

# Knowledge Extraction of Union Catalogue using Semantic and Ontology

Dharmeshkumar Shah<sup>(1, 4)</sup>, Harshal Arolkar<sup>2</sup> and Ashish Kumar Chauhan<sup>3</sup>

<sup>1</sup> Scientist-B (CS), Information and Library Network (INFLIBNET) Centre, Infocity, Gandhinagar-382007, India

<sup>2</sup> Professor & Head, PG Programme, FCT & FCAIT, GLS University, Ahmedabad, Gujarat-380006, India

<sup>3</sup> Library Officer (LS), Information and Library Network (INFLIBNET) Centre, Infocity, Gandhinagar-382007, India

<sup>4</sup> GLS University, Ahmedabad, Gujarat-380006, India

## Abstract

Information extraction and exploration are crucial tasks in the era of big data and knowledge-intensive applications. Traditional approaches often need help with data and to leverage semantic knowledge effectively. This paper proposes a semantic knowledge-based framework for information extraction and exploration using Simple Protocol and RDF Query Language (SPARQL) available in the Union Catalogue of Gujarat Colleges (GujCat). The framework incorporates semantic technologies and exploits the power of SPARQL queries to extract and explore structured information from diverse data sources. A semantic information retrieval system provides data using rule-based inference from the ontology. Users can query information from a database or any other data source that can be mapped to Resource Description Framework (RDF) using the "SPARQL Protocols as well as the RDF Query Language." The SPARQL standard, created and supported by the World Wide Web Consortium (W3C), enables developers and users to focus on anything they want to know rather than how a database is structured.

## Keywords

GujCat, Data Mining, Information Retrieval, GConTO, RDF, SPARQL

## 1. Introduction

The Online Union Catalogue for Indian Universities (IndCat) has been produced by the Information and Library Network (INFLIBNET) Centre. "IndCat is the largest free online union catalogue of books, theses, and serials from significant university/institute libraries in India. Inter-library loans (ILL), copy cataloguing, collection building, and metadata retro-conversion are the main uses of IndCat. Similar to IndCat, GujCat is a subset of INFLIBNET's Online Union Catalogue of Gujarat Colleges. It is a centralised online library catalogue that lists all the books that are accessible in the major college libraries in Gujarat state [10].

The task of information extraction from unstructured or semi-structured data sources is a challenging endeavour in the field of data processing and knowledge discovery. Existing approaches on the current GujCat website often need help to extract structured information from diverse data formats effectively and cannot leverage semantic knowledge for improved accuracy and relevance. Information extraction is the process of automatically extracting organised information from unorganised sources, such as entities, interactions among things, and

---


Joint Proceedings of Second International Workshop on Semantic Reasoning and Representation in IoT (SWIoT-2023) and Third International Workshop on Multilingual Semantic Web (MSW-2023), November 13-15, 2023, University of Zaragoza, Zaragoza, Spain

EMAIL: dashah@inlibnet.ac.in (D. Shah); harshal.arolkar@glsuniversity.ac.in (H. Arolkar); ashish01kc@gmail.com (A. K. Chauhan)  
ORCID: 0000-0001-9081-1998 (D. Shah); 0000-0003-0371-4466 (H. Arolkar); 0000-0003-0635-8815 (A. K. Chauhan)



© 2023 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

properties defining entities. This makes it possible to conduct far richer types of inquiries on the numerous unorganised resources than would be feasible with only keyword searches [2].

Union catalogues are essential for allowing information access across various collections housed by numerous libraries or repositories. Users can more effectively search for and retrieve materials thanks to these catalogues, which compile metadata from numerous sources. However, obtaining and analysing data from union catalogues can frequently be difficult due to the variability of data sources and the absence of standardised forms. Making a semantically based framework that uses SPARQL for information extraction and exploration can solve this issue. As a result, there is a need for a semantic knowledge-based framework that can overcome these limitations and provide robust information extraction capabilities. Although mining is an intriguing word to use, it is not a good metaphor to describe the overall knowledge discovery process [3] and what people really do in the field [5]. The key challenges faced in the context of union catalogues are heterogeneous data sources, semantic knowledge extraction, complex querying requirements, scalability and performance, information exploration and visualization.

