big data and cognitive computing

MDPI

Article

An Automatic Generation of Heterogeneous Knowledge Graph for Global Disease Support: A Demonstration of a Cancer Use Case

Noura Maghawry [1,*], Samy Ghoniemy [1,*], Eman Shaaban [2] and Karim Emara [2]

[1] Faculty of Informatics and Computer Science, The British University in Egypt, El-Sherouk City 11837, Egypt
[2] Faculty of Computer and Information Sciences, Ain Shams University, Abbasya, Cairo 11517, Egypt
* Correspondence: noura.elmaghawry@bue.edu.eg (N.M.); samy.ghoniemy@bue.edu.eg (S.G.)

Abstract: Semantic data integration provides the ability to interrelate and analyze information from multiple heterogeneous resources. With the growing complexity of medical ontologies and the big data generated from different resources, there is a need for integrating medical ontologies and finding relationships between distinct concepts from different ontologies where these concepts have logical medical relationships. Standardized Medical Ontologies are explicit specifications of shared conceptualization, which provide predefined medical vocabulary that serves as a stable conceptual interface to medical data sources. Intelligent Healthcare systems such as disease prediction systems require a reliable knowledge base that is based on Standardized medical ontologies. Knowledge graphs have emerged as a powerful dynamic representation of a knowledge base. In this paper, a framework is proposed for automatic knowledge graph generation integrating two medical standardized ontologies- Human Disease Ontology (DO), and Symptom Ontology (SYMP) using a medical online website and encyclopedia. The framework and methodologies adopted for automatically generating this knowledge graph fully integrated the two standardized ontologies. The graph is dynamic, scalable, easily reproducible, reliable, and practically efficient. A subgraph for cancer terms is also extracted and studied for modeling and representing cancer diseases, their symptoms, prevention, and risk factors.

Keywords: medical ontologies; ontology integration; knowledge graph construction; entity linking; semantic data integration

check for updates

Citation: Maghawry, N.; Ghoniemy, S.; Shaaban, E.; Emara, K. An Automatic Generation of Heterogeneous Knowledge Graph for Global Disease Support: A Demonstration of a Cancer Use Case. *Big Data Cogn. Comput.* **2023**, *7*, 21. https://doi.org/10.3390/10.3390/bdcc7010021

Academic Editors: Ahmed Abdeen Hamed and Maciej Malawski

Received: 2 December 2022
Revised: 17 January 2023
Accepted: 18 January 2023
Published: 24 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Semantic data integration is the process of combining data from heterogeneous resources and consolidating it into meaningful and valuable information [1,2]. Recent years have witnessed a massive rise in the production of large amounts of heterogeneous medical data for healthcare systems from a variety of sources, such as patient records, lab results, user experiences on social networks, and wearable devices. The nature of this data is characterized by its volume and variety making it difficult for data processing techniques to handle and manage it effectively in healthcare information systems.

Healthcare information systems must employ innovative methods and techniques for handling and processing such big data to extract usable information and knowledge due to the volume, velocity, and range of issues and challenges that face the process of managing such data in healthcare. Many intelligent healthcare systems have recently demonstrated enthusiasm for utilizing semantic web technologies with healthcare data to transform data into useful knowledge and intelligence [3–7]. Relying on semantic-based techniques to manage heterogeneous data in the healthcare domain assists medical doctors and users in early disease prediction if any disease ailment is detected. Intelligent healthcare systems need to rely on trustful resources and integrate these resources into a reliable knowledge base for such systems. One of these resources is standardized medical ontologies which

provide reliable and standard representation of the concepts of medical terminologies and the relation between them [8]. Currently, although each disease is diagnosed by certain defined symptoms, there are no existing ontologies or knowledge bases that link diseases with their symptoms. Despite the existing suggested models, most of them generate knowledge bases manually or automatically by focusing on a limited number of diseases. Integrating diseases and symptom ontologies into a knowledge base provides a strong base for any intelligent healthcare system. One of the forms of a knowledge base is building a knowledge graph for combining structured and unstructured medical data. This helps in visualizing the data, giving it the ability to analyze and infer new rules for enriching the knowledge base [9,10]. The following subsections discuss the need for constructing a knowledge graph of diseases and their symptoms, linking the knowledge graph entities to standardized ontologies, and thus integrating two medical standardized ontologies into a knowledge graph.

1.1. Knowledge Graph Construction

Knowledge graphs (KGs) are used mainly to describe real-world entities and their interrelations organized in a graph. A knowledge graph (KG) is considered a dynamically growing semantic network of facts about things [11]. Given a knowledge graph, G consists of a set of nodes N representing the entities, and a set of edges E representing the relationships between different entities, the graph is denoted as G = <N, E>.

The process of building a knowledge graph involves data acquisition from structured and unstructured resources and extracting relationships that currently require manual professional intervention. Knowledge graphs have proven to provide efficient and effective solutions to conceptualize a healthcare domain and thus be used for several healthcare systems [12]. Existing intelligent healthcare systems such as disease prediction and healthcare recommender systems lack a reliable knowledge base that has the power to represent heterogeneous data resources, having the ability to dynamically grow as more data is provided which could be visualized whenever needed [13–18]. The knowledge graph is an efficient representation of such a knowledge base. Figure 1 illustrates the basic components of an intelligent healthcare system and its dependency on a knowledge base of integrated knowledge graphs from different heterogeneous resources. In the reactive scenario, the healthcare system involves a knowledge base gathered from medical facts and could provide the user only with static information. Whereas in the proactive scenario, the system would be able to predict disease and give alerts using an inference engine that depends on a dynamically generated and updated knowledge graph from various resources.

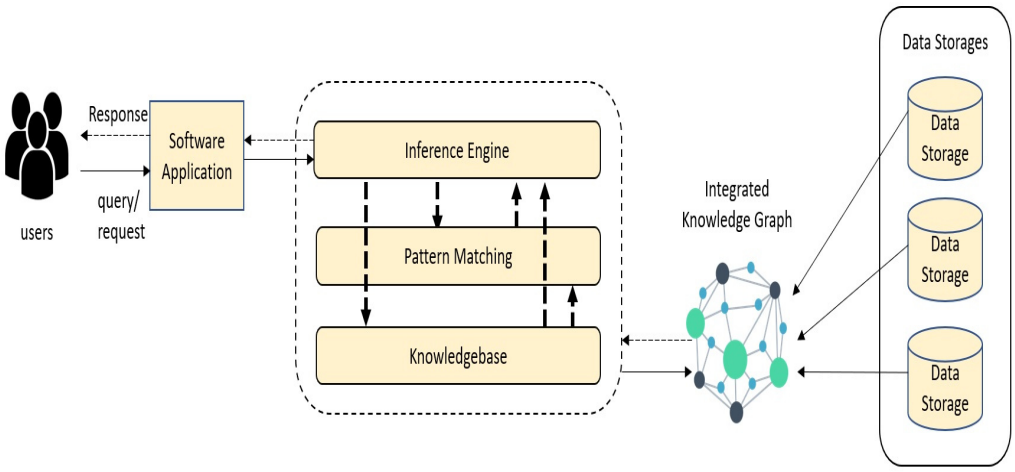

Figure 1. An Intelligent healthcare system.

The automatic construction of KGs from semi-structured and unstructured is still an open research problem, especially in the medical domain [19,20]. The challenge is to address medical domain texts given long and complex sentences that contain implicit or explicit

relations. Medical domain texts also involve abbreviations and different terminologies for the same medical concepts. The recognition of entities requires previous knowledge in the domain, turning it difficult for computational tools and natural language processing techniques (NLP) to perform such a task automatically [9,21].

1.2. Entity Linking and Integration to Standardized Ontologies

An ontology captures a set of concepts or facts, concepts' hierarchy, properties, and relationships between these concepts, especially the subsumption relationship. Thus, an ontology provides a rich, predefined vocabulary that serves as a stable conceptual interface to the data sources and is independent of the database schemas [22]. Domain Ontology facilitates knowledge sharing, with common vocabulary across independent software applications. A domain ontology also ensures the reusability and maintainability of domain-related information systems. KGs provide additional features than ontologies as they provide real-world instances and data. KGs add extra information and real-world experiences enriching the basic concepts extracted from a specific domain of interest [23,24]. Ontology entity linking is the process of identifying and associating entities mentioned to a unique concept identifier that best represents it in an ontology [25–28]. In our framework, the entities mentioned are the constructed KG entities. There are many challenges in named entity linking through an ontology, some of which are the variety and ambiguity problems of the named entities especially in the medical domain [29].

The medical domain involves many standardized online ontologies. The two standardized ontologies under study in this paper are the Human Disease Ontology (DO) [30,31], which is a project hosted at the Institute for Genome Sciences at the University of Maryland School of Medicine. The project was developed initially at Northwestern University to address the need for an ontology that covers the full spectrum of disease concepts. The ontology provides a unique disease ontology identifier for each disease, consisting of a prefix DOID followed by a number. For each DOID, there exists an Internationalized Resource Identifier (IRI) that provides information about the disease. The core relationship in this ontology is the subsumption relationship between disease concepts. However, the ontology does not provide the symptoms, causes, or any extra information except the diseases' concepts. The ontology also provides the synonyms terms for each disease concept and cross references with other medical ontologies. The second ontology is the Symptom Ontology (SYMP) [32] which is an ontology of disease symptoms, with symptoms encompassing perceived changes in function, sensations, or appearance reported by a patient indicative of a disease. The ontology provides a unique symptom ontology identifier for each symptom, which consists of a prefix SYMP followed by a number, and for each symptom there exists an IRI that provides information about the symptom. However, the ontology does not relate the symptoms to their respective diseases. The ontology provides the synonym terms for each symptom and cross references with other medical ontologies. Both DO and SYMP are ontologies provided by the OBO Foundry (Open Biological and Biomedical Ontology Foundry)—an organization for building and maintaining ontologies related to the life sciences, especially the biomedical field [33].

The two ontologies in the study are so specific containing axioms only. There is a need to automatically integrate these two ontologies for generating a knowledge base for any intelligent healthcare system. The integration of distinct concepts from different ontologies although having logical relationships between them is a challenging task that needs other trustful data sources. Online medical knowledge sources contain medicine and health-related information which is created and maintained by medical professionals. One of the most popular online encyclopedias for the medical domain nowadays is the MayoClinic website [34] which provides information about diseases, their causes, symptoms, risk, and prevention factors. The MayoClinic website provides comprehensive guides on hundreds of diseases and their conditions. It presents the same information provided by the Centers for Disease Control and Prevention (CDC) in more reachable, searchable, and user-friendly navigation options. One of the best features of this site is that

its data is always updated. It contains all the latest treatment options for patients based on scientific facts. The website provides healthcare information in different languages. MayoClinic ranks third in the most visited health websites ranking analysis of November 2022 [35]. It also ranks fourth in the most popular healthcare websites [36]. MayoClinic is an informative website for both professionals and medical care providers in different languages and based on reliable medical resources.

