UNIVERSITÉ PARIS SACLAY

DEPARTMENT OF COMPUTER SCIENCE

Semantic Web and Ontologies Project

Authors
Idrissa Mahamoudou Dicko
Tyler Marino

Introduction

In modern computing data has become of highest concern as it is used to make machines more intelligent. However the **World Wide Web**, is complex, full of links and nodes containing both various data types, private, public, ”structured” and ”unstructured”. Current web scraping methods used to train big models only use about 15% of the content that exists on the web. A proposed solution is the **Semantic Web**, a graphical structure, growing and pruning as people add and take away from the online environment that they engage with often.

On this Semantic Web, open data is accessible by common relationships as well as more specific traits/properties/characteristics that apply to domains and areas of special concern. It is a centralized place, similar to platforms like social networks and wikepedia, where traversing from any node to another should be possible assuming semantic understanding exists.

This linked data is designed in such a way that a standard protocol for retrieving information, or at least a well documented protocol is designed in conjunction with the addition for the data that is added to the data chain. This method will allow for data to be utilized by many different agents. There currently exists a challenge in sourcing and accessing data as we will discuss further later, so having open source well documented data will allow for more value to come from the data, more sharing of ideas and progress towards valuable discoveries all while expanding the web.

Going forward more informed searches, i.e. starting your search with a-priori information, will allow progressive ideas and methods to be applied to problems that were hard to test before.

Problem statement

The Exponential growth of the data on the web has led to the development of the **Semantic Web**, where information is structured and interconnected in a machine-readable form through **Linked Data.** While Large Language Models (LLMs) like ChatGPT are effective in answering many natural language queries, they often face limitations when dealing with highly structured, interlinked, and verifiable data across distributed sources. The problem, therefore, is to **explore how Semantic Web technologies** particularly **Linked Data and SPARQL querying** can complement LLMs in answering complex questions that require structured reasoning, factual verification, or **cross-dataset integration**.

This project specifically addresses the following challenges:

- Identifying questions that cannot be fully or accurately answered by LLMs but can be answered using Linked Data.

- Designing and executing SPARQL queries over Linked Data Sources such DBpedia, Wikidata, or Linked Data to obtain precise answers.

- Proposing and answering a complex natural language query using a hybrid approach that combines Linked Data reasoning with the interpretive and generative capabilities of LLMs.

- Evaluating the advantages, limitations, and justifications of purely Semantic Web–based answers versus hybrid approaches.

So we investigate these challenges while exploring two questions:

- What proteins are involved in signal transduction and are related to pyramidal neurons?

- What recipes from italian cuisine are safe to eat for people who feel bodily harm from pine nuts?

Methods

Unstructured data is a bountiful resource that to the untrained eye is not too valuable, however there is a great push to train models that can digest the data into a form that can be standardized or at least well documented. A natural way to do this is to find semantic similarities and store a set of data with associated counterparts.

The most natural form that this takes is a web or graphical strucute where objects(nodes) have relationships(edges) with other objects. This structure allows for clustering and centrality in the data for quick extraction and additions. These components not only help search, but infer about data based on existing relationships, thus allowing for easier interpolation and extrapolation.

Holding such properties that link data together, we may easily represent it in existing schema/form such as OWL/RDF/RDFS which allows us to easily preform queeries to find data of interest. The queeries are easily translated from pseudo code, i.e. natural language request, into SparQL queeries that can take parse through large amounts of data relatively quickly. So upon finding linked data that comes from soures like DBpedia[2], Wikidata[1], lod-cloud, etc. one may draft a natural language queery. The strengthen ones ability, large lange models such at ChatGPT[3] or Mistral are capable of finding data sources, converting natural language into computer readable queeries to find solutions to complex problems that even the strongest machine cannot find in all of its memory.

Solution

The problems stated above are great examples of challanging concepts for an LLM to parse through alone, but when we use valid SPARQL queeries we are able to find answers that contain the information we desire. Below we demonstrate the process, theory and application of the semantic web.

Question 1: What proteins are involved in signal transduction and are related to pyramidal neurons?