Addressing these challenges requires the development of a semantic knowledge-based framework for information extraction that leverages the power of SPARQL. SPARQL is a query language specifically designed for querying and manipulating data stored in RDF format. RDF is a standard for representing data on the web in a structured manner, where information is expressed as subject-predicate-object triples. SPARQL enables users to extract, explore, and reason over RDF data through powerful, flexible queries. Main key features and aspects of SPARQL include querying RDF data, pattern matching, variable binding, filtering and expressive functions, joins and optional patterns, aggregation and grouping, inferencing and reasoning, protocol for querying. Such a framework should enable the extraction of structured information from unstructured and semi-structured data sources by effectively incorporating semantic knowledge encoded in ontologies and knowledge graphs. It should provide robust querying capabilities, support the integration of heterogeneous data sources, and facilitate information exploration and visualization for improved understanding and decision-making.

Overall, the problem statement is that the need for a comprehensive framework that harnesses semantic knowledge and SPARQL to overcome the limitations of existing approaches, enabling more accurate, context-aware, and efficient information extraction from diverse data sources.

## **2. Related works**

The purpose of the semantic web is to enhance the current web of text-based documents/metadata with a layer that machines can understand. To accomplish this, specific requirements must be met. These include automating the processes of creating semantic annotations, connecting web content with ontologies, and developing and using ontologies for interaction [7]. Plugins are accessible for data processing in the following languages: French, German, Italian, Romanian, Arabic, Chinese, Hindi, and Russian. In some circumstances, the functionality of these plugins is classified as "basic," implying that they have valuable processing resources that can be used as a foundation for constructing applications [8].

Ontology learning techniques created an ontology by analysing unstructured and semi-structured material. The term was originally used by Maedche and Staab (1999) [9]. Information extraction relies heavily on ontologies, which are formal, explicit specifications of conceptualisation [11].

Cunningham et al. (2006) discovered that this is achieved by taking into account the relationships between concepts and extracting the information using a regulation expression.

This strategy uses the NLP framework and platform to apply linguistic rules for knowledge construction [4].

Shah and Jain (2014) observed that machines can now comprehend the meaning of data and information and make linkages between various entities and concepts to ontologies. The Semantic Web can facilitate more advanced information search, discovery, and integration across multiple domains and applications by annotating and linking data with ontologies [16].

Another study by Wimalasuriya, D.C., Dou, D. (2010), says that creating and using a semantic knowledge base is necessary for the knowledge-based framework's execution. Advanced reasoning techniques can be used to analyse and organise this knowledge base to draw new and intriguing facts from the provided domain data, which end users can then explore in an informed manner [18].

Afolabi et al. (2017), explains an ontology-based association rule mining approach for extracting knowledge from text. The methodology (Afolabi et al., 2017) incorporates keyword extraction and weighting based on the term frequency method as part of the data collecting and cleaning process [1].

Jiomekong & Tiwari (2023), explains that the research aimed to curate an Open Research Knowledge Graph (ORKG) with papers related to ontology learning and define an approach using ORKG as a computer-assisted tool to organize critical insights extracted from research papers. The ORKG was used to document the "epidemiological surveillance systems design and implementation" research problem and prepare related work. open research knowledge grapy as computer assistant tool consist of five tools like knowledge elicitation, Knowledge analysis and interpretation, Template creation, Knowledge representation, Verification and validation. [20].

