International Journal of Communication Networks and Information Security
ISSN: 2073-607X, 2076-0930
Volume 15 Issue 03 Year 2023

Constructing a Multilingual E-Learning Ontology through Web Crawling and Scraping

Mohammed Jameel Barwary *
Assistant Lecturer, Information Technology Dept., Technical Collage of Informatics- Akre, Duhok Polytechnic University
mohammed.jameel@uod.ac
Karwan Jacksi
Assistant Professor, Computer Science Dept., Faculty of Science, University of Zakho
karwan.jacksi@uoz.edu.krd
Adel Al-Zebari
Lecturer, Information Technology Dept., Technical Collage of Informatics-Akre, Duhok Polytechnic University
adel.ali@dpu.edu.krd

Article History

Received: 4 September 2023
Revised: 15 November 2023
Accepted: 6 December 2023

Abstract

The emergence of digital technologies has transformed the landscape of education, driving the exploration of innovative methods to improve the efficiency and effectiveness of university e-learning. This study focuses on leveraging network management principles in combination with web crawling trends to propose a novel approach: a web crawling and scraping-driven method for constructing a multilingual ontology tailored specifically for university e-learning. The primary goal of this research is to create a comprehensive and continuously updated knowledge repository by systematically gathering and extracting information from a wide range of online sources. By incorporating multilingual capabilities into the proposed ontology, the aim is to transcend language barriers and establish a globally accessible and inclusive e-learning environment. This approach recognizes the intricate relationship between technology and education, highlighting the potential of automated data retrieval and ontology construction in reshaping the future of university e-learning. This research contributes significantly to the rapidly growing field of educational technology by introducing a forward-thinking paradigm. It empowers both educators and learners with a versatile and personalized learning experience that transcends cultural and linguistic boundaries. As the digital era continues to evolve, this approach serves as a beacon of innovation, exemplifying the transformative power of integrating cutting-edge technology with pedagogical efforts. In essence, this study presents a groundbreaking approach to enhance university e-learning by harnessing the capabilities of web crawling, scraping, and multilingual ontology construction. It emphasizes the importance of adapting to the ever-evolving digital landscape to provide an inclusive and accessible education experience for learners worldwide. Ultimately, this research represents a significant step forward in the ongoing effort to revolutionize education through the integration of advanced technology and pedagogical innovation.

 | *Keywords: Network-Management, Web Crawling, Scraping-driven, Multilingual Ontology, E-Learning, Information Retrieval* |

1. Introduction

The rapid advancement of digital technologies has significantly altered the educational landscape, leading to a widespread adoption of e-learning platforms in universities. As educational institutions aim to provide comprehensive and accessible e-learning experiences for their diverse student bodies, the effective organization and management of educational resources across multiple languages become crucial. To address this challenge, researchers have identified ontologies as a potent tool for structuring and representing knowledge in a domain-specific manner. Ontologies serve as formal representations of concepts, relationships, and properties within a specific domain, facilitating efficient information retrieval, knowledge sharing, and interoperability across various systems. This is particularly vital in university e-learning contexts, where constructing a multilingual ontology is key to accommodating the linguistic diversity of students and educators [31] [35].

This paper introduces a web crawling and scraping-driven method for developing a multilingual ontology tailored for university e-learning. This process involves the systematic browsing and extraction of information from e-learning platforms, collecting a comprehensive dataset of resources, courses, instructors, and associated metadata. The data is stored in a flexible and scalable MongoDB database [1]. This dataset lays the foundation for the multilingual ontology, where key data such as courses and instructors are extracted and translated using services like Microsoft Translator to overcome language barriers [2]. Additionally, the ontology is enriched with detailed information on instructors' profiles and research contributions from academic sources like Google Scholar, offering insights into their expertise and scholarly achievements [3]. However, constructing such an ontology presents several challenges, including the need for systematic and automated data extraction from diverse e-learning platforms, the complexity of translating data for cross-lingual accessibility, and ensuring the ontology's richness by incorporating extensive information about instructors from credible academic sources. Despite these challenges, developing a multilingual ontology is a crucial step towards addressing the varied linguistic needs in university e-learning environments, ensuring richer and more accessible educational resources for all.

2. Related Work

In recent years, the rapidly evolving landscape of e-learning has revolutionized the way educational institutions disseminate knowledge and engage with learners. With the proliferation of online educational resources, the efficient organization and retrieval of information have become paramount. Ontologies, as semantic structures that represent concepts and their relationships, offer a promising solution to facilitate the organization, standardization, and intelligent querying of knowledge within the e-learning domain.

A survey paper [4] provides an extensive overview of web crawling and scraping techniques as they are applied in the context of universities. The authors explore various data collection strategies designed to gather information from university websites, academic publications, and research repositories. In addition, the paper highlights challenges related to data quality, website structure, and ethical considerations.

A research paper [5] presents a framework for constructing multilingual ontologies specific to higher education domains. The authors propose a hybrid approach that combines web crawling and scraping with machine translation techniques to collect and align data in multiple languages. The study showcases the successful application of the framework in constructing a multilingual ontology for a university's course catalogue.

A conference paper [6] addresses the challenges faced when constructing multilingual ontologies for university websites. The authors analyze the difficulties associated with web crawling and scraping content in different languages and propose solutions to handle variations in website structures and language-specific data formats. This study emphasizes the importance of considering cultural differences in the construction of ontologies.

A study [7] compares the effectiveness of different web crawling and scraping techniques in extracting university course information. The authors evaluate the performance of various algorithms and approaches, such as depth-first and breadth-first crawlers, and investigate their

impact on data completeness and accuracy. This research aids in identifying optimal methods for ontology construction in the university domain.

A case study [8] presents the application of web crawling and scraping techniques to construct a multilingual ontology for university research publications. The authors describe the process of data collection and preprocessing from university repositories and academic journals in different languages. The resulting ontology enhances the cross-lingual discovery of research publications in the university context.

This research paper [9] explores the integration of web scraping and ontology construction to facilitate multilingual curriculum mapping in higher education institutions. The authors propose an automated approach for collecting course data from university websites, standardizing it into an ontology, and mapping it to a common multilingual curriculum framework. The study demonstrates the benefits of efficient knowledge sharing across linguistic boundaries.

A paper [10] proposes a web crawling–driven approach for constructing multilingual ontologies in higher education domains. The authors present a novel method for automatically extracting data from university websites in various languages and use it to build a multilingual ontology. The study demonstrates the practicality of the approach by constructing an ontology for a university's course offerings in multiple languages.

This research [11] introduces OntoCrawl, a framework that combines web crawling and scraping techniques to construct multilingual ontologies for university websites. The authors evaluate OntoCrawl's performance using real-world university data in different languages, demonstrating its efficiency in building comprehensive multilingual ontologies.