We begin with a simple graphical representation of our data(Figure 1). In using the semantic web it is best to be able to visualize how key components are structured in order to access them accurately. Notice how ”Protien xx” may not directly be related to pyramidal neurons but it is related to an object that lives in a subclass of it and similar to the process it is involved with. This is why linked data can provide robust set of information. One note is the complexity of the graph. It can grow very fast with just a few key points and the associated data.

Having a better understanding of our graphical data, we may then parse the informationusing the SPARQL queery(Listing 1). Notice the structure of the queery. It begins by referencing where we are retrieving our information as well as the necessary ”packages” or predefined schema sets we will need to retrieve desired information. Following this, we select the values of interest and then we describe how to retrieve the information as well as filter for the data that meets our requirements.

Figure 2 shows the structured output of our search queery including the valuable information in computer and human readable forms.

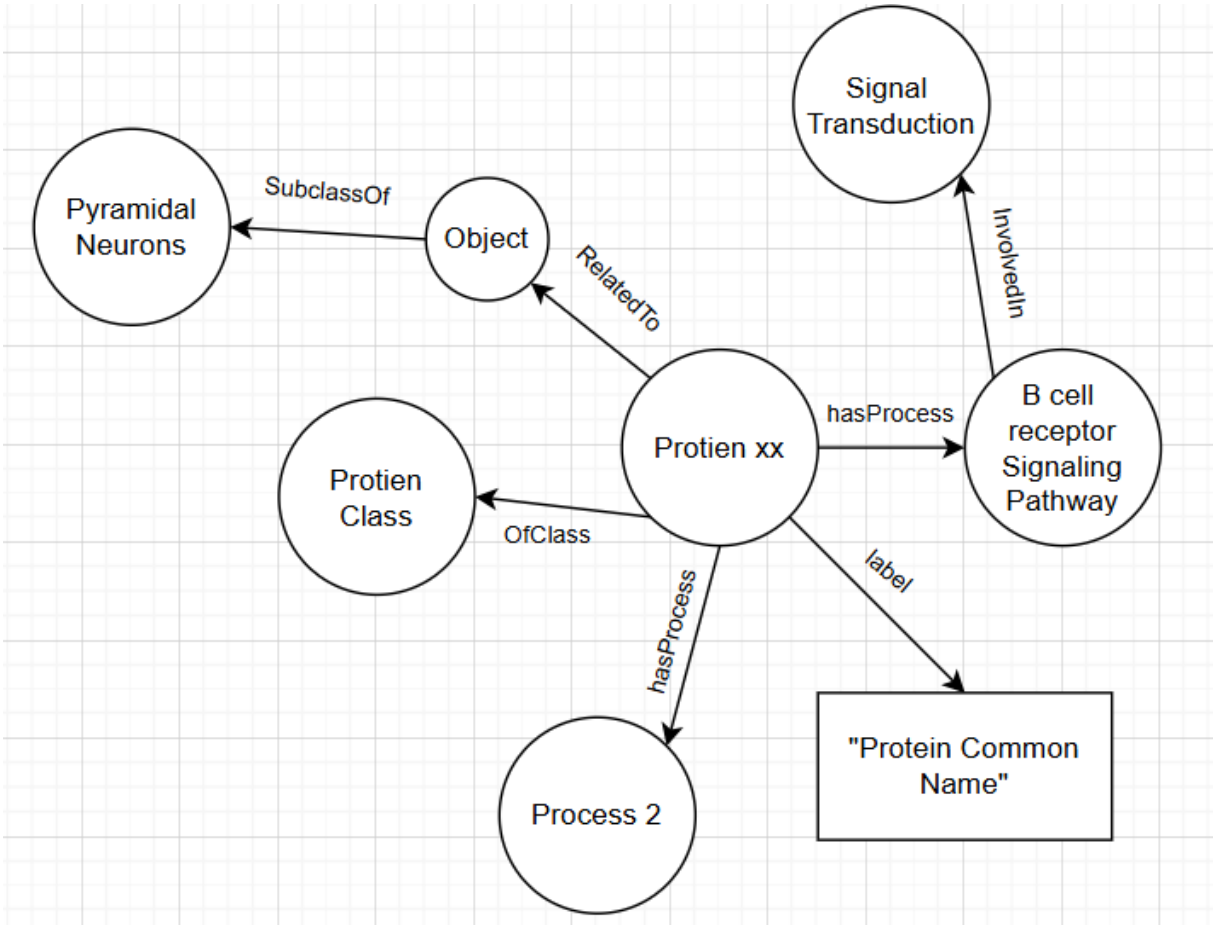

Figure 1: Example Graphical Representation of Protiens in the Database

Listing 1: Proteins in signal transduction related to pyramidal neurons

```
PREFIX wd:  <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX bd:  <http://www.bigdata.com/rdf#>

SELECT DISTINCT ?protein ?proteinLabel ?gene ?geneLabel ?processLabel ?
    cellLabel
WHERE {
  ?protein wdt:P31 wd:Q8054 .          # protein
  ?protein wdt:P682 ?process .         # biological process