In this paper, a framework is proposed for constructing an integrated disease symptom knowledge graph based on two standardized medical ontologies—Human Disease Ontology (DO) and Symptoms Ontology (SYMP). The integration methodology relies on the reliable medical encyclopedia website, MayoClinic, for linking diseases from the DO with their symptoms in SYMP. The outcome is a disease symptom knowledge graph where graph nodes are concepts of diseases, symptoms, causes, prevention, and risk factors. The graph edges represent the relationships between the disease and its related concepts. The integrated knowledge graph aims to be used as a knowledge base for any intelligent healthcare system that could predict diseases or give alerts whenever any disease ailments are detected. The system can be used by non-medical users and medical professionals as it is linked to standard medical ontologies. The paper also focused on a demonstration of a cancer use case, as cancer nowadays is increasingly a global health issue. Cancer is the second leading cause of death worldwide, where 10 million deaths in 2020 were attributed to cancer [37]. The cancer subgraph from the integrated knowledge graph describes a complete and standard hierarchy of cancer classes and subclasses adopted from DO, and the integration provides information about cancer symptoms, causes, risk, and prevention factors. Healthcare systems built on this knowledge graph could alert normal users of any early detected ailments. In addition, it provides the user with a standard language while communicating with medical professionals based on standard symptoms and diseases, thus bridging the gap between normal users and medical professionals.

This paper is presented in five sections. Section 1 is an introduction, and Section 2 discusses related work. The framework and methodologies developed are described in Section 3. The results achieved and experiments conducted are described in Section 4. Section 5 evaluates the overall generated knowledge graph. Finally, Section 6 presents the conclusion and future work.

2. Related Work

The fully automated generation of domain-specific knowledge graphs from unstructured or semi-structured text is still an open research problem in different domains and has special challenges in the medical field when constructing graphs of diseases with their related symptoms [17,38,39]. The recognition of medical-related entities usually requires previous knowledge in the domain. Gathering such knowledge and identifying related entities is a challenging entity recognition task that requires professional human intervention due to the complexity encountered by the diversity of biomedical domains that are not interrelated to each other. Another challenge to consider is the relatedness of such knowledge graph entities to standardized concepts reaching a knowledge graph with linked entities to medical ontologies, causing an integration between medical ontologies that were not integrated before. We classify related work into two categories. First, research in constructing a domain ontology, designed for the medical domain related to diseases and their symptoms, and existing research related to graph linking with standard ontologies to enrich knowledge graphs generated with medical vocabularies. Second, the existing research in the techniques for recognizing and identifying diseases and symptoms. Named entity recognition is a challenging task in the field of natural language processing (NLP), as diseases and symptoms need professional training data sets different from traditional NLP pipelines. Diseases and symptoms terms are multi-word terms with overlapping entities, involving abbreviations and synonyms for each term. These challenges need a specific NLP pipeline to be adopted.

In the research field of medical KG construction, came the work of generating Alzheimer's ontology system that was built in [40] where there was human intervention from domain experts during the development process. A research was proposed in [41] where an Open Information Extraction system based on unsupervised learning without a prebuilt dataset obtains a knowledge graph from a vast amount of text documents about the disease COVID-19 and the dataset that was used to generate a knowledge graph focused only on COVID-19. A computational framework was designed in [42] for detecting drug combinations, by extracting drug names from biomedical publications and treatment sections of clinical trial records, and a network model is constructed representing the drug names and their associations. The previous work was extended in [43] through an algorithm for constructing a knowledge graph from drugs, genes, and diseases mentioned in the biomedical literature are presented with two querying algorithms for searching the knowledge graph by a single drug or a combination of drugs.

A human disease symptom network was built based on the use of large-scale medical bibliographic records collected from PubMed [44], to generate a symptom-based network of human diseases—Human Symptoms Disease Network (HSDN). The work depended on the stated diseases in PubMed abstracts which do not cover all diseases and their symptoms, and in terminologies that are not suitable for an intelligent healthcare system that addresses normal users. There is also a knowledge database of disease-symptom built based on associations generated by an automated method based on information in textual discharge summaries of patients at New York Presbyterian Hospital admitted where the associations were applied to 150 frequent diseases from the hospital records [45]. DS-Ontology (Disease-Symptom Ontology) was proposed in [46], the research generated an ontology by manually linking a few diseases with their symptoms as a step to integrate DO and SYMP ontologies. In [38] a system is proposed that uses domain knowledge for enriching word embeddings used by NLP deep learning model to serve cancer phenotyping. The research used Unified Medical Language System to enrich the word representations. An intelligent health diagnosis technique is proposed in [39] where its inference engine exploits automatic ontology generation for answering a query by gathering information from different biomedical ontologies. This approach is gathering information based on the query and thus proposed ontology generated called HDDO is an upper-level ontology for personal health diagnosis, and it is used to identify possible diagnoses from the user queries and his personal data. In the work of [47], the disease symptoms relations were extracted by clustering algorithm based on structural disease symptoms relations' mentions from different medical ontologies. An automatic approach for constructing a knowledge base of symptoms in Chinese was discussed in [48] where a graph was built from Chinese data sources. The research adopted by [49,50] is based on extracting disease symptom relationships by analyzing text dependency and term occurrences from medical data sources. The work in [51] constructed a graph DSKG which added another dimension that was important to consider which is the use of social user discussion boards for the construction of a disease symptom knowledge graph. This emphasizes that normal users using any healthcare system need the diseases and their symptoms to be represented in a familiar yet standard language. A lightweight approach for extracting disease-symptom relation using an annotation tool using a few records of medical texts for the automatic generation of a disease knowledge base was discussed in [52]. Table 1 compares the studies related directly to our work, the methodologies, and approaches followed by each, the data sources used, and the number of diseases under each study stating whether the work is based on linking to standard medical vocabulary. The table also shows whether or not the constructed graph from each study if available is based on entity linking with DO and SYMP ontologies, and whether or not the graph integrated with the ontologies.

Table 1. Comparison between different studies.

Study	Methodology Used	Data Resources	Number of Diseases Covered	Entity Mentions Linked to Medical Vocabulary	Linked and Integrated with DO	Linked and Integrated with SYMP	Graph Generated
HSDN [44]	The term frequency-inverse document frequency	PubMed abstracts	4219	Yes	No	No	Yes
DS-Ontology [46]	Manual	Medical Experts intervention	200	Yes	Yes	Yes	Yes
HDDO [39]	Term co-occurrence analysis	PubMed abstracts and MedlinePlus website	1000	Yes	Linked but not integrated	Linked but not integrated	Yes
DSKG [51]	Word co-occurrence pattern	Medical web-board resources	70	Yes	No	No	Yes
Okumura, T. et al. [52]	Manual	Medical Texts	20	Yes	No	No	No
Ruan, T. et al. [48]	Fusing data extracted from Chinese data sources	healthcare websites and Chinese encyclopedia sites	32,956	Yes	No	No	Yes
Oberkampf, H. et al. [47]	Clustering based on relation mentions in different ontologies	Structural relations mentions in different ontologies	Limited	Yes	Linked but not Integrated	Linked but not Integrated	Yes
Hassan, M. et al. [49]	Patten learning from the text dependency graph	Abstracts of PubMed for rare diseases	457 rare diseases	Yes	No	No	No
Rotemensch, M. et al. [50]	Classification algorithms	Electronic medical records	Limited	Yes	No	No	Yes

Our paper focuses on a cancer case study where some researchers stated the importance of generating knowledge graphs in oncology research [53]. Further research made efforts to build cancer-related knowledge graphs, but they are either building it manually as in the work of [54] or are building it based on genes studies [55] which will be useful for biomedical engineers and medical professionals but not for normal users.

The second category of studies relevant to the paper is Disease Named Entity Recognition (DNER) [56–61] which has been a hot research topic for several years, and several approaches have been using machine learning techniques. However, more recent approaches have adopted deep learning approaches as these approaches have proved to achieve better performance in the field of DNER such as the use of Bidirectional Encoder Representations from Transformers (BERT) models which is a transformer-based machine learning technique for natural language processing [62–64] and Bidirectional Long Short-Term Memory (BILSTM) networks with a Conditional Random Field (CRF) layer models [65]. Based on the surveyed related work, the automatic construction of a disease-symptom knowledge graph that also considers disease synonyms, causes, risk, and prevention factors in addition to integrating it with standardized ontologies is still an open research issue. Our framework is one step towards constructing a disease-symptom knowledge graph integrated with standardized ontologies, based on trustful online medical resources and DNER models.

3. Proposed Ontology-Based Integration Framework

The framework adopts a methodology for automatically generating standardized-based medical knowledge graphs beneficial for any intelligent expert advisor healthcare system. The framework depends on four reliable resources for Medical and healthcare information. The first resource is the MayoClinic website which uses a scientific encyclopedia for diseases created by a nonprofit American academic medical center focused on integrated healthcare, education, and research. The second resource is the standard Human Disease Ontology (DO), an ontology of disease concepts. The third resource is the Symptom Ontology (SYMP) which is an ontology of symptoms, with symptoms encompassing perceived changes in function, sensations, or appearance reported by a patient indicative of a disease. Both DO and SYMP are ontologies provided by the OBO Foundry—an organization for building and maintaining ontologies related to the life sciences, especially the biomedical field. Taking into consideration that DO and SYMP ontologies are separate ontologies and are not linked. The fourth resource is the UMLS Meta thesaurus [66] which is a large biomedical thesaurus that is organized by concept or meaning and links synonymous names from over 200 different source vocabularies. The Meta thesaurus also identifies useful relationships between concepts and preserves the meanings, concept names, and relationships from each vocabulary. The UMLS concepts are needed as a reliable source for concept mapping. The framework relies on integrating the two ontologies on the medical facts provided in the online encyclopedia MayoClinic website.

The developed framework for generating the standardized-based knowledge graph is composed of three phases. First, the MayoClinic scraper phase where information is gathered about diseases, symptoms, causes, prevention, and risk factors from the website, thus the data is provided in an unstructured or semi-structured format. The second phase represents concept and graph extraction from Ontologies where concepts are extracted from the standardized disease and symptom ontologies and processed, and graphs representing each standardized ontology are extracted. The third phase is the Entity linking and Integration phase where the entity linking algorithm is adopted to link nodes from the online-based knowledge graph generated to the nodes from the standardized ontologies for constructing the integrated standardized-based medical knowledge graph. The framework for generating the integrated knowledge graph showing the three phases is shown in Figure 2.

