OntoCOVID

OntoCOVID: Ontology for Semantic Modelling of COVID19 Statistical Data

Shaukat Ali[1], Shah Khusro[2], Sajid Anwar[3], Abrar Ullah[4]

[1, 2] Department of Computer Science, University of Peshawar, Pakistan
[3] Centre of Excellence in IT, Institute of Management Sciences (IMSciences), Pakistan
[4] School of Mathematical and Computer Science, Heriot-Wat University, UK

Abstract Several COVID19 statistical datasets are provided to support stakeholders for better planning and decision-making in healthcare. However, the datasets are in heterogeneous proprietary formats that create data silos and compatibility issues and make data discovery and reuse difficult. Further, the data integration for analysis is difficult and is performed by the domain experts manually which is time consuming and error prone. Therefore, an explicit, flexible, and widely acceptable methodology is needed to represent, store, query, and visualize COVID19 statistical data in the datasets. In this paper, we have presented the design and development of OntoCOVID ontology for representing, organizing, sharing and reusing COVID19 statistical data in the datasets. The OntoCOVID is a lightweight ontology providing definitions of concepts, properties and axioms to semantically represent and relate information in the COVID19 statistical datasets. The OntoCOVID is evaluated to demonstrate its completeness and information retrieval for different use-case scenarios. The results obtained are promising and advocates for the improved ontological design and applications of the OntoCOVID.

Keywords Semantic Web, COVID19, Ontology, SPARQL, OWL, Dataset

1 Introduction

Information is the fundamental requirement in the public healthcare domain for effective decision-making regarding epidemiologic surveillance, disease outbreak, healthcare program planning and outcomes assessment, cost and expenditures, coverage and quality of services and policy analysis [1]. Like infectious diseases, the COVID19 pandemic has posed serious threats to the general public health and has resulted into significant morbidity, mortality, and economic damages worldwide. The COVID19 statistical data (i.e., patients' cases, clinical trials and drug efficiency) is constantly growing and different organizations are providing datasets under the initiatives of open data on the Web (e.g., ourworldindata, worldometer and kaggle) to support the stake holders. However, the COVID19

statistical datasets (CSDs) are in heterogeneous data formats, naming conventions, structures, encodings, data types and resides in distributed data repositories. Mostly, the CSDs are distributed in tabular formats (i.e., CSV and Excel) with small metadata; making data integration, comparison and reuse very difficult. In addition, the disparate CSDs are not linked despite of having related indicators and vocabularies and data formats used are inconsistent. This diversity, variety and veracity of CSDs poses serious time consuming and error-prone problems of data management (i.e., representation, encoding, storage and retrieval), sharing and integration for effective planning and decision-making by the stakeholders. The explosive growth of CSDs needs to be properly explored to reveal worthy information and convert them into worthy knowledge for improved COVID19 healthcare practices. Therefore, semantically enriched and flexible data representation, analysis, querying and visualization methods are needed for CSDs. Sharing semantically structured CSD would enable researchers to produce advanced methods for providing high quality information to revolutionize the healthcare sector.

The Semantic Web (SW) [2] extends the Web from interlinking documents only to revolutionize global data sharing, integration and analysis [3]. The Health Care and Life Science Interest Group (HCLSIG) is founded with the aim to develop, advocate and support the use of SW technologies in the healthcare and life science domains [4, 5]. The healthcare domain has already adopted the SW technologies principles for enhanced semantic representation, information retrieval (IR) and visualization, which are not possible with separate data stores [6, 7]. The ontology is SW technology which provides constructs to develop structured vocabularies comprising of terms to represent entities and relationships in human and computer interpretable format [8]. In the field of informatics, ontologies provides important platform for data standardization, representation, integration, sharing and analysis. Therefore, ontologies could provide semantic layer by annotating CSD with spatial, temporal and thematic metadata from the disparate sources to enhance semantic understanding, compatibility, interoperability and mapping of relationships between mismatching terms to improve performance of a system. Therefore, CSD ontology is required to provide a common and widely accepted language and dictionary of terminologies for understanding the structure of information regarding countries, patients cases, diagnostic tests, health facilities, population, etc., to provide highly expressive representations, advanced access, reuse domain knowledge, formal analysis of resources and make domain knowledge explicit without having knowledge of technical details regarding data management.