  ?process wdt:P279* wd:Q828130 .      # subclass of signal transduction (GO
      :0007165)
  OPTIONAL { ?protein wdt:P702 ?gene . }
  OPTIONAL { ?protein wdt:P5572 ?cell .
             ?cell wdt:P279* wd:Q2116409 . # pyramidal neuron or subclass
           }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],
      en". }
}
LIMIT 100
```

protein	proteinLabel	gene	geneLabel	processLabel
wd:Q14874000	C-abl oncogene 1, non-receptor tyrosine kinase	wd:Q14873999	Abl1	B cell receptor signaling pathway
wd:Q587961	ABL proto-oncogene 1, non-receptor tyrosine kinase	wd:Q14873998	ABL1	B cell receptor signaling pathway
wd:Q14863583	B cell leukemia/lymphoma 2	wd:Q14863437	Bcl2	B cell receptor signaling pathway
wd:Q425201	BCL2 apoptosis regulator	wd:Q14863436	BCL2	B cell receptor signaling pathway
wd:Q401536	BLK proto-oncogene, Src family tyrosine kinase	wd:Q17848770	BLK	B cell receptor signaling pathway
wd:Q21106693	BMX non-receptor tyrosine kinase	wd:Q17849815	BMX	B cell receptor signaling pathway

Figure 2: Proteins in signal Transduction related to pyramidal neurons

Question 2: What recipes from italian cuisine are safe to eat for people who feel bodily harm from pine nuts?

The flow for question 2 resembles the flow in question 1. Starting with a graph(Figure3) to understand pseudo strucutre, then generating a SPARQL queery(Listing 2) and retrieving a structured output(Figure 4).

We continue with the sturctured output and display the human readable key value pair from the machine readable code (Figure 5). Finally in Figure 6 we compare a LLM output. We see that yes the LLM is capable of generating a nice list but may not truly sift out all of the items containing pinenuts if it is not directly an ingredient, but a subingredient.