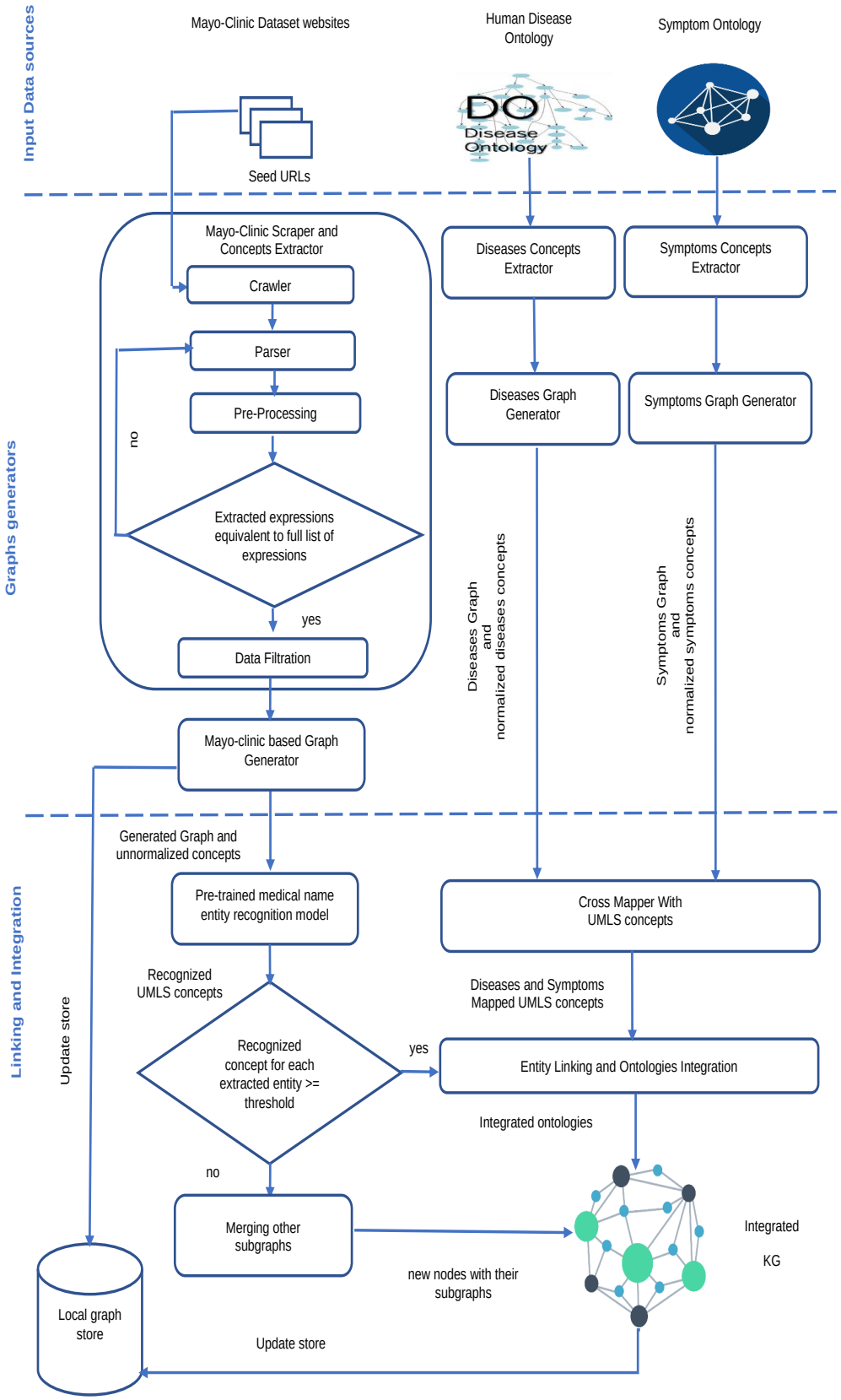

Figure 2. Framework for generating standardized-based integrated medical knowledge graph.

3.1. Mayoclinic Scraper

The Mayoclinic scraper initially starts by feeding seed URLs to the crawler representing the MayoClinic web page for diseases and conditions. The crawler extracts the pages of all diseases on the website starting alphabetically with diseases from the letter 'A' to the letter 'Z'. The parser parses the relevant information from pattern-matching techniques and regular expressions. The relevant information of interest on the page is the title of the disease, the list of its symptoms, the disease causes, prevention, and risk factors.

Text pre-processing techniques are applied during the parsing stage where all parsed information is pre-processed by removing punctuation, brackets, apostrophe-s, and s-apostrophe. This step is checked and compared with the website information until an acceptable accuracy level of parsed data is achieved. After the data filtration step, each disease, symptom, risk factor, cause, and prevention factor is given a unique number so that each item is represented once in the created online-based knowledge database. The data filtered is used for creating and extracting relationships, then generating Resource Description Framework RDF triples based on the core RDF for the knowledge graph showing four relationships as illustrated in Figure 3. The output of this stage is an online-based knowledge graph composed of triples based on the content parsed and processed from the MayoClinic encyclopedia. The graph resulted in 9,387 nodes of 5 labels—'Disease', 'Symptom', 'Cause', 'PreventionFactor', and 'RiskFactor'.

The graph involves 11539 relationships of 4 distinct types—'has_symptom', 'caused_by', 'prevented _by', and 'has _risk'. Figure 4 shows a subgraph of the disease 'lung cancer' from the generated knowledge graph with all its relationships with other nodes representing lung cancer disease symptoms, causes, risk factors, and prevention factors as stated on the MayoClinic website. The output of this stage is an online-based knowledge graph composed of triples based on the content parsed and processed from the MayoClinic encyclopedia. The graph resulted in 9387 nodes of 5 labels—'Disease', 'Symptom', 'Cause', 'PreventionFactor', and 'RiskFactor'.

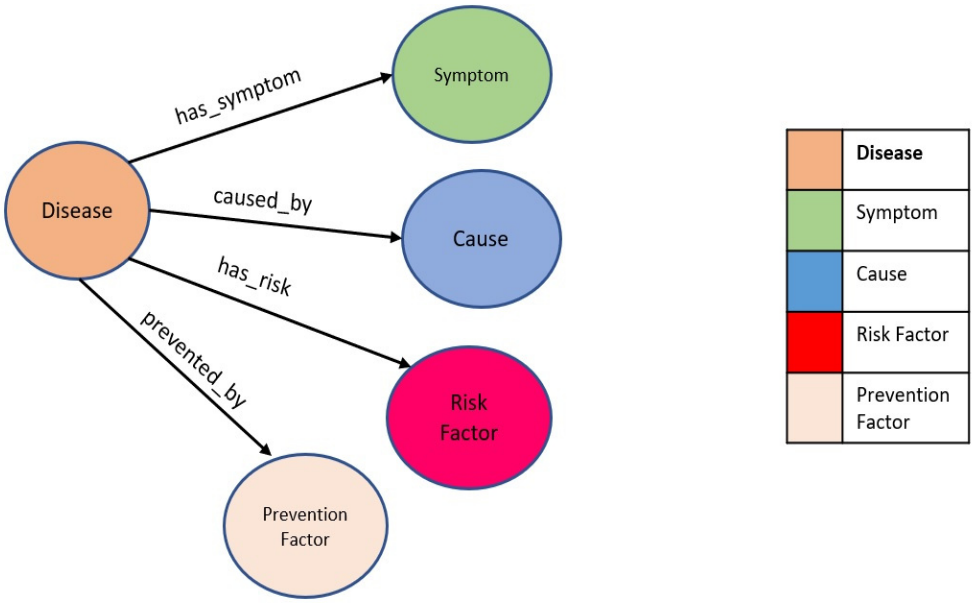

Figure 3. Phase 1 core RDF showing four relationships.



Figure 4. A subgraph of lung cancer disease node with all its four relationships with other nodes.

3.2. Concept and Graph Extraction from Ontologies

The second phase starts by extracting concepts from the DO and SYMP ontologies. Each concept in the ontology has its own properties involving the concept name with its other synonyms, the concept description, the unique identifier, the internationalized resource identifier (IRI) along with the cross-references to other medical ontologies. The most important relationship in these ontologies is the subsumption relationship, also known as a "hyponym-hypernym relationship". The subsumption relationship in semantic networks states that a concept is a subconcept of another concept.

Text pre-processing is then applied to the concepts' names and their synonyms where the list of synonyms for each disease or symptom is split and treated as a separate concept giving it the same unique identifier as their original concept. Diseases names and synonyms are preprocessed by removing punctuation, brackets, apostrophe-s, and s-apostrophe. Each of the disease and symptom concepts is a multi-term expression. Each of the disease ontology and Symptom ontology is extracted into a graph to be ready for the integration phase. Each node has data properties of name, unique identifier, and the concept IRI. Figure 5 shows a subgraph from the diseases graph focusing on the disease 'cancer', it shows its relationship with its superclass' disease of cellular proliferation and expands the subgraph of one of its subclasses 'organ system cancer'. The figure shows the subgraph corresponding subtree in the disease ontology. Figure 6 shows a subgraph from the symptoms graph focusing on symptom 'pain', it shows its relationship with its superclass 'sensation perception' and expands the subgraph showing its subclasses of specific types of pains. The figure shows the subgraph corresponding subtree in the symptom ontology. The resulting diseases' knowledge graph involved 13,381 nodes with 10,961 relationships. However, the symptoms knowledge graph resulted in 1004 nodes and 897 relationships.

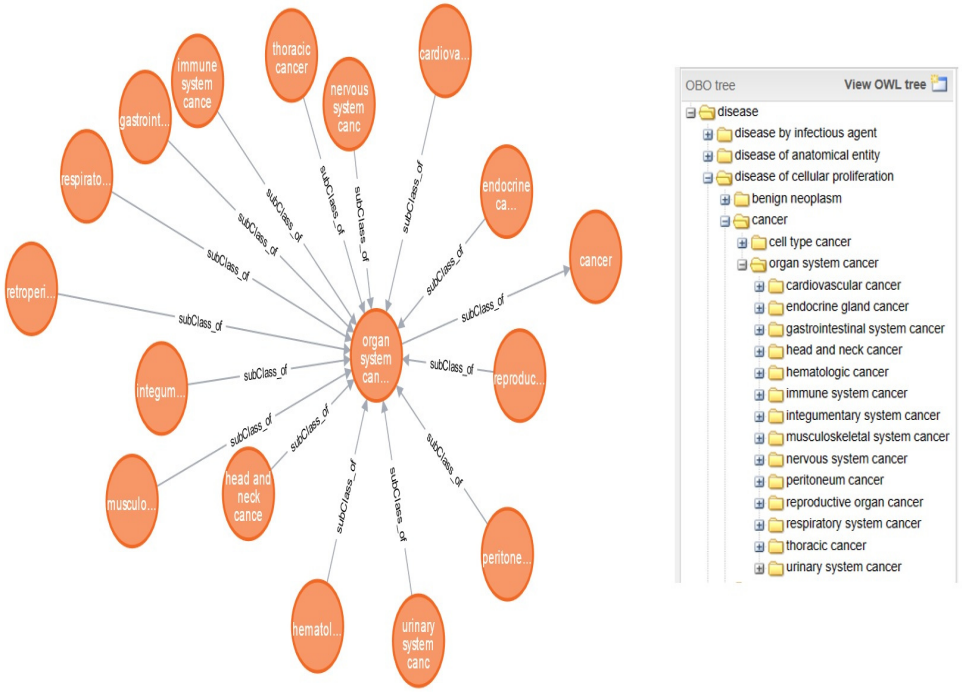

Figure 5. A subgraph of 'organ system cancer' concept with its superclass, subclasses, and the corresponding DO sub tree.