The CSDs heterogeneity is an open challenge. Therefore, an explicit and flexible solution is needed to exploit potential of the CSDs for effective planning and decision making by the stake holders. The objective of this paper is to design and develop OntoCOVID ontology consists of explicit and formal conceptualization of CSD from the heterogeneous COVID19 datasets. The OntoCOVID has numerous characteristics including: (1) semantic representation and annotation

of CSD to increase data interoperability; (2) fusion of multi-sources CSD to support intelligent health care monitoring and decision making; (3) description of CSD to increase data sharing and reusing; and (5) addition of geo-spatial information to support retrieval and visualization systems. The OntoCOVID architecture is kept conservative while keeping the option open for changes, potential reuse, extension and plugging into other COVID19 domain-specific heavy-weight ontologies. This paper presents a pragmatic approach for the development of OntoCOVID with no claim that orthogonal, complete, or universally acceptable CSD ontology is feasible. However, evaluation of OntoCOVID has shown promising results.

2 Related Work

The public healthcare systems and emergencies require data sharing in different domains and data systems to control and reduce the COVID19 impacts. However, this data sharing is hindered by representing and encoding relevant information in domain-specific terminologies and coding systems and resulting into loss of interpretability across data systems. The researchers have recognized the strength of using ontologies to provide solutions to the problems. The COVID19 surveillance ontology facilitates monitoring of COVID19 cases and relates respiratory conditions using data from multiple brands of computerized medical record systems [9]. The COVID19 ontology covers the role of molecular and cellular entities in virus-host-interactions, in the virus life cycle, as well as a wide spectrum of medical and epidemiological concepts linked to COVID19 [10].

The COVID19 is a relatively new disease with its unique prospects. The researchers have explored ontologies to semantically organize and share COVID19 data from medical, therapeutic and epidemiological aspects. However, none of the ontologies addresses CSD related to human cases and fatalities, planning and hospital systems, geography and demographics of countries, etc. As discussed earlier, heterogeneous CSDs are of limited significances for effective decision making and planning processes. Therefore, experiences from COVID19 ontology-based research can be applied to CSDs for semantically representation, organization, interlinking, sharing, reusing and visualization.

3 OntoCOVID Development

The ontology provides a solution to develop a domain data model by formally and explicitly defining its concepts and meaningful data linkages. Therefore, ontology is needed to semantically model and map CSDs for explicit, formal and uniform representation and using in applications. The existing ontologies could be helpful in certain extent to provide matching set of definitions that are fitting the data needs but not completely. Reusing existing ontologies could increase data interoperability but could increase complexity in retrieval by not efficiently

modelling the new data needs [11]. The bespoke ontology could reduce data interoperability but could allow to model data in an efficient manner [11]. Therefore, developing a bespoke ontology using 'in between' approach could be efficient solution by utilizing concepts and properties form well-known existing and accepted ontologies and vocabularies. However, new data needs are to be defined in ontology explicitly.

A bespoke reference "OntoCOVID" ontology is developed for CSDs. The OntoCOVID is developed by using 'in between' approach reusing ontologies and vocabularies (i.e., Schema, Friend of a Friend, DBPedia, Dublin Core and W3C Geospatial). However, the new concepts and properties are explicitly defined in the OntoCOVID to efficiently model the new data needs. A snippet of OntoCOVID is shown in Fig. 1. The OntoCOVID is modelled using concept of episodes where each episode represents CSD data of a particular country on a particular date. Each episode is associated with URIs and literal values. The values are important to numerically quantitate an episode's object, retrieving and reasoning process. Each episode is added with geographical coordinates for useful map-based tracking and visualization as location has become basic attribute for health data [12].

3.1 Goals and Design Principles

The goal and design principles of the OntoCOVID are:

- The ontology should represent CSD for using in variety of informatics systems.
- The ontology describes concepts, properties, relations and axioms in simple and understand manner to be easily and effectively used by the developers.
- The ontology is flexible and extendable to easily add new domain-specific concepts and relationships to enhance interoperability and knowledge sharing.
- The ontology is general by describing concepts, facets and relationships that are possibly present in a wide range of CSDs.
- The ontology supports memory-efficient and time-efficient inference methods.
- The concepts and properties are defined and arranged precisely for ease in access and produce acceptable values for quality and completeness metrics.
- The ontology can be easily used in any of SW framework and triplestore.