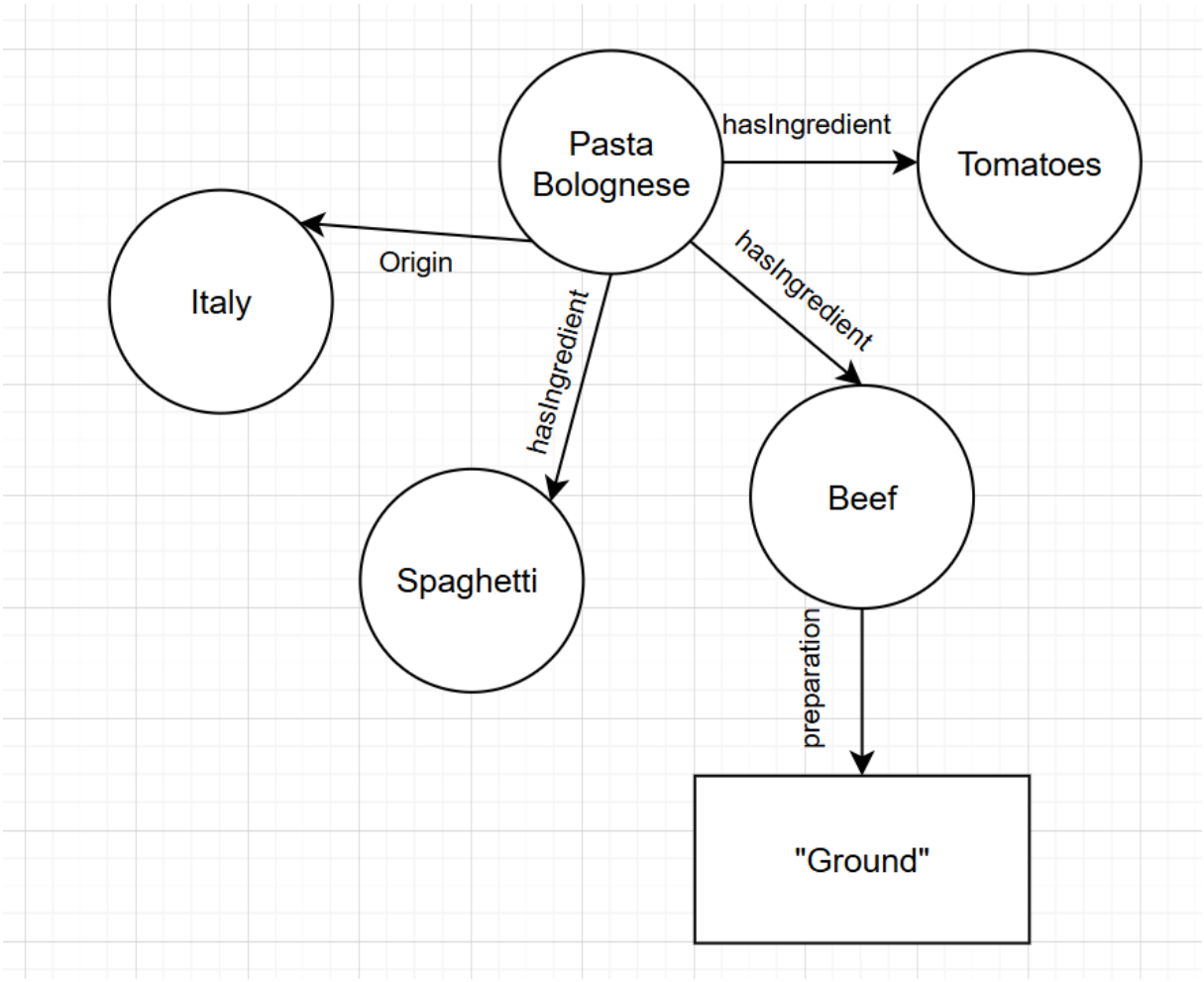

Figure 3: Example Graphical Represenation from the Recipes Database

A SPARQL query over Wikidata for Italian dishes can list recipes. We then exclude those containing pine nuts (ingredient `wd:Q40737`). The query returns dishes such as Agliata, Caprese salad, Castagnaccio, Fettuccine Alfredo, etc.

Listing 2: Italian dishes with ingredients that exclude pine nuts

```
PREFIX wd:  <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX bd:  <http://www.bigdata.com/rdf#>

SELECT DISTINCT ?dish ?dishLabel
WHERE {
  ?dish wdt:P31/wdt:P279* wd:Q746549 .
  { ?dish wdt:P2012 wd:Q1775756 } UNION { ?dish wdt:P495 wd:Q38 } .

  # only consider items that actually list some ingredient/material
  { ?dish wdt:P527 ?anyIng } UNION { ?dish wdt:P186 ?anyIng } .

  FILTER NOT EXISTS {
    VALUES ?pineNut { wd:Q40737 }
    { ?dish wdt:P527 ?i } UNION { ?dish wdt:P186 ?i } UNION { ?dish wdt:P4950 ?
        i } .
    FILTER (?i = ?pineNut)
  }

  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],
      en". }
}
ORDER BY ?dishLabel
LIMIT 50
```

dish	dishLabel
wd:Q3606603	Agliata
wd:Q21210955	Agnolini
wd:Q116842438	Amaretti di Sassello
wd:Q3646108	Brutti e buoni
wd:Q3647314	Buridda
wd:Q275508	Caesar salad
wd:Q61743260	Capra e fagioli
wd:Q568364	Caprese salad
wd:Q3661969	Cassatella di Sant'Agata
wd:Q3662049	Castagnaccio
wd:Q3676324	Ciauscolo
wd:Q3695911	Cotoletta alla bolognese
wd:Q2915859	Fettuccine Alfredo
wd:Q3753798	Frustingo

