Semantic Web 0 (0) 1
IOS Press

A Systematic Literature Review on RDF Triple Generation from Natural Language Text

André Gomes Regino [a,*], Anderson Rossanez [a], Ricardo da Silva Torres [b,c] and Julio Cesar dos Reis [a]
[a] *Institute of Computing, University of Campinas, Campinas, SP, Brazil*
E-mails: andre.regino@students.ic.unicamp.br, anderson.rossanez@ic.unicamp.br, jreis@ic.unicamp.br
[b] *Agricultural Biosystems Engineering and Wageningen Data Competence Center, Wageningen University & Research, Wageningen, The Netherlands*
E-mail: ricardo.dasilvatorres@wur.nl
[c] *Department of ICT and Natural Sciences, Norwegian University of Science and Technology, Ålesund, Norway*
E-mail: ricardo.torres@ntnu.no

Abstract. We live in a big data era of unstructured data expressed as natural language (NL) texts. As the volume of text-based information grows, effective methods for encoding and extracting meaningful knowledge from this corpus are of paramount relevance. A challenging task concerns transforming NL texts into structured and semantically rich data. Semantic web technologies have revolutionized the way we represent and access structured knowledge. Resource Description Framework (RDF) triples serve as a fundamental building block for this purpose, allowing the integration of diverse data sources. This survey examines methods for RDF triple generation and Knowledge Graphs (KGs) enhancement from natural language texts. This study area presents wide-ranging applications encompassing knowledge representation, data integration, natural language understanding, and information retrieval. Our systematic literature review addresses the understanding, characterization, and identification of challenges and limitations in existing approaches to RDF triple generation from NL texts and their inclusion into an existing KG. We retrieved, categorized, and analyzed 150 articles from several scientific databases. We provide a comprehensive overview of the field, identify research gaps, and provide directions for future research. We found the most commonly available study categories, especially considering the domain, the targeted language, the public availability of datasets, and real-world applications. Our results reveal a growing trend in this field in the last few years relate to the use of transformer-based machine learning methods for triple generation. Our study also drives innovation by highlighting open research questions and providing a roadmap for future investigations.

Keywords: Knowledge Graph Construction, RDF triples, Text-to-triple, Natural Language Processing

1. Introduction

In modern information systems, an unprecedented volume of text is generated daily, leading to large data sources from which meaningful knowledge can be derived. In particular, Natural Language (NL) texts (*e.g.*, web pages, social network posts, unstructured textual documents) have been generated substantially in the past few years [1].

A substantial portion of such textual data remains largely unprocessed and untapped [2], representing a reservoir of information that could be explored for deriving actionable insights. In this scenario, developing novel method-

*Corresponding author. E-mail: andre.regino@students.ic.unicamp.br.

ologies and software tools to convert large volumes of unstructured text into structured, computer-interpretable knowledge is of paramount relevance.

Unlocking the potential of unprocessed NL text can significantly contribute to informed decision-making grounded on knowledge representation. With Resource Description Framework (RDF) [3] triples at their core, Semantic Web technologies offer an approach to organizing this information. In this context, Knowledge Graphs (KGs) [4] play a key role in creating rich and interconnected RDF triples by structuring interlinked data meaningfully [5].

KGs serve as invaluable assets across diverse domains, ranging from applications related to improving search engine outcomes [6] to those targeting the enhancement of results of artificial intelligence models and applications [7]. Our motivation extends to the critical role KGs play in powering various applications and services [8, 9]. These structured knowledge repositories find applications in diverse fields, including but not limited to information retrieval systems [10], query and answering questions [11], and data integration platforms [12, 13].

The central problem addressed in this work is the effective generation of RDF triples from NL texts. This often encompasses tasks regarding entity recognition [14], relation extraction [15], and challenges in transforming textual information into a structured, semantically rich format. Our study also investigates existing methods from literature for incorporating generated RDF triples into an existing KG, respecting an underlying ontology [16].

The literature presents distinct approaches for generating RDF triples from NL texts based on Natural Language Processing (NLP) techniques. Some studies, for instance, consider rule-based methods like Open Information Extraction (OpenIE) [17] and Semantic Role Labeling (SRL) [18]. They aim to identify the triple elements (*subject*, *predicate*, and *object*) from textual sentences based on their grammatical structures. More recent studies employ Transformers [19] in identifying and generating RDF triples. These approaches target the generation of a KG from scratch.

We investigate additional challenges when enhancing an existing KG with new RDF triples. For instance, KGs require respecting existing intrinsic ontology statements. This often involves the identification of already-existing triples and whether they follow such ontology definitions. This integration is even harder when constructing a KG relies on assimilating a fully comprehensive knowledge set at once. KGs are constructed based on the information available at a given moment, making it arduous to capture the entirety of knowledge in a single instantiation comprehensively. The constantly expanding repository of NL text introduces a dynamic facet to this challenge. Textual data is not stagnant, as new knowledge is continuously generated. Thus, a mechanism is required for KGs to keep pace with this influx of information. In this sense, KGs need to constantly change and evolve, automatically or manually, to adhere to the evolution of knowledge in the domain they represent [20]. The "freshness" and "recency" of the information in KGs of any size or domain is critical to their usefulness [21].

Existing survey articles from the literature in the context of RDF triple generation emphasize the creation of new KGs [12, 13]. We found studies describing solutions for similar problems, which include: transforming table-formatted texts in RDF triples [22], transforming relational databases in RDF triples [23], and generating NL texts from RDF triples [24]. Nevertheless, to the best of our knowledge, there is no available comprehensive study addressing a systematic literature review that combines research related to RDF triple generation and KG enhancement from NL text.

This article presents a systematic literature review that provides insights into how unstructured textual data are transformed into RDF triples and how the produced knowledge is aggregated into existing KGs linked to an underlying ontology. Our specific objectives are:

- Investigate studies and techniques that identify relevant parts of the NL-produced text to transform it into RDF triples. Intent and entity discovery, relation extraction, and named entity recognition are examples of techniques;
- Investigate techniques that automatically or semi-automatically build RDF triples in a single domain (*e.g.*, biology, medicine) or multi-domain;
- Investigate methods that link the created triples to an existing set of classes and relations defined by one or more ontologies;
- Investigate techniques that check the added knowledge's consistency, validity, and semantic coherence.

Our study employs a systematic approach organized into three phases: Preparation, Execution, and Reporting. Each phase involves distinct steps, from formulating research questions to categorizing and analyzing relevant articles. Our methodology includes the definition of research questions, identification of query strings, definition of inclusion/exclusion criteria, categorization of studies and methods, generation of metadata, description and analysis of results, and reporting on open challenges in the field.

Our study deeply describes solution categories, challenges, advancements, and triple-generation trends from NL texts and KG enhancement. We offer researchers and practitioners valuable insights for future investigations. For those entering the study of text-to-triple transformation, this survey provides a structured overview, highlighting key considerations in the field. Our study also identifies the varied approaches and methodologies employed in the surveyed articles. We ended up with 15 articles, comprising RDF triple generation from unstructured NL texts with the final purpose of KG enhancement. These articles were, in turn, systematically categorized into 10 distinct categories originally defined in this study. Our systematic categorization offers a comprehensive, structured field overview of RDF triple generation and KG enhancement from NL texts. The proposed categorization is a foundation for synthesizing information and discerning common trends, challenges, and advancements within each thematic group. This structured approach may be relevant to researchers, practitioners, and enthusiasts interested in extracting targeted knowledge based on their specific areas of interest. Readers, especially those new to the field, can navigate these categories to understand specific aspects, such as technical methodologies, language specificity, and ontology construction.

The remaining of this article is organized as follows: Section 2 formalizes the problem and its components; Section 3 presents the methodology of how we systematically evaluated the state-of-art; Section 4 provides quantitative analyses of the retrieved articles. Section 5 categorizes and describes the key relevant articles filtered. Section 6 discusses relevant aspects of our findings and addresses the research questions. Finally, Section 7 draws conclusion remarks.

2. The Triple Generation Problem

This section defines the triple generation problem and highlights its inherent difficulties (Section 2.1), using a motivating example (Section 2.2).

2.1. Triple Generation: Formalization and Challenges

This section formalizes the key terms and concepts central to this research. Figure 1 presents the typical pipeline for triple generation and KG enhancement from NL texts.

The core elements of the pipeline are:

Ontology (O): A shared and formal conceptualization and representation of knowledge [16]. Formally, an ontology O is formed by the components (C, P, R), where C represents a set of classes, P is a set of properties, and R is a set of relationships. O is also named T-Box.

Knowledge Graph (KG): A graph-based representation constructed by linking RDF triples according to the structure defined in ontologies. Formally, a Knowledge Graph KG is a connected directed graph formed by V vertices (the entities) and their directed E edges (the connections).

Class (C_i): A category or type within the ontology representing a set of entities with common properties. Formally, C_i is an element of the set of classes C in the ontology.

Property (p_i): A characteristic or attribute associated with entities in the ontology. Formally, p_i is an element of the ontology's set of properties P.

Triple (T_i): A statement in the form of a subject-predicate-object, representing a relationship between entities in the KG. Formally, T_i is the triple (s, p, o) where s is a subject entity, p is a predicate (property), and o is an object entity.

Based on such formalization, the problem of generating RDF triples from NL text and appending them to an existing Knowledge Graph can be formally defined as follows:

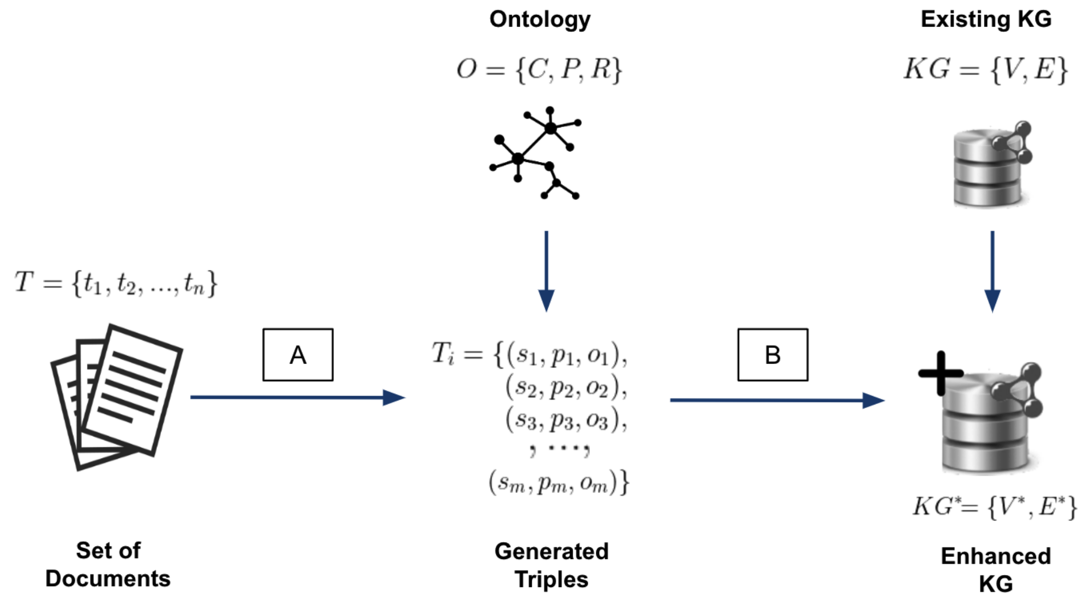

Fig. 1. Typical pipeline for triple generation and KG enhancement based on NL texts. The pipeline comprises five elements (set of textual documents, ontology, generated triples, existing KG, and enhanced KG) and two processing components. Boxes *A* and *B* represent the RDF Triple generation and the KG enhancement components, respectively.

Given a set of NL texts $T = \{t_1, t_2, \ldots, t_n\}$ and an existing Knowledge Graph KG ($KG \neq \varnothing$), the goal is to generate RDF triples $S_T = \{(s_1, p_1, o_1), (s_2, p_2, o_2), \ldots, (s_m, p_m, o_m)\}$ from T (component A in Figure 1), such that generated triples are inserted into KG leading to an enhanced knowledge graph KG', *i.e.*, $KG' = KG \cup S_T$ (component B).

Common challenging problems faced in triple generation refer to:

- **Ambiguity in natural language:** NL is inherently ambiguous, making it challenging to identify entities, relationships, and attributes precisely. Resolving this ambiguity is critical for accurate triple generation;
- **Diverse sentence structures:** NL exhibits a wide range of sentence structures, making it difficult to design an approach that fits all problems for extracting RDF triples. Handling diverse syntactic patterns adds complexity to the problem;
- **Contextual dependencies:** Extracting triples often requires understanding the context in a document or a broader knowledge context. Capturing contextual dependencies is key to accurately generating triples;
- **Named Entity Recognition (NER):** Identifying entities within the text is a subproblem that involves NER. Variability in entity types and expressions poses challenges in accurately recognizing entities in different contexts.