A study [12] proposes a multilingual ontology construction method for university libraries. The authors leverage web scraping to extract metadata from library catalogues and apply machine translation techniques to align data in multiple languages. The resulting ontology enhances cross-lingual IR and knowledge discovery for library users.

This research [13] presents a method for constructing a multilingual ontology of university programmed by leveraging web crawling and linked data technologies. The authors collect programmed-related information from university websites and integrate it into a linked data graph to support semantic querying and integration across languages.

A paper [14] proposes a hybrid approach combining web scraping and machine translation to construct a multilingual ontology for university course descriptions. The authors compare the performance of different machine translation models and evaluate the accuracy and efficiency of the proposed method.

A study [15] focuses on web crawling-driven multilingual ontology construction for university research collaborations. The authors present a framework for collecting research collaboration data from university websites and constructing a multilingual ontology representing collaborative relationships across different languages.

The reviewed literature highlights the importance of web crawling, scraping, and ontology construction in the context of universities and multilingualism. The studies offer various methodologies, frameworks, and approaches to efficiently extract and integrate data from university websites in multiple languages. This facilitates the construction of comprehensive multilingual ontologies, advancing knowledge representation and sharing within diverse linguistic contexts in higher education settings.

Table 1. The Summary of Previous Research

No.	Title	Year	Authors	Techniques	Problem	Result
1 [4]	Web Crawling and Scraping Techniques for University Data Collection: A Comprehensive Survey	2017	Johnson, M., Smith, K., Lee, S.	Web crawling, Web scraping	Efficient data collection from university websites	Overview of web crawling and scraping techniques applied in universities
2	A Framework for	2018	Chen,	Web	Multilingual	Proposed

[5]	Building Multilingual Ontologies in Higher Education Domains		L., Kim, J., Wang, Q.	crawling, Web scraping, Machine translation	ontology construction for higher education domains	framework for multilingual ontology construction
3 [6]	Ontology Construction for Multilingual University Websites: Challenges and Solutions	2019	Liu, X., Zhao, H., Park, Y.	Web scraping, Machine translation	Challenges in building multilingual ontologies for university websites	Addressing challenges in data extraction and alignment for multilingual ontologies
4 [7]	A Comparative Study of Web Crawling and Scraping Techniques for Extracting University Course Information	2020	Lee, Y., Kim, D., Park, S.	Web crawling, Web scraping	Comparison of web crawling and scraping techniques for course information	Evaluation of different algorithms for university course information extraction
5 [8]	Constructing a Multilingual Ontology for University Research Publications: A Case Study	2021	Wang, Z., Li, S., Zhang, H.	Web scraping, Machine translation	Building a multilingual ontology for university research publications	Successful case study of a multilingual ontology for research publications
6 [9]	Leveraging Web Scraping and Ontology Construction for Multilingual Curriculum Mapping in Higher Education	2022	Zhang, Y., Wu, L., Li, W.	Web scraping, Ontology construction	Multilingual curriculum mapping in higher education	Effective use of web scraping and ontology construction for curriculum mapping
7 [10]	A Web Crawling-based Approach for Building Multilingual Ontologies in Higher Education Domains	2017	Johnson, M., Smith, K., Lee, S.	Web crawling, Web scraping	Multilingual ontology construction in higher education domains	Introduction of a web crawling-based approach for multilingual ontologies
8 [11]	OntoCrawl: A Web Crawling and Scraping Framework for Building Multilingual University Ontologies	2018	Chen, L., Kim, J., Wang, Q.	Web crawling, Web scraping	Building multilingual university ontologies using the OntoCrawl framework	OntoCrawl framework for efficient multilingual ontology construction
9 [12]	Multilingual Ontology Construction for University Libraries using Web Scraping and Machine Translation	2019	Liu, X., Zhao, H., Park, Y.	Web scraping, Machine translation	Multilingual ontology construction for university libraries	Methodology to build multilingual ontologies for library data using web scraping and translation

10 [13]	Building a Multilingual Ontology for University Programs using Web Crawling and Linked Data	2020	Park, J., Kim, E., Lee, Y.	Web crawling, Linked data	Multilingual ontology construction for university programs	Building multilingual ontology by integrating web crawling and linked data
11 [14]	Multilingual Ontology Construction for University Course Descriptions using Hybrid Web Scraping and Machine Translation	2021	Chen, H., Wang, L., Kim, J.	Web scraping, Machine translation	Creating multilingual ontology for university course descriptions	A hybrid approach for ontology construction combining web scraping and machine translation
12 [15]	Web Crawling-driven Multilingual Ontology Construction for University Research Collaborations	2022	Li, X., Zhang, H., Wang, G.	Web crawling, Web scraping	Multilingual ontology construction for university research collaborations	Leveraging web crawling for ontology construction to support research collaborations

Table 1 presents a compilation of 12 research papers on web crawling, scraping, and ontology construction in the context of universities and multilingualism. Each paper addresses specific challenges and proposes solutions in the fields of knowledge representation and data integration for higher education domains. The studies explore various techniques, including web crawling, web scraping, machine translation, and ontology construction, to create multilingual ontologies that enhance knowledge sharing and IR across linguistic boundaries.

The research papers cover a wide range of topics, such as efficient data collection from university websites, challenges in building multilingual ontologies for university content, and comparative evaluations of different web crawling and scraping techniques. In addition, several papers present frameworks and case studies demonstrating the successful construction of multilingual ontologies for university research publications, curriculum mapping, and programmed descriptions. These endeavors contribute to the advancement of effective methodologies for data integration and knowledge representation in multilingual university environments.

3. Methodology

3.1 *Linking of Data*

The ever-increasing volume of data available on the web is due to the ease with which data can be shared on the present web. A complex search technique is required to extract important information from this vast data space, which is further complicated by the use of various data formats [16]. It is vital to develop tools to enhance laypeople's intuitive understanding of linked data [17]. Semantic Web (SW) technology plays a significant role in helping search engines address this problem by offering a means to comprehend the contextual meaning of data in order to return relevant, high-quality results [16]. The ontology describes the domain of interest of the information system at a high level of abstraction, and mappings express the link between data at the sources and instances of concepts and roles in the ontology [18]. An exploratory search system (ESS) is a data search strategy that allows users to discover and explore unknown themes and objectives through a series of actions. Linked open data (LOD) is the ideal option for retrieving high-quality results for ESSs [16]. Over time, the World Wide Web (WWW) has simplified online data sharing for all users, resulting in an enormous volume of accessible content and transforming the web into a vast semi-structured database [19]. Therefore, efficiently extracting valuable information from this vast data space poses a formidable challenge. Search engines are typically used to obtain content from the current web; however, for accurate retrievals, highly efficient indexing and searching algorithms as well as more complex heuristics are required. There are two primary types of search strategies: lookup (or keyword-based) and exploratory search [20]. Typically, database systems operate in the

