "Athena: An ontology-driven system for natural language querying over relational data stores." Proceedings of the VLDB Endowment 9.12 (2016): 1209-1220.
- **Winkler**, William E. "The state of record linkage and current research problems." Statistical Research Division, US Census Bureau. 1999.