Aggregating triples to an existing KG pose additional challenges, such as:

- **T-Box alignment:** A new RDF triple can be composed of entity and property types not described in the initial version of the KG. Such a new structure must be aggregated into the T-Box portion of the KG and the triple itself in the A-Box portion. In some situations, however, that specific triple may not be incorporated into the KG, depending on the context in which the KG should be used. If the KG's T-Box is aligned with a 3rd-party ontology, including such new knowledge representation may even affect such alignment.
- **A-Box alignment:** The entities (subject, object, or both) of the new triple may already be present in the graph. In this case, adding a triple will require identifying whether the URIs representing it are already available in the KG. If so, this could lead to appending a new property (*i.e.*, the new triple's predicate) to existing entities

or linking an existing entity to a new one. Also, neither of the elements of the new triple may already exist in the KG; in this case, those new elements may end up being disconnected.

- **URI alignment:** In case the URI creation for the new triples is automatic, adding such triples to an existing KG may require searching for existing URIs that represent the same entities. This would ensure consistency in the knowledge representation.

2.2. Motivating Example

Let us consider the following sentences:

- t_1 = "SpaceX, founded by Elon Musk, is known for pioneering space exploration and innovation;"
- t_2 = "The engine is incompatible with Ford Mustang 1990."

For the first example t_1, relevant parts of it are identified. Step A of Figure 1 identifies entities, such as "SpaceX," "Elon Musk," "space exploration," and "innovation" from the given text, transforming this information into RDF triples, producing relationships like:

- <SpaceX, founded_by, Elon Musk>
- <SpaceX, known_for, space_exploration>
- <SpaceX, known_for, innovation>

The created triples should be integrated into an existing KG, based on the application of ontology statements from O. In this context, the rules identify "SpaceX" as a company and "Elon Musk" as a founder, ensuring these classes are present in O. Given an existing KG, which initially comprises nodes for "SpaceX," "founder," and general relations, integrating new triples enriches this graph. The updated KG – enhanced knowledge graph KG' – (step B of Figure 1) now includes nodes specific to "space exploration" and "innovation," accompanied by their corresponding relations.

For t_2, entities like "engine," "incompatible," and "Ford Mustang 1990" are identified in T in step A of Figure 1. The generated RDF triples build relationships such as <engine, incompatible_with, Ford_Mustang_1990>. Exploring ontology elements helps ensure the meaningful integration of that into the KG. For instance, identify "engine" as a component, "incompatible" as a property, and "Ford Mustang 1990" as a specific model of O. Examining the state of KG before the integration of new information reveals existing nodes for "engine" and "incompatible." The subsequent integration of S_T enhances this KG by adding nodes specific to "Ford Mustang 1990" in step B of Figure 1 reaching knowledge graph KG'.

We list the following challenges for converting t_1 and t_2 into triples:

- **Ambiguity in natural language:** The phrase "founded by" may pose ambiguity. It could be interpreted as the founder of a company or the person behind a project;
- **Diverse sentence structures:** The sentence structure includes both the subject "SpaceX" and the additional information "founded by Elon Musk" requiring a mechanism to handle diverse syntactic patterns;
- **Contextual dependencies:** Understanding the context that associates Elon Musk with SpaceX is essential. What does "incompatible" mean? The whole engine? Without context awareness, incorrect triples may be generated;
- **Named Entity Recognition (NER):** NER is required to identify "SpaceX," "Elon Musk," "Ford," "Mustang," "1990," and "engine" as entities, introducing the challenge of recognizing entity types accurately.

3. Methodology

This section describes the methodology for systematically reviewing the literature on triple generation and KG enhancement from NL texts. We follow the work by Budgen and Bereton [25], which guides our methodology. Figure 2 presents our methodology with three distinct phases: Preparation, Execution, and Reporting. A series of ordered steps further delineate each of these phases. This phase-based structure provides a comprehensive overview of the entire procedure and highlights our experimental approach's systematic and controlled nature.

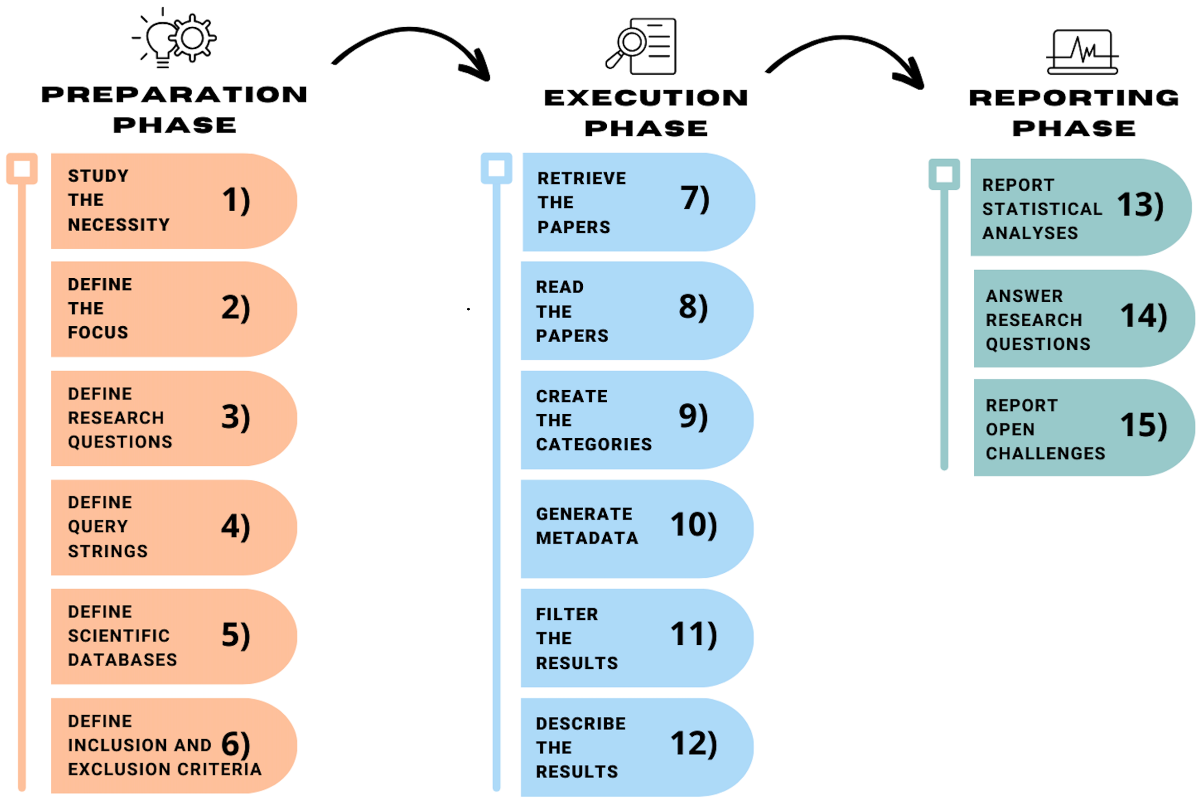

Fig. 2. Systematic literature review methodology comprising three main phases: Preparation, Execution, and Reporting. Each phase comprises steps represented by numbers from 1 to 15.

The “Preparation Phase” serves as the initial step, where we lay the groundwork for our investigation. During this phase, we formulate research questions, identify relevant scientific databases, and establish rigorous inclusion and exclusion criteria (steps 1 to 6 in Figure 2). The Preparation Phase aims to construct a process that guides this systematic literature review article.

We proceed to the “Execution Phase”, in which we pragmatically implement the process discussed in the Preparation Phase. This phase entails executing a sequence of planned tasks, including database queries, article retrieval, and applying inclusion and exclusion criteria. We synthesize data, categorize results, and perform preliminary analyses (steps 7 to 11 in Figure 2). The Execution Phase emphasizes data acquisition and initial assessments.

Finally, the conclusion of the research process occurs in the “Reporting Phase.”. We gather all our analyses, including bibliometric, quantitative, and qualitative assessments, into a cohesive and structured presentation. This phase is dedicated to synthesizing our findings, the systematic organization of results, and in-depth analytical examinations (steps 13 to 15 in Figure 2). We discuss our outcomes comprehensively, emphasizing their implications and significance within the context of our survey’s objectives.

Section 3.1, Section 3.2, and Section 3.3 describe each phase and their steps.

3.1. Preparation Phase

1. Study the Necessity: The initial step of the Preparation Phase was essential to thoroughly study the necessity of conducting this investigation. We understood the current landscape of RDF triple generation from NL texts and investigated the relevance of this topic in the context of semantic web technologies and knowledge representation. On this basis, we justify the significance of our study. The thorough study of such literature gaps not only justified

Table 1
Study Definition. Initial questions to define the reasons for conducting this study, as well as its targets.

Question	Definition
Why?	Investigate different methods of building knowledge graphs from natural language texts.
Where?	Via the existing literature related to the area of study
What?	The survey explores various approaches, algorithms, and techniques used for converting natural language texts into RDF triples.

Table 2
Overview of Research Questions (RQs). These focused inquiries form the basis for the exploration of the surveyed literature.

RQ	Question
RQ-01	What are the benefits and drawbacks of a method that generates RDF triples from texts?
RQ-02	Are there any patterns for texts used as input and for the triples generated as output?
RQ-03	What are the most used techniques? What are the most accurate ones?
RQ-04	Are there fully automated approaches to generate knowledge graphs from text?
RQ-05	What are the main applications that benefits from the text-to-triple approaches?
RQ-06	How do the methods explore the T-Box and A-Box in terms of text-to-triple generation?

the initiation of our investigation but also guided the subsequent phases by providing a clear rationale for their objectives and methodologies.

2. Define the Focus: After establishing the necessity, we defined the specific focus of our study. Here, we determined the boundaries and scope of our investigation. To guide this step, we answered three important questions (Table 1): What is the study? Where do we gather scientific data to answer the research questions? Why is the study important? We also made decisions regarding the scope of the study:

- We would accept any RDF triple generation method (*e.g.*, entity recognition, relation extraction). This choice aimed to embrace the diversity of methodologies prevalent in the field, ensuring a comprehensive exploration of the topic.
- We should not restrict the survey to scientific articles by the type of natural language texts (e.g., news articles, academic papers, social media posts). The rationale is that this decision allows to capture the breadth of RDF triple generation applications across different textual domains.

The motivation to review the literature on triple generation and KG enhancement derived from the lack of this type of research in the state-of-the-art based on our initial observations of result analyses. Among all the articles found on the subject of generating triples (in this study), only 10% of them referred to adding these triples to an already existing KG.

3. Define Research Questions: We defined clear and concise research questions in addition to the focus steps. These questions were tailored to address the specific aspects of RDF triple generation and KG enhancement from NL texts and are addressed through quantitative and qualitative analyses in Section 4 and Section 5. Those questions were defined to ensure that our study provides meaningful insights and answers to existing challenges or gaps identified.

Table 2 presents the research questions. The RQs address the benefits and drawbacks of the surveyed methods, patterns in input texts and generated triples, the most utilized and accurate techniques, the presence of fully automated approaches, applications benefiting from the methodologies; and the exploration of T-Box and A-Box in text-to-triple generation. These research questions not only serve as a guide for our analyses but also function as a means to benchmark the success of our survey.

4. Define Query Terms: To systematically retrieve relevant literature, we defined query strings used to search scientific databases. Effective search terms are essential for retrieving a comprehensive set of articles that align with our goals. We incorporated relevant keywords, terms, and Boolean operators to refine the search. Table 3 presents the employed query terms. These can be condensed in a unique query, as defined in Equation 1.

$$\begin{aligned} &(\text{ generation } OR \text{ extraction } OR \text{ construction } OR \text{ population }) \quad AND \\ &(\text{ triples } OR \text{ knowledge graph } OR \text{ knowledge base }) \quad AND \\ &(\text{ natural language text } OR \text{ unstructured text } OR \text{ textual data }) \end{aligned} \tag{1}$$

The query string represents a balance between completeness and specificity. The selection of terms reflects the multifaceted nature of the field, encompassing aspects such as entity recognition, relation extraction, and knowledge representation.

The iterative process of refining and validating the query string in Equation 1 involved a continuous dialogue among the research team (authors in this article), ensuring that the queries encapsulate emerging trends and diverse perspectives within the domain. The final set of search terms, detailed in Table 3, represents a distilled synthesis of our collective understanding.

5. Define Scientific Databases: At this step, we define in which scientific databases we could execute the queries listed in the previous step. We chose databases most likely to contain relevant articles related to Computer Science, including the domains of Text classification/Mining and Semantic Web. The chosen databases were ACM DL,[1] IEEE Xplore,[2] Springer Link,[3] Elsevier,[4] Scopus,[5] and ACL Anthology.[6]

We assume these databases contain highly ranked and cited articles relevant to our study. The chosen databases collectively offer a broad spectrum of publications, including conference proceedings and journals, ensuring a sound exploration of the academic landscape.

6. Define Inclusion and Exclusion Criteria: In the last step of the Planning Phase, we defined rigorous inclusion and exclusion criteria to determine which articles should be included in the survey. Inclusion criteria (Table 4) specify the characteristics an article must possess to be considered for the survey, while exclusion criteria (Table 5) identify reasons for excluding articles.