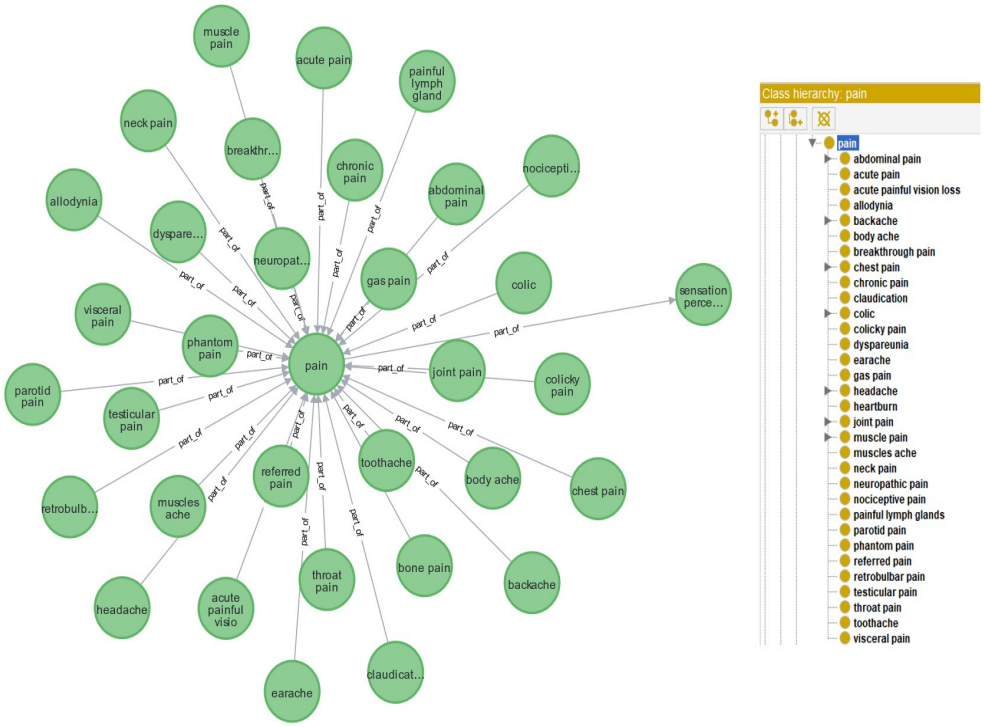

Figure 6. A subgraph of 'pain' symptom with its superclass, subclasses, and the corresponding SYMP subtree.

3.3. Entity Linking and Integration

During this phase, the three graphs—MayoClinic online-based knowledge graph, diseases graph, and symptoms graph should be prepared for the third phase which is the

Integration phase. Figure 7 shows the Phase 3 pipeline. Graph Alignment methodology needs to be applied, the objective of the graph aligner is to align two networks G and H by producing an alignment that consists of a set of pairs (x,y), where x is a node in G and y is a node in H. Graph Alignment methodology has been applied to works on property name on nodes from disease graph to be aligned with similar nodes from the online-knowledge graph. The aligner was also executed for the similar nodes of the symptom graph compared with the online-knowledge graph nodes. The similarity is measured by an exact phrase matcher for the first iteration, if the node x from graph G is similar to node y from graph H, both nodes are merged into one node z inheriting all its edges from both graphs. The merged node has a degree equal to the sum of the degree of node x and node y where deg(z) = deg(x) + deg(y).

The merged nodes have the properties from either the disease graph or symptom graph, including the property IRI. Thus, these merged nodes are linked to the standard ontologies and their internationalized resource identifier. The merged node has an additional property 'score' having a value of '1' indicating the certainty of semantic relatedness of the node to the standardized concept. Figure 8 shows the 'lung cancer' disease from the online-knowledge graph integrated with the disease graph, the merged node has the relationships from the disease graph (the superclass and its subclasses) together with the relationships from the online-knowledge graph 'has_symptom', 'caused_by', 'prevented_by', and 'has_risk' relationships. The figure also shows the node properties of the 'lung cancer' node having a property score of 1, IRI, and unique identifier of the disease 'lung cancer'. Figure 9 shows the node properties of one of its symptoms 'chest pain' node having a property score of 1, IRI, and a unique identifier of the symptom 'chest pain'.

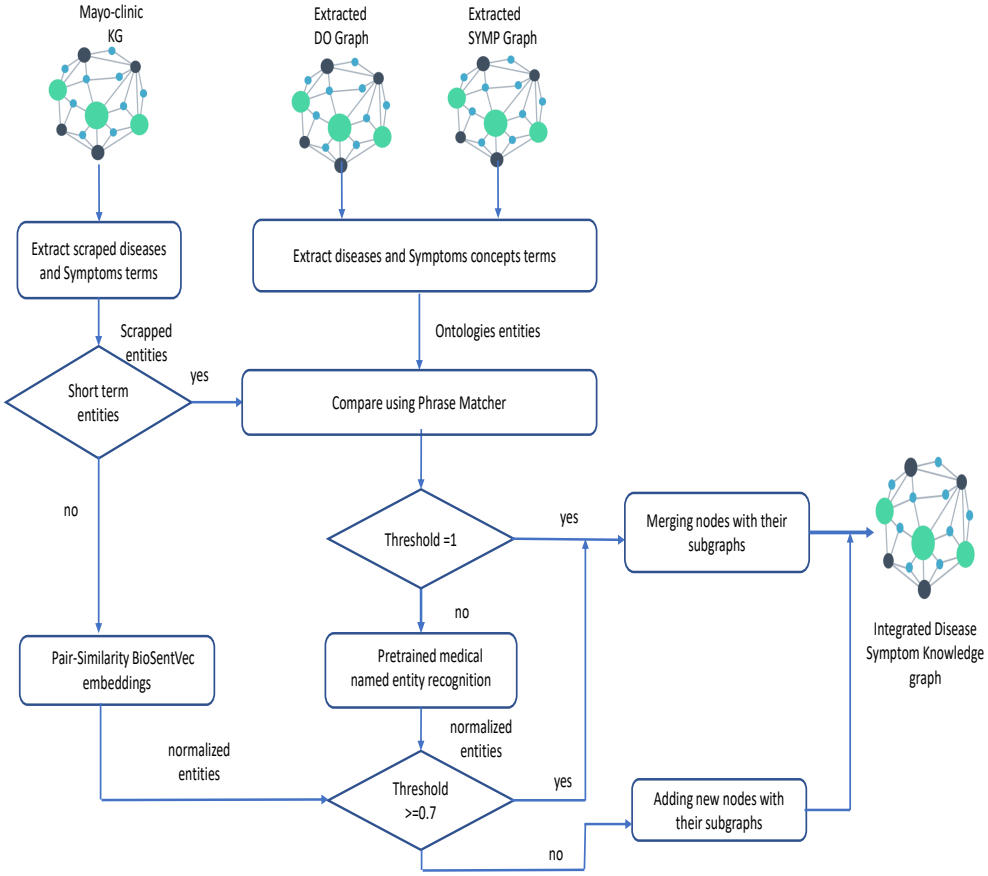

Figure 7. Entity linking and Integration pipeline.

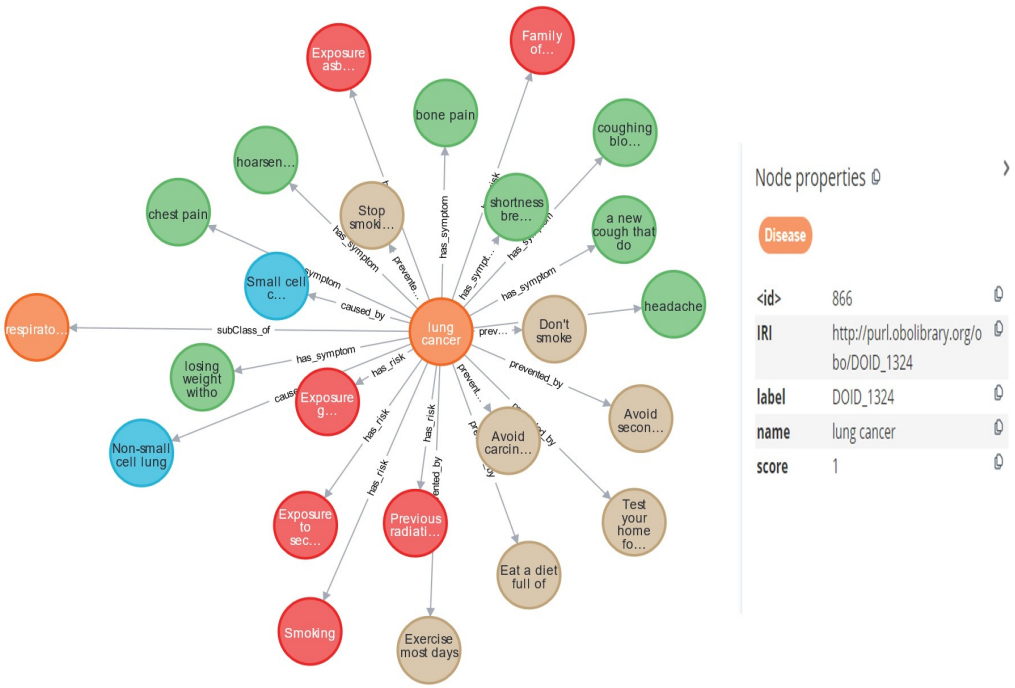

Figure 8. The subgraph of the integrated node 'lung cancer' with all its relationships and its node properties.

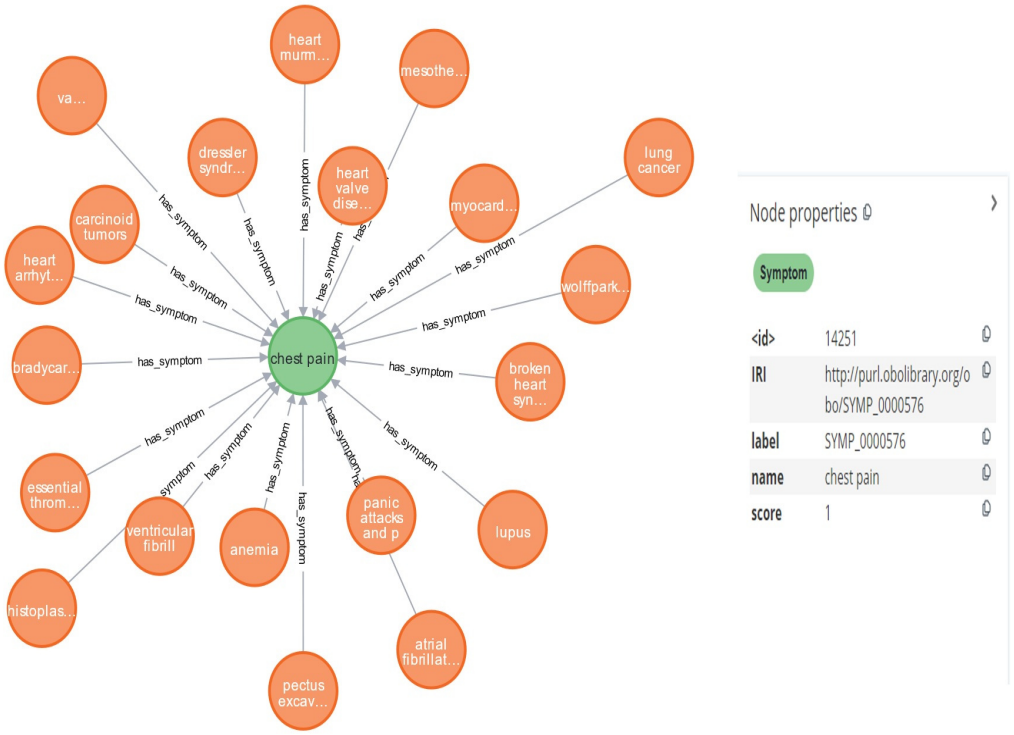

Figure 9. Properties of integrated node 'chest pain' as one of the symptoms of 'lung cancer' and other diseases.

