Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 4792909, 16 pages
https://doi.org/10.1155/2022/4792909

Research Article

Publishing and Interlinking COVID-19 Data Using Linked Open Data Principles: Toward Effective Healthcare Planning and Decision-Making

Shaukat Ali,[1] **Islam Zada**,[1] **Zahid Mehmood**,[2] **Amin Ullah**,[3] **Haider Ali**,[4] **and Mujeeb Ullah**[5]

[1]*Department of Computer Science, University of Peshawar, Peshawar 25120, Pakistan*
[2]*Department of Computer Engineering, University of Engineering and Technology, Taxila 47050, Pakistan*
[3]*Department of Computer Science, Swedish College of Engineering and Technology, Wah Cantt, Rawalpindi 47000, Pakistan*
[4]*Department of Pharmacy, University of Peshawar, Peshawar 25120, Pakistan*
[5]*Department of Zoology, Islamia College, Peshawar 25120, Pakistan*

Correspondence should be addressed to Zahid Mehmood; zahid.mehmood@uettaxila.edu.pk

Received 10 October 2021; Accepted 18 February 2022; Published 22 March 2022

Academic Editor: Giuseppe D'Aniello

The COVID-19 data is critical to support countries and healthcare organizations for effective planning and evidence-based practices to counter the pressures of cost reduction, improved coordination, and outcome and produce more with less. Several COVID-19 datasets are published on the web to support stakeholders in gaining valuable insights for better planning and decision-making in healthcare. However, the datasets are produced in heterogeneous proprietary formats, which create data silos and make data discovery and reuse difficult. Further, the data integration for analysis is difficult and is usually performed by the domain experts manually, which is time-consuming and error-prone. Therefore, an explicit, flexible, and widely acceptable methodology to represent, store, query, and visualize COVID-19 data is needed. In this paper, we have presented the design and development of the Linked Open COVID-19 Data system for structuring and transforming COVID-19 data into semantic format using explicitly developed ontology and publishing on the web using Linked Open Data (LOD) principles. The key motivation of this research is the evaluation of LOD technology as a potential option and application of the available Semantic Web tools (i.e., Protégé, Excel2RDF, Fuseki, Silk, and Sgvizler) for building LOD-based COVID-19 information systems. We have also underpinned several use-case scenarios exploiting the LOD format of the COVID-19 data, which could be used by applications and services for providing relevant information to the end-users. The effectiveness of the proposed methodology and system is evaluated using the system usability scale and descriptive statistical methods and results are found promising.

1. Introduction

Information is the fundamental requirement in the public healthcare domain for effective decision-making regarding epidemiologic surveillance, disease outbreak, healthcare program outcome assessment for performance and evaluation measurement, healthcare program planning, cost and expenditures, coverage and quality of services, and policy analysis [1]. Like infectious diseases, the COVID-19 pandemic has posed a serious threat to the public health and has resulted in significant morbidity, mortality, and economic damages worldwide. The WHO has reported 104, 370, 550 confirmed cases and 2, 271, 80 confirmed deaths by February 05, 2021. The COVID-19 data (i.e., patient records, clinical trials, and drug efficiency) is constantly growing and different organizations are providing Open Data (i.e., movement enabling sharing datasets) on the Web (e.g., OurWorldinData, Worldometer, and Kaggle) to support the caregivers and decision-makers. However, the COVID-19 datasets produced by the different organizations on the Web

are in heterogeneous data formats, naming conventions, structures, and resides in distributed data repositories; with no attention being paid to uniform data representation. Mostly, the COVID-19 datasets are represented and distributed in tabular formats (i.e., CSV, Excel, PDF, etc.) with minimum metadata, which makes COVID-19 data integration, comparison, and reuse very difficult. In addition, the organizations do not link disparate datasets despite having related indicators, and the vocabularies and data formats used in the datasets are inconsistent. This diversity, variety, and veracity of COVID-19 datasets pose serious time consuming, and error-prone problems of data management (i.e., representation, storage, and retrieval), sharing, and integration for effective planning and decision-making by the stakeholders. The explosive growth of COVID-19 datasets needs to be properly explored to reveal important information and convert them into effective knowledge for improved COVID-19 healthcare practices. Therefore, more flexible data representation, analysis, querying, and visualization methods are needed for COVID-19 datasets. In addition, the sharing of an enormous amount of structured COVID-19 data could provide opportunities for researchers to produce advanced methods for providing high-quality information to revolutionize the healthcare sector.

The COVID-19 datasets are produced under the initiatives of Open Data, which allows for the use, modification, and distribution of the freely available datasets generated and shared by third parties for any purpose [2]. The World Wide Web Consortium (W3C) has founded Health Care and Life Science Interest Group (HCLSIG) intending to develop, advocate, and support the use of Semantic Web (SW) technologies in healthcare and life sciences domains [3, 4]. The Linked Data (LD) uses SW technologies and provides practical solutions for publishing and interlinking structurally formatted data on the Web [5, 6]. The LD envisions the idea of SW by transforming the Web of documents into the Web of data and enables the Web as a distributed network of data accessible and useable by the software agents and machines [7]. To achieve the goal, LD works on the four basic rules: (1) using Uniform Resource Identifiers (URIs) for naming objects; (2) using HTTP URIs for names lookup; (3) using RDF for describing and relating data; and (4) using SPARQL Protocol and RDF Query Language (SPARQL) for querying RDF graphs [8]. The Linked Open Data (LOD) extends LD best practices and transforms the Web into a globally accessible data space of interconnected open data from the different areas (e.g., cross-domain, geography, government, life sciences, media, linguistics, social networking, scientific publications, scientific and statistical, etc.). The Web of data is formed into LOD Cloud. The LOD Cloud is an immense network of published and interconnected datasets using LD principles. In May 2020, the LOD cloud has 1255 datasets and 16174 links [9]. The LOD cloud is accessible using the available Web infrastructure using SW technologies [7].

The interlinked structure of the distributed datasets on the LOD Cloud provides end-users with new use-case scenarios to develop domain-specific or cross-domain applications, which are not possible over isolated datasets [7]. The universal hypothesis of the LOD is to extend World Wide Web (WWW) from interlinking documents only to revolutionize global data sharing, integration, and analysis [1]. The healthcare domain has already adopted the SW technologies and LOD principles to enhance information retrieval and visualization processes, which are not possible with separate data stores [7, 10]. However, to the best of our knowledge, there is no research demonstrating the LOD-based COVID-19 data management. The COVID-19 datasets heterogeneity is an open challenge, which needs attention from the research community to provide effective solutions to exploit COVID-19 datasets potentials for effective planning and decision making in the healthcare sector by the stakeholders.

The key motivation of this research is the evaluation of LOD technology as a potential option and application of the available SW tools (i.e., Protégé, Excel2RDF, Fuseki, Silk, and Sgvizler) for building LOD-based COVID-19 information systems. In this paper, we have presented the design and development of the Linked Open COVID-19 Data (LOCD) system emphasizing the exploitation of SW and LOD technologies for open COVID-19 datasets from data modeling to visualization. The LOCD will generate a semantically structured, interlinked, and machine-readable formatted RDF COVID-19 dataset from the disparate COVID-19 datasets available on the Web and the dataset will be published on the LOD cloud. A novel aspect of this paper is providing an ontology-based approach for LOCD to help in providing semantic description and representation to the COVID-19 data provided in heterogeneous formats, which will facilitate integration, exploration, and analysis of the semantically enriched COVID-19 dataset and will promote innovative ways to use semantically enriched COVID-19 data creatively. The output of the research could be used as input to the health informatics in the governments and organizations to overcome the issue of data silos for effective decision-making and develop applications and services to provide relevant COVID-19 information flexibly to the end-users. The main contributions of this article are as follows:

(1) To semantically model and map COVID-19 data from disparate heterogeneous datasets into RDF COVID-19 dataset using SW technologies.

(2) To present a methodology and system architecture by exploiting the SW and LOD potentials for the development of a LOD-based COVID-19 information system.

(3) To outline exemplary use-case scenarios from end users' perspectives for exploiting the LOD-based RDF COVID-19 data.

(4) To evaluate the proposed methodology and system from practicality, usability, availability, and information organization aspects.

The rest of the paper is organized as: Section 2 provides a quick overview of the LOD and SW technologies; Section 3 presents a comprehensive review of the related research; Section 4 presents an in-depth discussion of the proposed methodology and development; Section 5 portrays several

potential use-case scenarios and results; Section 6 presents evaluation and discussion, and Section 7 concludes this research and underpins future research directions.

2. Background, SW and LOD

The SW represents resources (i.e., documents and entities within the documents) on the Web scientifically to ease data representation, processing, storage, and linkage for supporting machines and users to function in cooperation [11]. The SW technologies (i.e., RDF, RDFs, OWL, and SPARQL) are based on W3C standards, which are aimed to materialize the SW initiatives by codifying formal and explicit definitions of the basic concepts and their relationships in a domain of interest in an easily understandable manner for the computer-based agents [7]. The SW technologies provide opportunities to deal with data heterogeneity, which could hinder the organization, interlinking, and sharing of datasets generated by different providers belonging to the same context [12, 13]. The Linked Data (LD) represents a set of best practices methods to take advantage of the SW by using SW technologies for publishing structured data as resources on the Web and interlinking them semantically with related datasets on the Web [14]. The primary concern of LD is to extend the Web usage from sharing documents only to publishing and interlinking individual data pieces. This initiative is fueled by the idea that applying SW technologies could revolutionize the data distribution, integration, and analysis on a larger scale like the Web revolutionized sharing and communication of the Web documents. The term LOD represents the use of LD principles to the Open Data [13]. A classical example of LOD is the DBPedia [15]. Technically, the basic idea of LOD is to unambiguously identify arbitrary resources and concepts using Uniform Resource Identifiers (URIs). Information about the resources and concepts are represented and encoded in standardized Resource Description Framework (RDF) language.

