Lopez-Rodriguez and Ceballos
Journal of Big Data (2022) 9:9
https://doi.org/10.1186/s40537-022-00562-x

Journal of Big Data

RESEARCH

Open Access

Modeling scientometric indicators using a statistical data ontology

Victor Lopez-Rodriguez* and Hector G. Ceballos

*Correspondence:
a00817161@tec.mx
School of Engineering and Sciences, Tecnologico de Monterrey, Monterrey, Nuevo Leon, Mexico

Abstract

Scientometrics is the field of study and evaluation of scientific measures such as the impact of research papers and academic journals. It is an important field because nowadays different rankings use key indicators for university rankings and universities themselves use them as Key Performance Indicators (KPI). The purpose of this work is to propose a semantic modeling of scientometric indicators using the ontology Statistical Data and Metadata Exchange (SDMX). We develop a case study at Tecnologico de Monterrey following the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. We evaluate the benefits of storing and querying scientometric indicators using linked data as a mean for providing flexible and quick access knowledge representation that supports indicator discovery, enquiring and composition. The semi-automatic generation and further storage of this linked data in the Neo4j graph database enabled an updatable and quick access model.

Keywords: Graph database, Ontology generation, CRISP-DM, Neo4j, Query evaluation

Introduction

Nowadays, the growth of Scientometrics in different contexts has an impact on the way of analyzing science information. Every year institutions and organizations generate an immense volume of information, becoming difficult to analyze it. Scientometric analysis studies the quantitative areas of the process of science, science policy, and communication in science by having a focus on the measure of authors, articles, journals, institutions and understanding citations related to them [1]. Braun and colleagues identify Scientometrics as focused on the study of scientific information, specifically in the analysis of the quantitative aspects of the generation, propagation, and utilization of scientific information to contribute to a better understanding of the mechanism of scientific research activities [2]. Vinkler refers to Scientometric Indicator as the measure of a single scientometric aspect of scientometric systems represented by a single scientometric set with a single hierarchical level also called gross indicators [3].

Case study

Currently, the Tecnologico de Monterrey's Research Office faces the problem of organizing statistical information about current and past research works of its research units. The actual process shown in Fig. 1, begins with the Research Office integrating information from diverse data sources such as Scopus and institutional databases. One example of a Scientometric Indicator is the Citation Count, which it is the sum of citations received to date by institutional outputs and answers the question of how much impact an institution's academic unit has [4]. This metric is defined by the Common European Research Information Formation (CERIF) that has been developed as a flexible model to describe any research data and information, both as a database model but also as a transfer method between repositories [5].

The research office receives the questions and interprets it with their knowledge and context to know what and where to search. A problem in the interpretation stage of the process is when a concept has several definitions, and it leads to wrong answers to the asked question, which would lead to bad decisions. After receiving the question and making an interpretation of it, the specialist performs statistical operations on scientific data following indicator formulas. Information is obtained from different sources and it is difficult to assure that it is up to date. Statistical operations are made and the specialist answers back to the requesting department. Scientometric Indicators are calculated each year in Tecnologico de Monterrey by the Research Office for decision making. Statistical information is gathered from heterogeneous and distributed sources to uncover insights, make predictions, and build smarter systems for the institution [6].

Semantic modeling

The Resource Description Framework (RDF) has formed a more systematic and comprehensive technical architecture in data and knowledge representation and processing. It has become one of the main forms for representing knowledge and ensures that the semantics of temporal data can be described accurately and flexibly, and also may help to realize the sharing in various applications [7].

Ontologies, on the other hand, provide an upper level of semantic representation. The main units of an ontology are concepts, relations between them, and their properties. Relations and properties are represented in the form of triplets (subject, predicate, object or literal value). Each concept has a universal resource identifier (URI) assigned to it [8].

The Statistical Data and Metadata Exchange (SDMX) ontology was initiated by seven international institutions to improve efficiency using technology for sharing and exchanging statistical data and metadata for interoperability. It represents data in

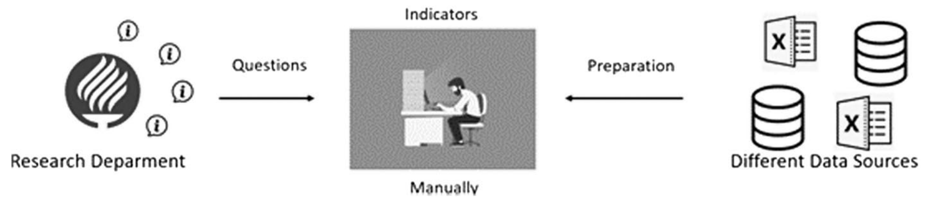


Fig. 1 Actual processing for scientometric data requests

flat files and as Extensible Markup Language (XML) to the definition of fact, dimension, and measure [9]. XML is a very flexible text format that plays an important role in the exchange of a wide variety of data on the Web and elsewhere [10].

Proposed approach

The main goal of our research is to build a model of Scientometric Indicators using RDF for the description of the resources by extending SDMX with a vocabulary appropriate for representing dimensions, attributes and values found in scientometrics indicators (e.g. schools, cites, papers). To evaluate our approach we will extract a sample of Scientometric Indicators used in the Research Office from Tecnologico of Monterrey. Data will be transformed manually to RDF and a tool will be constructed for automation. Another goal is to deploy this model in a graph database to evaluate queries and visualizations.

Our motivation is to have a reliable model that ensures data integration and interoperability by building a flexible, updatable, and easy-to-maintain model with quick access for several applications. Our approach is evaluated by performing scientometric indicator's discovery, enquiring and composition supported by the SDMX data model.

The research questions stated for this work are the following: How Scientometric Indicators can be represented using the SDMX model? Which dimensions need to be defined for modeling them? Which benefits do we obtain by using a semantic representation and storage?

Background

We revised previous works on which indicators are semantically modeled and enquired, as well as those where ontologies are used for representing science or scientometric data. Finally we introduce the semantic platform used for storing and enquiring linked data in our research. We conclude this section by identifying the gap our work contributes to close.