The Phrase matcher linked only entities with an exact match with the disease name or the symptom name. To improve the entity linking algorithm, the unlinked disease, and symptom names must be introduced to the natural language processing pipeline. The problem of overlapping concepts arises when disease names such as "lung cancer" and "breast cancer", if both passed by the traditional NLP pipeline, only the word "cancer" in

both cases will be considered and not the whole term. This is because traditional NLP pipelines are trained on English corpus not medical corpus and cannot accept overlapping words. There is also a need to be able to recognize diseases based on their concept with their unique identifier—also known as concept unique identifier, for example, terms such as 'breast tumor' and 'mammary cancer' should be identified as 'breast cancer' as they are synonyms for this disease. Disease entity name recognition is a field of interest today, automatic recognition of disease mentions is a challenging task due to the variations and ambiguities in disease names. In the following subsections, the details of applying entity name recognition on both diseases and symptoms on MayoClinic integrated KG are discussed to improve the number of linked entities with the nodes inherited from disease and symptom graphs.

3.3.1. Disease and Symptom Named Entity Recognition

Pre-trained models are using deep learning techniques such as Bidirectional Encoder Representations from Transformers (BERT) which is a transformer-based machine learning technique for natural language processing [62–64], another technique is using the bidirectional long short-term memory (BILSTM) networks with a Conditional Random Field (CRF) layer [65]. Currently, these are the best models for NLP systems that provide reliable annotating and mapping of the text of biomedical and medical terms and documents to UMLS concepts, which is a comprehensive resource of medically relevant concepts and relationships. The pretrained model adopted is based on BERT [67], as BERT-based models have proved to have better precision, recall, and F1 score in the biomedical domain, the BERT-based models are now considered one of the benchmarks for biomedical name entity recognition [68,69]. The model annotates the text using the concept unique identifier (CUI) of UMLS giving a similarity score.

The scraped entities for both diseases and short sentences representing symptoms from the MayoClinic knowledge graph are input to the Pre-trained NLP model and similarities are calculated as shown in Equation (1). Given two diseases x and y with their vectors dx and dy, n is the number of terms in a multiterm concept and i is the term iterator, Equation (1) shows the cosine similarity between the two diseases x and y.

$$\cos(d_x, d_y) = \frac{\sum_{i=1}^{n} d_{x,i} d_{y,i}}{\sqrt{\sum_{i=1}^{n} d_{x,i}^2} \sqrt{\sum_{i=1}^{n} d_{y,i}^2}} \quad (1)$$

The cosine similarity ranges from 0 (no shared terms) to 1 (identical concepts). The annotated concepts based on UMLS are cross-mapped with concepts of diseases and symptoms available in *xref* of the standardized ontologies, other annotated concepts not related to diseases and symptoms are ignored. Based on the similarities calculated, scraped diseases and short symptom sentences are normalized and merged with their corresponding nodes in the integrated graph and the nodes' scores properties are updated with the code similarity score value. According to our algorithm, a threshold greater than or equal to 0.7 is adopted.

3.3.2. Symptom Named Entity Recognition for Long Sentences

Some scraped symptoms are represented as long sentences or paragraphs on the MayoClinic website, these sentences need another algorithm to be adopted. In this paper, an algorithm for document-to-vector embeddings is adopted called BioSentVec [70]. BioSentVec embeddings are trained for calculating clinical sentence pair similarity. Long symptom sentences are input to the algorithm and comparing the pair similarities between the sentence and all symptoms' concepts from the standardized symptom ontology, the best match is considered, and the threshold of similarity considered is 0.7. The long sentences are merged and normalized with the symptom nodes to form an integrated graph and the

score is updated with the score of similarity. If we still have scraped entities that are not merged, they are added to the integrated graph with a score equal to 0.

By the end of this phase, an integrated disease-symptom knowledge graph is reached with 594 nodes linked with standardized disease ontology out of 994 scraped disease nodes and 661 nodes linked with standardized symptom ontology out of 4427 symptom nodes scraped.

4. Experimental Results

The knowledge graph generated succeeded in integrating both DO ontology and SYMP ontologies with all their concepts hierarchy based on trustful medical facts. This integrated knowledge graph was not addressed in any of the related research before. Each disease node in the graph is identified with the disease's unique identifier number, thus the disease name and its synonyms are all represented as one node with no redundancy. This has an impact on the graph size, instead of having more than 8000 nodes representing diseases in the DO ontology, they are represented in 14,010 nodes. The following discussion focuses on the graph nodes that are linked with their concepts from the standardized ontologies. Table 2 shows the number of nodes linked after each technique was applied. Figure 10 shows the percentage of scraped diseases linked with the standardized disease ontology using different methodologies. Figure 11 shows the percentage of scraped symptoms linked with the standardized symptom ontology using different adopted methodologies.

Further experiments are also conducted to evaluate the percentages of diseases and symptoms interlinked from the overall linked knowledge graph based on dictionary-based models (Phrase-Matcher) versus BERT models with different thresholds, and the BILSTM with CRF-based models with different thresholds. Figure 12 shows the comparison for diseases, whereas Figure 13 shows the comparison for symptoms. The results showed that BERT-based models achieved better results in nodes' linking, and thus it was chosen for generating the proposed knowledge graph.

Table 2. Number of interlinked nodes after each technique is applied.

Methodology	Disease Nodes	Symptom Nodes
Phrase Matcher	410	108
PreTrained NER with Threshold ≥ 0.8	580	588
PreTrained NER with Threshold ≥ 0.7	594	588
BioSentVec on Symptoms	594	661

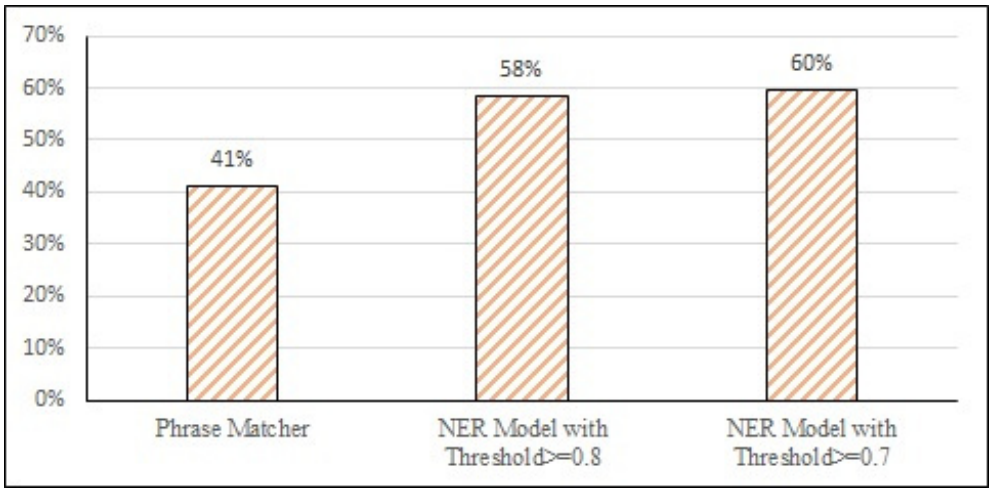

Figure 10. Chart showing the percentage of scraped diseases interlinked with disease ontology using different adopted methodologies.

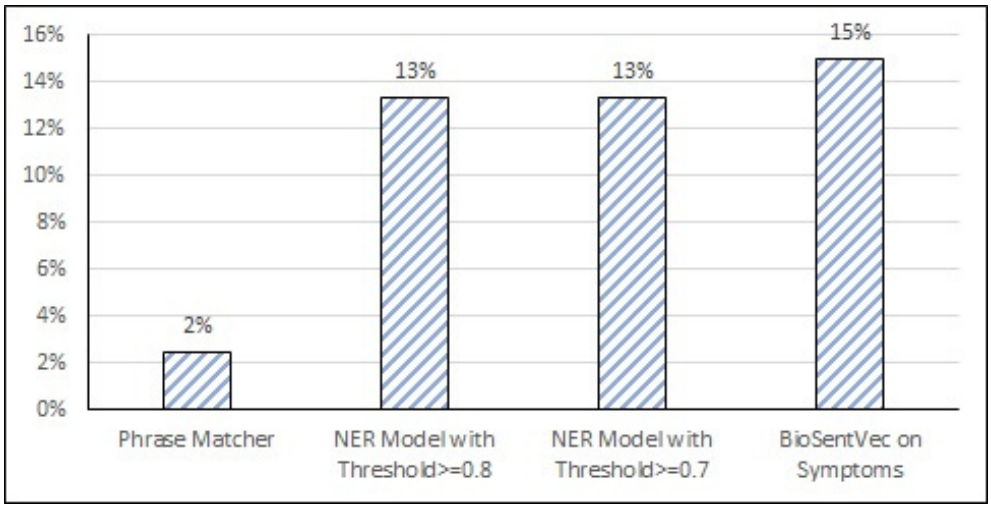

Figure 11. Chart showing the percentage of scraped symptoms interlinked with symptom ontology using different adopted methodologies.