The RDF uses a directed labeled graph data model to encode information and facts in triples format comprising of subject, predicate, and object. Ontology supports interoperability between the systems by providing globally agreed terminology to the terms, relationships, and rules [1]. Coupling RDF with ontology building languages (e.g., RDF Schema (RDFs) [16], and Web Ontology Language (OWL) [17]) would enable to develop structured semantic models to describe data in a domain. The RDFs and OWL uses classes and properties to build ontologies for describing knowledge in a domain. However, RDFs are used for building simple lightweight ontologies with limited reasoning support over data, and OWL is used for building heavy-weight ontologies with rich vocabulary for describing complex axioms for facilitating in-depth reasoning and consistency checking on data. The SPARQL [18] is query language for RDF data to retrieve specific information. Thus, the Web of data could be formed by using SW standards to represent information to univocally identify resources by unique identifiers, interconnect disparate resources and crawl the entire data space, and fuse data from different sources by linking, and provide expressive query capabilities over aggregated data to retrieve information [13]. An LOD dataset could be formed from RDF data in triples format using domain ontology and could be stored in databases called triplestores. The triplestores provide facility to efficiently store multiple LOD datasets and provide query-based access to the stored datasets via SPARQL endpoint. The SPARQL could be used for querying LOD datasets in a SPARQL endpoint like using SQL for querying tables in relational databases.

Nowadays, many COVID-19 datasets are available and accessible in human-readable proprietary formats. However, they require users to use specific proprietary software for their accessing and usage [13]. For example, a CSV file would require MS-Excel. Since the LD and LOD use open SW standards. Therefore, the domain users are needed to use generic software to create, store, retrieve, analyze, and visualize data. In addition, using ontologies gives formal definitions and meanings to the resources in the LOD datasets and increases the machines' capabilities to process the data automatically [13]. Therefore, the SW technologies and LOD principles could provide the necessary infrastructure for semantically representing, organizing, integrating, distributing, and visualization of COVID-19 data from the distributed data sources.

3. Related Works

Since the past few years, several organizations are publishing a large volume of heterogeneous healthcare open datasets. Most of them are publishing in their proprietary formats while the others have recognized and explored the application of SW technologies with the LOD widely accepted practices. The advantages these technologies could provide include describing and interlinking heterogeneous data using explicit semantics, relaxing cost of data integration, simplified annotation and sharing, easy reuse of data, and inferring additional information, and exposing data for accessing via HTTP. A top-level comparison of the relevant research works is shown in Table 1.

The W3C has founded the Semantic Web for Health Care and Life Sciences Interest Group (HCLS IG) intending to develop, advocate, and support the use of SW technologies for organizations from the healthcare, life sciences, clinical research, and translational medicines domains to their available data as LOD. The group focuses on using the capabilities of SW technologies to organize a collection of data uniformly to enhance decision-making both in primary care (electronic medical record) and clinical research. The subgroups are focusing on developing and maintaining biomedical ontologies to ensure the availability of biomedical data in RDF, drug safety, and efficient communication and help researchers to navigate and annotate a vast amount of relevant literature. The NeuroCommons project has presented a prototype knowledge base to demonstrate the feasibleness and advantages of SW technologies in the biomedical domain, scalability of the available SW tools and techniques for biomedical knowledge base creation, and assembling and querying biomedical knowledge base from multiple sources and disciplines for understanding a disease [19]. The CardioSHARE [20] project represents a generic,

Table 1: Swift comparison of related research works.

Research	Domain	Dataset	Architecture	SPARQL query	Evaluation	Publish	Size
[1]	Medical (disease cases)	WHO HIV	Yes	Yes	Yes	Yes	Not available
[13]	Health and fitness	Self-collected via IoT devices	Yes	Yes	No	Yes	Not available
[20]	Medial (bioinformatics)	BioMoby project	No	Yes	No	Yes	Not available
[21]	Medical (bioinformatics)	Not specified	Yes	Yes	No	Yes	Not available
[22]	Medical (bioinformatics)	Global health dataset (GHD)	No	Yes	No	Yes	8 million
[23]	Drugs	Web-based open databases	No	No	No	Yes	766, 000
[24]	Clinical trials	ClinicalTrials.gov	No	No	No	Yes	7, 011, 000
[25]	Drugs	DrugBank, DailyMed, ChEMBL, diseasome, TCMGeneDIT, SIDER, STITCH	No	No	No	Yes	>8 million
[27]	Health Insurance Fund of Republic of Macedonia	DrugBank	No	Yes	No	Yes	21,000

scalable, and decentralized web service framework that provides a SPARQL endpoint for querying transparently geographically distributed and independent resources in "deep web" with diverse ownership and formats and generates output in RDF format. The project is aimed for the clinical data on heart diseases analysis; however, the framework is general and can be used for different types of data. However, the CardioSHARE provides distributed semantic access to the bioinformatics web services of the BioMoby project but does not provide architectural, dataset, ontology, and access details. The Bio2RDF [21] project provides a framework to create on-demand knowledge discovery to create a mashup of data in the bioinformatics domain. The project provides a centralized largest LOD of 11 billion RDF triples from 35 diverse heterogeneously formatted datasets of multiple providers of the biomedical domain and provides a SPARQL endpoint for querying. The project has shown the strength of SW for automatic knowledge discovery and integration from billions of documents. However, the Bio2RDF is lacking with ontological details to represent and link an enormous amount of bioinformatics data and requires careful choosing and regular updating from the sources to decrease query time and curatorial efforts [20]. In addition, a custom methodology is used to create an RDF dataset with no compliance with the LOD principles. Zaveri et al. [22] have constructed a LOD dataset in RDF with about 8 million triples from WHO's GHO datasets using RDF Data Cube Vocabulary and published the dataset using the OntoWiki platform with a SPARQL endpoint for querying and browsing data in HTML format. The methodology used is complaint with LOD principals and provides a comprehensive dataset providing information about the number of deaths, health expenditure, disease prevalence, etc. However, it has problems related to data conversion, data linking and integration, data updating, and data exploration. The DrugBank project has constructed a LOD dataset about drugs (containing information about the chemical, pharmacological, and pharmaceutical) from a web-based open non-RDF database and the dataset consists of more than 766,000 RDF triples about 4,800 drugs [23]. The RDF dataset provides information about drugs only and does not provide information about diseases and prevalence. The LinkedCT project has converted the available clinical trials data from ClincalTrails.gov into RDF and discovered semantic links among the data internally and externally to other datasets on the LOD (i.e., PubMed and Even Though) and has associated each trial with a disease and a drug [24]. However, the RDF dataset does not provide information about the prevalence of disease in a specific country, which could be found in GHO. The LODD project has interlinked separated drug datasets already available on the web about the impact of drugs on gene expression during clinical trials and has collected LOD datasets of more than 8 million RDF triples interlinked with over 370,000 RDF links [25, 26]. The project is aimed to provide answers to scientific and industrial questions by linking distributed drug datasets. The project has converted several datasets including DrugBank, DailyMed, and SIDER. However, the RDF dataset provides drugs data only and does not provide information about diseases, number of deaths, health expenditure, and status of the health system in a country for a disease, which is provided by the GHO. Milos et al. used LOD principles and SW technologies to transform and publish drug data in the Health Insurance Fund (HIF) of the Republic of Macedonia into a 5-star LOD connected to the LODD and other LOD Cloud datasets through the DrugBank dataset [27]. However, the RDF dataset is restricted to drugs only and does not provide information about diseases, which is provided by the GHO. Tilahun et al. [1] have converted WHO's HIV data into an RDF dataset and have presented integration-based development of LOD-based health information, representation, querying, and visualization system by using SW tools and technologies. The research has developed an architecture following the LOD principles and shows the potential of the available SW technologies. However, the research is lacking ontology for data mapping. Reda has used LOD

principles and has presented the design and development of a web portal for accumulating, describing, integrating, and sharing heterogeneous IoT health and fitness datasets [13]. The methodology is developed using LOD principles. However, the RDF data contain physical health and fitness data and do not provide disease and drugs information.

The research efforts have presented valuable ideas of exploring the SW and LOD potential for semantically organizing and sharing drugs, healthcare, and clinical trials data. The COVID-19 is a relatively new disease with its unique prospects and significance. Realizing the importance of COVID-19, several organizations are producing COVID-19 datasets in their proprietary formats creating the issues of data silos and heterogeneity, which decreases the data significance for governments and organizations in their effective decision making and planning processes. To overcome the issue of COVID-19 data silos and heterogeneity, the experience from LOD-based research can be applied to COVID-19 datasets for semantical representation, organization, interlinking, and publishing using the LOD principles. Conclusively, the literature survey has shown that the available research can provide the necessary background and knowledge for transforming COVID-19 datasets into LOD.

4. Methodology

Keeping in view the severity of the morbidities and mortalities caused by the COVID-19, the COVID-19 datasets are published on the web in 2-star or 3-start quality (i.e., Excel spreadsheets or CSV files). They provide the opportunity to principally combine/link data into an individual flat file or RDBMS using spatial-temporal fields. However, the data will remain in the 2-star or 3-star quality. Considering the large volume and disparate nature of the COVID-19 data, a systematic RDF-based approach for the data integration would be more effective and will pave the way for transforming data into 5-star quality. This will allow data of different granularities to be recorded in their actual formats, helping to document their provenance. For example, the RDF will allow data (i.e., produced in different formats weekly, daily, and hourly) to be recorded in their actual formats, whereas, in an individual tabular file, human efforts and judgments would be needed to fill in the gaps [8]. In addition, the RDF approach will ease the distribution of the data to enable future research processes and applications development in the medical domain.