Semantic indicator modeling

Fox proposed modeling city indicators with a semantic approach in 2018 [11]. His approach includes key aspects such as membership extent, temporal extent, spatial extent, and measurement of populations. Fox uses the RDF Data Cube Vocabulary for specifying dimensions of city populations, but the SDMX standard was not incorporated as part of his solution. The evaluation of the ontology was divided into the representation of the population as the definition of indicators, consistency of indicator definitions against the interpretation of a city, and how it can be used to support data collection of a city.

A semantic approach can be implemented in several contexts, one of the most common scenarios is in statistical databases. In Thiry et al. [12] presented an interactive tool for a question answering system that accesses statistical databases and it follows the SDMX standard. In this research work, they take a look at understanding general dimensions from user questions and found that time and location were dimensions for this kind of data. This system was evaluated by testing queries and measuring the accuracy of the result in terms of detecting dimensions. Queries were tested using only one dimension. SDMX is used by many institutions and this research work represents a good approach for answering questions about the selected data.

Scientometric ontologies

Hu et al. converted data, originally stored in a relational database, collected from the Semantic Web Journal to RDF and published them as linked data [13]. This data contains an entire timeline for each paper along with metadata from the Semantic Web Journal (SWJ) unique open and transparent review process. This gives insights into scientific networks and new trends. The Bibliographic Ontology (BIBO) ontology was extended for capturing information about the paper's timeline. BIBO provides main concepts and properties for describing citations and bibliographic references (i.e. quotes, books, articles, etc.) on the Semantic Web [14].

In Osborne et al. [15] presented a novel approach for clustering authors according to their citation distribution. This work introduced the Bibliometric Data Ontology (BiDo) which allows an accurate representation of such clusters. BiDO is a modular ontology encoded using Ontology Web Language (OWL) 2, that allows the description of bibliometric data of people, articles, journals, and other entities described by Semantic Publishing and Referencing (SPAR) Ontologies in RDF [16]. BiDO has kinds of bibliometric data: numeric and categorical. Some measures such as citation count, e-index, and journal impact factor are available through BiDO's numeric property. Categorical data is for specifying categories describing the research career of authors.

Linked data platforms

A Linked Data Platform (LDP) provides a set of integration patterns for building RESTful HTTP services capable of reading and writing RDF data. Under this definition we can find applications like VIVO and Neo4j, that at some extent, enables the visualization of information stored in RDF format.

VIVO is an open source linked data platform that supports recording, editing, searching, browsing and visualizing scholarly activity. It encourages research discovery, expert finding, network analysis and assessment of research impact [17]. The VIVO ontology contains the schemas required for representing this information.

Neo4j is a native graph data store built from the ground up, to leverage not only data but also data relationships. It connects data as it is stored, enabling queries at high speed [18]. Neo4J uses native graph storage which provides the freedom to manage and store data in a highly disciplined manner. It is considered the most popular and used graph database worldwide, used in areas such as health, government, automotive production, military area, among others [19].

Stothers defines some advantages of using Neo4j graph database as follows [20]:

- Well suited to storing information structures that are not well suited to relational databases, such as ontologies or networks.
- Operational simplicity, especially in the use of relationships to avoid joining tables.
- Ability to include properties in relationships and nodes.
- Neo4j query language has the ability to present query results in multiple formats allowing creative insights in data interpretation.

- Efficient queries and attractive interface lead to ease of use and an intuitive user experience.

To load RDF triplets to the graph database the plugin called n10semantics from the Neo4j labs needs to be installed. This plugin enables the use of RDF and its associated vocabularies for data interchange (OWL, RDFS, SKOS and others). This plugin is also use to build integration with RDF generation and consuming components [21]. Some functionalities that are included by installing this plugin are the following:

- Import and Export RDF in multiple formats (Turtle, N-Triples, JSON, etc.)
- Model mapping on import and export
- Import and Export Ontologies in different vocabularies
- Graph validation
- Basic inference

Cypher is the graph-optimized query language incorporated in Neo4j. It understands and takes advantage of connections (relationships) between data. It is inspired by SQL, with the addition of pattern matching borrowed from SPARQL and uses simple ASCII symbols to represent nodes and relationships, making queries easy to read and understand [22].

Summary

In the literature review presented above, we found several methodologies that can be applied for representing and enquiring scientometric indicators, but that must be integrated in a single solution. On the first place, we must select a semantic representation of statistical data. Some authors used BIDO for the representation of numerical data about citations and other specific indicators. We found an area of opportunity because it is a problem to talk about numerical data in one way and when looking at other works they handle it differently. Using the SDMX will allow us to have a standardized way to represent any indicator. On the other hand, SDMX can be extended in terms of dimension, attributes and values appropriate for describing publications, citations, researchers, etc. This solves the problem of having data with a certain level of information such as properties of authors. In this way, we have flexibility for defining dimensions proper of Scientometric Indicators.

Finally, a semantic approach must be evaluated in order to demonstrate its advantages over traditional approaches. SDMX has demonstrated to provide a consistent representation of multidimensional data, but it must permit to capture particular differences between indicator definitions (e.g. annual versus quinquennial time periods). Besides, we must assure that any person familiar with SDMX is capable of discover, enquire and compose scientometric indicators encoded with this data model. We also must provide high-performance on query answering so we used the Neo4j platform for this purpose.

Methodology

In this work, we will follow a methodology commonly used in data science projects. The name of the methodology is CRISP-DM and it stands for Cross Industry Standard Process for Data Mining. It is a process model with six phases (Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment) that naturally describe the data science life cycle.

The Business Understanding phase focuses finding a common motivating business goal to maximize the uptime and efficiency of machines by using predictive analytics. The Data Understanding phase hypotheses for hidden information regarding the data mining project goal are formed based on experience and qualified assumptions. In the Data Preparation phase, the engineer collects the relevant data and prepares it for the actual data mining task. This includes the preprocessing, e.g. data reduction and filtering, as well as feature generation with respect to the data mining project goal.

The Modeling phase consists on a data mining workflow that is constructed to find the desired parameter settings for the selected algorithms and to execute the data mining task on the preprocessed data. In the Evaluation phase, the trained model is tested against real data sets within a production scenario and the data mining results are assessed according to the underlying business objectives. After successful evaluation of the trained model, it is deployed into production in the Deployment phase [23].