Figure 4: Italian Recipes free from pine nuts

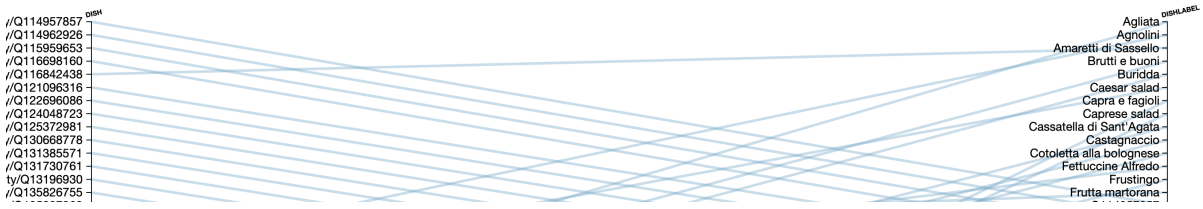

Figure 5: Italian Recipes free from pine nuts (dimensional view)

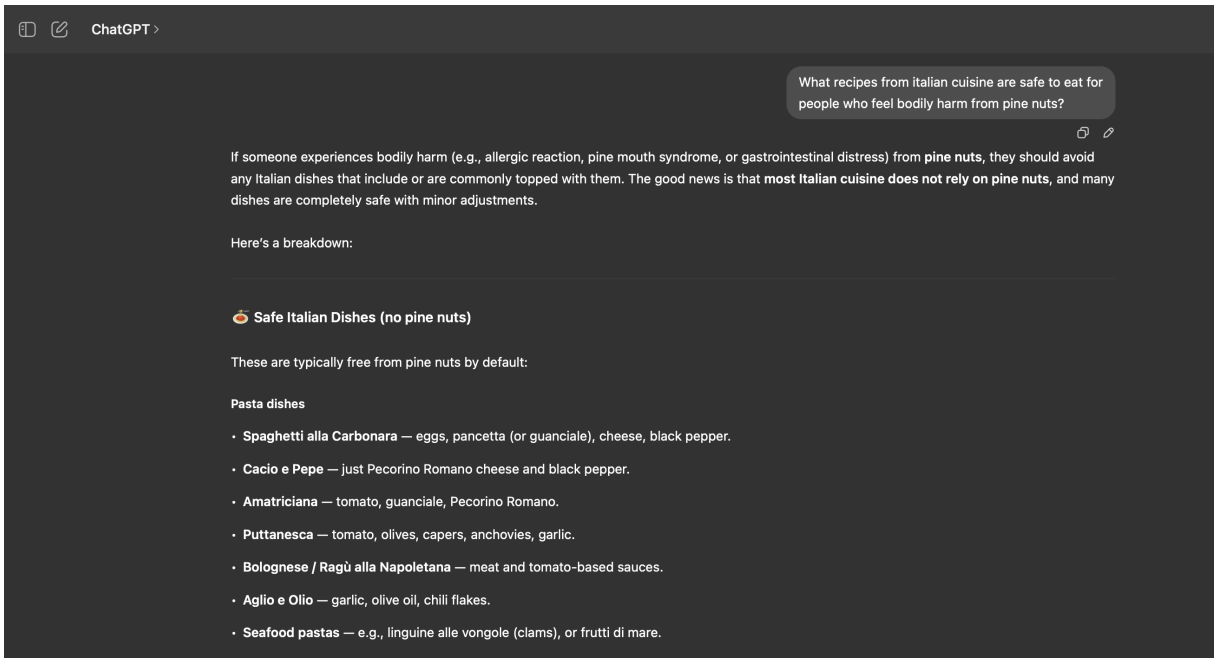

Figure 6: Answer Chatgpt

ChatGPT provides explanations: Safe pasta dishes include Carbonara, Cacio e Pepe, Amatriciana, Puttanesca, Ragù alla Napoletana, Aglio e Olio, and seafood pastas. However, **Pesto alla Genovese** should be avoided because pine nuts are a core ingredient.