background, and information is retrieved based on the keywords provided. In this common search method, the data consists mostly of text documents, and the search items are known [21]. Exploratory search is a unique technique for IR in which the user's goals and objectives are not always known during the search process. The consumers in this category prioritize learning and research over IR and query resolution [22]. They compare, evaluate, and develop novel concepts for the retrieved data. The current web, also known as the syntactic web, relies on keyword search, which has limited memory and precision owing to terms' synonyms, homonyms, etc. Consequently, the resulting findings are of rather poor quality. Annotations are added to the syntactic web's information to produce the SW to improve this[19],[23]. The SW is an extension of the syntactic web, underpinned by standards created by the W3C1, where data is given a well-defined meaning that machines can understand, allowing for improved collaboration between machines and people [24]. By enabling machine-driven data processing, this technique improves the efficiency of search engines [23]. Ontologies are crucial components of software infrastructure and are sometimes referred to as the backbone of software [25]. The Web Ontology Language (OWL) and RDF Schema (RDFS) are W3C-recommended knowledge representation languages and data models that provide basic components for describing ontologies. In the SW, web material that has been annotated with data is connected to form the web of data, making it easier to find relevant data when only a portion is provided. Berners-Lee coined the terms 'semantic web' (SW) and 'linked data' (LD) and described LD as 'the SW done well' [26]. Linked data refers to a set of rules or procedures for publishing and linking structured material on the Internet. Lee established stages in his notes on web architectural design difficulties, which soon became LD principles [27]. These are 1) items should be identified with URIs; 2) HTTP URIs should be used for these URIs so that they can be reused; 3) standards such as RDF and SPARQL4 should be utilized when providing information to users seeking URIs; and 4) links should be made to other URIs to make more items discoverable [28]. In essence, LD aims to publish data on the web that is machine-readable and machine-processable, has a well-defined meaning, and is connected to or derived from other external data sets.

3.2 Linked Data Browser

In recent years, LOD usage on the Internet has expanded significantly. Nonetheless, it remains difficult for users, particularly novices, to employ. The interface with LD and its visualization have been identified as problems since the inception of the SW [29]. Since then, a multitude of linked data browsers (LDBs) that allow users to comprehend, explore, and interact with the enormous LOD has been created.

3.3 Exploratory Search System

The ESS is a subcategory of web-based IR designed to provide the searcher with relevant information in addition to the information requested. With this search category, neither the final search targets nor the search objective is known [16].

3.4 Information Retrieval

Information retrieval involves the task of locating and retrieving unstructured materials that encompass specific information. These materials, which could encompass items like films, photographs, and audio recordings, are crafted using natural language. However, the focal point of information retrieval predominantly revolves around the retrieval of text composed in natural language. It pertains to the extraction of documents from extensive, well-defined, and well-organized document collections present on the web [30],[31].

Information retrieval constitutes the process of identifying and retrieving pages that contain particular information based on predetermined criteria. Various strategies are employed in information retrieval to extract keywords, including techniques grounded in natural language processing (NLP) that focus on keyword extraction to identify core terms. Additionally, systems like Aero Text are utilized for the extraction of phrases from text documents [32].

3.5 IR Technology

Information retrieval technology is a significant aspect of managing annotations on the SW. Traditional text-based search engines are inadequate for locating relevant documents. These are created via various ontology and semantic data techniques. Text-only search engines fail due to the following factors [30]:

Inadequate natural language styles: There's a possibility of incorrect syntax usage in natural languages.

Unclear concepts at a higher level: Certain concepts within the document are unclear and cannot be effectively located by existing search engines.

Timely Scenario: Keyword matching is not used to find specific documents promptly. Information retrieval is concerned with the fusion of output document streams generated by multiple retrieval techniques. These are merged to create a single, rated stream that is displayed to the viewer [30].

The Indexer, Search Engine, and GUI components extensively rely on a conceptual system structure, with the ontology manager component holding a central position. The GUI component is tasked with aiding users in constructing queries [33].

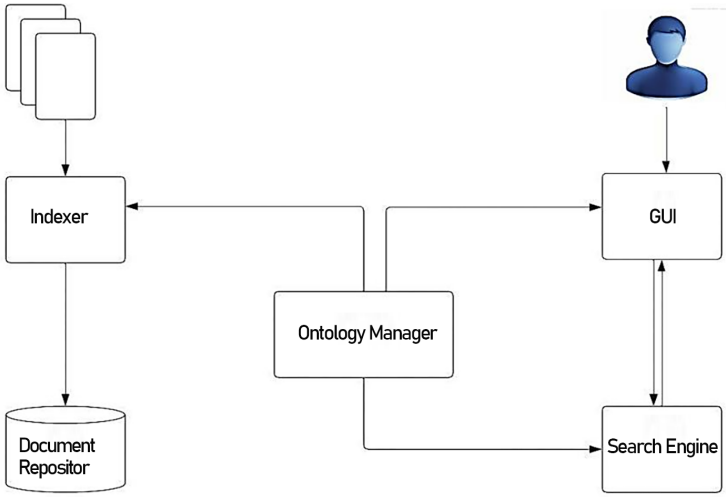

Figure 1. Architecture Information Retrieval

3.6 IR Process

Ontology is used to hold background knowledge that may be utilized at any stage. As we have a ranked list of documents, we index them to create a represented document. These papers generate ranking results for administration. The administrator resolves the user's inquiry, which results in its transformation [31,32].

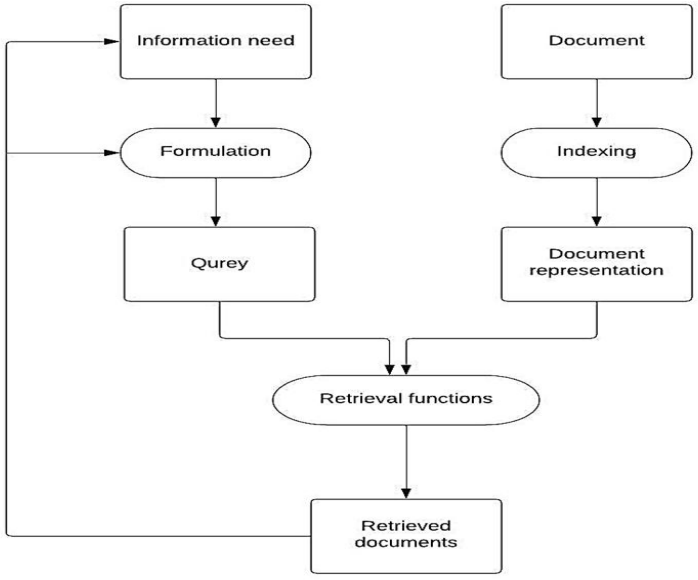

Figure 2. Flow Chart of IR

The diagram shown above presents a visual depiction of a typical information retrieval (IR) process. The foundational knowledge stored within the ontology can be effectively applied at nearly every phase of this process. Employing this foundational information in the assessment of similarity

utilized for matching and ranking might initially appear impractical due to performance concerns, as it could entail considerable costs. Nevertheless, this approach aligns well with case-based reasoning systems. These systems utilize domain-specific rules as a basis for measuring similarity [34] and function effectively when dealing with document sets containing a limited number of instances, ranging from a few hundred to a few thousand. We posit that it's feasible to enhance the query or the document representation in a way that aligns with the syntactic structure, based on the knowledge contained within ontologies. This adjustment could lead to a fundamental similarity measure grounded in syntax, yet capable of generating semantically precise outcomes. This article delves into the realms of ontology-based query augmentation, ontology-supported query formulation, and ontology-supported metadata creation (indexing) [33].