Data understanding

Data for this research work is taken from an official workbook of the Research Office that stores data in a tabular way. The file is frequently updated and we took the last version of March 2021, historical data stays the same unless an error is found in a past calculation. The list of Scientometric Indicators that appear on this worksheet is of more than 100 indicators and each one of them belongs to a category. Among these categories, we can find the following: Publications and Cites, Patents, Students, Researchers and Rankings.

The workbook lists each Scientometric Indicator with the person responsible for calculating it, historical data per year, if the indicator is evaluated by quinquennium it stores also the range of years evaluated and the actual value if it is already available. If the Scientometric Indicator also has a dimension such as school, level of education, researcher level, among others, the values of the indicator are described in terms of these dimensions.

Data preparation

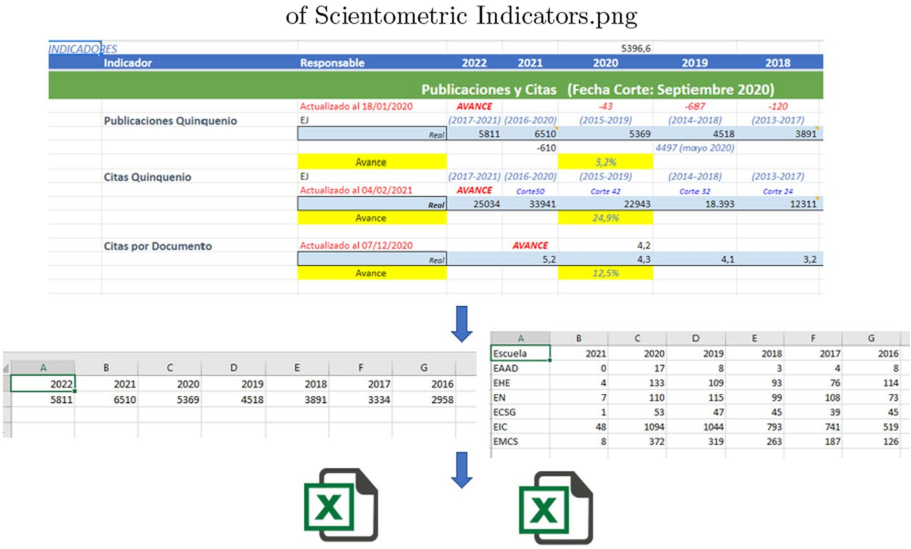
For experimentation purposes, we took a sample of 10 Scientometric Indicators shown in Table 1. These indicators were selected to illustrate that both unidimensional and multidimensional indicators can be represented.

- Unidimensional. The first kind of indicators associate a value to a time period.
- Multidimensional. The second kind of indicators associate a value to a time period and to another dimension(s) (e.g. Schools).

After selecting the sample of Scientometric Indicators to model, we will prepare the information as we need it to be able to automate its transformation. The first step is to make a manual conversion of the 10 Scientometric Indicators listed in the original worksheet and obtain as output 10 different CSV files with the values of each indicator stored in a tabular way. In Fig. 2 we can observe an example of the extraction of 3 Scientometric Indicators which are in turn stored in 3 different CSV files.

Table 1 Sample of 10 scientometric indicators for modeling

	Scientometric indicators	Category
1	Quinqueniall publications	Publications and cites
2	Quinqueniall cites	Publications and cites
3	Cites per document	Publications and cites
4	Annual publications scopus—Tec	Publications and cites
5	Annual publications per school	Publications and cites
6	Quinqueniall publications per school	Publications and cites
7	Quinqueniall cites per school	Publications and cites
8	Cites per document and school	Publications and cites
9	Number of researchers	Researchers
10	Number of PosDocs	Researchers



INDICADORES					5396,6		
	Indicador	Responsable	2022	2021	2020	2019	2018
					Publicaciones y Citas (Fecha Corte: Septiembre 2020)		
		Actualizado al 18/01/2020	AVANCE		-43	-687	-120
	Publicaciones Quinquenio	EJ	(2017-2021)	(2016-2020)	(2015-2019)	(2014-2018)	(2013-2017)
		Real	5811	6510	5369	4518	3891
				-610		4497 (mayo 2020)	
		Avance			3,2%		
	Citas Quinquenio	EJ	(2017-2021)	(2016-2020)	(2015-2019)	(2014-2018)	(2013-2017)
		Actualizado al 04/02/2021	AVANCE	Corte50	Corte 42	Corte 32	Corte 24
		Real	25034	33941	22943	18.393	12311
		Avance			24,9%		
	Citas por Documento	Actualizado al 07/12/2020		AVANCE	4,2		
		Real		5,2	4,3	4,1	3,2
		Avance			12,5%		

A	B	C	D	E	F	G
2022	2021	2020	2019	2018	2017	2016
5811	6510	5369	4518	3891	3334	2958

A	B	C	D	E	F	G
Escuela	2021	2020	2019	2018	2017	2016
EAAD	0	17	8	3	4	8
EHE	4	133	109	93	76	114
EN	7	110	115	99	108	73
ECSG	1	53	47	45	39	45
EIC	48	1094	1044	793	741	519
EMCS	8	372	319	263	187	126

Fig. 2 Extraction of scientometric indicators

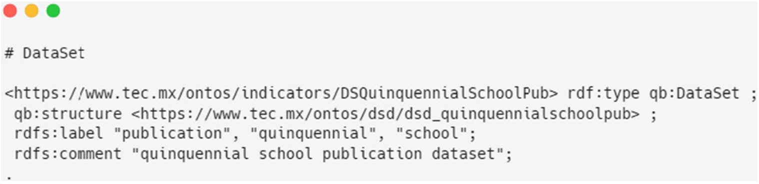
The next step is to manually write the head of the RDF files. We are building an RDF file for each Scientometric Indicator listed in our sample. As Scientometric Indicators are different, the head and observations of the RDF file will differ from others. The format of the RDF file would be Turtle as is one of the valid formats that the Neo4j graph database accepts. A Turtle file allows writing down an RDF graph in a compact textual form. The RDF model uses triples consisting of a subject, a predicate and an object <s, p, o> to represent data [24]. We divided the RDF file format into Vocabularies, Dataset, Data Structure Definition (DSD), Measure and Dimension Properties, Concept Scheme, and Observations.