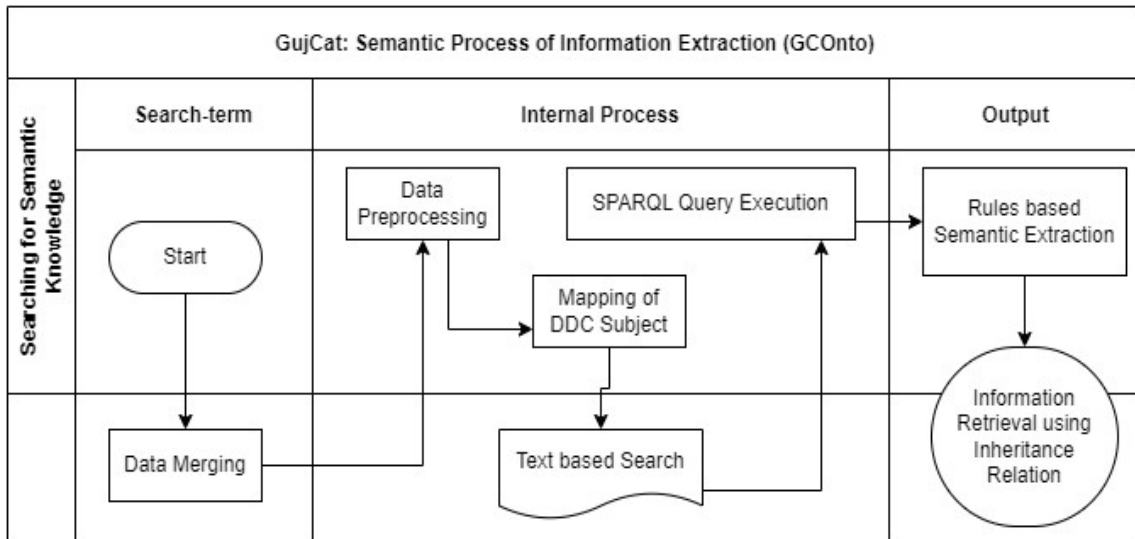
Amara et al. (2023), provides an in-depth analysis of semantic interoperability in Industry 4.0, highlighting its core concepts, problems, and implications for intelligent manufacturing. It explores the potential of semantic technologies like ontologies, linked data, and standard data models [19].

Khorashadizadeh et al. (2023), evaluates the use of ChatGPT in GPT-3.5, a large foundation model, to improve knowledge graph construction and completion. The qualitative analysis reveals ChatGPT's potential, but challenges like bias, hallucinations, and high computational costs must be addressed [21].

### **3. Proposed Research Framework of GCOnto**

The research effort reported in this paper claims that domain knowledge can significantly contribute to the activities of Information Extraction from unstructured data. This section presents a proposed process of information extraction as shown in Figure 1 for comprehensive, semantic-driven methodology for utilising that knowledge to improve the machine learning techniques employed in relation extraction.

- Collecting unstructured data and content detection.
- Preprocessing and merging the domain and construction of the knowledge map.
- Recognising the named entities.
- Subject mapping with the DDC schemes.
- Collecting the training datasets from structured datasets.
- Building rules based semantic extraction using SPARQL.
- Extracting Relations from the unstructured data by using the classification models.

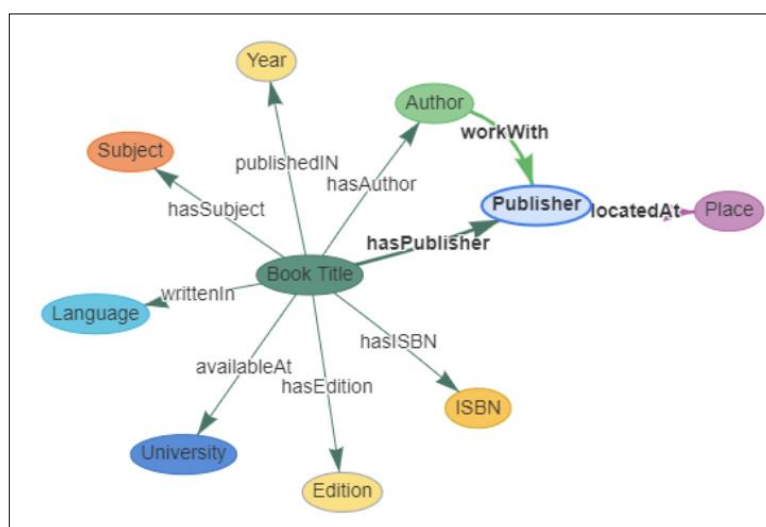


**Figure 1:** Proposed Process of Information Extraction from GCOnto

In this section, study discuss proposed schema of GCOnto model for searching for semantic knowledge as shown in Figure 2 and to achieve that researcher installed Apache Jena Fuseki and used Protege tool and for information extraction, researcher used a sample dataset with subject-wise metadata available in the Union Catalogue of Gujarat colleges (GujCat) using SPARQL query. A Semantic-based Framework for Information Extraction and Exploration of a Union Catalogue using SPARQL involves leveraging RDF and ontologies to represent the catalogue data and using SPARQL queries for extracting and exploring the information. The Web Ontology Language (OWL) is often used, assisted by ontology modelling tools like Protégé. [9]