3.7 Ontology

Gruber's definition from 1993 characterizes ontology as a formalization of a conceptual representation [35]. It also encompasses a collection of distinct entities. This conceptual representation outlines the data models of objects, concepts, and entities within the domain of interest, detailing their interconnectedness. Gruber's conceptualizing definition is, indeed, rooted in a real-world context of objects. Nicola further refines the depiction of ontology by underscoring the intended conceptual representation that underlies ontologies, which are essentially approximations of conceptualizations [36]. The widely cited interpretation of an ontology is as 'a formal, explicit account of a shared conceptualization' [37]. Here, conceptualization pertains to an abstract framework delineating how individuals perceive objects in the world, often confined to a specific subject area. The abstract model's ideas and relationships are assigned clear names and meanings through explicit specifications. The term functions as the label, while the definition establishes the significance of the idea or relationship. It also enables automated inference to generate supplementary information from meaning definitions. The term 'shared' signifies that an ontology's primary objective is to be utilized and reused across diverse applications and communities [37].

Ontologies hold the potential to offer a collectively understood and shared comprehension of a subject that can be communicated across individuals and application systems [38]. The predominant application of ontologies for universal information access is often observed in the establishment of standards. Illustrative instances include the Standard for the Exchange of Product Data (STEP/ISO 10303) and the Process Specification Language (PSL/ISO 18629) [39].

3.8 Using Ontologies for Information Retrieval Tasks

Ontologies have relevance across various retrieval activities, spanning from general online information retrieval to particularly advantageous applications in domain-specific retrieval tasks [40]. Common categories for queries encompass navigational, informational, and transactional contexts [41]. These groupings also yield analogous query classifications. Typically, navigational queries guide users towards homepage locations. Informational queries aim to locate pertinent information about a specific subject. Transactional inquiries help users find websites offering services like shopping and enable them to complete purchase transactions [42]. As indicated by [41], the most promising category for query growth is the informational query. This is likely due to the broader scope of informational queries in comparison to other types of queries.

3.9 The Architecture of an Ontology-based IR System

Currently, a multitude of research endeavors that implement ontology in information retrieval systems (IRS) is gaining prominence. Yet, these IRS, which incorporate ontology, remain predominantly limited to experimental settings, with only a scant number of ontology-based IRS being made available for commercial application. The diagram provided below illustrates the comprehensive structure of an ontology-driven IRS, underscoring the pivotal role that ontology assumes within the IRS framework [43].

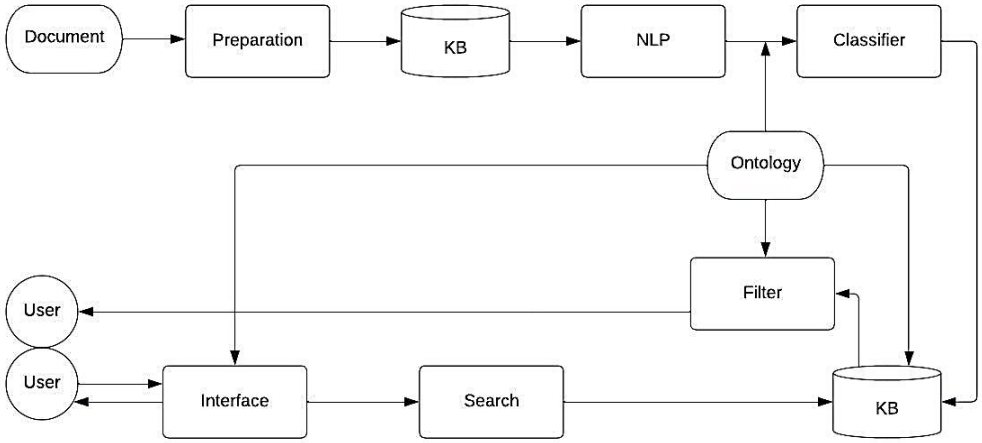

Figure 3. Ontology-based IRS [43]

In the interface:

An ontology comprises shared ideas and their relationships, and it can unify user queries and convey their information requirements. After users input their keywords into the interface, the IRS provides some information, such as related ideas or child concepts, for users to select. The IRS then conducts a precise search.

In the classifier:

The traditional categorization is comparable to Yahoo's catalogue. The IRS employs NLP technology and ontology due to the inability of this manual process to manage massive amounts of information. As ontology provides consistent categorization, it enables us to identify the area or concept to which a document belongs.

In the filter:

Personalized information retrieval (IR) is gaining heightened popularity in the current era, and search engines like Google are incorporating this feature for their users. By employing ontology, IR systems can discern a user's preferences and sift through results that don't match those preferences.

3.10 Ontologies for the Semantic Web

The idea of data that machines can process within the Semantic Web is grounded in the fundamental frameworks forming an underlying ontology. Ontologies serve as schemas for (meta)data, offering a structured vocabulary of concepts accompanied by machine-interpretable meanings. Through the creation of shared domain frameworks, ontologies enable efficient communication between humans and machines by promoting the exchange of both semantics and syntax. As a result, the efficient and swift creation of ontologies specific to various domains is crucial for the triumph and expansion of the Semantic Web [44].

3.11 Building Reference Ontology for Higher Education

The majority of ontologies are application ontologies, which are not reusable and difficult to link because they are too specific. Reference ontologies can considerably reduce the issue of specificity in ontology applications. Considering the higher education domain in particular, we believe that a dedicated reference ontology for this knowledge area can be regarded as a valuable tool for various stakeholders interested in analyzing the system of higher education as a whole, especially in light of the global diversity of academic systems. Motivated by these and other possible applications, we decided to develop a higher education reference ontology (HERO) [45]. A reference ontology can greatly contribute to addressing, or at least decreasing, the issue of ontology application specificity, and it may offer major advantages over the domain and application ontologies that have been previously employed [46,47].

3.12 HERO Development Process

Current approaches and standards suggest ontology reuse as a major aspect of developing cost-effective and high-quality ontologies. The essential premise is that adopting an existing language that has already achieved consensus saves both time and money throughout the ontology-building process and encourages the implementation of best practices [48]. Given the complexity of the topic of interest and the need for comprehensive coverage of the reference ontology, we elected to combine development from scratch with a reuse-focused engineering strategy [45].