Vocabularies

Additionally to standard vocabularies for representing RDF/XML (rdf, rdfs, xsd) and ontologies (owl, skos), we incorporated the ontologies shown in Table 2. The last

Table 2 Ontologies used for representing scientometric indicators

prefix	Ontology URI
qb	http://purl.org/linked-data/cube#
sdmx-attribute	http://purl.org/linked-data/sdmx/2009/attribute#
sdmx-dimension	http://purl.org/linked-data/sdmx/2009/dimension#
vivo	http://vivoweb.org/ontology/core
interval	http://reference.data.gov.uk/def/intervals/
tec	https://www.tec.mx/ontos/indicators/



```
# DataSet

<https://www.tec.mx/ontos/indicators/DSQuinquennialSchoolPub> rdf:type qb:DataSet ;
 qb:structure <https://www.tec.mx/ontos/dsd/dsd_quinquennialschoolpub> ;
 rdfs:label "publication", "quinquennial", "school";
 rdfs:comment "quinquennial school publication dataset";
.
```

Fig. 3 Example for the school's publications indicator

ontology (tec) was defined for publishing our definitions. Our objective is to create a model easy and flexible to expand in case of creating new scientometric indicators.

Dataset

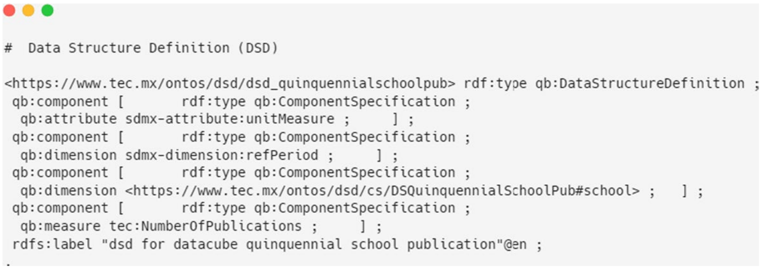
In this section of the RDF file, we created a Universal Resource Identifier (URI) for each Scientometric Indicator. The object can also be identified by a literal for the representation of simple values such as strings or numbers [25]. An example is shown in Fig. 3. We can observe that we use the RDF property label to identify it quicker for future tasks like queries. The Data Structure Definition (DSD) is defined in the data cube ontology.

Data structure definitions (DSD)

The Data Structure Definition of a dataset, describes the components such as dimensions, attributes and measures [26]. In this section, the structure of the dimensions and measures of the Scientometric Indicator is defined. Here is where we define the components including the SDMX attribute as a measure unit; for instance, number of publications, number of cites, number of researchers, etc. We also defined the component of the SDMX dimension, which includes Schools. In Fig. 4 we can observe how both of them are defined for defining the quinquennial number of publications per school.

We defined with a specific URI both measures and dimensions of the Scientometric Indicator. As we mentioned before, we have two kinds of Indicators, unidimensional (only time dimension) and multidimensional (time and other dimensions). We extended the property sdmx-measure:obsValue for representing measures such as the number of publications, citations, researchers and posdoctoral researchers in our indicators.

We also extended the property sdmx-dimension:refPeriod, used for representing temporal intervals. In our case we defined an annual and a quinquennial (5-years) time interval. Time interval instances were borrowed from the United Kingdom's reference data server (http://reference.data.gov.uk/).



```
#  Data Structure Definition (DSD)

<https://www.tec.mx/ontos/dsd/dsd_quinquennialschoolpub> rdf:type qb:DataStructureDefinition ;
 qb:component [       rdf:type qb:ComponentSpecification ;
  qb:attribute sdmx-attribute:unitMeasure ;     ] ;
 qb:component [       rdf:type qb:ComponentSpecification ;
  qb:dimension sdmx-dimension:refPeriod ;     ] ;
 qb:component [       rdf:type qb:ComponentSpecification ;
  qb:dimension <https://www.tec.mx/ontos/dsd/cs/DSQuinquennialSchoolPub#school> ;   ] ;
 qb:component [       rdf:type qb:ComponentSpecification ;
  qb:measure tec:NumberOfPublications ;     ] ;
 rdfs:label "dsd for datacube quinquennial school publication"@en ;
.
```

Fig. 4 Data structure definition

For indicating that an indicator is calculated for a School we extended the basic SDMX dimension property and constrained its value to instances of the School class provided by the VIVO ontology. School instances are available at the VIVO website of our institution (https://research.tec.mx/).

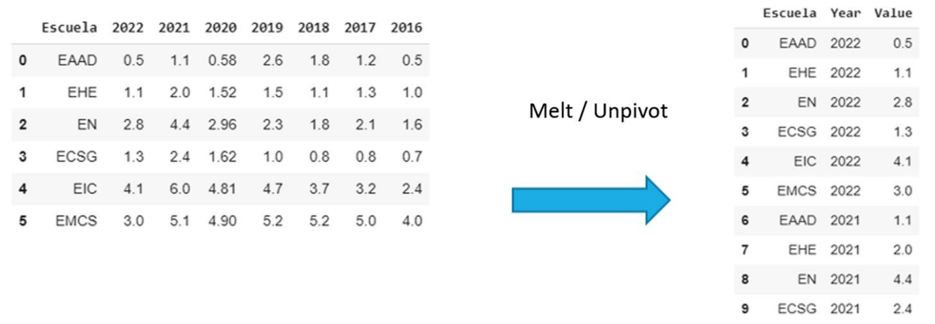
Concept scheme

A concept scheme provides labels to concepts and realizes both hierarchical and associative links [27]. For this purpose we used SKOS, a concept-centric data model based on RDF that identifies concepts using URIs to make already available knowledge organization systems public on the Web in machine-readable formats [28]. SKOS is devoted to developing specifications and standards that support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems, and taxonomies within the framework of the Semantic Web. It provides a standard way to represent knowledge organization systems using the Resource Description Framework (RDF) [29]. In our case we reused common concept schemes such as the list of current Schools.

Observations

In this section, automation is performed to generate the observations of each Scientometric Indicator. Multidimensional data are generally referred to datasets characterized by more than two dimensions [30]. We generated two interactive python notebooks, one for multidimensional indicators and another one for indicators with one dimension. In the first step of Data Preparation, we extracted manually in different files each Scientometric Indicator. This python code receives as input the CSV file and reads it. The next step is to melt or unpivot the data to be able to iterate through it and generate the observations. In Fig. 5 we can observe an example of this procedure.