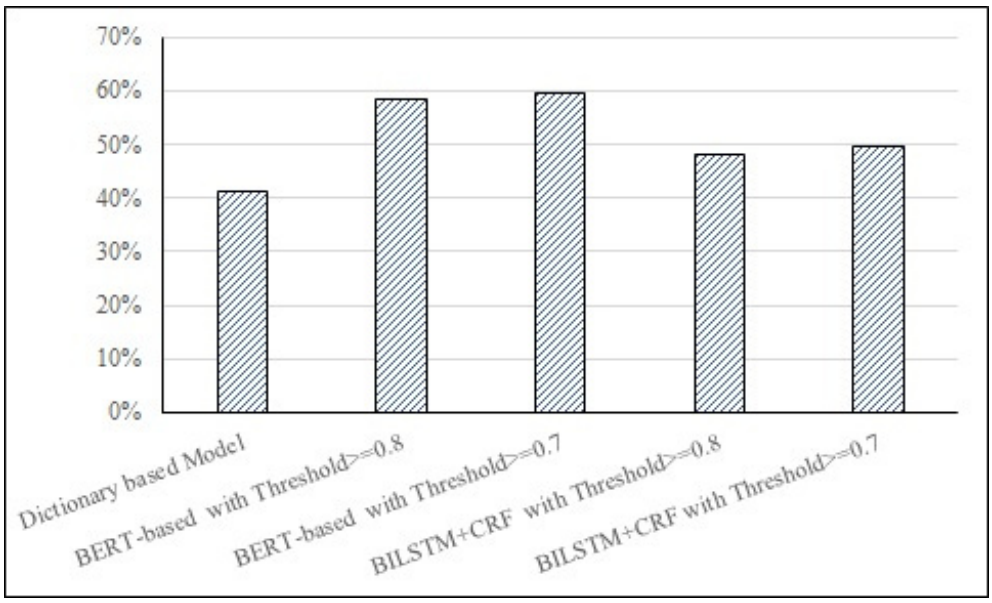

Figure 12. Chart showing the impact of the different entity-linked models on the percentage of scraped diseases interlinked with disease ontology.

For evaluation, the knowledge database of disease-symptom associations [45] is used where 150 frequent diseases with their symptoms from a hospital are represented and mentioned using UMLS concepts. The database has 1865 records stating diseases and their symptoms from patients' records. Figure 14 shows a chart stating the percentage of diseases and symptoms from the database that is linked with the standardized ontologies using different methodologies. The records in the database were tested using cypher queries for all records checking for the diseases-symptom relationship in the integrated graph. Table 3 shows the precision, recall, and F1 score values using different methodologies.

For the cancer case study, the subgraph representing the 'organ cancer terms' have been queried using cypher query language to discover the cancer terms and their symptoms. The results showed that 263 'has_symptom' relations have been created integrating cancer terms with their symptoms, from which 62 symptoms are linked to the SYMP ontology with a score of 1. The number of new relationships created representing 'caused_by' relationship is 162 edges. New relationships representing 'prevented_by' relationship is 66 edges, whereas 184 new edges are created representing 'has_risk' relationships between cancer

terms and their risk factors. The organ cancer subgraph could be extracted to serve as a knowledge graph for healthcare systems dedicated to cancer diseases.

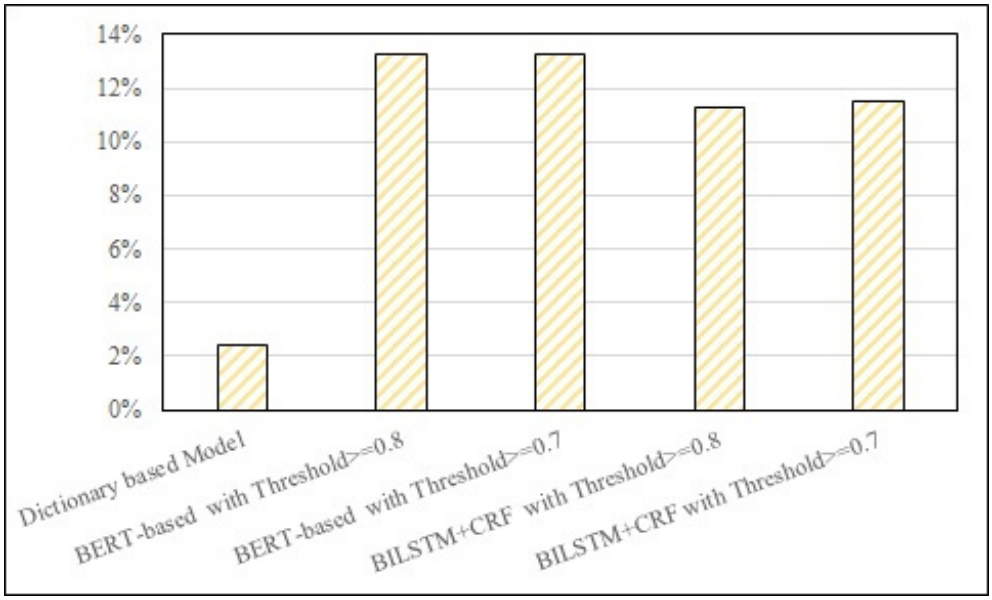

Figure 13. Chart showing the impact of the different entity-linked models on the percentage of scraped symptoms interlinked with symptoms ontology.

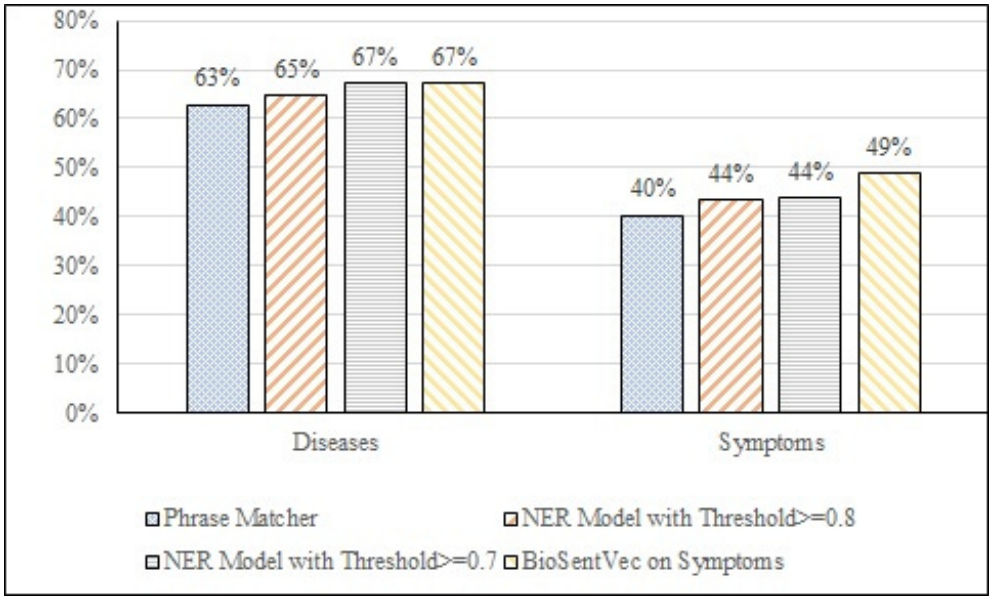

Figure 14. Chart showing the percentage of diseases and symptoms from the database that is interlinked with the standardized ontologies using different adopted methodologies.

Table 3. Precision, Recall and F1 Score values for different methodologies.

Methodology	Precision	Recall	F1 Score
Phrase Matcher	0.80	0.41	0.54
PreTrained NER with Threshold $\geq$ 0.8	0.85	0.51	0.64
PreTrained NER with Threshold $\geq$ 0.7	0.85	0.52	0.65
BioSentVec on Symptoms	0.88	0.54	0.67

Taking all together the experimental results, evaluation, and discussions; the proposed framework succeeded in fully integrating heterogeneous data sources and successfully

built a standard-based knowledge graph suitable as a knowledge base for healthcare applications with query results having a precision of 0.88, recall of 0.54 and F1 score of 0.67. The knowledge graph nodes covered all diseases and symptoms from the standard ontologies, and fully integrate two standardized ontologies. Taking into consideration that synonyms are represented by the same node, thus avoiding redundancy and unnecessary growth of the graph size. The smaller the graph size will have a positive impact on reducing the response time for any healthcare querying system. Each of the linked graph nodes has a unique identifier and IRI properties that are universal standards and independent of a specific language, thus the graph could easily be adjusted to serve any language. The proposed framework generated a knowledge graph that is fully integrated, dynamic, scalable, easily reproducible, reliable, and practically efficient. The cancer subgraph could serve as a separate graph for cancer-related healthcare systems.

5. Knowledge Graph Evaluation and Discussion

Frameworks were proposed for the systematic evaluation of knowledge graphs. The evaluation conducted is based on a proposed practical framework for evaluating the quality of knowledge graphs covered in [71,72] where the frameworks specified proposed quality dimensions for quantitatively and qualitatively comparing knowledge graphs. The evaluation covered here focused on ten dimensions from the evaluation studies and added three other dimensions- scalability, synonyms coverage, and types of relationships covered. The evaluation focused on comparing the proposed generated knowledge graph with the other studies from Table 1 having their output as a knowledge graph. In Table 4, the first ten dimensions discuss the accessibility where the proposed KG depends on available online resources, and thus reliable and trustful resources. The proposed KG covers all nodes from DO and SYMP, and this covers all standard diseases and symptoms having nodes linked to DO and SYMP with a score of 1. Linked nodes are self-descriptive in terms of URIs as they have property IRI. The linking is based on standard vocabulary which reinforces the interoperability of the KG. The proposed KG is built based on a variety of domain-specific up-to-date resources relevant to the field of interest. The generated graph is moderate in size compared to other KGs however, the graph covers all the disease concepts mentioned in DO ontology, it doesn't include redundancy in disease concepts as it considered diseases and their synonyms as exact nodes. The proposed KG and methodologies used are scalable leading to graph updates whenever additional data resources are considered. The proposed KG covers four types of relationships - a disease with its symptoms, causes, risk, and prevention factors.

Table 4. Quality dimensions-based comparison between different studies generated KGs.