In summary, we included articles that scientifically (IC-01) contribute to the understanding of RDF triple generation from NL texts (IC-02), and written in English (IC-03). By setting these standards, we aimed to ensure that the articles chosen were not only recent but also provided comprehensive insights into the current state of the art. We excluded articles that are too old (EC-01), duplicates (EC-02), short (EC-03), without abstract (EC-04), not written in English (EC-05), and without further explanation on how generating triples based on text (EC-06).

3.2. Execution Phase

7. Retrieve the Papers: In the first step of the Execution Phase, we retrieved the relevant articles based on the query strings defined in the Preparation Phase. We collected all articles that matched the search criteria in the predefined databases. To verify the criteria, we read the title and abstract of each article. We retrieved a total of 150 articles from our search queries. By relying on the query strings built in the Preparation Phase, we ensured that the retrieved articles aligned closely with our present study's specific focus and objectives.

8. Read the Papers: We read and reviewed each article systematically. This step involved a deep dive into the content of the selected articles. This in-depth review process involved a collaborative effort among the researchers, leveraging diverse perspectives to interpret and extract insights from the selected papers.

While reading, we took detailed notes and highlighted key information. The key information extracted is described in the following steps: "Create Categories," "Generate Metadata," and "Describe Results."

We employed a collaborative approach in the review process. To this end, we involved 15 graduate students, among Master's and Ph.D., enrolled in the Semantic Web course (Graduate Program in Computer Science, Institute

[1] https://dl.acm.org/ (As of Dec. 2023).
[2] https://ieeexplore.ieee.org/Xplore/ (As of Dec. 2023).
[3] https://link.springer.com/ (As of Dec. 2023).
[4] https://www.elsevier.com/ (As of Dec. 2023).
[5] https://www.scopus.com/ (As of Dec. 2023)
[6] https://aclanthology.org/ (As of Dec. 2023).

Table 3
Search queries. They represent a breakdown of the main query (*cf.* Equation 1), aligned with the study objectives.

Query	Terms description
Q-01	"generation" + "triples" + "natural language texts"
Q-02	"generation" + "triples" + "unstructured texts"
Q-03	"generation" + "triples" + "textual data"
Q-04	"generation" + "knowledge graphs" + "natural language texts"
Q-05	"generation" + "knowledge graphs" + "unstructured texts"
Q-06	"generation" + "knowledge graphs" + "textual data"
Q-07	"generation" + "knowledge base" + "natural language texts"
Q-08	"generation" + "knowledge base" + "unstructured texts"
Q-09	"generation" + "knowledge base" + "textual data"
Q-10	"extraction" + "triples" + "natural language texts"
Q-11	"extraction" + "triples" + "unstructured texts"
Q-12	"extraction" + "triples" + "textual data"
Q-13	"extraction" + "knowledge graphs" + "natural language texts"
Q-14	"extraction" + "knowledge graphs" + "unstructured texts"
Q-15	"extraction" + "knowledge graphs" + "textual data"
Q-16	"extraction" + "knowledge base" + "natural language texts"
Q-17	"extraction" + "knowledge base" + "unstructured texts"
Q-18	"extraction" + "knowledge base" + "textual data"
Q-19	"construction" + "triples" + "natural language texts"
Q-20	"construction" + "triples" + "unstructured texts"
Q-21	"construction" + "triples" + "textual data"
Q-22	"construction" + "knowledge graphs" + "natural language texts"
Q-23	"construction" + "knowledge graphs" + "unstructured texts"
Q-24	"construction" + "knowledge graphs" + "textual data"
Q-25	"construction" + "knowledge base" + "natural language texts"
Q-26	"construction" + "knowledge base" + "unstructured texts"
Q-27	"construction" + "knowledge base" + "textual data"
Q-28	"population" + "triples" + "natural language texts"
Q-29	"population" + "triples" + "unstructured texts"
Q-30	"population" + "triples" + "textual data"
Q-31	"population" + "knowledge graphs" + "natural language texts"
Q-32	"population" + "knowledge graphs" + "unstructured texts"
Q-33	"population" + "knowledge graphs" + "textual data"
Q-34	"population" + "knowledge base" + "natural language texts"
Q-35	"population" + "knowledge base" + "unstructured texts"
Q-36	"population" + "knowledge base" + "textual data"

Table 4
Inclusion Criteria. Articles must present the characteristics here defined to be included in the study.

EC	Type	Definition
IC-01	**Application**	Book chapters, conference papers (full and short articles), journals and thesis
IC-02	**Application**	Papers of type as listed at IC-01 that create, use or theoretically define a way to generate triples from any kind of textual data.
IC-03	**Language**	Articles written in English

Table 5

Exclusion Criteria. Articles with such characteristics are not included in the study.

EC	Type	Definition
EC-01	**Date**	Articles 10 years older than the execution of the queries in August 2023
EC-02	**Duplicate**	Articles' duplicate found in the scientific databases
EC-02	**Short Articles**	Articles with less than 4 pages
EC-03	**Abstract**	Articles without abstract
EC-04	**Language**	Articles not written in English
EC-05	**Application**	Articles that just mention a tool that generates triples from text, not explaining how and why it was used

of Computing, University of Campinas, Brazil) to actively contribute to this step. Each student, possessing the ability and expertise in reading and interpreting scientific articles (topic taught in the course), analyzed 10 articles each. Their task extended beyond reading; they were invited to categorize the articles based on predefined categories outlined in the subsequent step. The authors of this study double-checked the final assessment of each article reviewed by the students to ensure their quality further.

The rationale behind involving master's and Ph.D. students in this collaborative review lies in many reasons. First, the students' academic standing ensures an understanding and critical analysis necessary for interpreting the nuances present in scientific articles. Their familiarity with the Semantic Web and ontologies domain positions them as adept evaluators of the selected articles. Moreover, including these graduate students aligns with a commitment to diversity in perspectives. Drawing on their varied academic backgrounds and research interests within the Semantic Web, those students brought a breadth of insights that enriched our study's analytical depth. The collaborative approach contributed to a shared understanding of the reviewed literature among the research team, reducing risks associated with the use of a possible biased categorization.

In addition to the categorization, the students identified valuable aspects of each article's characteristics. These aspects included the availability of the data handled in the article, the type of evaluation conducted, and an assessment of the solution's applicability in real-world settings. These insights, derived from the students' discerning analysis, provided a holistic view of the articles beyond their immediate contributions to RDF triple generation and KG enhancement. After the students categorized the articles, the authors of this study reviewed the provided results, correcting potential erroneous categorizations when necessary.

9. Create the Categories: To organize the collected articles, we created categories that reflect the various aspects of RDF triple generation from NL texts. These categories are aligned with the research questions and focus areas. We defined such categories based on our screening of this literature review and the authors' background. The names were crafted to align with common terminology in the field, ensuring clarity and coherence in the categorization process.

Categorizing the articles facilitates the synthesis of information and helps identify common themes and trends in the literature. Table 6 presents the proposed 10-category set and a small description of each.

The categories for organizing the collected articles reflect a diverse landscape of methodologies in the context of RDF triple extraction and KG enhancement from NL texts. We grouped categories that describe similar aspects of the literature for the sake of improving comprehension:

– **Language specificity**: represented by "Multilingual," "English Specific," and "Non-English" categories. The rationale for choosing this category group (along with its categories) is that understanding how language is used in the text-triple-KG transformation directly impacts solutions created and whether there is a tendency to process multilingual texts generating KGs that are also multilingual;
– **Technical methodology**: like "Transformer" category. The rationale for studying models based on deep learning reveals the authors' interest in categorizing articles that follow the trend in using these techniques;
– **Ontology and KG enhancement**: represented by "T-Box Population" and "A-Box Population" categories. This group of categories aimed to group articles based on how the KG enhancement stage occurs: only via instances (A-box), or also through their classes (T-Box).

Table 6
Categories of the retrieved articles and their correspondent description, according to this study's focused areas.

Category	Definition
Transformer	Approaches that use the Transformer architecture to build the triples
Multilingual	Approaches that process and generate triples from and to more than one language
English Specific	Approaches that process and generate triples from texts written in English
Non-English	Approaches that process and generate triples from texts not written in English
T-Box Population	Approaches that define new classes and properties
A-Box Population	Approaches that insert new instances
Lexical Resources	Approaches that use any external lexical resources (e.g., WordNet)
Specific Transformation	Approaches that generate triples of a single domain (e.g., biomedical ontologies)
Generic Transformation	Approaches that generate triples of any type of domain (e.g., DBpedia)
Link Creation	Approaches that create links of the newly added resources to existing well-known Linked Open Data (LOD) datasets (e.g., Wikidata)

Table 7
List of all metadata used in the survey, their types, and descriptions.

Metadata	Type	Description
Year of Publication	quantitative	respecting ten-year-old exclusion criteria from Table 5
Country	quantitative	nationality of the authors
Category	quantitative	values found in Table 6
Objective	quantitative	3 possible values: extract knowledge from text; create KG; connect to KG with existing ontology rules
Degree of Automatism	quantitative	3 possible values: manual, automatic, semi-automatic
Methodology	qualitative	article's methodology
Domain	quantitative	10 possible values: health, technology, education, language, agriculture, commerce, arts, government, financial, no domain
Used in real-world	quantitative	the article generates a product used in real world (e.g software) or not
Dataset Available	quantitative	the article generates a dataset publicly available or not
Evaluation type	quantitative	5 possible values: quantitative, qualitative, user evaluation, gold standard comparison, case study
Strengths	qualitative	strong points of the methodology, results and evaluation
Weaknesses	qualitative	weak points of the methodology, results and evaluation
Validation Risks	qualitative	potential threats or problems that could compromise the validity and reliability of the results
Open Challenges	qualitative	research challenges that are still open according to the article

– **Resource utilization:** represented by “Lexical Resources” category. We understand that using external lexical resources can help create more rule-adherent triples and improve the overall quality of the final version of the KG;
– **Domain specificity**: represented by “Specific Transformation” and “Generic Transformation” categories. This group of categories concerns understanding whether there is a preference for enhancing KGs based on text from a specific domain, such as biomedical, with a large set of ontologies (*e.g.*, BioPortal[7]), or whether it is not related to any specific domain.

10. Generate Metadata: Metadata generation involves extracting essential data from each article to create a structured dataset for analysis. The collected metadata is further used in Section 4, generating relevant quantitative analyses. Figure 7 describes the metadata set.

[7]https://bioportal.bioontology.org/ (As of Dec. 2023).

11. Filter the Results: Upon completing the extensive literature review and categorizing the initial 150 articles across 10 categories, a critical refinement in our approach emerged during the subsequent analysis. We confirmed that not all the surveyed articles conformed to the specific focus of our study, which centered on the generation of triples from NL texts, and their insertion into an existing KG, adhering to predefined ontology specifications. We then narrowed our analysis to a subset of articles that align more closely with our study's primary goal. Specifically, we focused on investigations that actively process NL text, generate RDF triples, and integrate them into an existing KG. This refinement ensured a targeted exploration of methods and techniques aligned with our goals.

It is worth noting that while various surveys exist in the literature detailing techniques for generating triples from textual data [26–28], our present study aimed to fill a unique niche. To the best of our knowledge, there is a gap in the literature concerning literature review contributions specifically addressing the production of RDF triples that adhere to ontology specifications and are appended to an existing KG. In this sense, our study contributes to the broader understanding of RDF triple generation and finds a distinct space in the literature.

12. Describe the Results: We described the results of our literature review based on the generated metadata (*cf.* Section 4) and categorized articles (*cf.* Section 5). We summarized the key findings and insights from each category, highlighting trends, challenges, and advancements in the context of this survey.

3.3. Reporting Phase

13. Report Analyses: We reported analyses conducted on the collected data, creating quantitative insights into trends, correlations, and patterns within the literature. We analyzed publication trends, research challenge distribution, and the prevalence of specific keywords. We presented the findings using graphs, charts, and tables in Section 4.

14. Answer Research Questions: The first crucial step was systematically answering the research questions formulated in the Preparation Phase. Based on the findings from the Execution Phase, we provided clear and concise responses to each research question (*cf.* Section 4 and Section 5). We referenced the relevant articles and their key findings to support the answers. We ensured that the responses aligned with our study's scope and provided insights into the state of the field.

15. Report Open Challenges: We identified and reported open challenges and areas of future research within our study scope (*cf.* Section 6). We summarized the limitations and gaps in the existing literature and discussed the unresolved issues, methodological shortcomings, and opportunities for further investigation.