Before it enters the loop, it receives 3 parameters to build the observations. The parameters are the indicator name, measure name, and measure label. The label is parameterized in order to provide a self-contained description of the data point in natural language. Observations are built automatically and manually pass them to each file. An example of observations of the same Scientometric Indicator is shown in Fig. 6. After passing through all these steps for each Scientometric Indicator, we will have 10 RDF files ready for the modeling in a graph database.

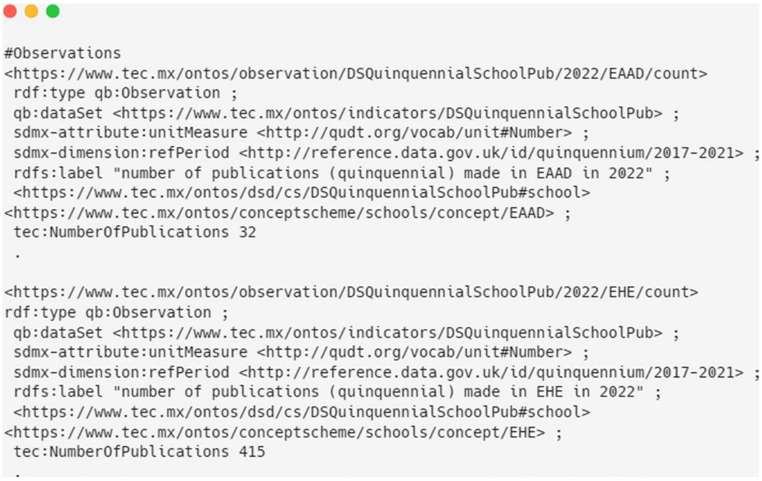


	Escuela	2022	2021	2020	2019	2018	2017	2016
0	EAAD	0.5	1.1	0.58	2.6	1.8	1.2	0.5
1	EHE	1.1	2.0	1.52	1.5	1.1	1.3	1.0
2	EN	2.8	4.4	2.96	2.3	1.8	2.1	1.6
3	ECSG	1.3	2.4	1.62	1.0	0.8	0.8	0.7
4	EIC	4.1	6.0	4.81	4.7	3.7	3.2	2.4
5	EMCS	3.0	5.1	4.90	5.2	5.2	5.0	4.0

Melt / Unpivot

	Escuela	Year	Value
0	EAAD	2022	0.5
1	EHE	2022	1.1
2	EN	2022	2.8
3	ECSG	2022	1.3
4	EIC	2022	4.1
5	EMCS	2022	3.0
6	EAAD	2021	1.1
7	EHE	2021	2.0
8	EN	2021	4.4
9	ECSG	2021	2.4

Fig. 5 Data melt/unpivot



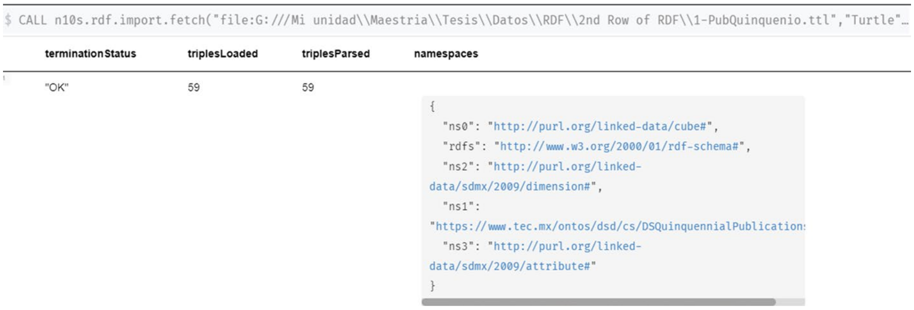
```
#Observations
<https://www.tec.mx/ontos/observation/DSQuinquennialSchoolPub/2022/EAAD/count>
 rdf:type qb:Observation ;
 qb:dataSet <https://www.tec.mx/ontos/indicators/DSQuinquennialSchoolPub> ;
 sdmx-attribute:unitMeasure <http://qudt.org/vocab/unit#Number> ;
 sdmx-dimension:refPeriod <http://reference.data.gov.uk/id/quinquennium/2017-2021> ;
 rdfs:label "number of publications (quinquennial) made in EAAD in 2022" ;
 <https://www.tec.mx/ontos/dsd/cs/DSQuinquennialSchoolPub#school>
<https://www.tec.mx/ontos/conceptscheme/schools/concept/EAAD> ;
 tec:NumberOfPublications 32
 .

<https://www.tec.mx/ontos/observation/DSQuinquennialSchoolPub/2022/EHE/count>
rdf:type qb:Observation ;
 qb:dataSet <https://www.tec.mx/ontos/indicators/DSQuinquennialSchoolPub> ;
 sdmx-attribute:unitMeasure <http://qudt.org/vocab/unit#Number> ;
 sdmx-dimension:refPeriod <http://reference.data.gov.uk/id/quinquennium/2017-2021> ;
 rdfs:label "number of publications (quinquennial) made in EHE in 2022" ;
 <https://www.tec.mx/ontos/dsd/cs/DSQuinquennialSchoolPub#school>
<https://www.tec.mx/ontos/conceptscheme/schools/concept/EHE> ;
 tec:NumberOfPublications 415
 .
```

Fig. 6 Examples of indicator data points (observations)

Modeling

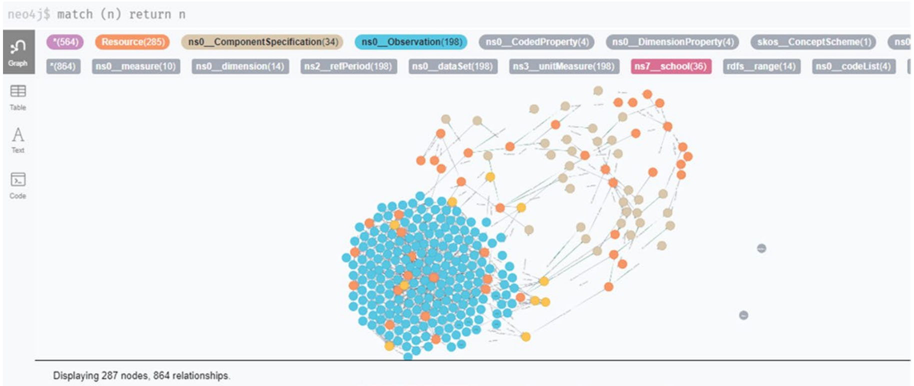
In this phase of the CRISP-DM methodology, we describe the load of these RDF files into the graph database Neo4j. We installed the plugin and proceed to initialize the graph with settings such as handle Vocab-Uris as shorten, overwrite multi-values, handle RDF types as labels and the remaining settings remained in their default value. After the graph initialization we proceed to use a store procedure from the n10s plugin for importing the first RDF file containing the Scientometric Indicator. This store procedure is called import.fetch and we call it from the browser terminal of the graph database. The procedure receive as parameters the location of the RDF file and the RDF format, in our case Turtle. An example of the load is shown in Fig. 7.