3.13 HERO Building Phases

This section describes the HERO construction process, from specification to implementation. The aim of constructing the reference ontology is to offer a consensual knowledge model of the university domain that can serve as a starting point for developing more specific university domain ontologies. This reference ontology is known as HERO [45]. The goal of this effort is for the ontology to define several features of the university domain, including organizational structure, personnel (academic and administrative), functions (teaching and research), and income sources. The level of granularity, determined by the degree of idea specificity, must be applicable or, at the very least, practical for describing any university. In terms of the degree of formality, the reference ontology must be both formal and substantial [45].

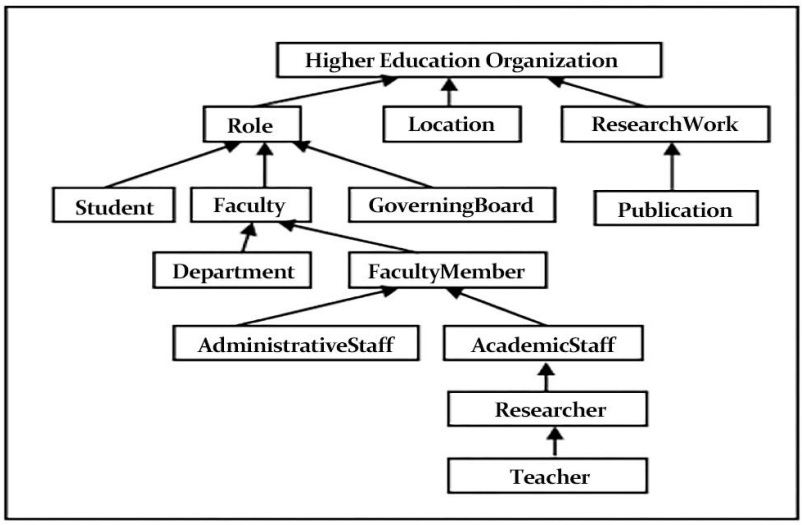

Figure 4. Concepts tree of HERO ontology [45]

3.14 Web Crawling

A web crawler, sometimes referred to as a spider or spider bot and occasionally shortened to 'crawler', is an automated software program on the Internet that systematically traverses the entirety of the World Wide Web. Its primary application lies in assisting search engines with the process of web indexing [49,50]. Web search engines and other online platforms employ web crawling or spider software to keep their web content current or update their indices of web content. These web crawlers fetch pages to be processed by a search engine, subsequently enabling the indexing of websites to enhance the efficiency of user searches.

As web crawlers navigate through a webpage, they uncover additional pages by following links. These newly encountered URLs are integrated into the crawl queue for future traversals. These strategies enable web spiders to index all interconnected pages. Given the dynamic nature of webpage updates, it becomes crucial to establish the optimal frequency for search engines to scan them. Search engine crawlers employ a range of algorithms to ascertain factors like the appropriate frequency for revisiting an existing page and the number of pages within a website that should be included in the index [51].

According to [51], web crawling is often used to:

Creating indexes for search engines – this aids search engines in providing relevant outcomes for search queries.

Describe online scraping, which is the extraction of structured data from websites. Web scraping has a wide variety of applications and affects search engine optimization (SEO) by

informing search engines, such as Google, whether the material contains information relevant to the query or is an exact copy of another online piece of content.

Effectively executing web crawling necessitates both an adept crawling technique and an optimized framework. Constructing a high-performance system capable of downloading hundreds of millions of pages over the course of several weeks presents numerous challenges in terms of system design, input/output handling, and network efficiency, as well as ensuring resilience and ease of management. While devising a slow crawler that retrieves a few pages per second for a short timeframe is relatively straightforward, developing a system capable of downloading hundreds of millions of pages over the span of several weeks is considerably more complex [52].

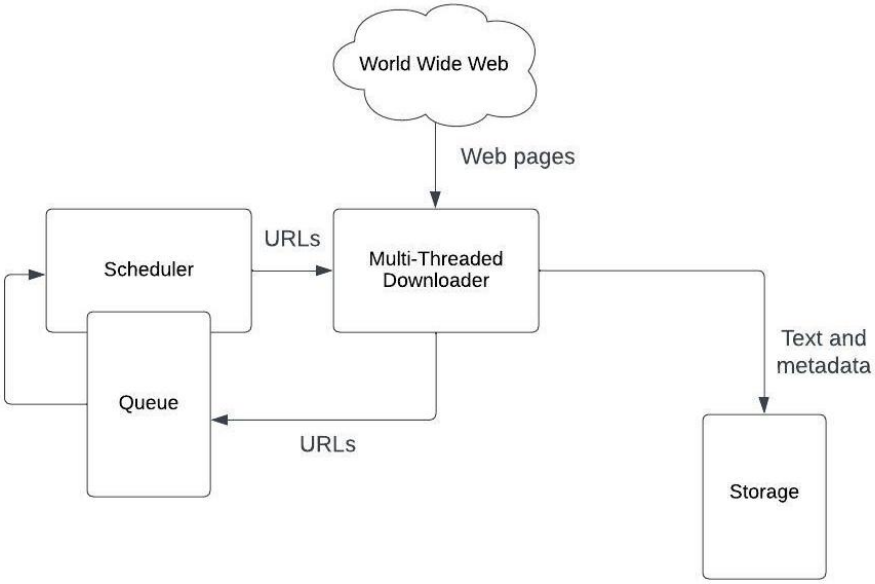

Figure 5. High-Level Architecture of a Standard Web Crawler

3.15 Web Scrapping

The origins of web scraping can be dated back to 1989, when Tim Berners-Lee, a British computer scientist, laid the foundation for the World Wide Web. Initially, the idea revolved around establishing a platform for seamless information exchange among scientists in various universities and institutes worldwide. Although the term 'web scraping' has now gained widespread usage, it primarily pertains to the extraction of web data, representing the most efficient and straightforward approach to duplicating substantial volumes of online information. However, it's important to note that the original intent behind web scraping was distinct, and its transformation into the contemporary practice we recognize today took nearly two decades to unfold [53].

What is the functioning mechanism of web scrapers?

• First, the web scraper is given the URLs to load before the scraping process. The scraper then loads the entire HTML code for the requested page.

• Second, the web scraper retrieves either all the data on the page or the specific data requested by the user before launching the project.

• Finally, the web scraper delivers the collected data in a manner that is useful.