4. Quantitative-based Literature Analysis Results

This section presents a quantitative analysis of the literature on the generation of RDF triples from NL texts. We aim to gain insights into the landscape of research in this domain. In this analysis, we used all 150 article to understand better how the state of the art dealt with the generation of triples with or without adding them to an existing KG (KG enhancement). We conducted a literature analysis based on the metadata, described in Step 10 of our methodology (cf. Figure 2 and Section 3). The list of used metadata and analyses and why we used them is as follows:

– **Number of articles per year:** provides insights into the temporal evolution of research, indicating trends over time (cf. Figure 3);
– **Number of articles per category:** enables an exploration of diverse approaches and methodologies, categorizing articles based on specific themes and focus areas (cf. Figure 4);
– **Number of articles by domain:** facilitates a domain-specific analysis of the literature, offering a deeper understanding of how RDF triple generation varies across different knowledge domains (cf. Figure 5);
– **Number of articles by scientific database:** identifies the distribution of relevant research across reputable scientific databases, revealing the popularity and impact of different platforms (cf. Table 8);
– **Number of articles by evaluation type:** identifies the methodologies used to assess proposed solutions, helping to discern the rigor and validity of the research within the surveyed literature (cf. Figure 6);

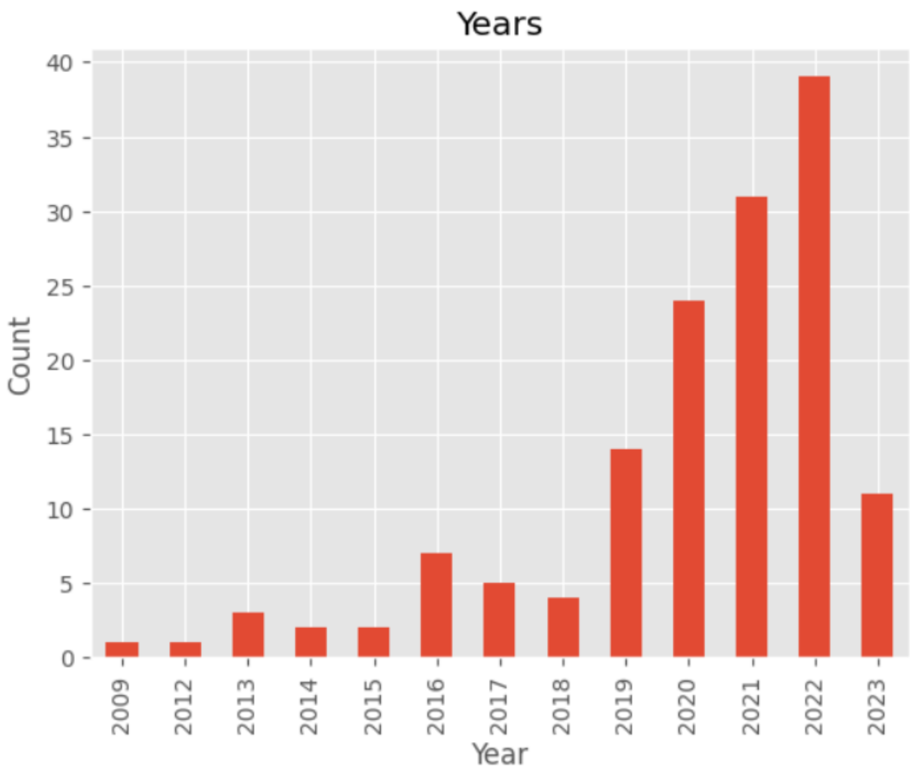

Fig. 3. Number of articles by publication year. We observe a growing trend in the later years, especially from 2019 to 2022. The numbers related to 2023 are from January to August.

- **Number of articles by the automation level:** classifies articles based on the degree of human intervention involved in the proposed solutions, providing insights into the level of automation achieved (cf. Figure 7);
- **Number of articles used in real-world applications:** distinguishes between papers with immediate practical implications and those whose contribution is more theoretical or conceptual, highlighting the potential impact of the proposed solutions (cf. Figure 7);
- **Number of articles that make data available:** promotes transparency and reproducibility in the scientific community by identifying articles whose study handled data that is publicly available (cf. Figure 7).

Figure 3 presents the number of articles published annually. One may notice a growing trend, especially considering the 2019–2022 interval. Such a finding helps to ensure the relevance of our work's subject in recent years. We further analyze and discuss our findings from this analysis in Section 6 (Discussion).

Figure 4 presents the number of articles per category. We included a category for survey-related studies. With a total of 10 categories. This plot helps us understand the distribution of research efforts across various subfields within the domain. This is crucial for identifying trends and emphasizing different aspects of the topic from the existing literature.

In addition, the distribution of articles across categories and evaluation approaches can indicate more mature areas (*e.g.*, A-Box Population, Generic Transformation, and English-Specific) or require further investigation (*e.g.*, Multilingual, Link Creation, and Non-English). In our analysis, each article may belong to one or more categories. We observe in Figure 4 that Transformers are one of the most used techniques in the literature (**RQ-03** question).

Figure 5 presents the most common domains in which RDF triples are extracted from NL texts for posterior KG enhancement. Most of the studies found in the literature target no specific domain. As for studies focusing on specific domains, *health* is the most frequent, followed by *technology*, and *science & education*. We understand a correlation between the domain of natural language text; the Knowledge Graph produced from these triples, and the

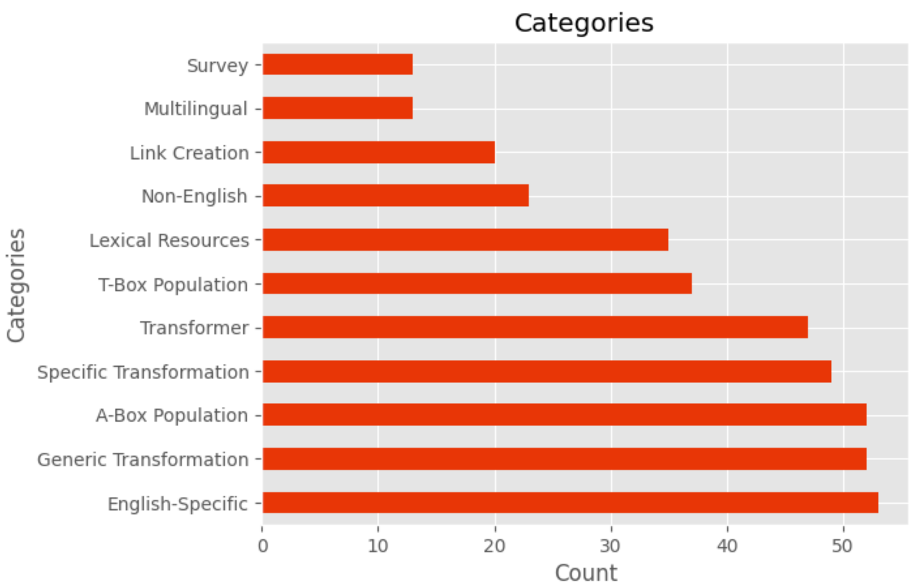

Fig. 4. Number of articles by category. Lower numbers of studies in the Multilingual or Link Creation categories present an opportunity for future research. On the other hand, studies focusing on the A-Box population or English-specific are more common.

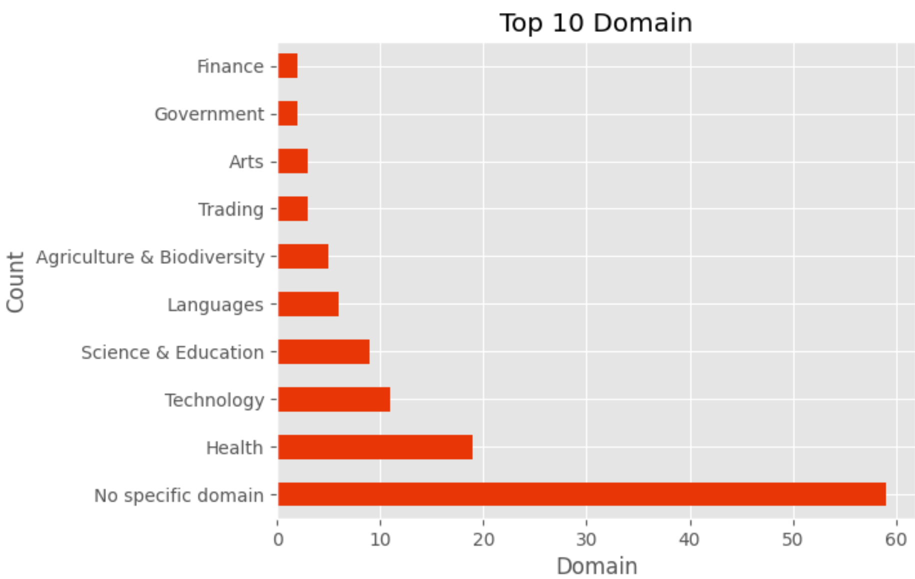

Fig. 5. Number of articles within the 10 most-populated domains. It helps to answer **RQ-05**. As observed, most studies are not specific to a particular domain.

application that benefits from this knowledge. Therefore, we believe that Figure 5 helps answering **RQ-05**. RQ-05 is related to "the main applications that benefit from the text-to-triple approaches."

Figure 6 presents the number of articles grouped by the type of evaluation they employ. We identified five distinct types of evaluation, such as quantitative analysis and case studies. This figure highlights the diversity of methodologies used in assessing the proposed solutions.

We investigated the extent to which the solutions are capable of automation, which is a fundamental factor in assessing these methods' scalability and real-world applicability. The first plot of Figure 7 categorizes articles based

Table 8
Number of studies found, according to the Scientific Database. *Elsevier Scopus* and *IEEE Xplore* have yielded more relevant studies.

Database	#
ACM DL	14
IEEE Xplore	43
SpringerLink	37
Elsevier Scopus	48
ACL Anthology	7

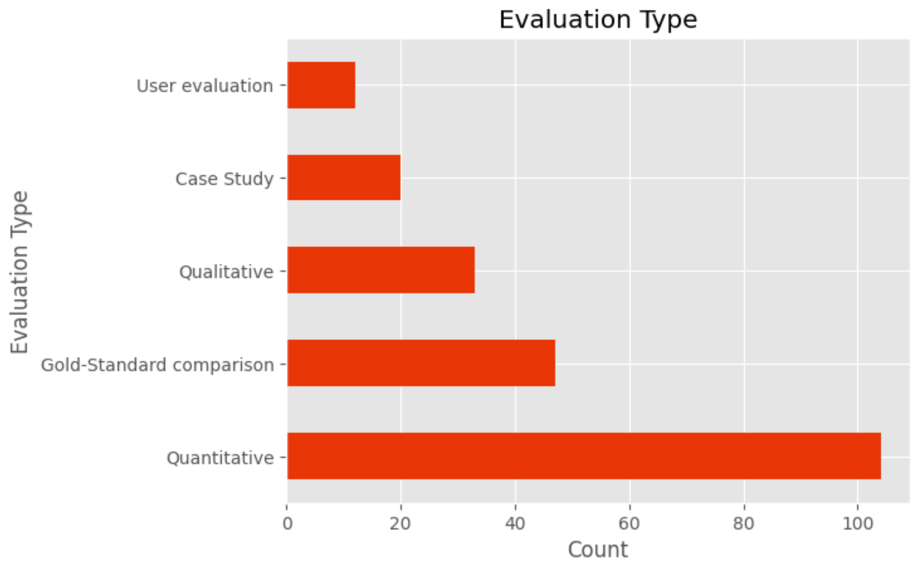

Fig. 6. Number of studies by each type of Evaluation. Studies evaluated by quantitative analysis (especially focusing on using gold standards) are more commonly observed in the literature.

on how they describe their solutions' level of "automaticity." We classify articles into three categories: automatic, semi-automatic, and manual. The level of "automaticity" in the proposed solutions can inform us about how ready for practical deployment they are.

We analyzed studies concerned with real-world applications and discussed data availability. Both of them have implications for the practicality and reproducibility of the research. Identifying how many articles have been applied in real-world scenarios and how many make their data publicly available guide researchers in selecting the most relevant and accessible resources for their work. In the second plot of Figure 7, we explore the practical relevance of the articles by counting how many of them describe applications of proposed methods in real-world settings. In the third plot of Figure 7, we show the number of articles that handled publicly available data. By observing the numbers concerning the automation aspect in Figure 7, we can positively answer our **RQ-04**. The number of studies in which no human intervention is required in the process overcomes those in which some level of manual interaction is required.

5. Category-based Literature Analysis Results

This section categorizes the discovered studies in the literature related to the generation of RDF triples from NL texts aiming to enhance Knowledge Graphs. The categorization of articles provides a navigational guide to help understand the area under study. In addition, this sets the stage for discussions, paving the way for advancements in this domain.