Dimension	HSDN [44]	DS-ontology [46]	HDDO [39]	DSKG[51]	Ruan, T. et al. [48]	Oberkampf, H. et al. [52]	Rotemensch, M. et al. [50]	Proposed KG
1. Accessibility	All data resources are available online	unavailable	All data resources are available online	All data resources are available online	All data resources are available online	All data resources are available online	Data resources available for Professionals	All data resources are available online
2. Appropriate amount	One resource not covering all diseases	Covering 200 diseases	Covering 1000 diseases	Covering 70 diseases	Covering 32,956 diseases	Covering limited diseases	Covering limited diseases	For Diseases mentioned in MayoClinic
3. Believability (Reliability)	Based on provenance of trustful information	Based on Medical Experts Intervention	Based on provenance of trustful information	Based on web-board resources	Based on provenance of trustful information	Based on provenance of trustful information	Based on provenance of trustful information	Based on provenance of trustful information
4. Completeness in terms of linkage to DO or SYMP	No linkage to DO or SYMP	Nodes linked to DO and SYMP	Nodes linked to DO and SYMP	No linkage to DO or SYMP	No linkage to DO or SYMP	Nodes linked to DO and SYMP	No linkage to DO or SYMP	Nodes linked to DO and SYMP
5. Cost-effective	small graph size	small graph size	small graph size	small graph size	Moderate graph size	small graph size	small graph size	Moderate graph size
6. Ease of understanding in terms of self-descriptive URIs	No URIs considered	Concepts URIs considered	Concepts URIs considered	No URIs considered	No URIs considered	Concepts URIs considered	No URIs considered	Concepts URIs considered
7. Interoperability	Use standard vocabularies	Use standard vocabularies	Use standard vocabularies	Use standard vocabularies	Use standard vocabularies	Use standard vocabularies	Use standard vocabularies	Use standard vocabularies

Table 4. *Cont.*

Dimension	HSDN [44]	DS-ontology[46]	HDDO[39]	DSKG[51]	Ruan, T. et al. [48]	Oberkampf, H. et al. [47]	Rotemensch, M. et al. [50]	Proposed KG
8. Relevancy	Yes, for specified diseases	Yes, for specified diseases	Yes, for specified diseases	Yes, for specified diseases	Yes, for specified diseases	Yes, for specified diseases	Yes, for specified diseases	Yes, for all DO diseases
9. Timeliness	Limited to study resource	Limited to study resource	Up to date	Limited to data resource	Up to date	Up to date	Limited to study resource	Up to date
10. Variety	Limited to study resource	Limited to medical experts' intervention	Use a variety of domain-specific resources	Limited to study resource	Use a variety of domain-specific resources	Limited to study resource	Limited to study resource	Use a variety of domain-specific resources
11. Scalability	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
12. Synonyms covered	No	No	No	No	Treated as separate nodes	Yes	No	Yes
13. Relationships covered	Disease-Symptom relationship	Disease-Symptom relationship	Disease-Symptom relationship	Disease-Symptom relationship	Disease-Symptom relationship	Disease-Symptom relationship	Disease-Symptom relationship	4 relationships covered

6. Conclusions and Future Work

The developed framework generated a knowledge graph based on medical facts from a reliable online medical encyclopedia integrated with standardized ontologies—Human Disease Ontology and Symptom Ontology where nodes are linked to their internationalized resource identifier. The integrated knowledge graph is a base for any intelligent expert advisor for professional disease prediction or for giving a normal user an alert for any ailments detected so they provide medical awareness. The framework also gives insights into the causes, prevention, and risk factors of diseases. The integrated knowledge graph resulted in 24,615 nodes with 29,165 relationships, with 594 nodes linked with disease concepts from standardized disease ontologies and 661 nodes linked with standardized symptom ontology. The graph also indicates nodes for the causes, prevention, and risk factors of diseases extracted from the MayoClinic online encyclopedia. Having a knowledge graph with linked nodes to standardized ontologies would be a standard for any intelligent healthcare system for disease prediction and symptom check systems. The graph will help have a common and standard language between normal users and medical professionals. The proposed framework provides an automated novel approach for generating a disease-symptom knowledge graph that is fully integrated, dynamic, scalable, easily reproducible, reliable, up-to-date, and practically efficient. Separate subgraphs could be extracted to serve as a separate knowledge base for specific-domain healthcare systems. A case study was performed to generate organ cancer diseases' subgraph connected with their symptoms, causes, risk, and prevention factors, with nodes linked with standardized ontologies.

Analyzing the knowledge graph would be useful for research in future work, as measuring the density of the graph helps in extracting the most shared symptoms among distinct diseases. The methodologies and techniques adopted within the framework are efficient for enriching the graph when more datasets are provided. More datasets are needed for building and training diseases and symptoms entity name normalization systems that would lead to an increase in the number of entities linked and normalized to standardized ontologies. In future work, more datasets from other medical resources would be considered to enrich the graph and affects both the score of certainty of normalizing the node and the weight of relationships between distinct nodes.

Author Contributions: Conceptualization: N.M., S.G.; methodology: N.M., S.G.; data collection: N.M.; analysis and interpretation of results: N.M., S.G.; draft manuscript preparation: N.M., S.G.; supervision: S.G., E.S., K.E. All authors reviewed the results and approved the final version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Human disease ontology analyzed during the current study is available online at https://disease-ontology.org/(accessed on 20 December 2022). The Symptom Ontology analyzed during the current study is available for download at https://obofoundry.org/ontology/symp (accessed on 20 December 2022). The dataset used for evaluation is available online at: https://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html (accessed on 17 December 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KG	Knowledge graph
DO	The Human Disease ontology
SYMP	The Symptoms Ontology
DOID	disease ontology identifier
IRI	Internationalized Resource Identifier
CDC	Centers for Disease Control and Prevention
UMLS	Unified Medical Language System

RDF	Resource Description Framework
DNER	Disease name entity recognition
BERT	Bidirectional Encoder Representations from Transformers
BILSTM	Bidirectional Long Short-Term Memory
CRF	Conditional Random Field

References

1. Hammad, R.; Barhoush, M.; Abed-Alguni, B.H. A Semantic-Based Approach for Managing Healthcare Big Data: A Survey. *J. Healthc. Eng.* **2020**, *20*, 8865808. [CrossRef] [PubMed]
2. Cheatham, M.; Pesquita, C. Semantic Data Integration. In *Handbook of Big Data Technology*; Springer: Cham, Switzerland, 2017; pp. 263–305. [CrossRef]
3. Panch, T.; Szolovits, P.; Atun, R. Artificial intelligence, machine learning and health systems. *J. Glob. Health* **2018**, *8*, 020303. [CrossRef] [PubMed]
4. Shaban-Nejad, A.; Michalowski, M.; Buckeridge, D.L. Health Intelligence: How Artificial Intelligence Transforms Population and Personalized Health. *NPJ Digit. Med.* **2018**, *1*, 53. [CrossRef] [PubMed]
5. Narayanasamy, S.K.; Srinivasan, K.; Hu, Y.C.; Masilamani, S.K.; Huang, K.Y. A Contemporary Review on Utilizing Semantic Web Technologies in Healthcare, Virtual Communities, and Ontology-Based Information Processing Systems. *Electronics* **2022**, *11*, 453. [CrossRef]
6. Sermet, Y.; Demir, I. A Semantic Web Framework for Automated Smart Assistants: A Case Study for Public Health. *Big Data Cogn. Comput.* **2021**, *5*, 57. [CrossRef]
7. Jagadeeswari, V.; Subramaniyaswamy, V.; Logesh, R.; Vijayakumar, V. A Study on Medical Internet of Things and Big Data in Personalized Healthcare System. *Health Inf. Sci. Syst.* **2018**, *6*, 14. [CrossRef] [PubMed]
8. Ferreira, J.D.; Teixeira, D.C.; Pesquita, C. Biomedical Ontologies: Coverage, Access and Use. In *Reference Module in Biomedical Sciences*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 382–395. [CrossRef]
9. Rossanez, A.; dos Reis, J.C.; da Torres, R.S.; de Ribaupierre, H. KGen: A Knowledge Graph Generator from Biomedical Scientific Literature. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 314. [CrossRef]
10. Tan, J.; Qiu, Q.; Guo, W.; Li, T. Research on the Construction of a Knowledge Graph and Knowledge Reasoning Model in the Field of Urban Traffic. *Sustainability* **2021**, *13*, 3191. [CrossRef]
11. Trouli, G.E.; Pappas, A.; Troullinou, G.; Koumakis, L.; Papadakis, N.; Kondylakis, H. SumMER: Structural Summarization for RDF S / KGs. *Algorithms* **2023**, *16*, 18. [CrossRef]
12. Abu-Salih, B.; L-Qurishi, M.A.; Alweshah, M.; L-Smadi, M.A.; Alfayez, R.; Saadeh, H. Healthcare Knowledge Graph Construction: State-of-the-Art, Open Issues, and Opportunities. *arXiv* **2022**, arXiv:2207.03771.
13. Kim, J.; Sohn, M. Graph Representation Learning-Based Early Depression Detection Framework in Smart Home Environments. *Sensors* **2022**, *22*, 1545. [CrossRef] [PubMed]
14. Qu, J. A Review on the Application of Knowledge Graph Technology in the Medical Field. *Sci. Program.* **2022**, *22*, 12. [CrossRef]
15. Shi, L.; Li, S.; Yang, X.; Qi, J.; Pan, G.; Zhou, B. Semantic Integration of Heterogeneous Medical Knowledge and Services. *Res. Artic. Semant. Health Knowl. Graph* **2017**, *2017*, 8–10.
16. Rajabi, E.; Kafaie, S. Knowledge Graphs and Explainable AI in Healthcare. *Information* **2022**, *13*, 459. [CrossRef]
17. Wu, X.; Duan, J.; Pan, Y.; Li, M. Medical Knowledge Graph: Data Sources, Construction, Reasoning, and Applications. *Big Data Min. Anal.* **2022**, *2022*. [CrossRef]
18. Zhang, Y.; Sheng, M.; Zhou, R.; Wang, Y.; Han, G.; Zhang, H.; Xing, C.; Dong, J. HKGB: An Inclusive, Extensible, Intelligent, Semi-Auto-Constructed Knowledge Graph Framework for Healthcare with Clinicians' Expertise Incorporated. *Inf. Process. Manag.* **2020**, *57*, 102324. [CrossRef]
19. Schriml, L.M.; Arze, C.; Nadendla, S.; Chang, Y.-W.W.; Mazaitis, M.; Felix, V.; Feng, G.; Kibbe, W.A. Disease Ontology: A Backbone for Disease Semantic Integration. *Nucleic Acids Res.* **2012**, *40*, D940–D946. [CrossRef] [PubMed]
20. Kirkpatrick, A.; Onyeze, C.; Kartchner, D.; Allegri, S.; An, D.N.; McCoy, K.; Davalbhakta, E.; Mitchell, C.S. Optimizations for Computing Relatedness in Biomedical Heterogeneous Information Networks: SemNet 2.0. *Big Data Cogn. Comput.* **2022**, *6*, 27. [CrossRef] [PubMed]
21. Gao, M.; Xiao, Q.; Wu, S.; Deng, K. An Improved Method for Named Entity Recognition and Its Application to CEMR. *Future Internet* **2019**, *11*, 185. [CrossRef]
22. Elnagar, S.; Yoon, V.; Thomas, M.A. An Automatic Ontology Generation Framework with an Organizational Perspective. *Proc. Annu. Hawaii Int. Conf. Syst. Sci.* **2020**, *2020*, 4860–4869. [CrossRef]
23. Postiglione, M. Towards an Italian Healthcare Knowledge Graph. In Proceedings of the 14th International Conference, SISAP 2021, Dortmund, Germany, 29 September–1 October 2021; Springer: Cham, Switzerland, 2021; pp. 387–394. [CrossRef]
24. Syed, M.H.; Huy, T.Q.B.; Chung, S.T. Context-Aware Explainable Recommendation Based on Domain Knowledge Graph. *Big Data Cogn. Comput.* **2022**, *6*, 11. [CrossRef]
25. Ruas, P.; Lamurias, A.; Couto, F.M. Linking Chemical and Disease Entities to Ontologies by Integrating PageRank with Extracted Relations from Literature. *J. Cheminform.* **2020**, *12*, 1–11. [CrossRef]