4.1. System Data Flow Design. The aim of this research is the design and development of a LOCD system to transform COVID-19 data collected from disparate sources into LOD and make the RDF COVID-19 LOD dataset freely available on the LOD Cloud. The system is developed using the design guidelines for generating and sharing LOD data sources in the Health Care and Life Sciences (HCLS) domain. The general data workflow for the LOCD system is shown in Figure 1. Briefly, the figure is composed of the following steps representing data flow of ontology-driven mapping of structured COVID-19 data (2-star and 3-start formats only) into RDF graph data model, linking with other LOD sources, and retrieval and visualization from a SPARQL endpoint using a Semantic Web application.

(1) Steps 1 and 2: selecting and preprocessing COVID-19 dataset in a structured format, mapping the data in COVID-19 dataset using concepts from a reference ontology, and transformation into RDF format using an automatic tool.

(2) Step 3: editing and customizing the RDF COVID-19 dataset manually for necessary changes and enhancements.

(3) Steps 4, 5, and 6: enriching the RDF COVID-19 dataset by discovering links and integration with related datasets in the LOD cloud using reference ontology for mapping concepts and automatic tools for linking and integration.

(4) Step 7: uploading and publishing the RDF COVID-19 dataset in an RDF triple store and making it available through a SPARQL endpoint.

(5) Step 8: sharing the RDF COVID-19 dataset on the LOD cloud using LOD principles.

(6) Step 9: create a Semantic Web application for querying and retrieving data from the RDF COVID-19 dataset through the SPARQL endpoint and information visualization through a dashboard in the application.

4.2. Architecture and Development. The application of LOD for healthcare and clinical trials research is relatively new. The researchers have proposed systems using their own experiences and methodologies such as [1, 13]. The systems vary from each other while fulfilling the fundamental requirements of a LOD system. Thus, resulting in individual islands and is the wastage of resources and time. To achieve this research's objective, design and development of Linked Open COVID-19 Data (LOCD) system for enhanced COVID-19 data monitoring and visualization, a multilayer architecture is proposed to provide flexibility, reusability, extendibility, and easy understandability (shown in Figure 2). The architecture is developed using the standard guidelines provided by LOD and HCLSIG for the Health Care and Life Science (HCLS) domain [3] and combines all of the data flow steps (shown in Figure 1) into a single holistic structure. The architecture is mainly emphasizing the data management process because effective data transformation and management is the backbone of LOD-based systems [28]. The architecture is a service oriented architecture (SOA) where each component extends the functionality of the architecture by capturing, aggregating, transform, and enriching COVID-19 data into high-level semantic formalism. The configurable modular and integration-oriented approaches are employed to enable and use the available SW technologies and tools for LOD-based system development and implementation: data transformation and mapping, integration and link discovery, organization and storage, and query and retrieval to enable numerous use-case

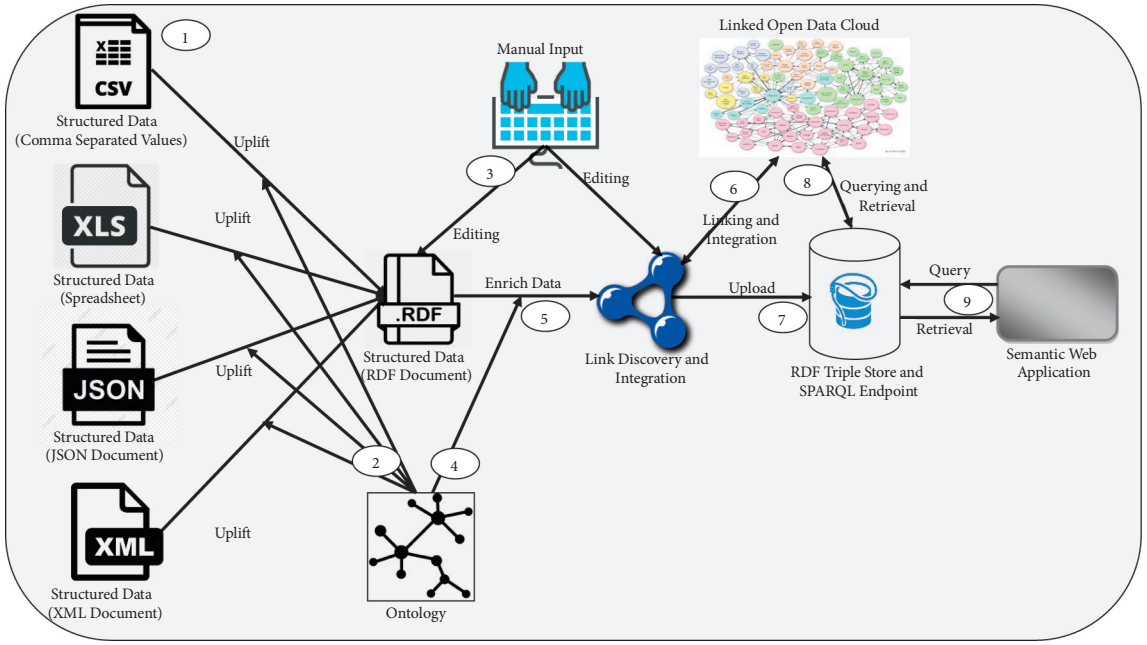

Figure 1: Data flow diagram of the proposed LOCD system.

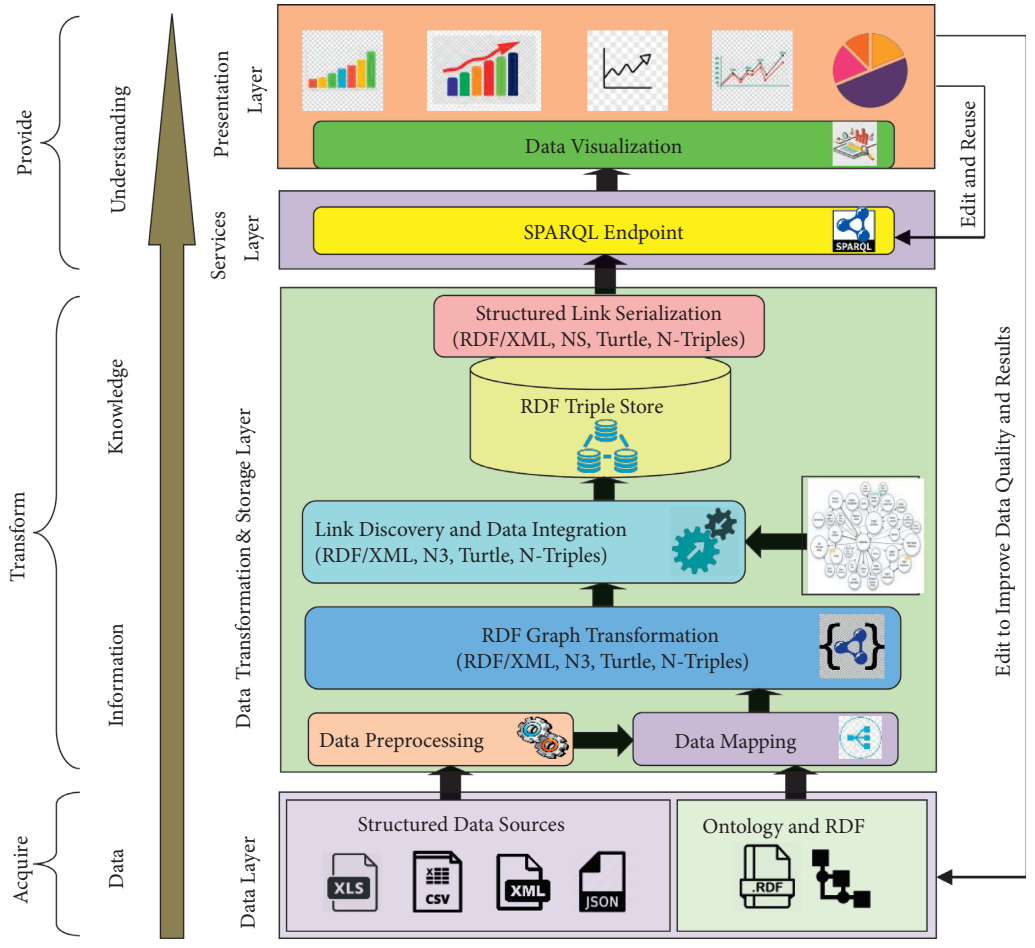

Figure 2: Multilayer architecture of the LOCD system.

scenarios. The LOCD is aimed to provide a reference point to collect, publish, and share COVID-19 datasets using LOD principles to enable accessing and reusing of COVID-19 datasets by the domain experts, scientists, and web developers without being restricted by any type of licensing or patent. The layers of the architecture and their implementation details are covered in the following subsections.

4.2.1. Data Layer. Collecting and integrating large amount of COVID-19 data in the homogenized pattern are a highly challenging job due to the data being provided by the disparate platforms in heterogeneous formats and organizations. One of the unappreciated and time-consuming challenges in data analytics projects is gaining ready and frequent access to the relevant datasets [8]. To carefully select the appropriate datasets, one needs to have knowledge and understanding of datasets availability, usage permission requirement, and capability of handling data in different formats. The severity of the COVID-19 pandemic has directed several organizations to produce COVID-19 data on regular basis to help governments, organizations, and researchers. However, the datasets providing COVID-19 data of different countries from January 22, 2020, in a single file are considered in this research. A summary of the available COVID-19 datasets is shown in Table 2.