Question 3

One advantage of Linked Data is that answers can be accompanied by explicit **justifications**. Instead of returning only values with a `SELECT` query, we can use `CONSTRUCT` to return the RDF triples that explain why an entity is included in the result set.
For example, consider the question:

> **What proteins are involved in signal transduction and are related to pyramidal neurons?**

This can be expressed as the following `CONSTRUCT` query over Wikidata:

Listing 3: Justifications with CONSTRUCT in SPARQL

```
PREFIX wd:  <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>

CONSTRUCT {
  ?protein wdt:P682 wd:Q828130 .
```

```
  ?protein wdt:P702 ?gene .
}
WHERE {
  ?protein wdt:P31 wd:Q8054 .
  ?protein wdt:P703 wd:Q15978631 .
  ?protein wdt:P682 wd:Q828130 .
  OPTIONAL { ?protein wdt:P702 ?gene . }
}
LIMIT 50
```

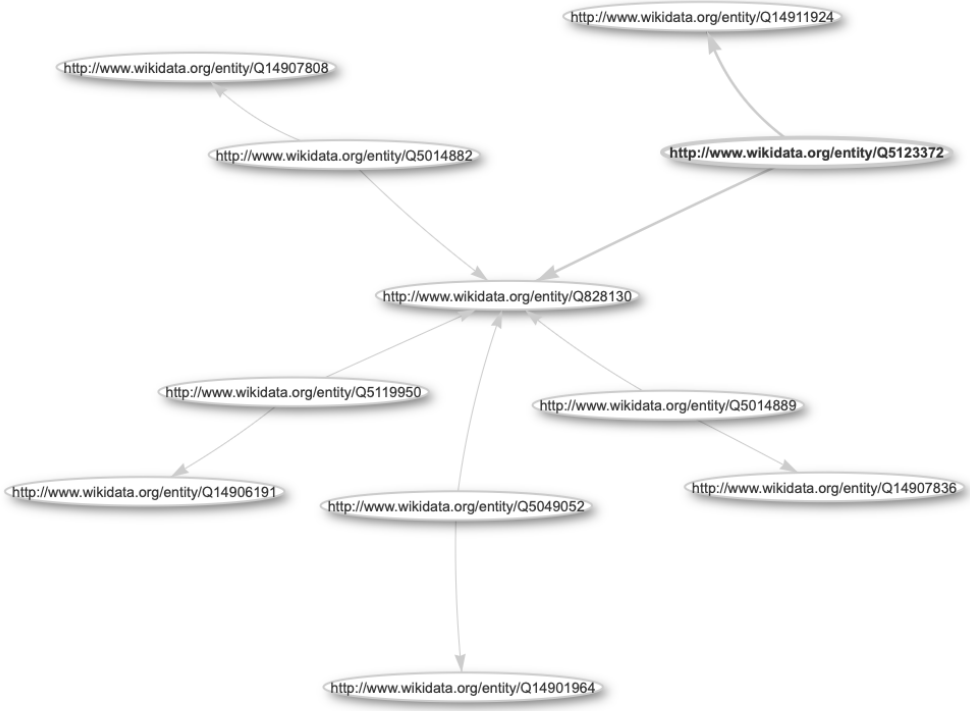

Figure 7: Graph from the construct Query

subject	predicate	object
wd:Q5014882	wdt:P682	wd:Q828130
wd:Q5014882	wdt:P702	wd:Q14907808
wd:Q5014889	wdt:P682	wd:Q828130
wd:Q5014889	wdt:P702	wd:Q14907836
wd:Q5049052	wdt:P682	wd:Q828130
wd:Q5049052	wdt:P702	wd:Q14901964
wd:Q5119950	wdt:P682	wd:Q828130
wd:Q5119950	wdt:P702	wd:Q14906191
wd:Q5123372	wdt:P682	wd:Q828130
wd:Q5123372	wdt:P702	wd:Q14911924
wd:Q5023958	wdt:P682	wd:Q828130

Figure 8: Construct query

This query does not only output identifiers of proteins, but the RDF statements themselves, i.e., the **justifications** for the answers.