This procedure is repeated for all Scientometric Indicators. We did it manually to make sure all the triplets were loaded correctly, but this procedure can be automated. After completing all the list of Scientometric Indicators our graph database is ready to be evaluated with some queries. In Fig. 8 we show all the nodes and relationships stored in our graph database of Scientometric Indicators. Blue nodes represent observation nodes, i.e. indicator data points, hence predominating in the graph.



```
$ CALL n10s.rdf.import.fetch("file:G:///Mi unidad\\Maestria\\Tesis\\Datos\\RDF\\2nd Row of RDF\\1-PubQuinquenio.ttl","Turtle"…

    terminationStatus    triplesLoaded    triplesParsed    namespaces
    "OK"                 59               59
                                                          {
                                                            "ns0": "http://purl.org/linked-data/cube#",
                                                            "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
                                                            "ns2": "http://purl.org/linked-
                                                          data/sdmx/2009/dimension#",
                                                            "ns1":
                                                          "https://www.tec.mx/ontos/dsd/cs/DSQuinquennialPublication
                                                            "ns3": "http://purl.org/linked-
                                                          data/sdmx/2009/attribute#"
                                                          }
```

Fig. 7 Example of RDF file load to Neo4j



```
neo4j$ match (n) return n
```

Displaying 287 nodes, 864 relationships.

Fig. 8 Scientometric indicators loaded in Neo4j

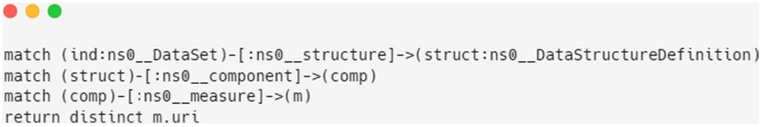
Results

In this section, we will proceed to evaluate our approach by demonstrating that knowing the SDMX data model is enough for enabling indicator discovery, enquiring and composition. To do so we define and test five queries using the Cypher language. At the end of this section, we also compare the number of nodes and relationships in our Neo4j graph database against the number of triplets uploaded to a graph database using RDF files in an Apache Jena Fuseki server to analyze the complexity of the data structure.

Indicator discovery

By knowing that indicators are encoded as SDMX datasets we can ask for what is measured, as shown in the Cypher query in Fig. 9. In this query is inspected the structure of all the indicators in order to list the available metrics (measures). The result of this query, shown in Table 3, has the four measures available in the indicator sample: counts of publications, citations, researchers and posdoc researchers.

Once we select an indicator we can investigate how it is reported, i.e. how it is broken down. For this we need to investigate along which dimensions is reported the measure. The query shown in Fig. 10 extends the previous query by selecting only those indicators that report publication counts and adding the dimensions associated

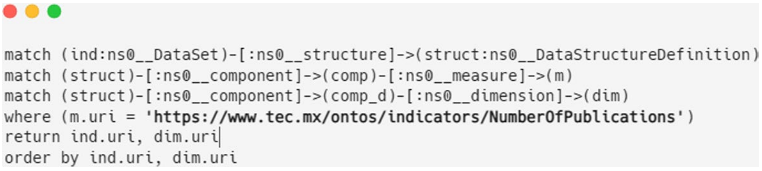


```
match (ind:ns0__DataSet)-[:ns0__structure]->(struct:ns0__DataStructureDefinition)
match (struct)-[:ns0__component]->(comp)
match (comp)-[:ns0__measure]->(m)
return distinct m.uri
```

Fig. 9 Cypher query: what is measured?

Table 3 Results for query: what is measured?

	m.uri
1	tec:NumberOfPublications
2	tec:NumberOfCites
3	tec:NumberOfResearchers
4	tec:NumberOfPosdocs



```
match (ind:ns0__DataSet)-[:ns0__structure]->(struct:ns0__DataStructureDefinition)
match (struct)-[:ns0__component]->(comp)-[:ns0__measure]->(m)
match (struct)-[:ns0__component]->(comp_d)-[:ns0__dimension]->(dim)
where (m.uri = 'https://www.tec.mx/ontos/indicators/NumberOfPublications')
return ind.uri, dim.uri
order by ind.uri, dim.uri
```

Fig. 10 Cypher query: how is measured the number of publications?

Table 4 Results for query: how is measured the number of publications?

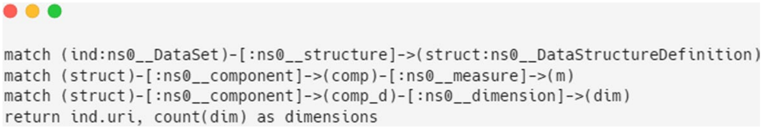
	ind.uri	dim.uri
1	tec:DSAnnualPubScopusTec	tec:Annual
2	tec:DSQuinquennialPublications	tec:Quinquennial
3	tec:DSQuinquennialSchoolPub	tec:Quinquennial
4	tec:DSQuinquennialSchoolPub	tec:School
5	tec:DSSchoolPub	tec:Annual
6	tec:DSSchoolPub	tec:School

to it. The results show that publications are reported in annual or quinquennial periods, and at institutional level or broken down by school (see Table 4).

Next we asked for the number of dimensions used for describing each indicator (see Fig. 11). In this way we can distinguish unidimensional from multidimensional indicators (see Table 5).