(1) Identifying and Obtaining Source Datasets. The availability of granular and authentic data is essential to support countries in managing their COVID-19 pandemic initiatives. However, the 2-star datasets produced by the different organizations are varying in their granularities and none of the COVID-19 datasets provides all this information in a single file. Therefore, multiple datasets will be used and uplifted into RDF and integrated using link discovery. The OurWorldinData COVID-19 dataset provides most of the COVID-19 information. The dataset is published in 2-start (.XLSX) and 3-star (.CSV) data with the date, country, and COVID-19 positive cases and tests performed information of the different countries. The dataset is updated daily and freely available to use. However, the dataset is lacking with geospatial, WHORegion, active COVID-19 positive cases, and recovered COVID-19 cases information of the different countries. The dataset also provides an extensive additional set of information (e.g., total_cases_per_million, new_cases_per_million, total_tests_per_thousand, new_cases_per_thousand, etc.), which could provide insight information and would be more helpful in advanced analysis and visualization. The Kaggle COVID-19 dataset provides missing COVID-19 data in the OurWorldinData COVID-19 dataset. The dataset is published in 3-star (.CSV) data, updated daily, and freely available to use. Since the entries in both of the datasets are not linked directly and use different terminologies. Therefore, data from the selected datasets will be linked by matching the countries' names. The structured datasets would be cleaned, transformed into RDF, and integrated by link discovery using available ontologies. Our published linked data would respect the original data publisher's license and terms of use for the data license.

(2) OntoCOVID Ontology. The availability of a vast amount of necessary data makes it essential to find the best methodology to organize the data landscape into a widely agreed and explicit format to enable representation and understanding by the users and machines for reliable and scalable statistical analysis and visualization. The ontology provides a solution to develop a domain data model by formally and explicitly defining its concepts and meaningful data linkages [29]. The COVID-19 datasets not only vary in data formats heterogeneity but also the number of data items provided and their headings (terminologies). However, all data items provided by the COVID-19 datasets belong to the COVID-19 domain. Therefore, ontology is needed to semantically model and map all data items from COVID-19 datasets for uniform representation and effective transformation into RDF.

The reference OntoCOVID [30] ontology is used to map the data in the COVID-19 datasets. The ontology reuses relevant concepts and properties from the existing widely accepted ontologies and vocabularies to improve interoperability. However, the new concepts and properties are explicitly declared and defined in the ontology to meet and efficiently model the new data needs. The ontology is aimed to provide a common understating of COVID-19 datasets to increase information value and reusability for the LOD-based application development and sharing. The ontology is developed in Protégé ontology editor using the POEM methodology [31] to ensure its coverage, validity, and usability. The ontology is modeled around the concept of episodes where each episode represents a collection of COVID-19 data of a particular country on a particular date. Each episode is associated with the date and numerical values of different aspects. The data are important to numerically quantitate an episode's object and assign it temporal information effective for retrieving and reasoning process. In addition, each episode is added with geographical coordinates for useful map-based tracking and visualization as the location has become a basic attribute for health data [32]. A snippet of the OntoCOVID ontology classes and properties is shown in Figure 3.

4.2.2. Data Transformation and Storage Layer. To share the heterogeneous COVID-19 datasets using the LOD principles, the data is needed to be unified and formally semantically represented. The data transformation and storage layer is a composite layer containing components from data cleansing to semantically storing.

(i) Data Processing and Data Mapping: The data processing and data mapping receives a potential COVID-19 dataset and reference ontology from the data layer, and transforms it into structured machine-readable format (i.e., RDF) for further transformation into 4-star or 5-star data. The data processing pre-processes and cleans a COVID-19 dataset using MS-Excel. The pre-processing and cleansing include conversion of date into standard XML Schema date format, creating URIs, declaring namespaces, and filtering and filling missing data

Table 2: Comparative analysis of the COVID-19 datasets available on the Web.

Dataset	Temporal data level	Geospatial data level	Statistical data levels	Format	Sources	Freely available online?
OurWorldinData	Complete-day wisely	Not available	Confirmed cases, active cases deaths cases, tests	.XSLX,.CSV,.JASON	Not available	Yes
Kaggle	Complete-day wisely	Geographic coordinates	Confirmed cases, active cases, death cases, recovered cases	.CSV	Not available	Yes
Github	Complete-day wisely	Geographic coordinates	Confirmed cases, recovered cases, death cases	.CSV	Not available	Yes
Worldometer	Day wisely only	Not available	Confirmed cases, recovered cases, death cases, critical cases, tests	.CSV	Official websites	No–requiring formal request/agreement

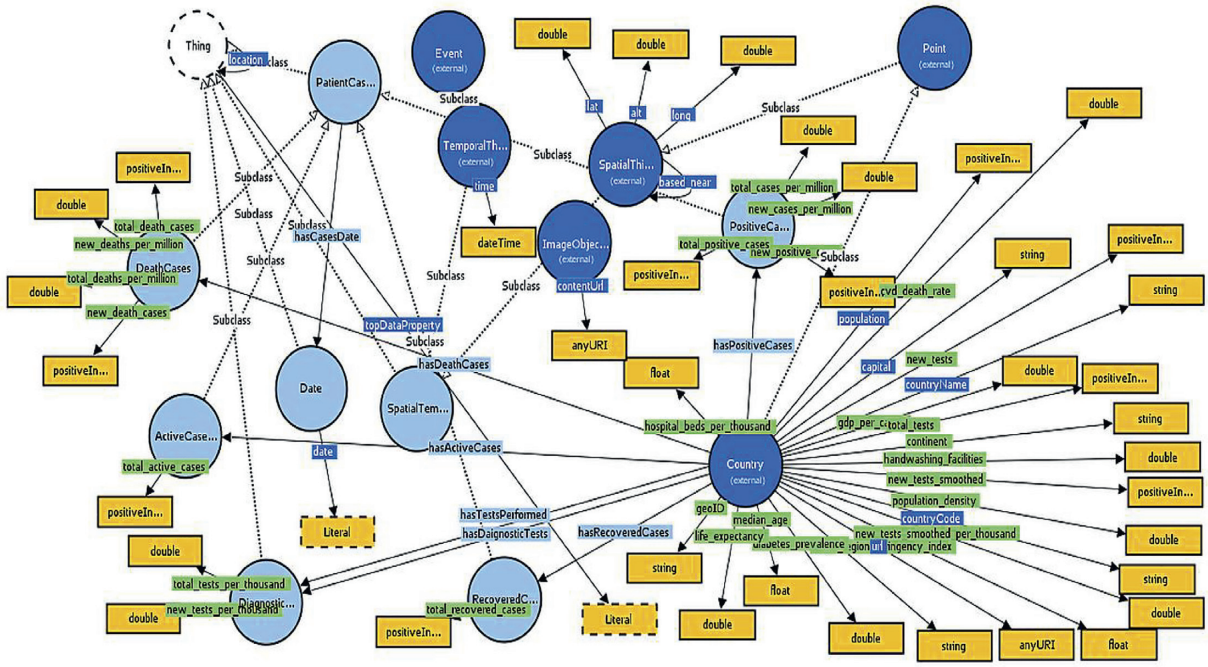

Figure 3: A snippet of OntoCOVID classes and properties [30].

values. Similarly, fully qualified URIs are created and given to the entities in a COVID-19 dataset. Thus, each row in the COVID-19 dataset represents an instance and each column in a COVID-19 dataset represents a property of an instance whose value could be either a fully qualified URI of another instance or literal. Since the data in the candidate COVID-19 datasets is not directly interlinked; therefore, we are needed to link the data in the datasets by matching the countries' names and geographic codes. The structured datasets are normalized using OpenRefine for link discovery and integration further. The data mapping maps the structure of a COVID-19 dataset following the reference OntoCOVID ontology. The column headers in the COVID-19 dataset are mapped and annotated using the concepts and properties of the reference ontology.

(ii) RDF Graph Transformation: The RDF Graph Transformation receives the well-structured and annotated COVID-19 dataset from the data processing and data mapping layer and transforms it into a 4-start RDF graph format using python's Excel3RDF [33] package. The Excel2RDF will take a well-structured and annotated COVID-19 dataset in Excel or CSV format as input and will produce the RDF COVID-19 dataset in Turtle serialization format as output. The RDF COVID-19 dataset is in RDF graph format where COVID-19 data will be connected in subject, predicate, and object triple format. The RDF triples will be labeled with the information given in data processing and data mapping. In each RDF triple, subject and predicate will be represented by fully qualified names and the object will be represented either by a fully qualified URI or literal.

(iii) Link Discovery and Data Integration: To transform the generated RDF COVID-19 dataset into 5-star LOD, links within the entities of the RDF COVID-19 dataset and entities in other related datasets must be made. This will enable new applications to discover and link data from our source as well as other different sources. One of the best practices for creating a LOD is linking it to different sources [14]. Interlinking an RDF COVID-19 dataset with other datasets available on the LOD cloud is a challenging task involving painful processes of identifying similar and matching links types among internal and external datasets [1]. The link discovery and data integration enrich the RDF COVID-19 dataset by finding and establishing the links with the LOD cloud to guarantee the public availability of the dataset, allow federated SPARQL queries, facilitate data integration and data analysis, and links with the LOD cloud. After analyzing the existing health care LOD datasets, the Silk Link Discovery Framework [32] is used for automatic link discovery and data integration. The Silk framework is flexible and provides tools to generate time-efficient RDF link discovery between entities within existing different Web data sources using user-provided link specifications and conditions [34]. These RDF links can be published along with the original RDF COVID-19 dataset on the LOD cloud. For example dbpprop: countryName owl:sameAs skos:preflabel, dbpprop: countryName owl:sameAs madsrdf:authoritativeLabel, etc. The RDF COVID-19 dataset is enriched by establishing owl:sameAs relations with the related LOD data sources (e.g., Bio2RDF, PubMed, and LinkedCT). These interconnections will enable use-case scenarios to provide information about a country and its healthcare, which are otherwise not available in the RDF COVID-19 dataset.