3.2 Materials and Methodology

The OntoCOVID is developed using Protégé 5.5.0. The POEM methodology [3] is used, which emphasizes on searching, reusing, reengineering and merging ontological and non-ontological resources and ontology design patterns, which are the main designing rationales of OntoCOVID. The OntoCOVID emphasizes on leveraging upper-level ontologies and vocabularies for providing a common

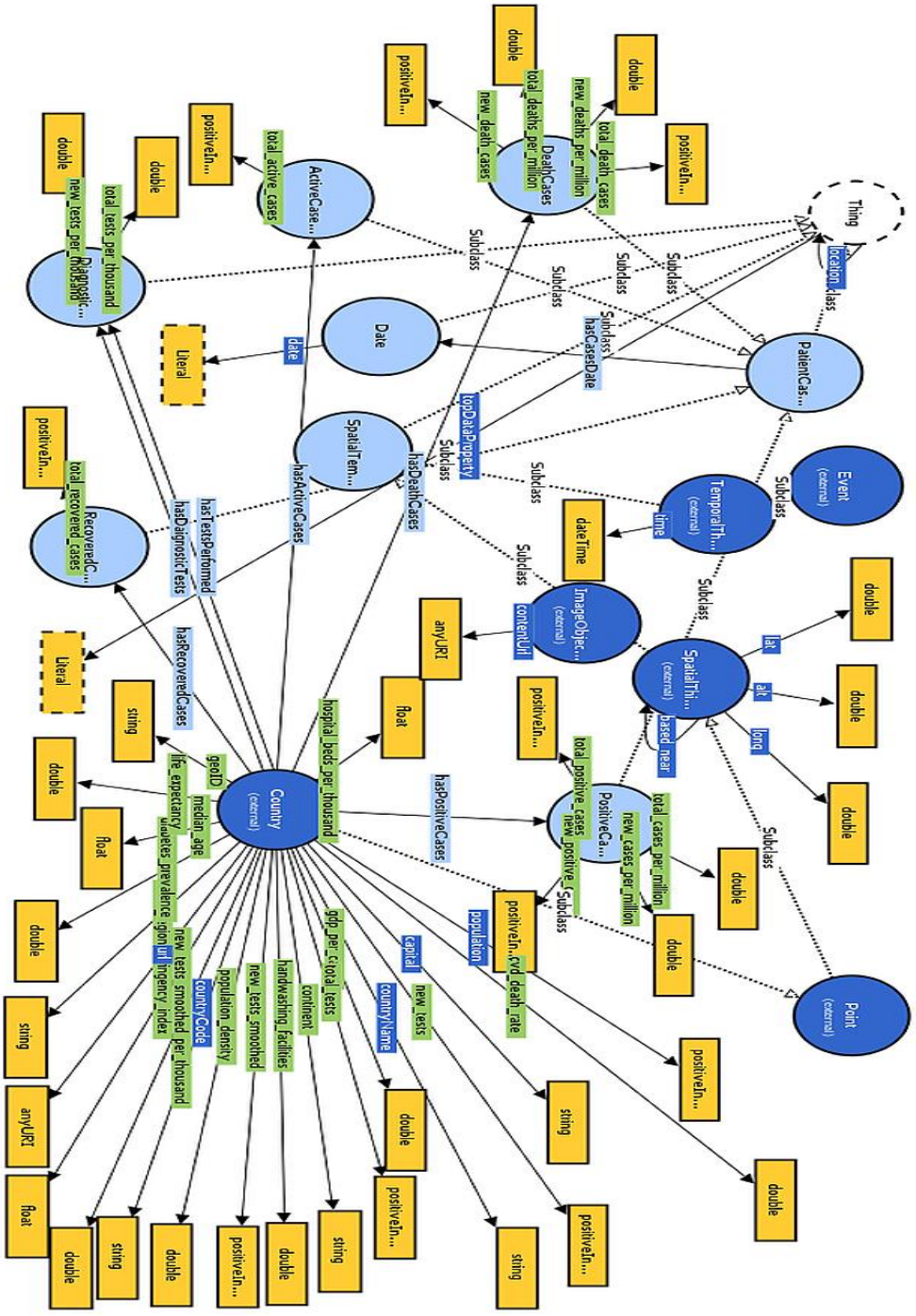

Fig. 1. Snippet of OntoCOVID concepts and properties in Protégé

specific ontology to achieve shorter development cycles, universalism, initial requirements set identification, easier and faster integration with other contents and more stable and robust knowledge systems [3]. The Content Ontology Design pattern [13] is used where contents in OntoCOVID are conceptual and instantiated from logical upper-level ontologies and provide explicit non-logical vocabulary for the domain. Furthermore, Componency Pattern is used for representing concepts and properties as either proper parts of other concepts and properties or having proper parts.