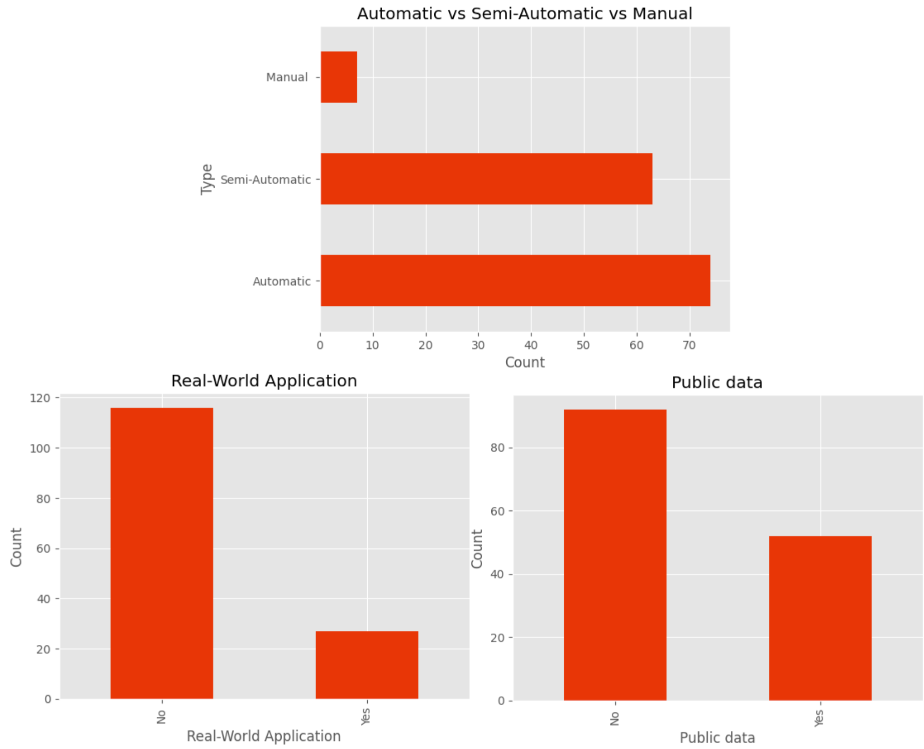

Fig. 7. Three plots describing how automatic the solution is, how applicable it is in the real-world settings, and how research data are publicly available. Most studies present a fully automated method, do not relate to a real-world usage scenarios, and do not disclose their data publicly.

From the total of 150 articles obtained after the first screening (*cf.* Section 3), we narrowed them down to 15 after applying filters described in Step 11 of Section 3 for the key aspects considered in this study. Section 4 presents the considered categories' descriptions.

The upcoming sections describe in further detail the categories and the studies that best fit such categories (relying on our procedure conducted for this purpose). Figure 8 presents our proposed organization of categories relying on our observations from the analyzed articles (in white, described in Table 6) and group of categories (in black, described in the ninth item of Section 3). Some categories were created by overlapping two larger groups of categories (*e.g.*, the Transformers category is the intersection of the Technical Methodology and Language Specificity group of categories). Figure 8 aims to visually represent how the categories relate, highlighting what is common and different among them.

5.1. *Language Specificity*

We considered the language specificity addressed in the study. More specifically, we considered whether the method applies to English texts (*cf.* Section 5.1.1), other languages than English (*cf.*, Section 5.1.2), or even if applicable to texts in multiple languages (*cf.*, Section 5.1.3). The Language Specificity is represented by the green circle in Figure 8.

Multilingual KGs created over multilingual texts provide a versatile solution by handling multiple languages within a single model, enabling broader language coverage. However, they may face challenges in achieving language-specific precision. Language-specific KGs, while resource-intensive, can offer superior accuracy and depth for individual languages. The choice between multilingual and language-specific depends on the application's language requirements and the trade-off between broad coverage and language-specific proficiency.

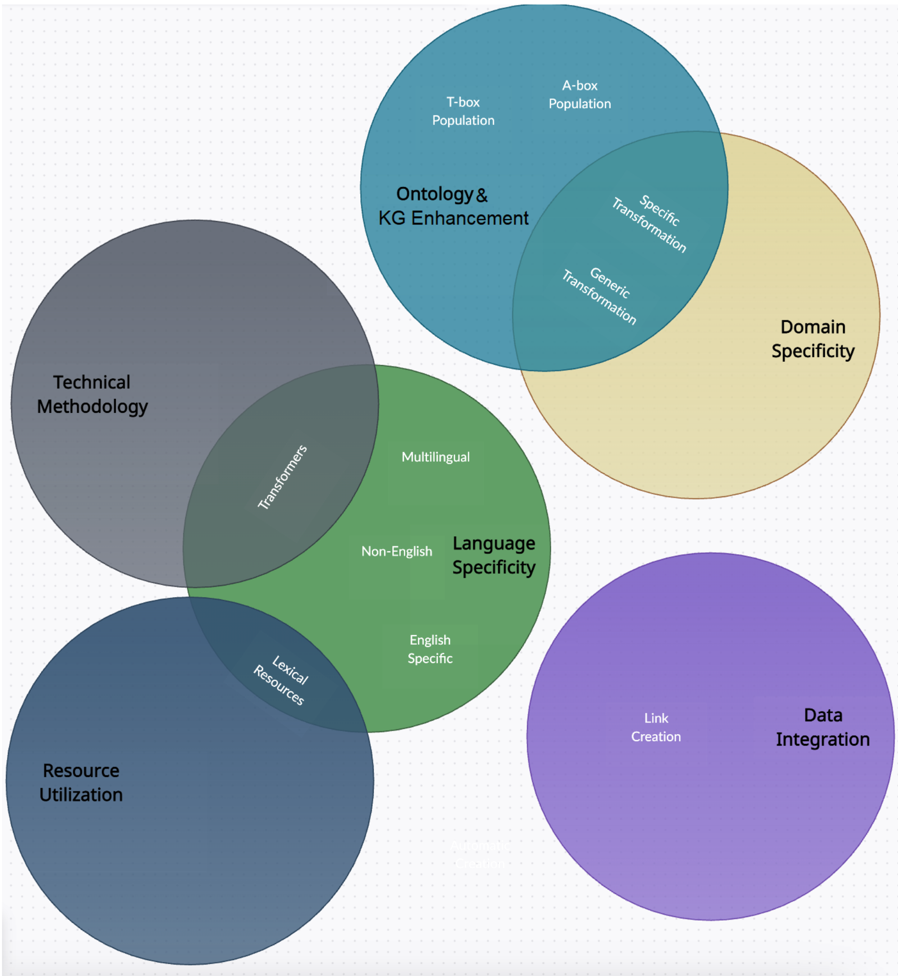

Fig. 8. Venn Diagram showing the articles' categories [15 articles] (in white), their set of categories (in black) in addition to their intersection according to the findings from our result analysis.

5.1.1. *English-Specific*

The work by Zhang and Nguyen [29] focused on generating triples from a text in the English language. Their primary objective was to create KGs and connect them with ontology classes and properties. The methodology involved text preprocessing, including filtering by English, removal of special words, and capitalization. The following step was done by applying distinct relation and entity extraction methods, including rule-based, OpenNRE, and OpenIE [30]. The extracted relations were inserted into a graph database, and a software tool was created to explore the results. The specific textual domain is health-related (COVID-19), and although the application has not been used in real-world settings, the dataset is public. The conducted evaluations are quantitative. The solution's strengths lie in its automation, independence from domain experts, and generalizability. On the other hand, the solution may yield limited results because it sometimes employs models trained in unrelated domains. Open challenges include using advanced extraction techniques like BERT [31] and connecting to external vocabularies for additional information, such as DBpedia [22].

Similarly, the work by Yu *et al.* [32] focused on generating triples from texts in the English language. Their primary objective was establishing connections with classes and properties in an existing ontology. The methodology lies in constructing a network encyclopedia classification system and KG. Domain experts defined the classification system as a KG with concepts and instances. A web crawler algorithm extracted classification information, capturing upper and lower concept relationships. Co-occurrence analysis was employed to mine implicit associations between concepts, facilitating the extraction of upper and lower relationships during KG construction. Their algorithm quantifies co-occurrence probabilities based on the occurrence frequency of classifications, identifying relationships if certain conditions are met. Their solution utilized classification data from encyclopedia entries to ensure a reliable and rich KG, demonstrating the interplay between web crawling, co-occurrence analysis, and semantic analysis in constructing a comprehensive classification system. The work by Yu *et al.* [32] specifically constructed a KG related to food. Their method was applied in a project using a public dataset. The conducted evaluation was quantitative. The solution's strengths lie in its comprehensive approach to relationship extraction in a domain KG, through various knowledge sources, including structured, semi-structured, and unstructured data. Their method employed a convolutional residual network for extracting lower relationships from web texts and stored the resulting KG in *Neo4j*[8], demonstrating practical applicability. Concerning their investigation weaknesses, the solution presented dependencies on available knowledge sources, and its effectiveness is contingent on the quality of training data, which is highly sensitive to extraction errors. Risks to the validation included the potential imperfections and errors in the Wikipedia data used for experiments and the necessity for parameter tuning in NLP and semantic analysis algorithms. Open challenges involved developing different relationship extraction models for diverse forms of knowledge existence.

5.1.2. Non-English

The work by Rios-Alvarado *et al.* [33] investigated the triple generation from documents written in languages other than English, specifically Spanish. Their main goal was the extraction of knowledge from text, the creation of KGs, and the connection with classes and properties from an existing ontology. The methodology involved text segmentation using NLP, word tagging, knowledge extraction through lexical analysis, and the construction of the KG. The application domain was technology. Evaluations included quantitative analysis and comparison with a gold standard. The strength of such a solution appears in enabling knowledge extraction from Spanish language text, which is not common in the literature. However, potential weaknesses included the manual validation of extracted entities using a small sample, which can impact accuracy and reliability. Open challenges included exploring alternative methods for feature extraction and implementing the KG for question-answering applications.

The study by Stern and Sagot work [34] belongs to the non-English category as they focused on extracting entities from French texts. The methodology employed Named Entity Recognition (NER) with the support of well-known databases, such as Wikipedia,[9] Aleda [35], and Geonames,[10] specifically tailored for news-related content. The automated process links these entities to ontologies and assigns URIs, subsequently compiling a database containing all identified entities and their occurrences in news articles. The solution's strengths lie in its reliance on widely acknowledged databases for NER. Their work addressed texts in the French language exclusively. The investigation publicly evaluated its methodology through quantitative measures and a comparison with a Gold Standard, emphasizing transparency in its assessment.

5.1.3. Multilingual

Kertkeidkachorn and Ichise [36] constructed a multilingual framework to map predicates from NL text to KG triples, named *T2KG*. The methodology combined rule-based and similarity-based approaches and includes five steps: 1) entity mapping, linking entities in the text to corresponding entities in the KG; 2) coreference resolution, detecting chains of entities and pronouns referring to the same entity; 3) triple extraction, extracting relation triples from the text using open information extraction techniques; 4) triple integration, generating a text triple by combining results from entity mapping, coreference resolution, and triple extraction; and 5) predicate mapping, mapping a predicate of a text triple to a predefined predicate in other KGs. Their solution extracts information from textual documents and aims to integrate it into a KG. The goal was to incorporate new information into preexisting knowledge structures (*e.g.*, ontologies and KGs). The solution applies to various domains. The authors report achieving high precision, recall, and F1 scores, effectively mapping their generated triples to DBpedia's, and integrating significant new knowledge into the existing KG. The major drawback is the difficulty in mapping complex predicates. Risks to validation include potential errors in graph generation depending on the input data. Open challenges involve im-

[8] https://neo4j.com/ (As of Feb. 2024).
[9] https://www.wikipedia.org/ (As of Feb. 2024).
[10] https://www.geonames.org/ (As of Feb. 2024).

proving triple extraction accuracy, handling complex predicates, integrating multiple sources of information, and developing effective methods for assessing KG quality, completeness, correctness, and consistency, as well as error identification and correction.

5.2. Ontology and KG Enhancement

In this category, we investigate whether the study comprises the generation of triples from NL texts, aiming to either enhance the Ontology that describes an existing KG, *i.e.*, populating T-Box (*cf.* Section 5.2.1), or enhancing the instances portion of the knowledge representation, *i.e.*, populating A-Box (*cf.* Section 5.2.2). The Ontology and KG Enhancement are represented by the blue circle in Figure 8.

5.2.1. T-Box Population

The work by Sordo *et al.* [37] focused on extracting knowledge from unstructured text sources to generate structured data for music recommendations. The methodology involved identifying relevant entities in texts, extracting meaningful relationships between them, and connecting this knowledge to existing ontology classes and properties. The methodology of the knowledge graph construction comprised several key steps. It begins with text input preprocessing, segmenting the input text into sentences, and tokenizing. This work employed Named Entity Recognition (NER) using *DBpedia Spotlight* [38] to identify music-related entities, focusing on types like song, band, person, album, and music genre. Simultaneously, Dependency Parsing (DP) generates trees for sentences, aiming to find relationships between multi-word music-related entities. These processes were integrated in a subsequent step, combining the results of NER and DP by merging nodes in the dependency tree corresponding to recognized entities. The subsequent stages involved Relation Extraction (RE), in which relations between recognized music-related entities are extracted based on paths in the dependency trees. Empirical rules were introduced to filter out irrelevant relations linguistically. Finally, a graph representation was constructed, encapsulating the music-related entities as nodes and their relationships as edges. This graph, composed of five distinct entity types, provided a structured representation of the knowledge extracted from the input text, facilitating a comprehensive understanding of relationships within the music domain.