26. Batbaatar, E. Ontology-Based Healthcare Named Entity Recognition from Twitter Messages Using a Recurrent Neural Network Approach. *Int. Environ. Res. Public Health* **2019**, *16*, 3628. [CrossRef]
27. Sboev, A.; Rybka, R.; Gryaznov, A.; Moloshnikov, I.; Sboeva, S.; Rylkov, G.; Selivanov, A. Adverse Drug Reaction Concept Normalization in Russian-Language Reviews of Internet Users. *Big Data Cogn. Comput.* **2022**, 6, 145. [CrossRef]
28. Makris, C.; Simos, M.A. Otnel: A Distributed Online Deep Learning Semantic Annotation Methodology. *Big Data Cogn. Comput.* **2020**, *4*, 31. [CrossRef]
29. Karadeniz, I.; Özgür, A. Linking Entities through an Ontology Using Word Embeddings and Syntactic Re-Ranking. *BMC Bioinform.* **2019**, *20*, 1–13. [CrossRef] [PubMed]
30. Schriml, L.M.; Munro, J.B.; Schor, M.; Olley, D.; McCracken, C.; Felix, V.; Baron, J.A.; Jackson, R.; Bello, S.M.; Bearer, C.; et al. The Human Disease Ontology 2022 Update. *Nucleic Acids Res.* **2022**, *50*, D1255–D1261. [CrossRef]
31. Disease Ontology Project. Available online: https://disease-ontology.org/ (accessed on 20 December 2022).
32. Symptom Ontology. Available online: http://purl.obolibrary.org/obo/symp.owl (accessed on 20 December 2022).
33. OBO Foundary. Available online: https://obofoundry.org/ (accessed on 15 December 2022).
34. Mayo Clinic Diseases and Conditions. Available online: https://www.mayoclinic.org/diseases-conditions (accessed on 22 December 2022).
35. Health Websites Ranking. Available online: https://www.similarweb.com/top-websites/category/health/ (accessed on 27 December 2022).
36. Top 15 Most Popular Health Websites. Available online: https://escapingthehealthcareprison.org/consumer-information-navigator/top-15-popular-health-websites/ (accessed on 27 December 2022).
37. Global Burden of Disease Cancer Collaboration. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived with Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study. *JAMA Oncol.* **2019**, *5*, 1749–1768. [CrossRef] [PubMed]
38. Alawad, M.; Gao, S.; Shekar, M.C.; Hasan, S.M.S.; Christian, J.B.; Wu, X.C.; Durbin, E.B.; Doherty, J.; Stroup, A.; Coyle, L.; et al. Integration of Domain Knowledge Using Medical Knowledge Graph Deep Learning for Cancer Phenotyping. *arXiv* **2021**, arXiv:2101.01337.
39. Kim, G.W.; Lee, D.H. Intelligent Health Diagnosis Technique Exploiting Automatic Ontology Generation and Web-Based Personal Health Record Services. *IEEE Access* **2019**, *7*, 9419–9444. [CrossRef]
40. Cahyani, D.E.; Wasito, I. Automatic Ontology Construction Using Text Corpora and Ontology Design Patterns (ODPs) in Alzheimer's Disease. *J. Ilmu Komput. dan Inf.* **2017**, *10*, 59. [CrossRef]
41. Kim, T.; Yun, Y.; Kim, N. Deep Learning-Based Knowledge Graph Generation for Covid-19. *Sustainability* **2021**, *13*, 2276. [CrossRef]
42. Hamed, A.A.; Fandy, T.E.; Tkaczuk, K.L.; Verspoor, K.; Lee, B.S. COVID-19 Drug Repurposing: A Network-Based Framework for Exploring Biomedical Literature and Clinical Trials for Possible Treatments. *Pharmaceutics* **2022**, *14*, 567. [CrossRef] [PubMed]
43. Hamed, A.A.; Rey, M.; Rey, M. Mining Literature-Based Knowledge Graph for Predicting Combination Therapeutics: A COVID-19 Use Case. *Preprints* **2022**. [CrossRef]
44. Zhou, X.; Menche, J.; Barabási, A.L.; Sharma, A. Human Symptoms-Disease Network. *Nat. Commun.* **2014**, *5*, 4212. [CrossRef]
45. Disease-Symptom Knowledge Database. Available online: https://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html (accessed on 17 December 2022).
46. Mhadhbi, L.; Akaichi, J. DS-Ontology: A Disease-Symptom Ontology for General Diagnosis Enhancement. In Proceedings of the ICISDM'17: 2017 International Conference on Information System and Data Mining, Charleston, SC, USA, 1–3 April 2017; pp. 99–102. [CrossRef]
47. Oberkampf, H.; Gojayev, T.; Zillner, S.; Zühlke, D.; Auer, S.; Hammon, M. From Symptoms to Diseases–Creating the Missing Link. In *European Semantic Web Conference*; Springer: Cham, Switzerland, 2015; pp. 652–667.
48. Ruan, T.; Wang, M.; Sun, J.; Wang, T.; Zeng, L.; Yin, Y.; Gao, J. An Automatic Approach for Constructing a Knowledge Base of Symptoms in Chinese. *J. Biomed. Semant.* **2017**, *8*, 71–79. [CrossRef]
49. Hassan, M.; Makkaoui, O.; Coulet, A.; Toussaint, Y. Extracting Disease-Symptom Relationships by Learning Syntactic Patterns from Dependency Graphs. In Proceedings of the BioNLP 15, Beijing, China, 26–31 July 2015; pp. 71–80. [CrossRef]
50. Rotmensch, M.; Halpern, Y.; Tlimat, A.; Horng, S.; Sontag, D. Learning a Health Knowledge Graph from Electronic Medical Records. *Sci. Rep.* **2017**, *7*, 5994. [CrossRef]
51. Pechsiri, C.; Piriyakul, R. Applied Sciences Construction of Disease—Symptom Knowledge Graph from Web—Board Documents. *Appl. Sci.* **2022**, *12*, 6615. [CrossRef]
52. Okumura, T.; Tateisi, Y. A Lightweight Approach for Extracting Disease-Symptom Relation with MetaMap toward Automated Generation of Disease Knowledge Base. In Proceedings of the International Conference on Health Information Science, HIS 2012, Beijing, China, 8–10 April 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 164–172. [CrossRef]
53. Silva, M.C.; Eugénio, P.; Faria, D.; Pesquita, C. Ontologies and Knowledge Graphs in Oncology Research. *Cancers* **2022**, *14*, 1906. [CrossRef]
54. Gong, M.; Wang, Z.; Liu, Y.; Zhou, H.; Wang, F.; Wang, Y.; Hong, N. Toward Early Diagnosis Decision Support for Breast Cancer: Ontology-Based Semantic Interoperability. *J. Clin. Oncol.* **2019**, *27*, e18072. [CrossRef]

55. Gogleva, A.; Polychronopoulos, D.; Pfeifer, M.; Poroshin, V.; Ughetto, M.; Martin, M.J.; Thorpe, H.; Bornot, A.; Smith, P.D.; Ben Sidders, B.; et al. Knowledge Graph-Based Recommendation Framework Identifies Drivers of Resistance in EGFR Mutant Non-Small Cell Lung Cancer. *Nat. Commun.* **2022**, *13*, 1667. [CrossRef] [PubMed]
56. Patel, H. Bionerflair: Biomedical named entity recognition using flair embedding and sequence tagger. *arXiv* **2020**, arXiv:2011.01504.
57. Weber, L.; Sänger, M.; Münchmeyer, J.; Habibi, M.; Leser, U.; Akbik, A. HunFlair: An Easy-to-Use Tool for State-of-the-Art Biomedical Named Entity Recognition. *Bioinformatics* **2021**, *37*, 2792–2794. [CrossRef] [PubMed]
58. Abulaish, M.; Parwez, M.A.; Jahiruddin. DiseaSE: A Biomedical Text Analytics System for Disease Symptom Extraction and Characterization. *J. Biomed. Inform.* **2019**, *100*, 103324. [CrossRef] [PubMed]
59. Cho, H.; Choi, W.; Lee, H. A Method for Named Entity Normalization in Biomedical Articles: Application to Diseases and Plants. *BMC Bioinform.* **2017**, *18*, 451. [CrossRef]
60. Soshnikov, D.; Petrova, T.; Soshnikova, V.; Grunin, A. Analyzing COVID-19 Medical Papers Using Artificial Intelligence: Insights for Researchers and Medical Professionals. *Big Data Cogn. Comput.* **2022**, *6*, 4. [CrossRef]
61. Gates, L.E.; Hamed, A.A. The Anatomy of the SARS-CoV-2 Biomedical Literature: Introducing the Covidx Network Algorithm for Drug Repurposing Recommendation. *J. Med. Internet Res.* **2020**, *22*, e21169. [CrossRef]
62. Zongcheng, J.; Wei, Q.; Xu, H. Bert-based ranking for biomedical entity normalization. *Amia Summits Transl. Sci. Proc.* **2020**, *20*, 269–277. [CrossRef]
63. He, Y.; Zhu, Z.; Zhang, Y.; Chen, Q.; Caverlee, J. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. *arXiv* **2020**, arXiv:2010.03746.
64. He, Y.; Chen, J.; Antonyrajah, D.; Horrocks, I. BERTMap: A BERT-Based Ontology Alignment System. *Proc. Conf. AAAI Artif. Intell.* **2022**, *36*, 5684–5691. [CrossRef]
65. Xu, K.; Yang, Z.; Kang, P.; Wang, Q.; Liu, W. Document-Level Attention-Based BiLSTM-CRF Incorporating Disease Dictionary for Disease Named Entity Recognition. *Comput. Biol. Med.* **2019**, *108*, 122–132. [CrossRef] [PubMed]
66. UMLS Metathesaurus. Available online: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html (accessed on 20 December 2021).
67. Neumann, M.; King, D.; Beltagy, I.; Ammar, W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 1 August 2019; pp. 319–327. [CrossRef]
68. Cariello, M.C.; Lenci, A.; Mitkov, R. A Comparison between Named Entity Recognition Models in the Biomedical Domain. In Proceedings of the Translation and Interpreting Technology Online Conference, Online, 6–7 July 2021; pp. 76–84. [CrossRef]
69. Abdurxit, M.; Tohti, T.; Hamdulla, A. An Efficient Method for Biomedical Entity Linking Based on Inter-and Intra-Entity Attention. *Appl. Sci.* **2022**, *12*, 3191. [CrossRef]
70. Zhang, Y.; Chen, Q.; Yang, Z.; Lin, H.; Lu, Z. BioWordVec, Improving Biomedical Word Embeddings with Subword Information and MeSH. *Sci. Data* **2019**, *6*, 1–9. [CrossRef] [PubMed]
71. Chen, H.; Cao, G.; Chen, J.; Ding, J. A Practical Framework for Evaluating the Quality of Knowledge Graph. In *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding 4th China Conference, CCKS 2019, Hangzhou, China, 24–27 August 2019*; Springer: Singapore, 2019. [CrossRef]
72. Huaman, E. Steps to Knowledge Graphs Quality Assessment. *arXiv* **2022**, arXiv:2208.07779.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.