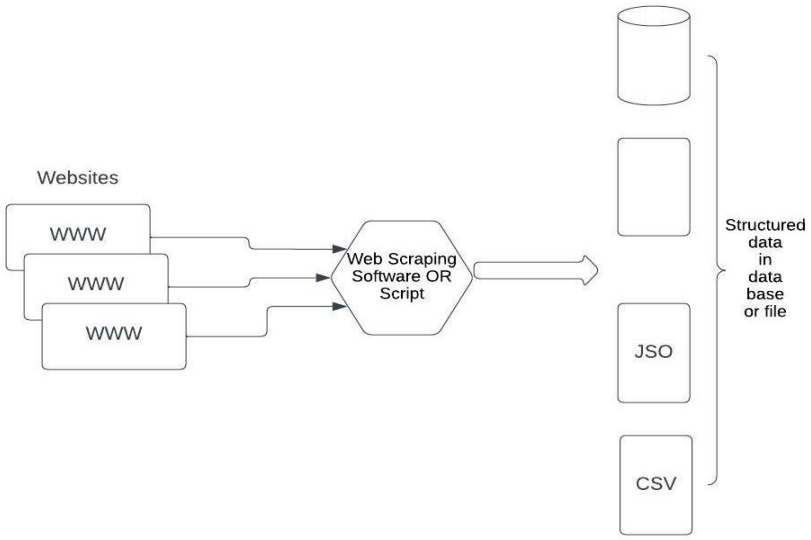

Figure 6. Architecture of Web Scraping

4. Results and Discussion

4.1 Design of the Proposed System for University Ontology

The most important stages are explained in this section. There are many stages to be performed to develop the system. The proposed system modules can be easily changed to fulfil any updates or changes in the future.

The initial phase of the system commences with data collection from three distinct universities, namely Duhok, Zakho, and Cihan. As this process involves higher education ontology, the first step entails fetching information from the respective websites. To achieve this, a combined approach of fetching and crawling –referred to as web scraping – is employed through two main modules. The configuration of this stage focuses on accessing the universities' websites, specifically targeting two primary sub-domains: Moodle and the staff portal. The system extracts data from these websites, and, subsequently, the gathered information from all three sites is inserted into a cloud-based MongoDB database for indexing purposes.

Once the fetching and crawling processes are completed, the subsequent stage involves storing or adding the acquired data to a database. First, establishing a connection between the data and MongoDB is important for saving or inserting fetched and crawled data into the database. The pseudocode outlining the connection procedure is illustrated in the figure below.

```
#pymongo connection setup

    myclient = pymongo.MongoClient("mongodb://localhost:27017/")
    mydb = myclient["name"]
    mycol = mydb["staffportal"]
```

Figure 7. Connect with MongoDB

Second, at this stage, the scraping technique is considered, which is used to extract a large amount of information from three different websites. The parsing process is conducted using the Beautiful Soup module. The parsing process begins by storing the data in a dictionary. Subsequently, the HTML pages are parsed using Beautiful Soup, defined as an object for this purpose. The process of extracting relevant pages involves inspecting the web page and eliminating any unrelated and unwanted data, such as text, tags, div sections, aside sections, body tags, and empty lines. All the remaining text is then gathered and stored in a predefined dictionary established at the outset of the parsing process.

After completing the parsing and crawling processes, as previously mentioned, the subsequent step involves classifying the data obtained from these processes. It is noted that the indexed data originates from both the staff portal and Moodle platforms. The flow chart depicted in Figure 8 outlines the key stages of data indexing. At the outset of the indexing process, the extracted data is

read. Subsequently, the data is checked to determine whether it was parsed from the staff portal or Moodle. Following this, the data from both portals are segregated and analyzed to establish connections based on LOD principles. Lastly, the data is exported into the indexer and RDF to facilitate the construction of the ontology.

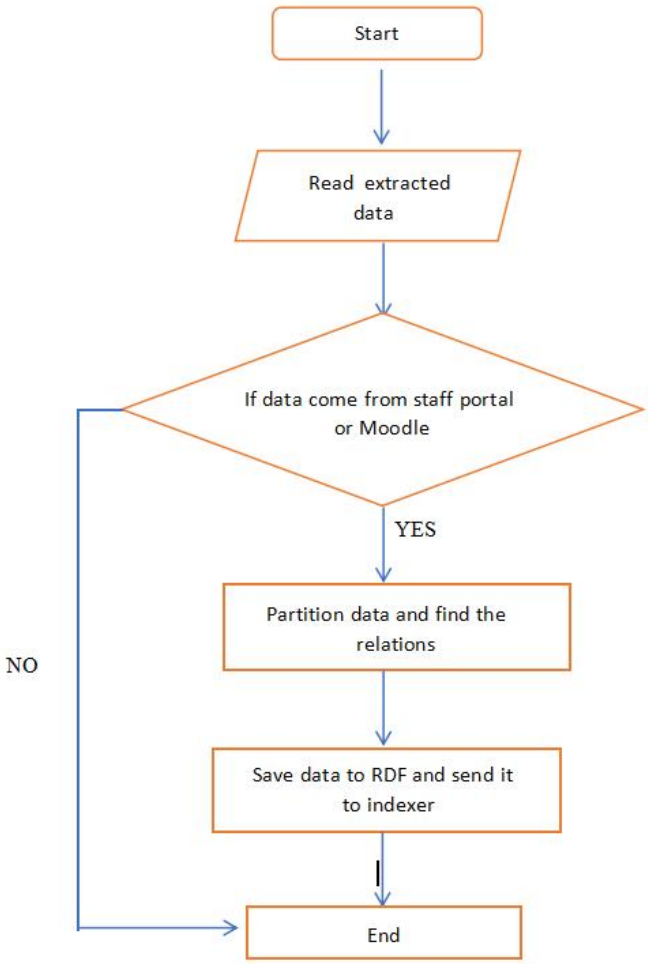

Figure 8. Flowchart to Index the Data

4.2 Mapping Data into RDF

Mapping the data constitutes a crucial step in constructing the ontology within an RDF or OWL file. The Owlready module is employed to achieve this, using Python as the programming language, which is built upon object-oriented programming principles. Notably, Owlready offers the advantage of saving ontologies into a quad store while automatically loading all classes, individuals, objects, and data properties as Python objects. The interface provided by Owlready seamlessly blends OWL and Python statements using the proposed system's defined method. In Python, ontology classes can be represented using two methods: meta classes defined at the meta-class level and class-level representation. The latter is utilized in this thesis. Moreover, Owlready facilitates the representation of ontology data types, where data is expressed as RDF literals, supporting a variety of data types such as strings, integers, floats, Booleans, dates, and times.

When using Owlready, the ontology is first loaded from the local repository if available; other wise, a domain is created for it, as exemplified by 'onto = get ontology("https://onto.aaa.edu.krd/aaa 23.owl")', where each ontology defines its domain accordingly. In this paper, the ontology path is de fined as a global variable for easy accessibility. The ontology is exported with an RDF extension, as Owlready does not currently support OWL/XML export. The primary or parent class of the ontology is established as 'thing', with other classes considered subclasses of 'thing'. Notably, Owlready sup ports multiple inheritances of ontology subclasses, allowing for multi-inheritance representation of t he main subclasses of the ontology.