The domain of the work by Sordo *et al.* [37] was in the arts, particularly music. The application of their work relates to a real-world project, and the dataset used was publicly available. The evaluations include quantitative and qualitative analyses, user evaluations, and comparisons with a gold standard. The solution's strengths include using natural language for user recommendations, which enhances the user experience, and comparing the extracted knowledge with a gold standard to assess the quality of the relationships. Its weaknesses include low recall in extracting relationships, potentially leading to the loss of relevant information, and the need for a prior syntactic simplification step to handle potentially noisy relationships. Potential risks in validation include limited generalization due to a single dataset and the possibility of noisy relationships extracted from text variability. Open challenges relates to improving relationship extraction system recall and evaluating the method on larger, more diverse datasets for generalization assessment.

5.2.2. A-Box Population

The work by Dutta *et al.* [17] generated KGs and connected them to classes and properties of an existing ontology. The methodology employed in their study involved data clustering before mapping the relationships between phrases and clusters. The employed methodology concerns constructing a KG by converting Open Information Extraction (OIE) [30] facts into assertions within a target knowledge base (KB). The process involved four key components: Instance Matching (IM), Lookup (LU), Clustering (CL), and Property Mapping (PM) modules. The domain of this work referred to language, and although it was not used in a real-world setting, the dataset is publicly available. Their investigation described an evaluation protocol that includes quantitative assessments and comparisons with a gold standard. One of their solution's strengths is its focus on simplifying the mapping process for knowledge bases, with clustering being a valuable addition. No weaknesses were identified explicitly, but potential validation risks include dealing with identical phrases that might have different meanings. An open challenge is creating a T-box when certain relationships present in OIE do not exist in the target knowledge base.

5.3. Underlying Domain

In this category, we considered studies that are either applicable to a specific domain, *e.g.*, biomedical (*cf.* Section 5.3.1), or those which are domain-agnostic (*cf.* Section 5.3.2). The Domain Specificity is represented by the yellow circle in Figure 8.

5.3.1. Specific Transformation

Specific Transformation refers to methodologies for generating triples for a specific domain. The intersection emphasizes the alignment of domain-specific approaches with ontology construction (*cf.* Figure 8), highlighting how these methodologies often tailor the generation process to the intricacies of a particular domain.

Rossanez *et al.* [18] presented a specialized method for generating KGs specifically in the biomedical domain. The methodology comprised four main steps: preprocessing, triple extraction, ontology linking, and graph generation. The focus was on extracting knowledge from biomedical texts, connecting it to existing ontologies, and ultimately generating a KG. The domain was in the field of healthcare. While the application has not been used in a real-world application, the dataset used is public, and the evaluations included quantitative and qualitative assessments. The strengths of their solution include proposing a semi-automatic method for KG generation and relating it to existing ontologies. The method identifies primary and secondary relationships. Validation relies on a ground truth defined by medical experts, albeit not domain specialists. The results show promise, especially regarding the Jaccard coefficient. The weaknesses include the method not being entirely automatic, domain limitations due to the vast internal vocabulary of various medical subfields, and the relatively small number of samples (e.g., articles in the medical domain) evaluated by medical experts. Identified potential validation risks include the involvement of non-specialist medical experts, which might have introduced bias, and the use of abstracts rather than full-text articles, which hinders the assessment of the generalizability of the results. Open challenges included the development of an automatic approach for generating RDF triples more akin to those created by domain specialists, leveraging logical inferences to capture implicit relationships in texts, exploring the use of other KGs and ontologies to enrich the main set of triples, comparing knowledge graphs linked to different ontologies using UMLS CUIs [39], and involving more domain-specific experts to establish a baseline for comparison.

5.3.2. Generic Transformation

Generic Transformation represents methodologies that generate triples agnostic to a specific domain. This intersection shows how even domain-agnostic approaches often involve ontology construction (*cf.* Figure 8), showcasing the relevance of aligning the generated triples with a predefined ontology.

Lin *et al.* [40] introduced a method that generates triples across various domains. They aimed to extract knowledge from a text by generating KGs and their connection with existing ontology classes and properties. Their methodology included two key stages: knowledge extraction and knowledge linking. The first stage extracts information from documents, including entities and triples. The second stage constructs a graph from this data, and the linkage between entities and predicates is determined using similarity measures. Notably, their study was applied in real projects and leverages public datasets, making it suitable for generalized application. Their evaluation involved quantitative analysis and a comparison with a gold standard. Lin *et al.* [40] solution's strengths included the population of incomplete and outdated knowledge bases, extraction and organization of information from unstructured documents, utilization of a semantic graph approach, efficient integration of entities and relations, and effectiveness demonstration compared to other reference techniques. Their study also handled the coherent creation of new entities. However, the proposed method may face limitations in scenarios with very large or complex datasets, and its linking efficacy can depend on the quality of reference KB data. Challenges include dealing with ambiguity and polysemy in extracted information, potential biases in datasets, erroneous input data, and matching failures between entities and relations in reference datasets. Open challenges involve improving entity and relation linking in lengthy and complex documents, adapting the method for different domains and languages, exploring techniques for handling noisy and ambiguous data, and investigating the scalability of the method for larger datasets.

The article entitled "A Task-Agnostic Machine Learning Framework for Dynamic Knowledge Graphs", authored by Sendyk *et al.* [21], falls under the category of Generalized Transformations. Their study described the development of a generic framework capable of generating triples in any domain. Their methodology involved a series of steps, including web data extraction, training NLP models based on synthetic data, page classification, sentence segmentation, and graph generation. Notably, the framework was designed for a specific domain, allowing its application across various fields. While their defined process is semi-automatic, incorporating manual classification, the study addressed potential biases by implementing steps to avoid user biases during classification and introducing synthetic data to enhance the available dataset. The solution's strengths lie in its attempt to mitigate user biases and augment data availability through synthetic data. However, challenges include the need for a robust manual classification step and potential risks associated with biased classification and imbalanced base texts. Validation aspects included the assessment of the framework through quantitative measures and a case study, demonstrating its real-world applicability and emphasizing the need for well-defined manual classification steps in open challenges.

5.4. Transformers

Transformers [41] are neural network models that rely on the attention mechanism to draw global dependencies between their inputs and outputs. They are often applied to text-to-text applications, such as translation. In the literature, several studies employ such models to identify entities and relations from text to build triples.

This category represents methodologies that leverage technical approaches, specifically the Transformer architecture, for triple generation. This category intersects with Language Specific in Figure 8 because it emphasizes applying these techniques to handle linguistic variations in multilingual, English-specific, or non-English contexts.

The work by Xu *et al.* [42] described how they used BERT-based and Bootstrapping methods to construct a constantly evolving KG in the domain of Traditional Chinese Medicine (TCM). BERT is a Transformer-based deep learning model by Google used in various NLP tasks [31]. The authors state that the KGs in the TCM domain are static, not fully representing the evolving characteristics of the medicine domain. To overcome this problem, they proposed a methodology that generates a dynamic growth of the proposed KG. It starts by collecting data based on user input keywords and then employs the BERT-CRF [43] method to identify entities and Bootstrapping to identify relations and obtain structured data. Finally, the structured data was integrated into an existing KG. Bootstrapping is applied to extract the relationship between symptoms of diseases and their causes. The entity recognition and relation extraction result is merged into an existing KG. While the method allows the continuous and dynamic growth of the TCM KG through user interactions, its main difficulty lies in merging KGs, especially in handling equivalent entities and term ambiguity. Also, the authors [42] do not describe how to assess the qualitative aspects of the added information to the KG. In addition, the methodology was only applicable to English-based knowledge and the TCM domain [42].

The work by Fei *et al.* [44] relied on few-shot Relational Triple Extraction (RTE) to construct triples. The authors state that traditional triple-construction approaches are not aware of the semantics and coherence of the generated triples. The proposed methodology, Perspective-Transfer Network (PTN), included a multi-perspective approach to constructing a KG. It operates on episodes comprising support and query sets. The framework designed by Fei *et al.* [44] used three perspectives: Relation, Entity, and Triple. In the Relation Perspective, the query detects potential relations by marking entity pairs sentences. A binary classifier predicts if a relation exists among them. If a relation is identified, the Entity Perspective extracts entity spans in the query, combining relation and entity information into triples. The Triple Perspective then validates these triples, utilizing labeled query sentences and a binary classifier. Although it does not have a specific domain, the solution can be applied generally. It has not been used in a real project, but the dataset is public, and the evaluations involve quantitative analysis and a comparison with the gold standard. The main strength of their solution [44] is the ability to handle a few training examples, which is a common limitation in many NLP tasks. Additionally, the utilization of an efficient and scalable neural network architecture that can be trained on modern GPUs is explored in their study. Also, the solution is fully automatic and does not require human intervention. As a downside, the solution may be computationally intensive, requiring substantial hardware resources. Its performance can be sensitive to the quality of the input data, mainly affected by annotation errors and data noise [44].

Liu *et al.* [45] proposed an application named *Seq2RDF*, which uses the Transformer architecture to construct triples. The Seq2RDF was one of the first applications of Transformers to build KGs, in 2018. The methodology involves applying Transformers and generating embeddings for triple generation. It uses DBpedia [22] as input to train models, with the encoder processing NL sentences and the decoder producing triples in the <subject, predicate, object> format. Their solution [45] is generic and not limited to a single domain. It was applied to a real-world project, but their dataset is public. Evaluations included quantitative analysis and comparison with a gold standard. The authors state that the differential of *Seq2RDF* is its simplicity and efficiency for generating triples from NL texts. The downside of *Seq2RDF* remains that it can only generate a single triple per sentence.

5.5. Lexical Resources

External lexical resources, such as WordNet [46], PropBank [47], or VerbNet [48], for instance, have been explored to assist the generation of RDF triples. Examples of their application include identifying verbs and their parameters in sentences as candidates for subjects or objects. Resources like Yago [49], BabelNet [50], or SpaCy [51], are often employed for identifying named entities in texts. This category, represented by an intersection between Resource Utilization and Language Specificity (*cf.* Figure 8), emphasizes the reliance on linguistic resources in language-specific contexts, showcasing how the utilization of such resources plays a crucial role in these methodologies.

The work by Yan and Gao [52] fits this category. Their work used the Baidu Encyclopedia to aid in the KG construction. Their main objective was to generate KGs and connect them to existing ontology classes and properties, focusing on the domain of biology, specifically water. Their methodology involved information extraction, knowledge fusion, and knowledge processing. Information extraction encompasses extracting entities (concepts), attributes, and relationships between entities from the data source, forming the foundation for ontological knowledge representation. After acquiring new knowledge, the following step, called knowledge fusion, integrates such new knowledge to eliminate contradictions and ambiguities. The method was applied in a project and the used dataset is public. The conducted evaluation [52] was based on a case study. The solution's strengths lie in customizing the wrapper for extracting entity attributes and values from semi-structured water entry data in the encyclopedia. On the other hand, its weakness lies in the dependence on data sources, as the quality and quantity of data can impact the comprehensiveness of content extraction. The main validation risk is the lack of comparisons with other methods. Open challenges involve expanding the scope of extraction to acquire more knowledge and conducting further research to identify potentially better methods for KG construction in the water domain.

5.6. Link Creation

This category focuses on methodologies that create links to existing well-known Linked Open Data (LOD) datasets. LODs often adhere to standardized ontologies and vocabularies, such as RDF and OWL. This adherence ensures consistency and interoperability between different datasets, reducing ambiguity and enhancing the overall quality of information in the KG. Figure 8 presents the Link Creation as the purple circle.

The article titled "AliMe KG: Domain Knowledge Graph Construction and Application in E-commerce," by Li *et. al* [53] was categorized under Link Creation due to its primary focus on connecting newly created triples with other existing datasets. Their proposal involved key components, such as phrase mining, named entity recognition, relation extraction, and knowledge fusion. The authors described a semi-automated process for knowledge acquisition and validation, incorporating human annotation and feedback cycles to enhance the precision and completeness of the Knowledge Graph. Their study [53] applied in the e-commerce domain and showcases real-world applications of the AliMe KG in pre-sales conversation scenarios. The evaluation combines both quantitative and qualitative assessments. The solution's strengths include providing a systematic methodology with semi-automated processes for mining structured knowledge from natural language texts applicable to multiple languages and domains. Identified challenges include the significant need for human annotations and feedback cycles, which can be time-consuming and expensive. The solution also relies on external knowledge sources, introducing potential issues of reliability and currency, leading to biases or errors in the Knowledge Graph. Risks to validation involve biases in data sources and potential inaccuracies in external knowledge, impacting the representation and generalization of results. Open challenges include expanding the AliMe KG coverage in the Alibaba e-commerce platform and exploring its application in various domains beyond e-commerce.

6. Discussion

This section describes key findings (cf. Subsection 6.1) and common themes and revisits our research questions (cf. Subsection 6.2). We discuss open research challenges and potential venues for future investigations (cf. Subsection 6.3).