(iv) RDF Triplestore: The RDF triplestore is a database system for the data modeled using RDF triples format. The triplestore provides tools to efficiently store and access RDF triples via Application Programming Interfaces (APIs) and query languages. The LOD-based system requires a triple store as the main storage [1]. For this work, the RDF COVID-19 dataset and RDF links dataset are loaded and stored in the Fuseki triplestore [35].

4.2.3. Services Layer. The services layer will control the RDF data access and will provide a bridge between the users and the data server via service protocol. The services layer will expose and share the RDF COVID-19 dataset and RDF links dataset transformed into a well-organized LOD structure through a SPARQL endpoint. The SPARQL is a Declarative Query Language (like SQL) for data manipulation and definition of data represented in RDF triples. A SPARQL endpoint is a point of presence on the Web (identifiable by a URL) that can receive and process SPARQL Protocol requests (federated SPARQL queries) from applications to query RDF data and retrieve data. The representation of data in RDF format and exposure through a SPARQL endpoint makes the type of storage used unimportant from the SPARQL query perspective. For this work, the Fuseki [35] is used as SPARQL endpoint along with triplestore.

4.2.4. Client Layer. The client layer enables users to interact with the system using retrieval and visualization tools. The coherent LOD visualization is an essential part of any LOD-based system to enable users to utilize the Web of Data and enhance the accessibility and usability of a system [36–38]. The LOD-based visualization is aimed to transform and present RDF COVID-19 data into visual presentations to enable users to explore and use the data [13]. In addition, a flexible methodology is needed for healthcare data visualization to address users with unique backgrounds and requirements. The RDF data model can provide opportunities by enabling to attach data to presentation in unanticipated and dynamic ways [36]. To take advantage of the flexibility provided by the RDF, the system permits users to visualize RDF COVID-19 data via a web-based visualization interface called Sgvizler [39]. The Sgvizler is a JavaScript library for rendering results of the SPARQL queries. The Sgvizler requires users to specify target SPARQL endpoint URL, dataset, and write federated SPARQL queries to define the information to be displayed. The Sgvizler supports some visualization methods to render the retrieved data in different modalities such as area chart, timeline, pie chart, bubble chart, geo map, treemap, etc. A user must select the visualization method before executing a SPARQL query and the LOCD system support all e visualization methods available in Sgvizler.

5. Use-Cases and Results

As discussed earlier, the LOCD system integrates COVID-19 data from disparate providers and makes it available in a structured RDF format on the LOD Cloud to enable users to access and query uniformly using a structured query language (i.e., SPARQL). The concept of LOD provides additional information and services to enable end-users to solve complex use-case scenarios, which are previously impossible using isolated datasets. The COVID-19 datasets from the selected sources are used and passed through several on-hand SW technologies for transformation into structured RDF format and linked with LOD Cloud. Our system is expected to support several potential use-case scenarios in the different applications. However, in the following subsections, we will present some example visualization use-case scenarios to illustrate the capabilities of the LOD nature of the RDF COVID-19 dataset in our system.

5.1. Time Line Visualization. A potential use-case scenario could be to monitor and perform a comparative analysis of the different COVID-19 cases (e.g., morbidity, mortality, etc.) statistical information of a country in timeline visualization. The timeline visualization could be helpful to show

patterns and trends, which could not be readily apparent in the numbers themselves. The timeline visualization from traditional databases is cumbersome and time-consuming due to the requirement of explicit external applications. However, the Sgvizler provides users to visualize datasets in the LOD cloud in a timeline by specifying SPARQL query and chart type. Figure 4 shows the SPARQL query to retrieve different USA COVID-19 cases statistics (i.e., total positive cases, new positive cases, total death cases, new death cases) from 01/05/2020 to 01/10/2020 (prefixes are omitted for brevity). Figure 5 shows timeline visualization of the SPARQL query.

5.2. Percentage-Wise Comparative Visualization. Another potential use-case scenario could be to find the percentage proportion of a particular COVID-19 statistical information (i.e., the total number of positive cases, the total number of deaths, total number of recovered, the total number of hospitals, total number of beds in hospitals, etc.) among the countries in a particular continent. The pie chart is used to display percentage or proportional data of the different categories in different slices of pie. Figure 6 shows the SPARQL query to retrieve the total number of COVID-19 positive cases up to 20/01/2021 in the different countries of Europe (prefixes are omitted for brevity). Figure 7 shows the pie chart visualization of the data retrieved by the SPARQL query.

5.3. Column Chart Comparative Visualization. Another possible use-case scenario could be to compare and analyze the number of new positive COVID-19 cases detected or number of COVID-19 death cases or number of COVID-19 recovered cases in the different countries on a particular date. The column chart visualizes data by representing each category by a rectangle bar and the height of a rectangle bar is proportional to the value. Figure 8 shows the SPARQL query to retrieve the total number of new COVID-19 positive cases detected in Russia, France, India, the USA, and Great Britain on 10/01/2021 (prefixes omitted for brevity). Figure 9 shows the column chart visualization of the data retrieved by the SPARQL query.

5.4. Map-Based Visualization. Another possible use-case scenario could be location-based COVID-19 impacts visualization using a map. Location-based visualization is gaining prime importance in health data [31]. However, location-based visualization in traditional database systems requires exporting data into Geographic Information Systems (GIS) for analysis. The location-based visualization is facilitated by our LOCD system. The Sgvizler enables a user to input SPARQL query and select the map visualization and displays the query results on a map using Google Map services. Figure 10 shows the SPARQL query to retrieve location information of countries in Europe and their COVID-19 impacts statistical information of 27/06/2020 in a chart image and URL of the related web page (prefixes omitted for brevity). Figure 11 shows a snippet of the

```
SELECT ?mydate (xsd:integer(?newcases) as ?New_Cases_USA)
          (xsd:integer (?totalcases) as ?Total_Cases_USA)
          (xsd:integer (?newdeaths) as ?New_Deaths_USA)
          (xsd:integer (?totaldeaths) as ?Total_Deaths_USA)
WHERE {
      ?node dbpprop:countryCode ? countrycode.
      ?node dc:date ?mydate.
      ?node covid:total_positive_cases ?totalcases.
      ?node covid:new_positive_cases ?newcases.
      ?node covid:total_death_cases ?totaldeaths.
      ?node covid:new_death_cases ?newdeaths.
      FILTER (?mydate > "2020-05-01" &&
                ?mydate <= “2020-10-01”).
      FILTER regex (?countrycode, "USA") .
} ORDER BY ?mydate
```

Figure 4: SPARQL query to retrieve different USA COVID-19 statistics.

location-based map visualization of the data retrieved by the SPARQL query. Clicking the location marker of a country displays a popup with a country's some basic information, COVID-19 impacts statistics in graphical format, and a hyperlink to the related web page.

6. Evaluation and Discussion

To date, no widely agreed methodology or automatic tools are proposed by the researchers to evaluate LOD-based systems. Therefore, the effectiveness of the LOCD methodology and system is evaluated quantitatively by conducted an empirical study to take feedback about its practicality, usability, information organization, and availability aspects. The practicality is defined in the availability and conceptuality of the required tools. Usability is defined in usage, simplicity, system structure, and user satisfaction. The information organization is defined in information structure, the semantic definition of concepts and relationships, flexibility, explicitness, and fine-grained retrieval. The availability is defined in omnipresent and easy access, multiple use-cases, and swift retrieval. A sample of 20 volunteer participants with technical background is selected consisting of officials working on the COVID-19 pandemic, professional software developers from industry, and users. However, the threshold for sample selection is that participants selected must have substantial knowledge of computer science. Out of the 20 participants selected, 7 (35%) are the officials responsible for COVID-19 pandemic monitoring and policymaking, 9 (45%) are professional web and information systems, developers, with LOD experience and knowledge, and 4 (20%) are the health professionals and web content generators. The sample size is kept small due to constraints of time and lack of resources availability. However, the sample size is large enough to meet requirements of the tests conducted to derive conclusions. In addition, to perform a justified study, a comprehensive 3-days training workshop is conducted to educate the participants about the LOCD methodology and system proposed in this study.

A questionnaire is developed to collect data from the participants for the empirical study. The questionnaire and descriptive analysis methodology are developed using the

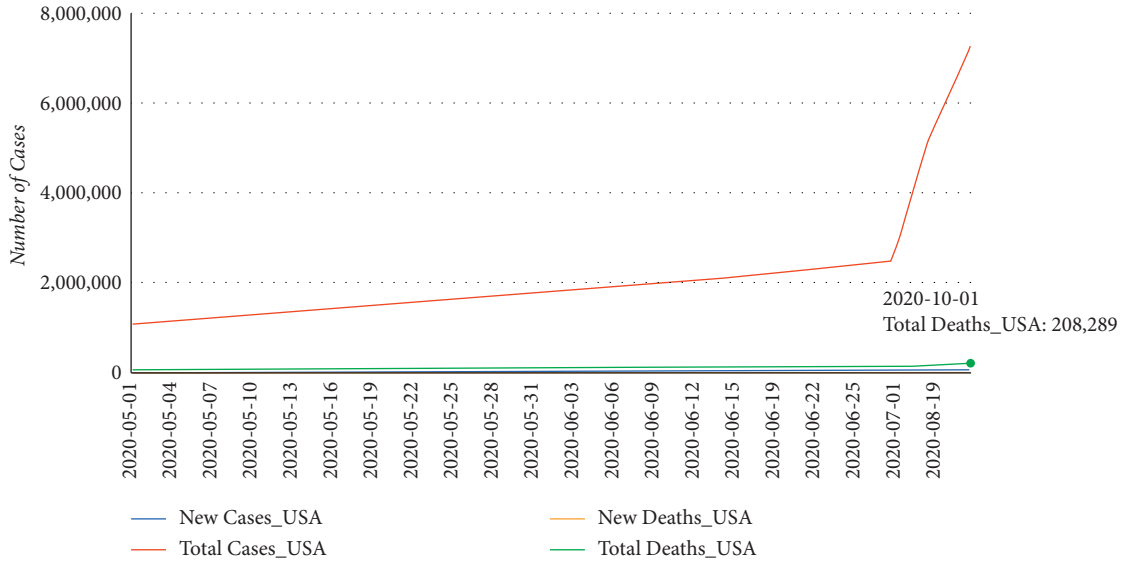

Figure 5: Timeline visualization of the SPARQL query to retrieve different USA COVID-19 statistics.

```
SELECT ?countryname xsd:integer (?totalPositiveCases)
WHERE {
        ?node dbpprop:countryName ?countryname.
        ?node covid:total_positive_cases ?totalPositiveCases.
        ?node covid:continent ?continent.
        ?node dc:date ?mydate.
        FILTER regex (?mydate, "2021-01-20") .
        FILTER regex (?continent, "Europe").
        FILTER (?totalPositiveCases > 300000).
}
```

Figure 6: SPARQL query to retrieve the total number of COVID-19 cases in different countries of Europe.