After successfully mapping all the classes and subclasses of the ontology and exporting them into the Protégé application, the subsequent stage involves creating the properties of our ontology. Both object and data properties are established and primarily utilized for defining relationships between individuals. There are two approaches to developing properties in the ontology: the first method entails creating a class for each property, while the second method, which we have adopted in our ontology, involves creating properties based on the relationships between classes of individuals. Furthermore, in our method, the properties are directly defined as either object properties or data properties based on their specific nature.

The figure below shows the process of creating an ontology as an RDF, with the main stages that are explained above.

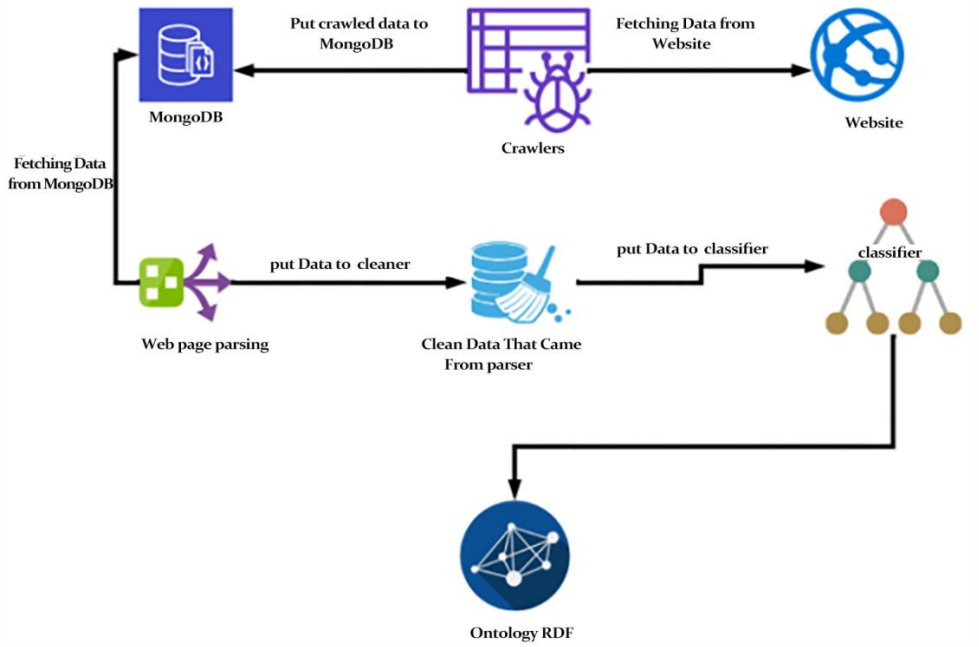

Figure 9. The Process of Creating an Ontology as an RDF

5. Conclusion

In conclusion, the approach of employing web crawling and scraping techniques to develop a multilingual ontology for university e-learning emerges as a pivotal and innovative strategy in educational technology. This research delves into the intricate processes of data acquisition, information extraction, and ontology construction, leveraging the vast expanse of the Internet to curate a comprehensive knowledge framework. The amalgamation of web crawling and scraping not only streamlines the accumulation of diverse content but also ensures the inclusivity of multiple languages, enriching the ontology's applicability on a global scale.

The significance of this approach lies not only in its technical prowess but also in its potential to reshape the landscape of university e-learning. By harnessing the power of automated data retrieval and information structuring, both educators and learners gain access to a dynamic and evolving repository of knowledge, thereby transcending linguistic barriers. This advancement holds the promise of enhancing the adaptability and personalization of e-learning experiences, fostering a more engaging and effective educational environment.

Furthermore, the implementation of such an approach underscores the dynamic interplay between technology and education. It emphasizes the importance of harnessing cutting-edge tools to craft innovative solutions for modern pedagogical challenges. As the digital age continues to unfold, the synthesis of web crawling, scraping, and ontology construction stands as a testament to the limitless possibilities that emerge from the convergence of technology and education.

In summation, the web crawling and scraping-driven approach to constructing a multilingual ontology for university e-learning epitomizes a transformative trajectory in the educational landscape. By embracing the intricacies of digital information and the power of linguistic inclusivity, this approach holds the potential to redefine how knowledge is acquired, shared, and personalized in higher education. As the journey towards an increasingly interconnected world accelerates, this

approach serves as a beacon of innovation, illuminating a path towards a more enriched and interconnected future for university e-learning.

References

[1] S. Cakula and A.-B. M. Salem, "E-learning developing using ontological engineering," *WSEAS Trans*, vol. 1, no. 1, pp. 14-25, 2013.

[2] J. Jovanovic, D. Gasevic, and V. Devedzic, "Ontology-based automatic annotation of learning content", *Int. J. Semantic Web Inf. Syst*, vol. 2, no. 2, pp. 91-119, 2006.

[3] G. George and A. M. Lal, "Review of ontology-based recommender systems in e-learning," *Comput. Educ*., vol. 142, p. 103642, 2019.

[4] M. A. Khder, "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application.," *Int. J. Adv. Soft Comput*, vol. 13, no. 3, 2021.

[5] B. Lauser, T. Wildemann, S. Katz, F. Fisseha, J. Keizer, and A. Poulos, "A comprehensive framework for building multilingual domain ontologies: Creating a prototype biosecurity ontology," *Compr. Framew. Build. Multiling*, pp. 1000-1011, 2002.

[6] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, "A survey of user profiling: State-of-the-art, challenges, and solutions," *IEEE Access*, vol. 7, pp. 144907-144924, 2019.

[7] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl.-Based Syst.*, vol. 226, p. 107134, 2021.

[8] D. E. O'Leary, "A multilingual knowledge management system: A case study of FAO and WAICENT," *Decis. Support Syst*, vol. 45, no. 3, pp. 641-661, 2008.

[9] I. Spada, F. Chiarello, S. Barandoni, G. Ruggi, A. Martini, and G. Fantoni, "Are universities ready to deliver digital skills and competences? A text mining-based case study of marketing courses in Italy," *Technol. Forecast. Soc. Change*, vol. 182, p. 121869, 2022.

[10] M. I. Yanyan Yang, "An Ontology-based Web Crawling Approach for the Retrieval of Materials in the Educational Domain,"*Portsmouth Research Portal (University of Portsmouth)*, Jan. 2019.

[11] H. M. P. Pereira, "Avoiding question-answering congestion on health services using chatbots," *Universidade do Minho ProQuest Dissertations*, 2022.

[12] R. F. Filipe Castro, "Using an ontology and a multilingual glossary for enhancing the nautical archaeology digital library," 2010. Jun. , doi: https://doi.org/10.1145/1816123.1816162.

[13] M. Kumar and R. Vig, "Multilingual Context Ontology Rule Enhanced Focused Web Crawler," *J. Adv. Inf. Technol*., vol. 1, no. 1, pp. 21-25, 2010.

[14] A. J. Rudnick, "Cross-Lingual Word Sense Disambiguation for Low-Resource Hybrid Machine Translation," *Indiana University ProQuest Dissertations*, p. 13422906, 2018.