3.3 Requirements Specifications

The OntoCOVID requirements are mainly concerned with functional and non-functional requirements. The non-functional requirements comprise terminological requirements and the naming convention used for the terms. The terminological requirements of OntoCOVID can be broadly divided into: (1) the terms representing concepts and sub-concepts of entities in the domain; (2) the terms representing properties of concepts in the domain, where some properties represent relationship between concepts (object properties) and other represent characteristics of concepts (datatype properties). For example, Country, Positive Cases, Event, Diagnostic Tests, etc., are representing concepts and Country Code, Total Positive Cases, Population, Latitude, Longitude, etc., are representing properties. A lexicon representing relevant set of terminologies is identified and collected from the domain-specific and application-specific resources. The functional requirements at present consists of 43 competency questions representing intended tasks that ontology should perform by executing SPARQL queries such as What is frequency of active cases in 'a country' on a 'date or month'?; What is total number of tests conducted in 'a country/countries' on a 'date or month'?; What is total number of hospital beds available in 'a country' and how many are used on a 'date/month'?; What is relationship of death cases and diabetic patients in 'a country'? The competency questions would help in determining correctness, completeness, verifiability and understandability of ontology requirements [3].

3.4 Concepts Hierarchy

The OntoCOVID concepts are constructed from concepts of countries, patients' cases, spatial and temporal, diagnostic tests, medical facilities, and demographics. The identified concepts are hierarchically arranged by determining their relationships that whether a concept is a sub-concept of another concept or not. The required concepts from upper-level ontologies/vocabularies (e.g., schema:Country, dc:Date, etc.) are directly reused. New concepts are created either parent concepts or sub-concepts of the concepts in the upper-level ontologies/vocabularies.

The Fig. 2 shows a snippet of OntoCOVID taxonomic concepts hierarchy. The OntoCOVID at present contains 23 concepts, where each concept is formed by keeping in mind the unique requirements of the domain. The Onto-COVID:PatientCases is declared to define the COVID19 cases and Onto-COVID:ActiveCases, OntoCOVID:DeathCase, OntoCOVID:PostiveCases, and OntoCOVID: RecoveredCases are declared sub-concepts of Onto-COVID:PatientCases to represent different types of COVID19 cases in a country on a particular date. The dc:Date is used to represent date information and On-toCOVID:DiagnosticTests is declared to represent the COVID19 diagnostic tests conducted in a country. The schema:Country is declared sub-concept of wgs84:Point to represent country information. The wgs84:Event is declared to represent related COVID19 event information and schema:ImageObject repre-sents COVID19 cases graphical information. Apart from using sub-concept axi-oms, concepts are coupled with other axioms to facilitate the creation of individ-uals (i.e., objects) unambiguously and semantically. For example, the disjoint axiom is defined for concepts belonging to the same generation level to restrict individuals' behaviours such that OntoCOVID:ActiveCases and Onto-COVID:DeathCase are disjointed that an individual can be an instance of only one of them.

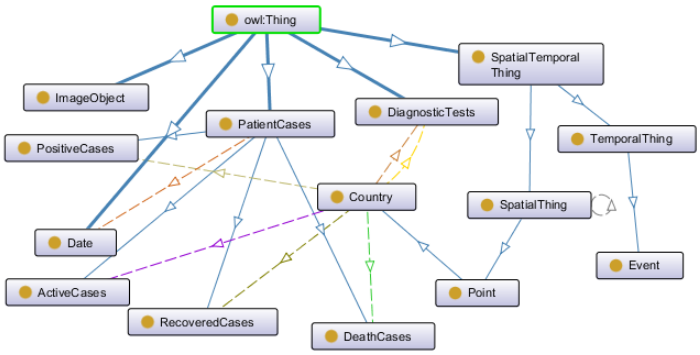

Fig. 2. Snippet of OntoCOVID concepts hierarchy.