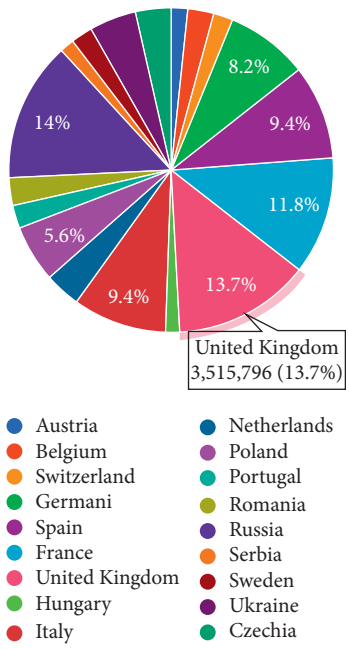

Figure 7: Pie chart visualization of the SPARQL query.

```
SELECT ?countryname (xsd:integer(?newPositiveCases) as
        ?New_Positive_Cases)
WHERE {
      ?node dbpprop:countryName ?countryname.
      ?node dbpprop:countryCode ?ccode.
      ?node covid:new_positive_cases ?newPositiveCases.
      ?node dc:date ?mydate.
      FILTER regex (?mydate, "2021-01-10").
      FILTER (regex (?ccode, "RUS") || regex (?ccode, "FRA") ||
            regex (?ccode, "IND") || regex (?ccode, "USA") ||
            regex (?ccode, "GBR")).
}
```

Figure 8: SPARQL query to retrieve a total number of new COVID-19 cases in Pakistan, France, India, Bangladesh, and the United States.

knowledge and help from prior research in the data acquisition and elicitation domain, and data analysis experts. In addition, the questionnaire is designed using the guidelines proposed by Pew Research Center. Firstly, the questionnaire is composed of 25 questions (both positive and negative) consisting of 10 questions from the pre-defined System Usability Scale (SUS) [40] templet and 15 questions explicitly developed keeping in view the aim and goals of the research. However, the questions are comprehensive enough and fairly distributed to evaluate the effectiveness of the methodology and system. Secondly, the questions in the questionnaire are propositions whose answers are selected from a 5-level Liker-Scale, ranging from 1 = "strongly disagree" to 5 = "strongly agree." The closed-ended questions are used because of being economical to gather a large amount of research data at relatively lowest costs, easy conversion of the collected data from respondents into quantitative data for statistical analysis, and standardized questions by asking respondents the same questions in the same order. Thirdly, the questions are phrased in easily understandable words and arranged in a logical order,

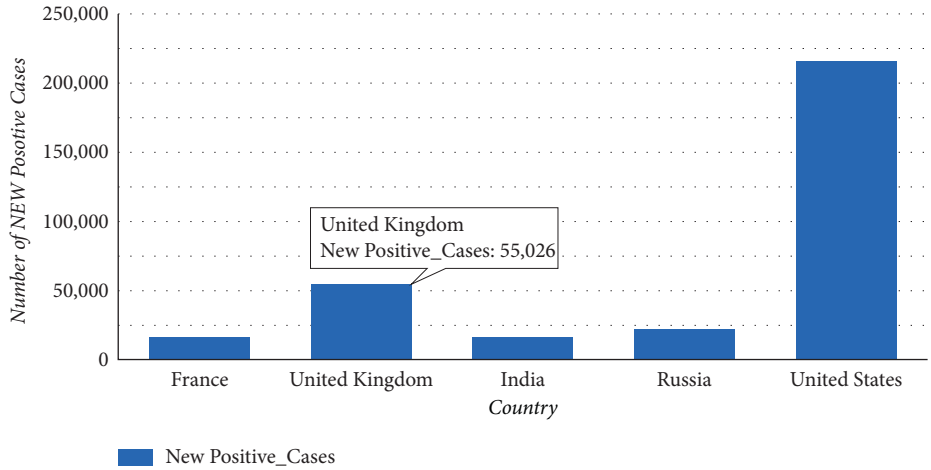

Figure 9: Column chart visualization of the SPARQL query.

```
SELECT    xsd:decimal (?lat) xsd:decimal (?long) ?countryname ?label ?url ?image
WHERE {
          ?node dbpprop:countryName ?countryname.
          ?node covid:continent ?continent.
          ?node dc:date ?mydate.
          ?node wgs84:lat ?lat.
          ?node wgs84:long ?long.
       OPTIONAL{
                ?node covid:hasLabel ?label.
                ?node schema:contentUrl ?image.
                ?node schema:url ?url.
                }
          FILTER regex (?mydate, "2020-06-27").
          FILTER regex (?continent, "Europe").
}
```

Figure 10: SPARQL query to retrieve location information of countries in Europe and their COVID-19 statistics.

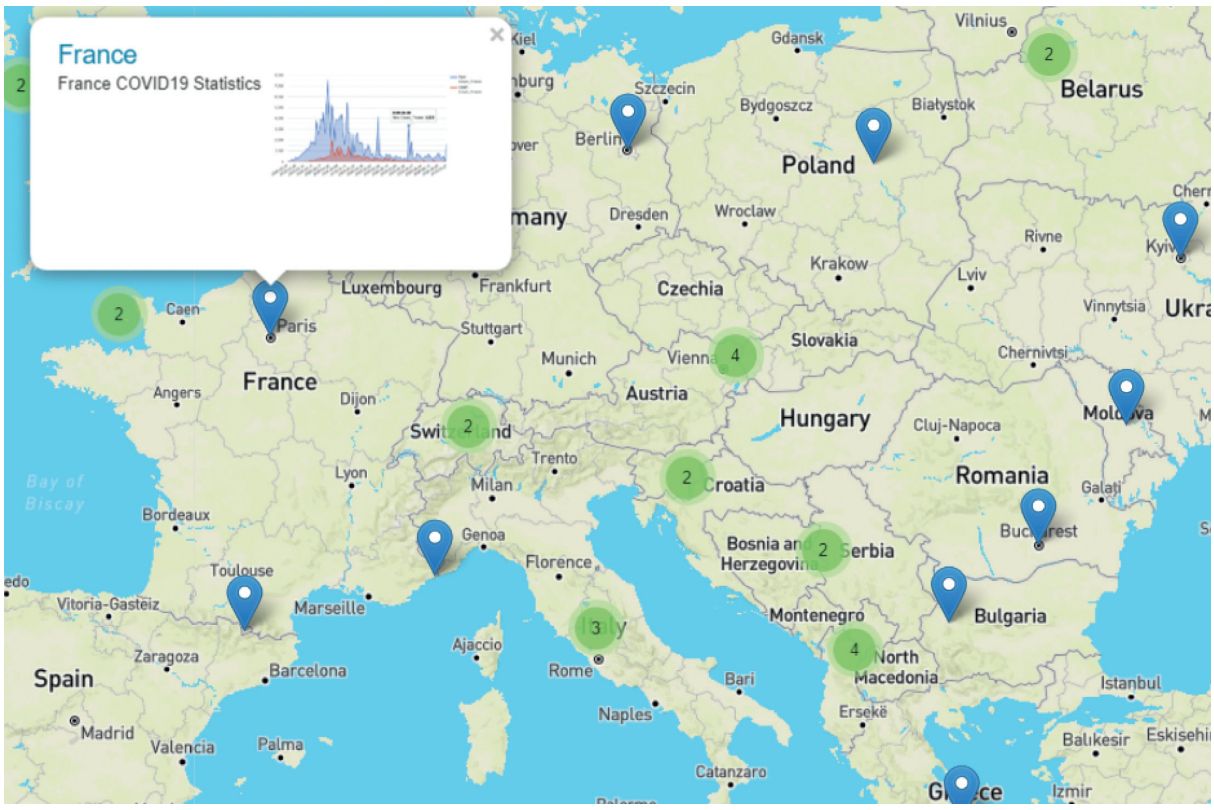

Figure 11: Map-based visualization of the SPARQL query.

while keeping in view the cognitive and sensory competencies of the participants to avoid any misunderstanding and ambiguities. Fourthly, a pilot test is performed to identify and solve issues in readability, understanding, and overall management of the questionnaire before pushing it to the participants. Finally, to evaluate the internal consistency of the questionnaire, Cornbach's alpha test is performed using the collected data. Cornbach's alpha coefficient is 0.726 which is higher than an acceptable threshold (i.e., 0.7) and suggesting that items in the questionnaire have a relatively acceptable level of internal consistency. The data is collected from the participants using the questionnaire with the best efforts of keeping transparency and fairness and avoid conflicting interests. Using responses of the participants, the effectiveness of the methodology and system is evaluated using the SUS and descriptive statistics analysis methods, which are discussed in the following sub-sections.