Ontologies need a well-designed language and rigorous logical reasoning to be an effective method. Ontological knowledge bases come with the T-box, a terminological formalism, and the A-box, a declarative formalism. [12]. Ontology consists of Tbox, Abox and Graph (linking the relationships to ideas) as per below formula.

$$\text{Ontology GCOnto} = (\text{Terminological Formalism} + \text{Assertional Formalism}, \text{Graph})$$



**Figure 2:**

Proposed Schema of GCOnt

### 3.1. Inferencing and Reasoning

Formulate SPARQL queries to explore the relationships between entities and discover patterns and insights.

```
SELECT *
WHERE {
  ?book rdf:type :Book .
  ?book :hasTitle ?title .
  ?book :hasAuthor ?author .
  ?book :hasSubject ?subject .
  ?relatedBook rdf:type :Book .
  FILTER (?book != ?relatedBook) .
}
```

### 3.2. Information Retrieval based on Inheritance Relation

The primary semantic relationship among concepts is inheritance. When querying a parent class, it should encompass all its subclasses or child classes. Figure 3 shows the computer science subject-wise metadata exploration from GCOnto using SPARQL. For instance, when searching subjects related to "Computer Science, Information & General Works", the query statement would be:

```
SELECT ?all_sub_class
FROM GCOnto
WHERE { f: Computer science, knowledge and systems
f: value ?all_sub_class}
<rdfs:Class rdf:ID="Computer science, information & general
works"></rdfs:Class>
<rdfs:Class rdf:ID="Computer science, knowledge and systems"></rdfs:Class>
<rdfs:SubClassOf rdf:resource="#Computer science, information & general
works"/>
</rdfs:Class>
<rdfs:Class rdf:ID="General Knowledge of Computer"></rdfs:Class>
<rdfs:SubClassOf rdf:resource="#Computer science, knowledge and systems"/>
</rdfs:Class>
<rdfs:Class rdf:ID="The book (writing, libraries, and book-related
topics)"></rdfs:Class>
<rdfs:SubClassOf rdf:resource="#Computer science, knowledge and systems"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Systems"></rdfs:Class>
<rdfs:SubClassOf rdf:resource="#Computer science, knowledge and systems"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Data processing & computer science"></rdfs:Class>
<rdfs:SubClassOf rdf:resource="#Computer science, knowledge and systems"/>
</rdfs:Class>
<rdfs:Class rdf:ID="Computer programming, programs & data"></rdfs:Class>
<rdfs:SubClassOf rdf:resource="#Computer science, knowledge and systems"/>
</rdfs:Class>
```

## 4. Proposed Research Framework of GCOnto

- A semantic-based framework that enables efficient information extraction and exploration from GCOnto using SPARQL queries.
- Enhanced semantic interoperability and data integration among heterogeneous sources.
- Improved query performance and response times, even with large-scale data.
- A foundation for future research and development in the realm of semantic-based information extraction and exploration.