Indicator retrieval

Now we formulated a query for retrieving the values stored in a specific indicator. Figure 12 shows the query used for asking the number of publications made by the institution, reported in quinquennial periods. The results, shown in Table 6, include the quinquennial periods and the corresponding number of publications. Quinquennial

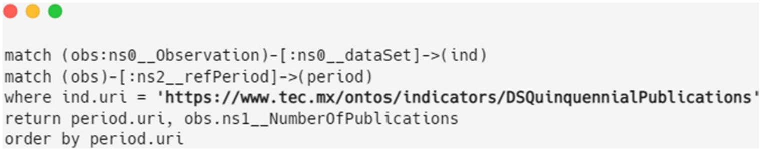


```
match (ind:ns0__DataSet)-[:ns0__structure]->(struct:ns0__DataStructureDefinition)
match (struct)-[:ns0__component]->(comp)-[:ns0__measure]->(m)
match (struct)-[:ns0__component]->(comp_d)-[:ns0__dimension]->(dim)
return ind.uri, count(dim) as dimensions
```

Fig. 11 Cypher query: how many dimensions has every indicator?

Table 5 Results for query: how many dimensions has every indicator?

	ind.uri	Dimensions
1	tec:DSSchoolPub	2
2	tec:DSQuinquennialSchoolPub	2
3	tec:DSQuinquennialSchoolCites	2
4	tec:DSDocumentSchoolCites	2
5	tec:DSResearchers	1
6	tec:DSPosDocs	1
7	tec:DSQuinquennialPublications	1
8	tec:DSQuinquennialCites	1
9	tec:DSDocumentCites	1
10	tec:DSAnnualPubScopusTec	1



```
match (obs:ns0__Observation)-[:ns0__dataSet]->(ind)
match (obs)-[:ns2__refPeriod]->(period)
where ind.uri = 'https://www.tec.mx/ontos/indicators/DSQuinquennialPublications'
return period.uri, obs.ns1__NumberOfPublications
order by period.uri
```

Fig. 12 Cypher query: how many papers published the institution (quinquennial periods)?

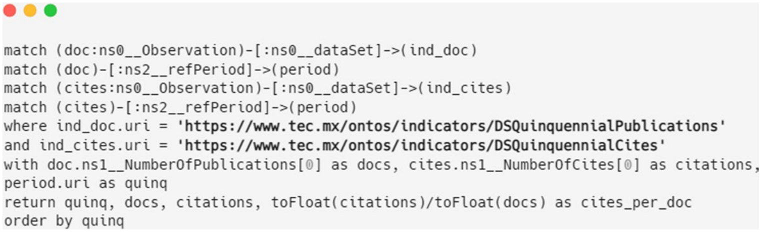
Table 6 Results for query: how many papers published the institution (quinquennial periods)?

	period.uri	NumberOfPublications
1	interval:id/quinquennium/2011–2015	2958
2	interval:id/quinquennium/2012–2016	3334
3	interval:id/quinquennium/2013–2017	3891
4	interval:id/quinquennium/2014–2018	4518
5	interval:id/quinquennium/2015–2019	5369
6	interval:id/quinquennium/2016–2020	6510
7	interval:id/quinquennium/2017–2021	5811

intervals are defined according to the specification provided by the United Kingdom's reference data server (http://reference.data.gov.uk/).

Indicator composition

Finally we evaluated the capability of our approach for calculating a new indicator from those currently stored. Figure 13 shows how both quinquennial publications and quinquennial citations are retrieved for calculating a scientific impact indicator: citations per



```
match (doc:ns0__Observation)-[:ns0__dataSet]->(ind_doc)
match (doc)-[:ns2__refPeriod]->(period)
match (cites:ns0__Observation)-[:ns0__dataSet]->(ind_cites)
match (cites)-[:ns2__refPeriod]->(period)
where ind_doc.uri = 'https://www.tec.mx/ontos/indicators/DSQuinquennialPublications'
and ind_cites.uri = 'https://www.tec.mx/ontos/indicators/DSQuinquennialCites'
with doc.ns1__NumberOfPublications[0] as docs, cites.ns1__NumberOfCites[0] as citations,
period.uri as quinq
return quinq, docs, citations, toFloat(citations)/toFloat(docs) as cites_per_doc
order by quinq
```

Fig. 13 Cypher query: calculate cites per publication?

Table 7 Results Cypher query: calculate cites per publication?

	quinq	docs	citations	citer_per_doc
1	interval:id/quinquennium/2011–2015	2958	6682	2.2580
2	interval:id/quinquennium/2012–2016	3334	9759	2.9271
3	interval:id/quinquennium/2013–2017	3891	12311	3.1639
4	interval:id/quinquennium/2014–2018	4518	18393	4.0710
5	interval:id/quinquennium/2015–2019	5369	22943	4.2732
6	interval:id/quinquennium/2016–2020	6510	33941	5.2136
7	interval:id/quinquennium/2017–2021	5811	25034	4.3080

Table 8 Neo4j graph database

Scientometric indicators	Nodes	Relationships
10	287	864

Table 9 RDF graph

Scientometric indicators	Triplets
10	1578

publication. The results (see Table 7), show the time interval, the number of publications (docs), the number of cites (citations), and the number of citations per publication (cites_per_doc). We had to cast both publications and citations to float for making a correct calculation.

Storage optimization in Neo4j

In order to observe the behavior of our graph structure in Neo4j, we decided to make a comparison of the number of nodes and relationships in Neo4j against the number of triplets in a default graph database created in Apache Jena Fuseki server with all the Scientometric Indicator RDF files created for this work.

Both graphs received as input the 10 RDF files of Scientometric Indicators. In Table 8 we can observe that the Neo4j graph database has 287 nodes and 864 relationships and in Table 9 we observe that an RDF graph has 1578 triplets. The Neo4j is 45% smaller than the original RDF graph in terms of relationships/triplets. The reduction in the number of relationships in Neo4j is due to the absorption of RDF literal values into nodes.

Discussion

Previous ontologies such as VIVO and BIBO provide semantic definitions for researchers, institutions, schools and publications, but they do not provide an appropriate representation of scientometric indicators. Whereas BiDo defines some scientometric indicators, using a vocabulary for statistical data such as SDMX allows to describe each indicator and compose new ones using their description. By extending SDMX with definitions borrowed from the VIVO ontology we provide a mean for representing scientometric indicators. In this way, the detailed information used for calculating these indicators could be accessed and verified in the institutional VIVO instance.