6.1. System Usability Scale Evaluation. The System Usability Scale (SUS) [40] is commonly regarded as a "quick and dirty" method that provides a quick and easy way for usability assessment of any system and has been evaluated as valid and reliable by a large number of studies (greater than 600) [1, 41]. The SUS is a common method used for web sites usability evaluation; therefore, selected for LOCD as being a web-based application. The SUS questions template is included in the questionnaire to collect data from the participants to evaluate the usability of the LOCD methodology and system. Table 3 shows the participants' responses to the SUS questions in the questionnaire and Table 4 shows the SUS score calculation using the participants' responses. Using the SUS scoring criterion describe in [40] and shown in equations 1, 2, and 3 respectively, the final calculated average SUS score is 70.3, which is higher than the average score ($70.3 > 68$) and ranked in grade "B" with adjective rating "Good." Conclusively, the LOCD methodology and system are found acceptable to the participants from the usability aspect.

$$X = \text{Sum of the points for all odd numbered questions} - 5, \tag{1}$$

$$Y = 25 - \text{Sum of the points for all even numbered questions}, \tag{2}$$

$$SUS\ \text{Score} = (X + Y) * 2.5. \tag{3}$$

6.2. Descriptive Statistics Evaluation. To evaluate the effectiveness of the proposed methodology and system from other aspects (i.e., practicality, organization, and availability), the questionnaire is included 15 questions. An example question is "I think the reference OntoCOVID ontology used provides enough constructs (i.e., concepts and properties) to model COVID-19 data of the different countries?," and "I felt it easy to define new use-cases using the LOCD system." Table 5 represents cumulative statistical information of the participant's responses to the questions in percentage, mean, and standard deviation. Responses analysis has shown that participants responded to the questions with 12.3% strongly disagree, 34.3% disagree, 2.0% neutral, 35.3% agree, and 16.0% strongly agree. Concerning the options numerical values, the overall mean value is 3.08, and the standard deviation value is 1.352. However, due to the lack of availability of standard methods for evaluating a LOD-based system from the above-mentioned aspects, the LOCD methodology and system are evaluated using the descriptive statistical analysis method.

The Likert-Scale data is ordinal, where the order of the values is significant but the exact difference between the values is not known. The Chi-Square is an important descriptive statistical method for the analysis of the categorical data [29, 31]. To analyze the ordered scale 5 levels Likert-scale responses data using Chi-Square descriptive statistics: (1) the negative questions are transformed into positive questions and the participants' responses are adjusted accordingly for uniformity (2) the five response categories (i.e., strongly disagree, disagree, neutral, agree, and strongly agree); are breakdown into two nominal categories (i.e., disagree and agree) by combining the lower level three categories and upper level two categories respectively. The Chi-Square test is executed on the nominal categories using Statistical Package for the Social Sciences (SPSS) 16.0 to show the effectiveness of the LOCD methodology and system. The null and alternative hypotheses are:

H_0: The LOCD methodology and system of using the LOD approach to represent and visualize COVID-19 data is not significantly effective for the healthcare planners and decision-makers.

H_1: The LOCD methodology and system of using the LOD approach to represent and visualize COVID-19 data is significantly effective for the healthcare planners and decision-makers.

Results collected after performing the Chi-Square test on the dataset are shown in Table 6. The top row in the table shows Pearson Chi-Square statistics $\chi^2 = 8.000E2$ and p-value <0.001. Since the p-value is less than the level of significance ($\alpha = 0.05$); therefore, the null hypothesis (H_0) is rejected. The alternative hypothesis (H_1) stands true, which signifies that the LOCD methodology and system of using the LOD approach to represent and visualize COVID-19 data has significant importance for healthcare planners and decision-makers.

Table 3: Participants' responses to the SUS questions for usability evaluation.

Q#	SUS evaluation question	Strongly Disagree-1, *n* (%)	Disagree-2, *n* (%)	Neutral-3, *n* (%)	Agree-4, *n* (%)	Strongly Agree-5, *n* (%)
Q1	I think that I would like to use the LOCD system frequently.	0 (0)	1 (5)	0 (0)	10 (50)	9 (45)
Q2	I found LOCD system unnecessarily complex.	5 (25)	9 (45)	0 (0)	5 (25)	1 (5)
Q3	I thought the LOCD system was easy to use.	2 (10)	6 (30)	0 (0)	7 (35)	5 (25)
Q4	I think that I would need the support of a technical person to be able to use the LOCD system.	6 (30)	8 (40)	0 (0)	6 (30)	0 (0)
Q5	I found the various functions in the LOCD system were well integrated.	0 (0)	5 (25)	1 (5)	7 (35)	7 (35)
Q6	I thought there was too much inconsistency in the LOCD system.	7 (35)	10 (50)	0 (0)	3 (15)	0 (0)
Q7	I would imagine that most people would learn to use the LOCD system very quickly.	0 (0)	2 (10)	1 (5)	12 (60)	5 (25)
Q8	I found LOCD system very cumbersome to use.	6 (30)	11 (55)	0 (0)	3 (15)	0 (0)
Q9	I felt very confident using LOCD system.	1 (5)	4 (20)	1 (5)	10 (50)	4 (20)
Q10	I needed to learn a lot of things before I could get going with the LOCD system.	5 (25)	8 (40)	2 (10)	5 (25)	0 (0)

Table 4: SUS score calculation using the participants' responses.

Participant	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	X	Y	SUS score
P1	4	2	4	1	4	1	4	2	4	2	15	17	80
P2	5	1	5	2	5	2	4	2	4	3	18	15	82.5
P3	4	2	4	1	4	1	5	2	5	4	17	15	80
P4	5	2	5	1	5	2	4	1	2	1	16	18	85
P5	4	1	5	1	5	2	4	1	5	2	18	18	90
P6	4	4	2	2	4	4	2	4	4	4	11	7	45
P7	4	2	1	1	4	2	4	2	4	2	12	16	70
P8	5	4	2	4	5	2	4	2	4	3	15	10	62.5
P9	4	2	4	2	2	1	5	2	4	4	14	14	70
P10	2	2	4	2	3	4	2	2	3	4	9	11	50
P11	4	5	2	4	4	2	4	1	2	2	11	11	55
P12	4	2	1	4	2	2	4	4	1	1	7	12	47.5
P13	5	4	2	4	2	1	4	2	4	1	12	13	62.5
P14	5	1	5	1	4	2	5	2	2	2	16	17	82.5
P15	5	1	5	2	5	4	3	1	4	1	17	16	82.5
P16	5	2	4	2	5	2	5	1	4	1	18	17	87.5
P17	4	2	4	2	2	1	5	2	5	2	15	16	77.5
P18	4	4	2	4	4	2	4	4	2	2	11	9	50
P19	5	1	4	2	5	1	4	2	5	4	18	15	82.5
P20	5	4	2	4	2	1	4	1	4	2	12	13	62.5
Average SUS score													70.3

Table 5: Cumulative participant's responses to the questions in the questionnaire.

Questions ∗ sample	Likert-scale options	Frequency	Percentage (%)	Mean	Standard deviation
15 ∗ 20	Strongly disagree = 1	37	12.3%	3.08	1.352
	Disagree = 2	103	34.3%		
	Neutral = 3	6	2.0%		
	Agree = 4	106	35.3%		
	Strongly agree = 5	48	16.0%		
	Total	300	100.0		

Table 6: Results of the Chi-Square test using SPSS 16.0.

Descriptive statistics	Value	df	*p*-Value
Pearson chi-square (χ^2)	3.000E2[a]	4	0.000
Likelihood ratio	415.675	4	0.000
Linear-by-linear association	261.028	1	0.000
N of valid cases	300		

[a] 2 cells (20.0%) have an expected count of less than 5. The minimum expected count is 2.92.

7. Conclusion and Future Work

One of the most promising fields of research in recent years is data representation, storage, and retrieval. The availability of many web-based datasets has demanded the realization of data management techniques for the available scattered datasets on the Web. The LOD paradigm provides a set of practices for publishing and interlinking open data on the Web using SW technologies and enables the transformation of the Web of documents into a Web of data. The interlinked nature will enable the Web as a distributed network for datasets and access, useable by software agents and machines, and provides end-users with a new spectrum of use-case scenarios.

The COVID-19 pandemic has produced serious impacts socially and economically. This has posed serious challenges for governments and organizations in their planning and decision-making processes. The data is critical for the decision-makers to counter the pressures of cost reduction, improved coordination and outcome, and produce more with less. The collection and integration of COVID-19 data from the distributed sources are performed manually by the domain experts, which is difficult, time-consuming, and error-prone. The healthcare domain has already adopted the SW technologies and LOD principles to enhance information retrieval and visualization processes, which are not possible with separate data stores. In this research work, we have identified the designing and development of the LOCD system for transforming and publishing 2-start disparate COVID-19 datasets into 5-star RDF COVID-19 datasets using the available SW technologies and LOD principles to provide promising opportunities to the end-users. Reference ontology OntoCOVID is used for COVID-19 data modeling and RDF is used for data representation using the ontological model. The Silk framework is used for link discovery and integration of data from disparate COVID-19 datasets. The Fuseki triple store is used for data storage and the SPARQL endpoint. The Sgvizler is used for querying and visualization of the RDF COVID-19 dataset using the SPARQL query interface. A few potential use-case scenarios are presented to demonstrate the validity and application of the system by the stakeholder while developing applications and services. The system is evaluated empirically using the system usability scale and Chi-Square descriptive statistical analysis method, where both have resulted in the effectiveness of the LOCD methodology and system from practicality, usability, organization, and availability aspects. Conclusively, it if found that SW and LOD technologies could be a potential option for building LOD-based COVID-19 information systems.