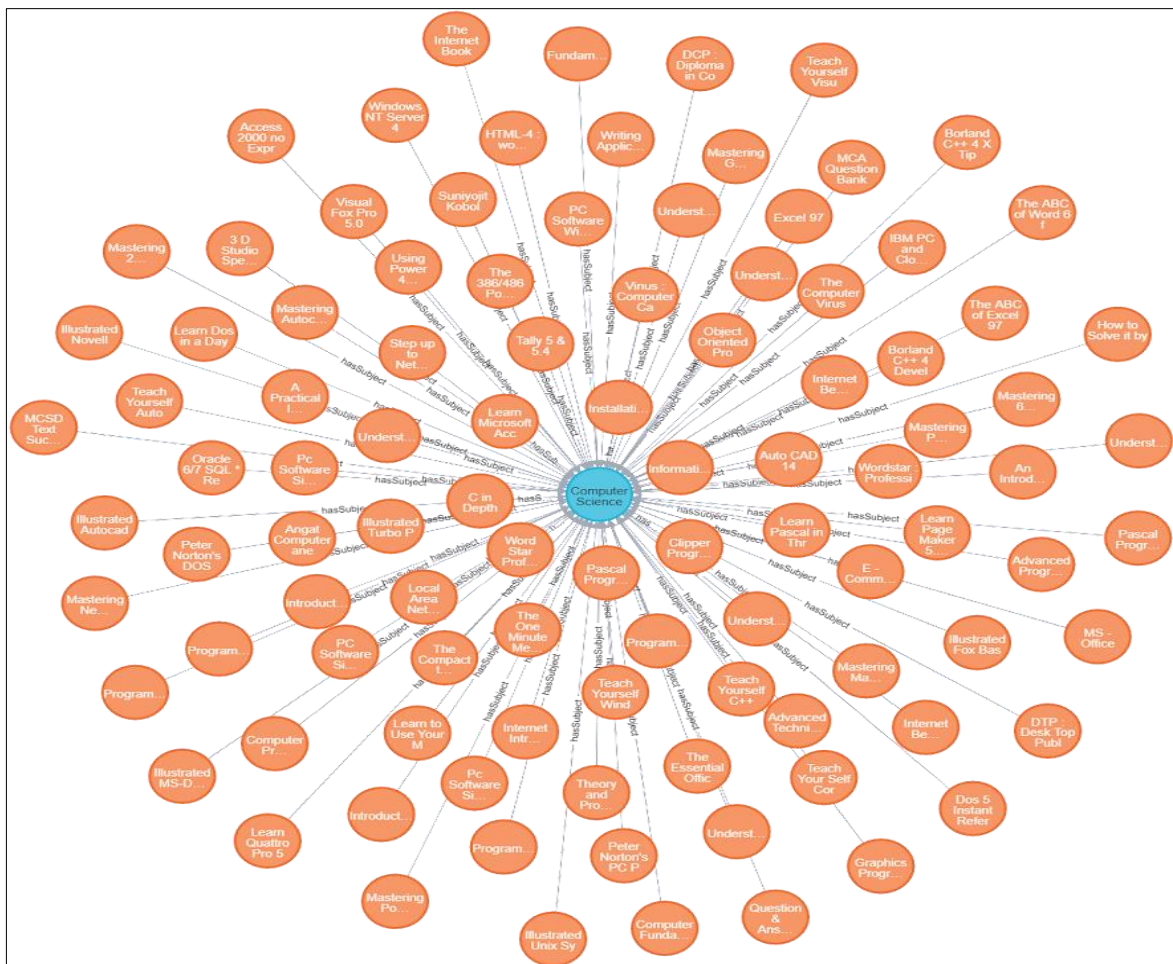


Figure 3: Computer Science subject-wise metadata exploration from GCOnto using SPARQL

## 5. Conclusion and Future Work

In this paper, study is reviewing information extraction from unstructured and semi-structured data in an accurate and efficient way using a semantic based framework. For semantic based information retrieval, it has proposed an ontology concept for union catalogue and SPARQL query to retrieve the result in an efficient way. Inference based on inheritance relation in subject hierarchy increases the precision and recall of result set data. Generating the semantic content from the proposed framework makes it a more interesting concept. GCOnto system can be used for identifying the semantic content for the semantic web and also implemented Ontology Based web service for results.

While searching the content in ontology, it returns a large number of results. Using the advanced techniques of NLP, analyse the outcome and find the most relevant subset of data to improve the relevancy of search text with the result.