6.1. Findings and Limitations

The analysis of the publication years (Figure 3) reveals a growing trend, especially from 2019 to 2022. This trajectory suggests an increasing interest in and importance of the surveyed topic recently. It justifies the relevance of our study. We believe that the values for 2023 did not follow the trend because many accepted articles have not yet entered the scientific databases (Table 8). In 2024, data related to 2023 will be more consolidated. Given the increasing adoption of generative models for various NL processing tasks, including those related to KGs, we expect an acceleration in the growth in the coming years. Adopting the term RAG (Retrieval Augmented Generation) [54], aligns with our thinking, highlighting that KGs can and should be used with Large Language Models (LLMs) for information retrieval tasks.

The categorization of studies in this study served as a venue for understanding the diverse methodologies employed in generating RDF triples from NL texts and enhancing KGs. Each category addresses specific aspects, such as language specificity, ontology and KG enhancement, domain specificity, the use of Transformers, lexical resources, and link creation.

The clustering of articles across categories (Figure 4) provides a comprehensive overview of research efforts. These results align with the subset of 15 articles described in Section 5. Our findings highlight some facts that warrant reflection. Despite thousands of spoken and computationally represented languages, there is still the dominance of only one—English—as a language used for NL texts and the KGs resulting from text processing. A more in-depth study as future work is needed to identify how non-English languages use KGs generated in English for information representation.

Regarding transformations and domains categories of Section 5, we identified the proximity between specific single-domain transformations and generic transformations across multiple domains. We understand that creating domain-specific KGs can benefit companies, government agencies, or any other entities interested in using specific KGs, given that their maintenance involves domain experts to ensure information consistency and less maintenance. On the other hand, KGs spanning multiple domains may face maintenance and accelerated growth challenges due to the multiplicity of domains and, consequently, the continuous creation of RDF triples. Based on these assumptions, we believe that, for example, creating a specific and local KG about the health sector — the domain with the most articles based on our results (cf. Figure 5) — offers more benefits in terms of creation and maintenance than a domainless Knowledge Graph.

The two categories with less representation but of great relevance to this research are Multilingual and Link Creation. In our view, more solutions should somehow support the creation of links with other ontologies or existing KGs. The rationale is that such solutions would facilitate information reuse, connecting nodes from local KGs with larger and more established KGs. In this sense, they would help to contribute to the LOD movement, transforming a Web with some disconnected data islands into an interconnected archipelago of data [20].

The distribution of articles based on evaluation types (Figure 6) highlights the methodological diversity in assessing existing proposed solutions. Quantitative methods were the most widely used ones. This type of assessment can use a large set of texts and triples as input, which facilitates adapting the methodology and refactoring tests. With a large dataset, understand that quantitative analysis enables researchers to iteratively adjust their methodologies and improve testing procedures, ensuring a more comprehensive and informed evaluation of proposed solutions.

Figure 7 presents that most of the 150 articles present automatic or semi-automatic solutions, which, to some extent, streamlines the production of RDF triples and improvements in existing KGs. In this way, quantitative methods dealing with large amounts of data are necessary for an overall evaluation of the results of these solutions. However, qualitative methods are valuable allies that can assist in identifying semantically incoherent RDF triples. Even though the triples are generated in the <subject, predicate, object> pattern and can be connected to an existing KG, such triples may be redundant, repeated, or inconsistent with those in the KG. Qualitative methods can identify such issues. Despite the potential limitations in volume and speed associated with evaluation assessments using quantitative methods, we understand that combining both quantitative and qualitative approaches is crucial for generating more consistent and comprehensive evaluations. This is particularly important as it allows for a multi-faceted assessment of results from different perspectives.

Several proposed solutions used a gold standard to evaluate the generated triples. We did not find a single standard used by the 15 studies in the literature described in Section 5. Constructing a "generic gold standard" to evaluate the generation of Knowledge Graphs from texts would be a valuable asset, helping future solutions to be compared using a unified evaluation basis.

The surveyed articles across various categories exhibit several limitations that provide opportunities for future research and refinement of NL-based KG generation methodologies. First, a common limitation revolves around the challenge of domain specificity. Many studies, particularly those in the Specific Transformation category, struggle with the adaptability of their approaches to diverse domains. Biomedical applications, for instance, present specific requirements that may not be addressed adequately by generic NL-based KG generation methods. In addition, several articles lack extensive evaluation in real-world scenarios, relying on relatively small datasets or limited case studies. This limitation hinders a comprehensive understanding of the scalability and robustness of the proposed solutions, making it crucial for future work to address this gap, *i.e.*, validating solutions in diverse and realistic settings.

Second, while Transformer-based approaches showcase remarkable capabilities, as demonstrated in the Transformer category, they are not without limitations. Many studies employing transformer architectures focus on English-centric applications, raising concerns about the generalizability of these methods to non-English languages. The effectiveness of transformer models can be contingent on the availability of large and high-quality training datasets, limiting their applicability to less-resourced languages. Additionally, the computational intensity associated with transformers is a common challenge, potentially restricting their deployment in resource-constrained environments. These limitations demand research efforts to enhance the multilingual applicability, dataset diversity, and computational efficiency of transformer-based KG generation methods to foster broader advancements. Third, Figure 7 indicates that fewer articles explicitly mention real-world applications, suggesting a potential gap between research and practical applications. The availability of publicly accessible data remains limited, impacting the reproducibility and accessibility of the proposed studies.

6.2. Answer Summary of Research Questions

Our literature review enables us to answer the defined research questions in Table 2. Each research question was answered throughout this review. In the following, we provide a summary of them:

1. **RQ-01** – "What are the benefits and drawbacks of a method that generates RDF triples from texts?" as highlighted in Section 5, these approaches contribute to KG enrichment by extracting structured information from unstructured texts, fostering a more comprehensive understanding of various domains. These methods' versatility is evident across applications, including healthcare, e-commerce, and the arts. However, drawbacks include language dependency, which limits the applicability of some methods to specific languages, and potentially excludes valuable information from texts in other languages. Domain sensitivity is another challenge, as certain methods are tailored to specific domains, making them less adaptable across diverse knowledge domains. The dependence on external lexical resources introduces challenges related to the consistency, coverage, and real-time updates of those resources.
2. **RQ-02** – "Are there any patterns for texts used as input and for the triples generated as output?" we identified that methods preprocess texts to extract entities, relationships, and contextual information, leading to diverse patterns in generated triples. The Venn Diagram in Figure 8 showcases the intersection of different categories, emphasizing the variety in approaches and potential patterns. Generally, texts undergo preprocessing through filtering, entity extraction, and relation extraction, resulting in patterns like <*entity, relation, entity*>.
3. **RQ-03** – "What are the most used techniques? What are the most accurate ones?", we observed that most studies, especially the most recent ones, rely on Transformers techniques in either a step of the process or the overall method. Previous research relied mostly on NLP techniques combined with rule-based approaches to

identify and extract RDF triples from text [5][7]. The advent of transformers showed greater accuracy when considering those used in past research. Due to their ease of use, transformer-based solutions are becoming more popular in diversified applications [55].

4. **RQ-04** – "Are there fully automated approaches to generate knowledge graphs from text?", we identified that few studies present a completely manual method. Most studies employ a fully-automated method, *i.e.*, the KG is generated with no human intervention or a semi-automatic method. As such, human intervention is required to correct the outcomes of the automated processes or as part of the overall triple-generation procedure.
5. **RQ-05** – "What are the main applications that benefit from the text-to-triple approaches?", we observed that a portion of the studies in our survey are not currently employed in real-world applications. This suggests an existing gap between research advancements and practical implementation. Despite this, for the subset of studies that are applied in real-world scenarios, we verified a direct impact on information retrieval-based applications across diverse domains. Specifically, these applications span various sectors such as health, technology, education, and more. In the healthcare domain, for instance, text-to-triple approaches contribute to the construction of KGs that aid in medical research, diagnosis, and treatment recommendations. In the education domain, the application of text-to-triple techniques facilitates the creation of educational KGs, supporting adaptive learning systems and personalized content delivery.
6. **RQ-06** – "How do the methods explore the T-Box and A-Box in terms of text-to-triple generation?", we identified some relevant approaches. T-Box population methods enhance/enrich existing ontologies by extracting knowledge from unstructured texts and connecting it to ontology classes and properties. On the other hand, A-Box population methods focus on populating the instances portion of knowledge representation with relationships derived from texts. The challenge lies in balancing both aspects for a comprehensive and accurate KG construction. As discussed in Subsections 5.2.1 and 5.2.2, methods explore T-Box and A-Box differently, emphasizing the relevance of understanding and addressing ontology and instance-related aspects for effective text-to-triple generation.

6.3. Open Research Challenges

This subsection describes the challenges still open in the state of the art, arising from the analyses and findings described in Section 4 and Section 5. We add challenges identified by us (the authors) and not identified in the reviewed articles in our study.

Multilingual NL-Based KG Generation: Developing methodologies that can effectively generate RDF triples from NL texts in multiple languages remains a significant challenge. The inherent linguistic variations, syntactic structures, and semantic nuances across different languages pose obstacles. Moreover, the availability of large and diverse training datasets for less-resourced and less-spoken languages is limited, complicating the training of robust multilingual models. Addressing this challenge requires innovative approaches to handle linguistic diversity, improve cross-language generalization, and explore techniques for effective knowledge transfer between languages.

Real-World Applicability and Scalability: Many NL-based KG generation methods lack extensive validation in real-world scenarios, impacting their practical applicability and scalability. Validation methods applied to diverse domains and large-scale applications are challenging due to the complexity and variability of real-world data. Researchers face difficulties using comprehensive gold-standard datasets for various domains and ensuring their proposals can scale to handle large, dynamic datasets. Overcoming these challenges demands the development of evaluation frameworks that simulate real-world conditions and the exploration of scalable NL processing techniques suitable for diverse application contexts.

Evaluation Metrics for Quality and Completeness: The lack of standardized evaluation metrics to calculate the quality and completeness of generated RDF triples represents a challenge. Existing metrics often focus on quantitative aspects, such as precision and recall, but are insufficient to capture the generated knowledge's semantic accuracy, coherence, and relevance. Designing comprehensive evaluation metrics that consider both quantitative and qualitative aspects is challenging due to the subjective nature of semantic accuracy and the lack of well-defined benchmarks.

Handling Ambiguity and Polysemy: Ambiguity and polysemy in NL introduce complexities in accurately identifying entities and relationships. Resolving the ambiguity arising from multiple interpretations of words or phrases and disambiguating between polysemous meanings is challenging. Existing NL-based KG generation methods often struggle in contexts where entities have diverse meanings or where words can be used in multiple contexts. Mitigating this challenge requires the development of context-aware models, advanced disambiguation techniques, and strategies to incorporate contextual information for accurate entity and relationship identification.

Ethical and Bias: NL-based KG generation methods are susceptible to biases present in training data, which may perpetuate biases. Addressing ethical considerations and mitigating biases in the generated knowledge is a pressing challenge. The difficulty lies in identifying and addressing implicit biases in training data, ensuring fair representation of diverse perspectives, and establishing guidelines for responsible NL-based KG generation. Overcoming this challenge requires interdisciplinary collaboration and ethical frameworks to promote fairness in knowledge representation.

Dynamic Knowledge Graph Evolution: Ensuring the dynamic growth and temporal evolution of NL-based KGs, such as the Temporal Knowledge Graphs (TKGs) [56] to capture real-time changes in knowledge domains is complex. Many existing methods focus on static KGs, limiting their ability to adapt to emerging information. Addressing this challenge involves designing frameworks that allow continuous knowledge acquisition, integration, and evolution while considering the computational complexity and potential risks associated with real-time updates.

User Involvement: Incorporating user feedback, domain expert insights, and human-in-the-loop interactions in NL-based KG generation can be meaningful to the overall quality of the KG. Many solutions lack mechanisms for effectively involving users and domain experts in the generation process, leading to potential gaps in understanding contextual nuances. Overcoming this challenge requires the development of interactive NL-based KG generation interfaces, feedback loops, and collaborative frameworks that help users contribute domain-specific knowledge, validate generated triples, and enhance the overall quality of the knowledge representation.

7. Conclusion

A considerable portion of textual data remains unprocessed, representing substantial amount of information that holds the potential to provide actionable insights. The motivation behind this stems from the need to develop novel methodologies and software tools capable of transforming these large volumes of unstructured text into structured, computer-interpretable knowledge. Semantic Web technologies, particularly with the core use of RDF triples and Knowledge Graphs, offer an approach to organize this information. This article presented the outcome of a systematic literature review on studies dealing with the generation of RDF triples from unstructured NL texts, aiming to enhance existing KGs. We identified from the literature the most prominent approaches to the extraction of RDF triples from text, especially concerning such triples' inclusion in existing KGs. We provided a comprehensive overview of the domain pointing out the main challenges, as well as the limitations on the current state-of-the-art on such research. Our study systematically surveyed a diverse set of 150 articles from distinct scientific databases. From such, we identified 10 categories and presented a detailed description of the most key studies from those, resulting in 15 articles discussed in detail. We contributed by outlining opportunities for future research, *i.e.*, the identification of the categories, approaches, challenges, open research questions, and research gaps.