3.5 Properties and Restrictions

The properties represent relationships/links and can be either object property or datatype property. For an object property, both the domain and range should be instances of concepts, whereas, for a datatype property domain should be an in-stance of a concept and range should be an instance of a datatype. The object properties are either used from upper-level ontologies/vocabularies or explicitly created. The object properties available in upper-level ontologies/vocabularies are inherited and specialized for using with OntoCOVID concepts. For example, foaf:based_near relates an instance of wgs84:SpatialThing with itself. In addi-

tion, several domain-specific object properties are defined to relate concepts in OntoCOVID in more meaningful and subtle ways such as OntoCOVID:hasDeathCase and OntoCOVID: hasCasesDate, etc., relates instance of schema:Country with instances of respective concepts. The Table 1 shows an excerpt of OntoCOVID object properties. Similarly, the datatype properties in OntoCOVID are either used form upper-level ontologies/vocabularies or explicitly created. For example, wgs84:lat, dbpprop:countryCode, schema:url, OntoCOVID:new_death_ cases, OntoCOVID:new_positive_cases, OntoCOVID:total_tests, dbpprop: population, etc. Table 2 shows an excerpt of OntoCOVID datatype properties. The concepts are refined by using properties to superimpose constraints and axioms for describing their individuals. The property restrictions are used to describe number of occurrences and values of a property for an individual to be an instance of a concept. For example, an individual to be instance of schema:Country, should have at least one occurrence of wgs84:lat and wgs84:lat datatype properties.

Table 1. Excerpt of OntoCOVID object properties.

Domain	Object Property	Range	Quantifier	Cardinality
schema:Country	OntoCOVID:hasActiveCases	OntoCOVID:ActiveCases	Existential	Exactly 1
schema:Country	OntoCOVID:hasRecoveredCases	OntoCOVID:RecoveredCases	Existential	Exactly 1
schema:Country	OntoCOVID:hasTestsPerformed	OntoCOVID:DiagnosticTests	Universal	Max 1
Wgs84:SpatialThing	Foaf:based_near	Wgs84:SpatialThing	Exis & Unil	Min 1
schema:Country	OntoCOVID:hasDataDate	dc:Date	Existential	Max 1
schema:Country	OntoCOVID:hasEvent	Wgs86:Event	Universal	Nil

Table 2. Excerpt of OntoCOVID object properties.

Domain	DataType Property	Range	Quantifier	Cardinality
schema:Country	Wgs84:alt	Double	Existential	Exactly 1
schema:Country	dbo:countryName	String	Existential	Exactly 1
OntoCOVID:DeathCases	OntoCOVID:New_Death_Cases	Positive Integer	Existential	Exactly 1
OntoCOVID:PositiveCases	OntoCOVID:New_Positive_Case	Positive Integer	Existential	Exactly 1
OntoCOVID:DiagnosticTests	OntoCOVID:NewTests	Positive Integer	Universal	Max 1
schema:Country	dbpprop:Population	Positive Integer	Existential	Exactly 1

4 OntoCOVID Evaluation

The OntoCOVID is evaluated using feature-based ontology quality and analysis tool OntoQA [14]. The OntoQA uses different metrics to evaluate quality of ontology from different aspects [15]. The overall OntoCOVID quality results are shown in Table 3. Since OntoCOVID is the first attempt to ontologically model CSDs; therefore, the results cannot be compared with any other ontology. However, the results obtained shows improved OntoCOVID ontological design with rich potential to represent knowledge in the domain.

Table 3. Quality analysis of OntoCOVID.

Ontology	Classes	Relationships	Classes Richness	Relationships Richness	Inheritance Richness	Tree Balance
OntoCOVID	19	28	87.34	73.86	2.17	1.15

The accuracy of OntoCOVID is evaluated to check that the asserted knowledge agrees with domain experts' knowledge. Ontology with correct and complete definitions of concepts, properties, and individuals will result into high accuracy [15]. Recall and precision are the IR measures for evaluating accuracy of ontology and are shown in Equations (1) and (2) [15]. The recall and precision are measured using functional requirements by executing SPARQL queries on OntoCOVID knowledge base. The knowledge base is formed by instantiating OntoCOVID with relevant information from the CSDs using Protégé 5.5.0 built-in Cellfie plugin and MappingMasterDSL language. The knowledge base is hosted on Fuseki triplestore and SPARQL endpoint to query for retrieving results using web-based JavaScript library Sgvizler. The Fig. 3 shows proof-of-concepts SPARQL query to retrieve the total number of COVID19 positive cases up to 20/01/2021 in the different countries of Europe for total number of positive cases greater than 300000. The Fig. 4 shows pie chart of data retrieved by the SPARQL query in Sgvizler. The query has resulted into acceptable precision and recall values by retrieving relevant and required information. Thus, OntoCOVID effectively organize disparate CSDs into a semantic model to answer high-level queries.

$$Recall = \left(\frac{Number\ of\ relevent\ items\ retrieved}{Total\ number\ of\ relevant\ items}\right) \quad (1)$$

$$Precision = \left(\frac{Number\ of\ relevent\ items\ retrieved}{Total\ number\ of\ items\ retrieved}\right) \quad (2)$$

```
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).
}
```

Fig. 3. SPARQL query to retrieve total number of COVID19 cases of countries in Europe.