This research work has some limitations, which will be covered in our future research. Firstly, the system has shown satisfactory results with the available data size and sample queries, but the COVID-19 data size is increasing day by day. Therefore, more robust and rigorous evaluations in different aspects (i.e., performance and query results, etc.) are needed before its widespread adaptations. As the SW technologies reportedly work poorly with large datasets. Secondly, although the OntoCOVID ontology provides extensive domain-specific concepts and properties to robustly model data in the disparate COVID-19 datasets. However, it is recommended to extend OntoCOVID ontology with additional concepts and properties and interlinking with relevant ontologies and vocabularies. Finally, the evaluation is done with a relatively small set of participants due to limitations of time and resources; therefore, evaluation with a large set of participants is needed to derive more accurate and reliable results.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

All the authors contributed equally.

Acknowledgments

This research study was carried as a joint research collaboration of the University of Peshawar Pakistan, University of Engineering and Technology, Taxila, Pakistan, Islamia College, Peshawar, Pakistan, and Swedish College of Engineering and Technology, Wah Campus, Pakistan. The authors received no specific funding for this study.

References

[1] B. Tilahun, T. Kauppinen, C. Keßler, and F. Fritz, "Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation," *JMIR medical informatics*, vol. 2, no. 2, p. e31, 2014.

[2] S. Bonacina, "Linked open data in health and clinical care A review of the literature," *European Journal of Biomedical Informatics*, vol. 12, no. 2, 2016.

[3] W3C, "W3C Semantic Web Health Care and Life Sciences Interest Group Charter," 2020, https://www.w3.org/2011/09/HCLSIGCharter.

[4] X. Zenuni, B. Raufi, F. Ismaili, and J. Ajdari, "State of the art of semantic web for healthcare," *Procedia - Social and Behavioral Sciences*, vol. 195, pp. 1990–1998, 2015.

[5] T. Heath and C. Bizer, "Linked data: evolving the web into a global data space," *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 1, no. 1, pp. 1–136, 2011.

[6] S. Khusro, S. Ali, A. Rauf et al., "Unleashing sensor data on linked open data - the story so far," *Life Science Journal*, vol. 10, no. 4, pp. 1766–1786, 2013.

[7] M. Jovanovik, B. Najdenov, G. Strezoski, and D. Trajanov, "Linked open data for medical institutions and drug availability lists in Macedonia," *Advances in Intelligent Systems and Computing*, vol. 312, pp. 245–256, 2015.

[8] B. P. Reddy, B. Houlding, L. Hederman et al., "Data linkage in medical science using the resource description framework: the AVERT model," *HRB Open Res*, vol. 59, 2018.

[9] Data Cloud, "The Linked Open Data Cloud," 2020, https://lod-cloud.net/.

[10] K.-H. Cheung, E. Prud'hommeaux, Y. Wang, and S. Stephens, "Semantic web for health care and life sciences: a review of the state of the art," *Briefings in Bioinformatics*, vol. 10, no. 2, pp. 111–113, 2009.

[11] T. Berners-Lee, J. Hendler, and O. Lassila, "The semantic web," *Scientific American*, vol. 284, no. 5, pp. 34–43, 2001.

[12] F. Shen and Y. Lee, "Biobroker: knowledge discovery framework for heterogeneous biomedical ontologies and data," *Journal of Intelligent Learning Systems and Applications*, vol. 10, no. 1, pp. 1–20, 2018.

[13] R. Reda, "Carbonaro A design and development of a linked open data-based web portal for sharing IoT health and fitness datasets," in *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, pp. 43–48, Bologna Italy, November, 2018.

[14] C. Bizer, T. Heath, and T. Berners-Lee, *Semantic Services, Interoperability and Web Applications*, pp. 205–227, IGI Global, Hershey, Pennsylvania, 2011.

[15] D. Abián, F. Guerra, J. Martínez-Romanos, and T.-L. R Wikidata, "DBpedia: A comparative study," in *Semanitic Keyword-Based Search on Structured Data Sources*, pp. 142–154, Springer, New York, NY, USA, 2017.

[16] D. Brickley and R. V. Guha, "RDF Schema 1.1," 2014, https://www.w3.org/TR/rdf-schema/.

[17] W. C. O. W. Group, "OWL 2 Web Ontology Language Document Overview," 2012, https://www.w3.org/TR/owl2-overview/.

[18] E. Prud'hommeaux and A. Seaborne, "SPARQL Query Language for RDF," 2008, https://www.w3.org/TR/rdf-sparql-query/.

[19] A. Ruttenberg, J. A. Rees, M. Samwald, and M. S. Marshall, "Life sciences on the semantic web: the neurocommons and beyond," *Briefings in Bioinformatics*, vol. 10, no. 2, pp. 193–204, 2009.

[20] B. Vandervalk, L. McCarthy, and M. Wilkinson, "Cardio-SHARE: web services for the semantic web," *Semantic Web Challenge*, vol. 8, 2008.

[21] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2RDF: towards a mashup to build bioinformatics knowledge systems," *Journal of Biomedical Informatics*, vol. 41, no. 5, pp. 706–716, 2008.

[22] A. Zaveri, J. Lehmann, S. Auer, M. M. Hassan, M. A. Sherif, and M. Martin, "Publishing and interlinking the global health observatory dataset," *Semantic Web*, vol. 4, no. 3, pp. 315–322, 2013.

[23] V. Law, C. Knox, Y. Djoumbou et al., "DrugBank 4.0: shedding new light on drug metabolism," *Nucleic Acids Research*, vol. 42, no. D1, pp. D1091–D1097, 2014.

[24] O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang, "LinkedCT: A Linked Data Space for Clinical Trials," 2009, https://arxiv.org/ftp/arxiv/papers/0908/0908.0567.pdf.

[25] A. Jentzsch, J. Zhao, O. Hassanzadeh, K.-H. Cheung, M. Samwald, and B. Andersson, "Linking open drug data," *I-SEMANTICS*, 2009.

[26] M. Samwald, A. Jentzsch, C. Bouton et al., "Linked open drug data for pharmaceutical research and development," *Journal of Cheminformatics*, vol. 3, no. 19, pp. 19–26, 2011.

[27] M. Jovanovik, B. Najdenov, and D. Trajanov, "Linked Open Drug Data from the Health Insurance Fund of Macedonia," in *Proceedings of the 10th Conference for Informatics and Information Technology (CIIT)*, Bitola Macedonia, April, 2013.

[28] N. Gür, L. Díaz, and T. Kauppinen, "Gi systems for public health with an ontology based approach," in *Proceedings of the 15th AGILE International Conference on Geographic Information Science (AGILE2012)*, pp. 86–91, France, April, 2012.

[29] S. Ali, S. Khusro, I. Ullah, A. Khan, and I. Khan, "SmartOntoSensor: ontology for semantic interpretation of smartphone sensors data for context-aware applications," *Journal of Sensors*, vol. 2017, pp. 1–26, Article ID 8790198, 2017.

[30] S. Ali, S. Khusro, S. Anwar, and A. O.C. O. V. I. D. Ullah, "Ontology for semantic modeling of COVID19 statistical data," in *Proceedings of the 15th International Conference on Information Technology and Applications (ICITA 2021)*, Dubai, UAE, November, 2021.

[31] S. Ali and S. Khusro, "POEM: practical ontology engineering model for semantic web ontologies," *Cogent Engineering*, vol. 3, no. 1, Article ID 1193959, 2016.

[32] N. Andes and J. E. Davis, "Linking public health data using geographic information system techniques: alaskan community characteristics and infant mortality," *Statistics in Medicine*, vol. 14, no. 5-7, pp. 481–490, 1995.

[33] M. L. Pesce and K. K. Breitman, "Casanova MA Surfacing scientific and financial data with the Xcel2RDF plug-in," in *Proceedings of the 2012 Second International Workshop on Developing Tools as Plug-Ins (TOPI)*, pp. 73–78, IEEE, Zurich, Switzerland, June, 2012.

[34] A. Jentzsch, R. Isele, and C. Bizer, "Silk-generating rdf links while publishing or consuming linked data," in *Proceedings of the 9Th International Semantic Web Conference (ISWC'10)*, Citeseer, Shanghai, China, November, 2010.

[35] J. A. A. Jena, "Jena Fuseki," 2020, https://jena.apache.org/documentation/fuseki2/.

[36] J. M. Brunetti, S. Auer, R. García, and J. Klímek, "Nečaský M Formal linked data visualization model," in *Proceedings of the International Conference on Information Integration and Web-Based Applications & Services*, pp. 309–318, New York, NY, USA, November, 2013.

[37] G. Kopanitsa, C. Hildebrand, J. Stausberg, and K. H. Englmeier, "Visualization of medical data based on EHR standards," *Methods of Information in Medicine*, vol. 52, no. 01, pp. 43–50, 2013.

[38] J. M. Brunetti, S. Auer, and R. García, "The linked data visualization model," in *Proceedings of the International Semantic Web Conference (Posters & Demos)*, Boston, USA, January, 2012.

[39] S. M. G. Sgvizler, "A javascript wrapper for easy visualization of sparql result sets," in *Proceedings of theExtended Semantic Web Conference*, pp. 361–365, Heraklion Crete Greece, April, 2012.

[40] J. Brooke, "SUS: a "quick and dirty'usability"," *Usability Evaluation in Industry*, vol. 189, no. 3, p. 189, 1996.

[41] S. Mazumdar, D. Petrelli, and F. Ciravegna, "Exploring user and system requirements of linked data visualization through a visual dashboard approach," *Semantic Web*, vol. 5, no. 3, pp. 203–220, 2014.