References

[1] V. Kríž, B. Hladká, M. Nečaský and T. Knap, Data extraction using NLP techniques and its transformation to linked data, in: *Human-Inspired Computing and Its Applications: 13th Mexican International Conference on Artificial Intelligence, MICAI 2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings, Part I 13*, Springer, 2014, pp. 113–124.

[2] R. Egger and E. Gokce, *Natural Language Processing (NLP): An Introduction*, in: *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, R. Egger, ed., Springer International Publishing, Cham, 2022, pp. 307–334. ISBN 978-3-030-88389-8. doi:10.1007/978-3-030-88389-8_15.

[3] K.S. Candan, H. Liu and R. Suvarna, Resource Description Framework: Metadata and Its Applications, *SIGKDD Explor. Newsl.* **3**(1) (2001), 6–19–. doi:10.1145/507533.507536.

[4] L. Ehrlinger and W. Wöß, Towards a Definition of Knowledge Graphs, in: *12th International Conference on Semantic Systems (SEMANTiCS2016)*, 2016, pp. 14–17. http://ceur-ws.org/Vol-1695/paper4.pdf.

[5] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, *Semantic Web* **8** (2017), 489–508.

[6] C. Xiong, R. Power and J. Callan, Explicit semantic ranking for academic search via knowledge graph embedding, in: *Proceedings of the 26th international conference on world wide web*, 2017, pp. 1271–1279.

[7] P. Schneider, T. Schopf, J. Vladika, M. Galkin, E. Simperl and F. Matthes, A decade of knowledge graphs in natural language processing: A survey, *arXiv preprint arXiv:2210.00105* (2022).

[8] X. Zou, A Survey on Application of Knowledge Graph, *Journal of Physics: Conference Series* **1487**(1) (2020), 012016. doi:10.1088/1742-6596/1487/1/012016.

[9] S. Ji, S. Pan, E. Cambria, P. Marttinen and P.S. Yu, A Survey on Knowledge Graphs: Representation, Acquisition, and Applications, *IEEE Transactions on Neural Networks and Learning Systems* **33**(2) (2022), 494–514. doi:10.1109/tnnls.2021.3070843. https://doi.org/10.1109%2Ftnnls.2021.3070843.

[10] H. Arnaout and S. Elbassuoni, Effective searching of RDF knowledge graphs, *Journal of Web Semantics* **48** (2018), 66–84. doi:10.1016/j.websem.2017.12.001. https://www.sciencedirect.com/science/article/pii/S1570826817300677.

[11] M. Yasunaga, H. Ren, A. Bosselut, P. Liang and J. Leskovec, QA-GNN: Reasoning with Language Models and Knowledge Graphs for Question Answering, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty and Y. Zhou, eds, Association for Computational Linguistics, Online, 2021, pp. 535–546. doi:10.18653/v1/2021.naacl-main.45. https://aclanthology.org/2021.naacl-main.45.

[12] F. Suchanek and G. Weikum, Knowledge harvesting from text and Web sources, in: *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 2013, pp. 1250–1253. doi:10.1109/ICDE.2013.6544916.

[13] A. Bytyçi, L. Ramosaj and E. Bytyçi, Review of automatic and semi-automatic creation of knowledge graphs from structured and unstructured data (2023).

[14] A. Mansouri, L.S. Affendey and A. Mamat, Named entity recognition approaches, *International Journal of Computer Science and Network Security* **8**(2) (2008), 339–344.

[15] Z. Nasar, S.W. Jaffry and M.K. Malik, Named Entity Recognition and Relation Extraction: State-of-the-Art, *ACM Comput. Surv.* **54**(1) (2021). doi:10.1145/3445965.

[16] T.R. Gruber, Toward principles for the design of ontologies used for knowledge sharing, *International Journal of Human-Computer Studies* **43**(5) (1995), 907–928. doi:https://doi.org/10.1006/ijhc.1995.1081.

[17] A. Dutta, C. Meilicke and H. Stuckenschmidt, Enriching Structured Knowledge with Open Information, in: *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2015, pp. 267–277–. ISBN 9781450334693. doi:10.1145/2736277.2741139.

[18] A. Rossanez, J.C. Dos Reis, R.d.S. Torres and H. de Ribaupierre, KGen: a knowledge graph generator from biomedical scientific literature, *BMC medical informatics and decision making* **20**(4) (2020), 1–24.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser and I. Polosukhin, Attention is All you Need, in: *Advances in Neural Information Processing Systems*, Vol. 30, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[20] A.G. Regino, J.C. dos Reis, R. Bonacin, A. Morshed and T. Sellis, Link maintenance for integrity in linked open data evolution: Literature survey and open challenges, *Semantic Web Journal* **12**(3) (2021), 517–541. doi:10.3233/SW-200398.

[21] N. Sendyk, C. Davies, T. Priscu, M. Sutherland, A. Madi, K. Dick, H. Khalil, A. Abu Alkheir and G. Wainer, A Task-Agnostic Machine Learning Framework for Dynamic Knowledge Graphs, in: *Proceedings of the 32nd Annual International Conference on Computer Science and Software Engineering*, 2022, pp. 22–31.

[22] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, Dbpedia: A nucleus for a web of open data, in: *international semantic web conference*, Springer, 2007, pp. 722–735.

[23] P.T.T. Thuy, N.D. Thuan, Y. Han, K. Park and Y.-K. Lee, RDB2RDF: completed transformation from relational database into RDF ontology, in: *Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication*, 2014, pp. 1–7.

[24] D. Moussallem, T.C. Ferreira, M. Zampieri, M.C. Cavalcanti, G.B. Xexéo, M. Neves and A.-C.N. Ngomo, RDF2PT: Generating Brazilian Portuguese Texts from RDF Data, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[25] D. Budgen and P. Brereton, Performing systematic literature reviews in software engineering, in: *Proceedings of the 28th international conference on Software engineering*, 2006, pp. 1051–1052.

[26] M. Masoud, B. Pereira, J. McCrae and P. Buitelaar, Automatic construction of knowledge graphs from text and structured data: A preliminary literature review, in: *3rd Conference on Language, Data and Knowledge (LDK 2021)*, Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

[27] F. Suchanek and G. Weikum, Knowledge harvesting from text and Web sources, in: *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, IEEE, 2013, pp. 1250–1253.

[28] E. Nismi Mol and M. Santosh Kumar, Review on knowledge extraction from text and scope in agriculture domain, *Artificial Intelligence Review* **56**(5) (2023), 4403–4445.

[29] W.E. Zhang and Q. Nguyen, Constructing covid-19 knowledge graph from a large corpus of scientific articles, in: *2021 IEEE International Conference on Big Knowledge (ICBK)*, IEEE, 2021, pp. 237–244.

[30] A. Fader, S. Soderland and O. Etzioni, Identifying Relations for Open Information Extraction, in: *Proceedings of the Conference of Empirical Methods in Natural Language Processing (EMNLP '11)*, Edinburgh, Scotland, UK, 2011.

[31] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *North American Chapter of the Association for Computational Linguistics*, 2019. https://api.semanticscholar.org/CorpusID:52967399.

[32] H. Yu, H. Li, D. Mao and Q. Cai, A relationship extraction method for domain knowledge graph construction, *World Wide Web* **23** (2020), 735–753.

[33] A.B. Rios-Alvarado, J.L. Martinez-Rodriguez, A.G. Garcia-Perez, T.Y. Guerrero-Melendez, I. Lopez-Arevalo and J.L. Gonzalez-Compean, Exploiting lexical patterns for knowledge graph construction from unstructured text in Spanish, *Complex & Intelligent Systems* **9**(2) (2023), 1281–1297.

[34] R. Stern and B. Sagot, Population of a knowledge base for news metadata from unstructured text and web data, in: *AKBC-WEKEX 2012-The Knowledge Extraction Workshop at NAACL-HLT 2012*, 2012.

[35] B. Sagot and R. Stern, Aleda, a free large-scale entity database for French, in: *LREC 2012 : eighth international conference on Language Resources and Evaluation*, Istanbul, Turkey, 2012, p. 4 pages. https://hal.science/hal-00699300.

[36] N. Kertkeidkachorn and R. Ichise, An Automatic Knowledge Graph Creation Framework from Natural Language Text, *IEICE Transactions on Information and Systems* **101**(1) (2018), 90–98.

[37] M. Sordo, S. Oramas and L. Espinosa-Anke, Extracting relations from unstructured text sources for music recommendation, in: *Natural Language Processing and Information Systems: 20th International Conference on Applications of Natural Language to Information Systems, NLDB 2015, Passau, Germany, June 17-19, 2015, Proceedings 20*, Springer, 2015, pp. 369–382.

[38] P.N. Mendes, M. Jakob, A. García-Silva and C. Bizer, DBpedia spotlight: shedding light on the web of documents, in: *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, Association for Computing Machinery, New York, NY, USA, 2011, pp. 1–8–. ISBN 9781450306218. doi:10.1145/2063518.2063519.

[39] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic acids research* **32 Database issue** (2004), D267–70.

[40] X. Lin, H. Li, H. Xin, Z. Li and L. Chen, KBPearl: A Knowledge Base Population System Supported by Joint Entity and Relation Linking, *Proc. VLDB Endow.* **13**(7) (2020), 1035–1049–. doi:10.14778/3384345.3384352.

[41] H. Balabin, C.T. Hoyt, C. Birkenbihl, B.M. Gyori, J. Bachman, A.T. Kodamullil, P.G. Plöger, M. Hofmann-Apitius and D. Domingo-Fernández, STonKGs: a sophisticated transformer trained on biomedical text and knowledge graphs, *Bioinformatics* **38**(6) (2022), 1648–1656.

[42] T. Xu, C. Guo, L. Du, J. Xu, P. Zhang, X. Feng and M. Li, A Method for Traditional Chinese Medicine Knowledge Graph Dynamic Construction, in: *Proceedings of the 5th International Conference on Big Data Technologies*, ICBDT '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 196–202–. ISBN 9781450396875. doi:10.1145/3565291.3565323.

[43] S. Wu and Y. He, Enriching Pre-trained Language Model with Entity Information for Relation Classification, 2019, pp. 2361–2364. ISBN 978-1-4503-6976-3. doi:10.1145/3357384.3358119.

[44] J. Fei, W. Zeng, X. Zhao, X. Li and W. Xiao, Few-Shot Relational Triple Extraction with Perspective Transfer Network, in: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 488–498–. ISBN 9781450392365. doi:10.1145/3511808.3557323.

[45] Y. Liu, T. Zhang, Z. Liang, H. Ji and D.L. McGuinness, Seq2rDF: An end-to-end application for deriving triples from natural language text, in: *CEUR Workshop Proceedings*, Vol. 2180, CEUR-WS, 2018.

[46] G.A. Miller, WordNet: a lexical database for English, *Communications of the ACM* **38**(11) (1995), 39–41.

[47] P. Kingsbury and M. Palmer, Propbank: the next level of treebank, in: *Proceedings of Treebanks and lexical Theories*, Vol. 3, Citeseer, 2003.

[48] M. Palmer, K.L. Kipper et al., VerbNet, *The Oxford Handbook of Cognitive Science* (2004).

[49] F.M. Suchanek, G. Kasneci and G. Weikum, Yago: a core of semantic knowledge, in: *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 697–706.

[50] R. Navigli and S.P. Ponzetto, BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, *Artificial intelligence* **193** (2012), 217–250.

[51] M. Honnibal and I. Montani, spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2017.

[52] J. Yan and K. Gao, Research and exploration on the construction method of knowledge graph of water field based on text, in: *2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE)*, IEEE, 2019, pp. 71–77.

[53] F.-L. Li, H. Chen, G. Xu, T. Qiu, F. Ji, J. Zhang and H. Chen, AliMeKG: Domain knowledge graph construction and application in e-commerce, in: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020, pp. 2581–2588.

[54] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in Neural Information Processing Systems* **33** (2020), 9459–9474.

[55] S. Casola, I. Lauriola and A. Lavelli, Pre-trained transformers: an empirical comparison, *Machine Learning with Applications* **9** (2022), 100334. doi:https://doi.org/10.1016/j.mlwa.2022.100334. https://www.sciencedirect.com/science/article/pii/S2666827022000445.

[56] A. Rossanez, J.C. dos Reis and R. da Silva Torres, Representing Scientific Literature Evolution via Temporal Knowledge Graphs, in: *6th Managing the Evolution and Preservation of the Data Web (MEPDaW) Workshop, International Semantic Web Conference (ISWC)*, 2020, pp. 33–42. http://ceur-ws.org/Vol-2821/paper5.pdf.

[57] Z. Wei, J. Su, Y. Wang, Y. Tian and Y. Chang, A Novel Cascade Binary Tagging Framework for Relational Triple Extraction, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter and J. Tetreault, eds, Association for Computational Linguistics, Online, 2020, pp. 1476–1488. doi:10.18653/v1/2020.acl-main.136. https://aclanthology.org/2020.acl-main.136.