Hybrid approach: LLMs contribution

In a hybrid system, ChatGPT (or another large language model) complements Linked Data by:

- **Turning structured output into natural language:** It takes the RDF-based results and crafts readable explanations (e.g. “Pesto includes pine nuts, so avoid it; but Carbonara or Aglio e Olio are safe”).
- **Adding domain context and suggestions:** It can recommend alternatives (e.g. nut-free pesto recipes) or explain why certain dishes are risky or safe.
- **Handling missing data / uncertainty:** If some dishes lack ingredient records in the graph, the LLM may infer their safety status based on culinary knowledge or suggest additional dataset sources or checks.
- **Interactive Q&A and clarification:** Users can ask follow-ups or ambiguous cases, which LLMs can address seamlessly, referencing both the structured data and general knowledge.

Thus the hybrid architecture ensures correctness via Linked Data and understandability via LLMs, offering both verifiable evidence and human-friendly explanations.

Figure 9

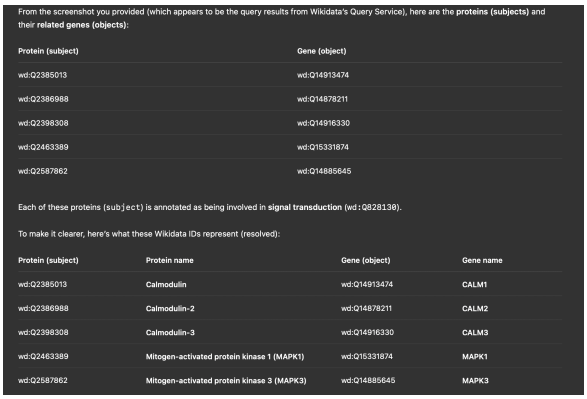

Figure 10

Figure 11: Hybrid method answer

Benefits and Limitations of the Techniques

Benefits

- **Verifiability and Justification:** Linked Data queries (e.g., SPARQL with `CONSTRUCT`) provide explicit RDF triples as justifications. This means results are not only reproducible but can be traced back to the source dataset, increasing transparency and trust.

- **Cross-Dataset Integration:** Semantic Web technologies allow queries that integrate multiple heterogeneous datasets (e.g., DBpedia, Wikidata, Linked Life Data, FoodOn). This makes it possible to answer complex questions that span biology, medicine, and cultural domains.

- **Precision in Retrieval:** Since data is structured in RDF/OWL ontologies, queries can directly filter for specific classes, properties, and relationships (e.g., "proteins involved in signal transduction" or "dishes without pine nuts"), reducing ambiguity.

- **Hybrid Enhancement with LLMs:** When combined with ChatGPT (GPT-5), Linked Data results become more interpretable. LLMs can explain the meaning of the RDF triples in natural language, suggest alternatives, handle uncertainty, and provide contextual insights for end users.

Limitations

- **Incomplete Coverage:** Many entities in Wikidata or DBpedia lack detailed annotations (e.g., missing ingredient information for recipes, or missing cell-type

associations for proteins). Queries often return empty results due to incomplete modeling.

- **Complex Query Design:** SPARQL requires specialized knowledge, and queries can be verbose. Small mistakes in properties or ontology mappings often result in no answers.
- **Heterogeneity Across Sources:** Different datasets may use slightly different ontologies or property conventions (e.g., `has part`, `ingredient`, `material used`). Aligning these across knowledge bases is non-trivial.
- **LLM Limitations:** While ChatGPT can provide context and explanations, it cannot guarantee factual correctness without grounding in Linked Data. It may "hallucinate" missing details if used alone.
- **Scalability Issues:** Very large or federated queries may be slow to execute or may time out when accessing remote SPARQL endpoints.

Summary

Overall, Semantic Web technologies excel at providing structured, verifiable, and cross-domain answers by analyzing multiple levels of relationships to extract and filter correct information.

References

[1] Wikidata: A free and open knowledge base. `https://www.wikidata.org/`. Accessed: 2025-10-07. pages 3

[2] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2015. pages 3

[3] OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, 2025. Accessed: 2025-10-07. pages 3