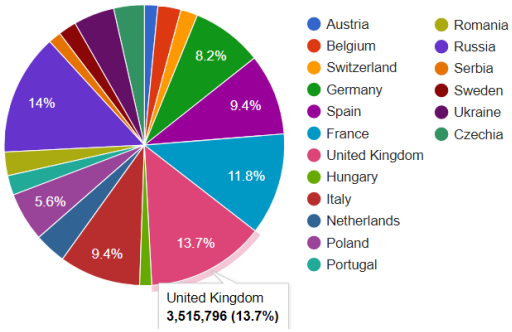

Fig. 4. Pie chart visualization of the SPARQL query.

5 Conclusion

The availability of heterogeneous CSDs demands for the realization of explicit, uniform and flexible data management techniques. The healthcare domain has already used SW technologies to enhance information management, retrieval and visualization. In this paper, we have presented a novel approach for the design and development of OntoCOVID ontology to define an explicit, uniform and flexible semantic data model to represent, organize, share and reuse heterogeneous CSDs. The OntoCOVID is developed and evaluated using state-of-the-art SW technologies to demonstrate its utility and value. The OntoCOVID includes definitions of concepts, properties, relationships and axioms. The test results indicate that OntoCOVID has an improved extendable ontological design and is complete to answer ontology-related competency questions. The OntoCOVID can be easily extended by defining new concepts and properties to represent and accommodate new data needs. The OntoCOVID can have a number of potential application areas including government and healthcare decision making and Linked Open Data.

References

1. Tilahun B, Kauppinen T, Keßler C et al (2014) Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation. JMIR medical informatics 2 (2):e31
2. Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Scientific american 284 (5):34-43
3. Ali S, Khusro S (2016) POEM: Practical ontology engineering model for semantic web ontologies. Cogent Engineering 3 (1):1193959
4. W3C Semantic Web Health Care and Life Sciences Interest Group Charter. https://www.w3.org/2011/09/HCLSIGCharter. Accessed 06/06/2020 2020
5. Zenunia X, Raufia B, Ismailia F et al (2015) State of the Art of Semantic Web for Healthcare. In: Procedia-Social and Behavioral Sciences, 195, pp 1990-1998

6. Jovanovik M, Najdenov B, Strezoski G et al (2015) Linked Open Data for Medical Institutions and Drug Availability Lists in Macedonia. In: New Trends in Database and Information Systems II vol 312. Springer, Cham, pp 245-256
7. Cheung K-H, Prudhommeaux E, Wang Y et al (2009) Semantic Web for Health Care and Life Sciences: a Review of the State of the Art. Briefings in Bioinformatics 10 (2):111-113
8. He Y, Yu H, Ong E et al (2020) CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. Scientific data 7 (1):1-5
9. McGagh D, HarshanaLiyanage, Lusignan Sd et al (2020) COVID-19 Surveillance Ontology. https://bioportal.bioontology.org/ontologies/COVID19. Accessed 24/04/2021 2021
10. Sargsyan A, Kodamullil AT, Baksi S et al (2020) The COVID-19 Ontology. Bioinformatics 36 (24):5703-5705
11. Reddy BP, Houlding B, Hederman L et al (2018) Data linkage in medical science using the resource description framework: the AVERT model. CASI, pp.59
12. Andes N, Davis JE (1995) Linking public health data using geographic information system techniques: Alaskan community characteristics and infant mortality. Statistics in medicine 14 (5-7):481-490
13. Presutti V, Gangemi A (2008) Content ontology design patterns as practical building blocks for web ontologies. In: International Conference on Conceptual Modeling, 2008. Springer, pp 128-141
14. Tartir. S, Arpinar. B (2007) Ontology Evaluation and Ranking using OntoQA. In: Proceedings of the ICSC, pp. 185-192
15. Ali S, Khusro S, Ullah I et al (2017) Smartontosensor: ontology for semantic interpretation of smartphone sensors data for context-aware applications. Journal of Sensors 2017:1-26