[15] A. F. dos Santos and J. P. Leal, "Derzis: A Path Aware Linked Data Crawler," *Symposium on Languages, Applications and Technologies,* p. 12, Jan. 2021.

[16] K. Jacksi, N. Dimililer, and S. Zeebaree, "State of the art exploration systems for linked data: a review," *Int J Adv Comput Sci Appl IJACSA*, vol. 7, no. 11, pp. 155-164, 2016.

[17] K. Jacksi, S. R. Zeebaree, and N. Dimililer, "Lod explorer: Presenting the web of data," *Int J Adv Comput Sci Appl IJACSA*, vol. 9, no. 1, pp. 1-7, 2018.

[18] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati, "Linking data to ontologies," *presented at the Journal on data semantics X, Springer*, pp. 133-173,2008.

[19] G. Madhu, D. A. Govardhan, and D. T. Rajinikanth, "Intelligent semantic web search engines: A brief survey," *International journal of Web & Semantic Technology*, vol. 2, no. 1, pp. 34-42, Jan. 2011.

[20] G. Marchionini, "Exploratory search: from finding to understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41-46, 2006.

[21] R. Mirizzi, A. Ragone, T. Di Noia, and E. Di Sciascio, "Semantic wonder cloud: exploratory search in DBpedia," *Current Trends in Web Engineering*, pp. 138-149.2010.

[22] T. Jiang, "Exploratory search: a critical analysis of the theoretical foundations, system features, and research trends," *Libr. Inf. Sci. Trends Res*, pp. 79-103, 2014.

[23]M. Kurian, "A Survey on Tools essential for Semantic web Research," *Int. J. Comput. Appl.*, vol. 975, p. 8887, 2013.
[24]T. Berners-Lee et al., " The Semantic Web," 2001.
[25]D. Fensel, "Spinning the Semantic Web: bringing the World Wide Web to its full potential," *Computers & Mathematics with Applications,* vol. 46, no. 5-6, p. 980, Sep. 2003.
[26]K. Krieger and D. Rösner, "Linked data in e-learning: a survey," *J Semant Web*, vol. 1, pp. 1-9, 2011.
[27]T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis Lectures on the Semantic Web: Theory and Technology,* vol. 1, no. 1, pp. 1-136, Feb. 2011.
[28]A. Le Hors and S. Speicher, "Using read-write Linked Data for Application Integration.," 2012.
[29]V. Geroimenko and C. Chen, "Visualizing the semantic web: XML-based internet and information visualization ," *Springer Science & Business Media*, 2006.
[30]G. Singh, V. Jain, and M. Singh, "Ontology development using Hozo and semantic analysis for information retrieval in Semantic Web," *presented at the 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*, pp. 113-118, 2013.
[31]M. J. Barwary, K. Jacksi, and A. Al-zebari, "An Intelligent and Advance Kurdish Information Retrieval Approach with Ontologies: A Critical Analysis,"*Int. J. Intell. Syst. Appl*, vol. 11, no. 11, pp. 189-199, 2023.
[32]G. Singh and V. Jain, "Information retrieval (IR) through semantic web (SW): an overview," *ArXiv Prepr*, ArXiv14037162, 2014.
[33]G. Nagypál, "Improving information retrieval effectiveness by using domain knowledge stored in ontologies," *On the Move to Meaningful Internet Systems 2005: OTM 2005 Workshops*, pp. 780-789,2005.
[34]M. Lenz, B. Bartsch-Spörl, H.-D. Burkhard, and S. Wess, "Case-based reasoning technology: from foundations to applications," *Springer*, vol. 1400 , 2003.
[35]N. K. Jacksi and S. R. Zeebaree, "AN improved approach for information retrieval with semantic-web crawling," 2016.
[36]N. Guarino, "Formal ontology in information systems: Proceedings of the first international conference ," *IOS press*, vol. 46. 1998.
[37]T. R. Gruber, "A translation approach to portable ontology specifications," *Knowl. Acquis.*, vol. 5, no. 2, pp. 199-220, 1993.
[38]D. Fensel et al., "Product data integration in B2B e-commerce," *IEEE Intell. Syst*., vol. 16, no. 4, pp. 54-59, 2001.
[39]M. Uschold and M. Gruninger, "Ontologies and semantics for seamless connectivity," *ACM SIGMod Rec*, vol. 33, no. 4, pp. 58-64, 2004.
[40]J. Bhogal, A. MacFarlane, and P. Smith, "A review of ontology based query expansion," *Inf. Process*., vol. 43, no. 4, pp. 866-886, 2007.
[41]A. Broder, "A taxonomy of web search," *presented at the ACM Sigir forum, ACM New York*, pp. 3-10, 2002.
[42]I.-H. Kang and G. Kim, "Query type classification for web document retrieval," *presented at the Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 64-71,2003.
[43]W. Dan and W. Hui-Lin, "Role of ontology in information retrieval," *J. Electron. Sci. Technol*., vol. 4, no. 2, pp. 148-154, 2006.
[44]A. Maedche and S. Staab, "Ontology learning for the semantic web," *IEEE Intell. Syst.*, vol. 16, no. 2, pp. 72-79, 2001.
[45]L. Zemmouchi-Ghomari and A. Ghomari, "Process of building reference ontology for higher education," *presented at the Proceedings of the world congress on engineering*, pp. 1595-1600, 2013.
[46]B. Smith and C. A. Welty, "FOIS introduction: Ontology-towards a new synthesis.," *presented at the FOIS*, pp. 3-9, 2001.
[47]B. Keltoum, N. Nabila, and M. Djamel, "Towards a Reference Ontology in Islamic Finance and Banking," *presented at the 2018 International Conference on Information and Communication Technology for the Muslim World*, pp. 74-79, 2018.

[48] L. Zemmouchi-Ghomari and A. Ghomari, "Terminologies versus Ontologies from the perspective of ontologists," *Int. J. Web Sci*., pp. 315-331, 2013.

[49] "web-browsers-intro-and-its-settings," 2022. https://mycstutorial.in/web-browsers-intro-and-its-settings/ (accessed Sep. 15, 2023).

[50] M. J. Sadeeq and S. R. Zeebaree, "Semantic Search Engine Optimisation (SSEO) for dynamic websites: A review," *Int. J. Sci. Bus.*, vol. 5, no. 3, pp. 148-158, 2021.

[51] C. Dilmegani, "Web Crawler: What It Is, How It Works & Applications in 2023," 2023.

[52] V. Shkapenyuk and T. Suel, "Design and implementation of a high-performance distributed web crawler," *presented at the Proceedings 18th International Conference on Data Engineering*, pp. 357-368, 2002.

[53] A. James, "Explore The History of Web Scraping," 2023, https://scrape.do/blog/explore-the-history-of-web-scraping/ (accessed Sep. 15, 2023).