## 6. References

- [1] I. Afolabi, O. Sowunmi, O. Daramola, Semantic association rule mining in text using domain ontology, *Int. J. Metadata, Semant. Ontol.* 12.1 (2017) 28. doi:10.1504/ijmso.2017.087646.
- [2] A. Aljamel, T. Osman, G. Acampora, Domain-Specific Relation Extraction - Using Distant Supervision Machine Learning, in: 7th International Conference on Knowledge Discovery and Information Retrieval, SCITEPRESS - Science and Technology Publications, 2015. doi:10.5220/0005615100920103.
- [3] K. BONTCHEVA, V. TABLAN, D. MAYNARD, H. CUNNINGHAM, Evolving GATE to meet new challenges in language engineering, *Nat. Lang. Eng.* 10.3-4 (2004) 349–373. doi:10.1017/s1351324904003468.
- [4] R. Studer, P. Warren, D. John, *Semantic Web Technologies: Trends and Research in Ontology-Based Systems*, Wiley & Sons, Limited, John, 2006.
- [5] N. Borisova, An Approach for Ontology Based Information Extraction, *Inf. Technol. Control* 12.1 (2014) 15–20. doi:10.1515/itc-2015-0007.
- [6] O. Etzioni, The World-Wide Web, *Commun. ACM* 39.11 (1996) 65–68. doi:10.1145/240455.240473.
- [7] The University of Sheffield, GATE General Architecture for Text Engineering. URL: <https://gate.ac.uk/>.
- [8] T. R. Gruber, A translation approach to portable ontology specifications, *Knowl. Acquis.* 5.2 (1993) 199–220. doi:10.1006/knac.1993.1008.
- [9] M. A. Hearst, Untangling text data mining, in: the 37th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, 1999. doi:10.3115/1034678.1034679.
- [10] INFLIBNET Centre, GujCat Online Union Catalogue of Gujarat Colleges. URL: <https://gujcat.inflibnet.ac.in/>.
- [11] Y. Kodratoff, About Knowledge Discovery in Texts: A Definition and an Example, *Comput. Sci., Philos.* (2000). URL: <https://api.semanticscholar.org/CorpusID:1510629>.
- [12] R. Kosala, H. Blockeel, Web mining research, *ACM SIGKDD Explor. Newsl.* 2.1 (2000) 1–15. doi:10.1145/360402.360406.
- [13] A. Maedche, S. Staab, Ontology learning for the Semantic Web, *IEEE Intell. Syst.* 16.2 (2001) 72–79. doi:10.1109/5254.920602.
- [14] M. A. Musen, The protégé project, *AI Matters* 1.4 (2015) 4–12. doi:10.1145/2757001.2757003.
- [15] S. Sarawagi, Information Extraction, *Found. Trends in Databases* 1.3 (2007) 261–377. doi:10.1561/1900000003.
- [16] R. Shah, S. Jain, Ontology-based Information Extraction: An Overview and a Study of different Approaches, *Int. J. Comput. Appl.* 87.4 (2014) 6–8. doi:10.5120/15194-3574.
- [17] ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD, M. Vanni, A. Neiderer, General Architecture for Text Engineering (GATE) Developer for Entity Extraction: Overview for SYNCOIN, ARL-TR-7000, 2017. URL: <https://apps.dtic.mil/sti/citations/ADA607573>.
- [18] D. C. Wimalasuriya, D. Dou, Components for information extraction, in: the 19th ACM international conference, ACM Press, New York, New York, USA, 2010. doi:10.1145/1871437.1871444.
- [19] Amara, F. Z., Djezzar, M., Hemam, M., Tiwari, S., Hafidi, M. M. (2023). Unlocking the Power of Semantic Interoperability in Industry 4.0: A Comprehensive Overview. In F. Ortiz-Rodriguez, B. Villazón-Terrazas, S. Tiwari, & C. Bobed (Eds.), *Knowledge Graphs and Semantic*

Web. Springer Nature Switzerland. 14382 (2023) 82–96. [https://doi.org/10.1007/978-3-031-47745-4\\_7](https://doi.org/10.1007/978-3-031-47745-4_7)

[20] Jiomekong, A. Tiwari, S. An approach based on Open Research Knowledge Graph for Knowledge Acquisition from scientific papers. (2023) <https://doi.org/10.48550/ARXIV.2308.12981>

[21] Khorashadizadeh H. Mihindukulasooriya N. Tiwari S. Groppe J. Groppe, S. Exploring In-Context Learning Capabilities of Foundation Models for Generating Knowledge Graphs from Text (2023). <https://doi.org/10.48550/arXiv.2305.08804>