Unlike Fox's approach [11], we used the well-known SDMX data model, which would facilitate the adoption of our approach. On the other hand, we extended the work of Thyri et al. [12] by evaluation the representation of multidimensional indicators modeled with SDMX.

Furthermore, we evaluated the capability for discovering available indicators, retrieving the corresponding values and calculating new indicators using the existing ones. In contrast to a relational data model, a user only needs to know the basic structure of a SDMX model to discover which metrics are defined and how they are broken down.

In our deployment to Neo4j we only imported the basic definitions for time intervals and VIVO instances (e.g. Schools). Nevertheless, it is possible to import additional information available in the corresponding linked data platforms. This information includes, sequence order between time intervals, and the list of schools' faculty. The former can be used for calculating the increment from 1 year to another in a given indicator (e.g. publications), whereas the latter can be used for counting the number of faculty members and build a new indicator.

Conclusions

Scientometric Indicators are important for universities in terms of decision making because they are used for university rankings. By modeling scientometric indicators extending the Statistical Data and Metadata Exchange (SDMX) ontology we enabled indicator representation, discovery, enquiring and composition. We showed how to semi-automatically generate these indicators and link them to data published in institutional (VIVO) and international platforms (UK Government), which in turn can be used for making additional inferences. Using the Neo4j graph database we deployed an efficient solution for data access.

Future work of our approach include the development of a chatbot that answers natural language questions about Scientometric Indicators. A chatbot is are conversational agents that allow the user access to information and services through natural language dialogue, including text and voice [31]. The chatbot will provide natural language processing to obtain relevant information of the question such as the intent and its entities. The chatbot must be capable of identifying the indicator been asked for, extract the parameters of the query (dimensions) and build the Cypher query. The chatbot must support the incorporation of new indicators and new data points.

Despite SDMX only supports the representation of statistical metrics calculated over the entire population, it could be extended to support approximate calculations made on Big Data. Approaches such as Gapprox [32] make use of clustering and sampling techniques for obtaining statistical metrics with 95% of confidence. The remaining 5% could be annotated as an uncertainty attribute in the indicators built with this method.

Acknowledgements
Authors would like to thank to the Research Office by the data provided for developing this research and for providing access to the VIVO instance at http://research.tec.mx/.

Authors' contributions
VL: software, formal analysis, data curation, visualization, writing—original draft. HC: conceptualization, methodology, validation, writing—review and editing. Both authors read and approved the final manuscript.

Authors' information
Victor Lopez-Rodriguez has a bachelor degree in Computer Science by Universidad Autonoma de Nuevo Leon, and he is currently enrolled in a masters program in Computer Science at Tecnologico de Monterrey at Monterrey. He has worked in private industry in Software Development and Business Intelligence.
Hector G. Ceballos has a master and a PhD on Intelligent Systems, both by the Tecnologico de Monterrey at Monterrey. He is Director of the Living Lab & Data Hub at the Tecnologico de Monterrey's Institute for the Future of Education (IFE). He collaborates with the Research Group on Intelligent Systems of the School of Science and Engineering at Tecnologico de Monterrey where he is also advisor of master and PhD students on Computer Science. He is also member of the Mexican Researcher System (SNI Level 1) and adherent member of the Mexican Society on Computing (AMEXCOMP). He has also worked as expert consultant for bank and IT companies, and promoted the adoption of Semantic Web technologies in academy, government and industry. He counts with more than 30 scientific papers published in Journals, Conferences and books.

Funding
This research is partially funded by CONACYT and Tecnologico de Monterrey.

Availability of data and materials
Linked data generated in this case study is available on demand by sending an email to ceballos@tec.mx.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 27 August 2021 Accepted: 10 January 2022
Published online: 24 January 2022

References

1. Zakka WP, Lim NHAS, Khun MC. A scientometric review of geopolymer concrete. J Clean Prod. 2021;280:124353.
2. Vitanov N. Science dynamics and research production: indicators, indexes, statistical laws and mathematical models. Bulgaria: Springer; 2016.
3. Vinkler P. The evaluation of research by scientometric indicators. Abington: Elsevier; 2010.
4. Colledge L. Snowball metrics recipe book. Montreal: Elsevier; 2017.
5. Engelman A, Enkvist C, Pettersson K. A FAIR archive based on the CERIF model. Procedia Comput Sci. 2019;146:190–200.
6. Capadisli S, Auer S, Riedl R. Towards linked statistical data analysis. In: 1st international workshop on semantic statistics (SemStats 2013); 2013. p. 61–72.
7. Zhang F, Wang K, Li Z, Cheng J. Temporal data representation and querying based on RDF. IEEE Access. 2019;7:85000–23.
8. Shachnev D, Karpenko D. Using subject area ontology for automating processes in sphere of scientific investigation and education. Program Comput Softw. 2018;44:15–22.
9. Wisnubhadra I, Baharin SSK, Herman NS. Modeling and querying spatiotemporal multidimensional data on semantic web: a survey. J Theor Appl Inf Technol. 2019;97:3608–33.
10. W3C. Extensible markup language (XML). 2021. https://www.w3.org/XML/.
11. Fox MS. The semantics of populations: a city indicator perspective. J Web Semant. 2018;48:48–65.
12. Thiry G, Manolescu I, Liberti L. A question answering system for interacting with SDMX databases. In: The 6 natural language interfaces for the Web of Data (NLIWOD) workshop (in conjunction with ISWC). HAL; 2020.

13. Hu Y, Janowicz K, McKenzie G, Sengupta K, Hitzler P. A linked-data-driven and semantically-enabled journal portal for scientometrics. In: The semantic web—ISWC 2013; 2013. p. 114–29.
14. The Bibliographic Ontology. Bibliographic ontology specification; 2021. http://bibliontology.com/.
15. Osborne F, Peroni S, Motta E. Clustering citation distributions for semantic categorization and citation prediction. In: Proceedings of the 4th international conference on linked science, vol. 1282. CEUR-WS.org; 2014. p. 24–35.
16. Peroni S, Shotton D. The SPAR ontologies. In: The semantic web—ISWC 2018. Springer International Publishing; 2018. p. 119–36.
17. Conlon M, Woods A, Triggs G, O'Flinn R, Javed M, Blake J, et al. VIVO: a system for research discovery. J Open Source Softw. 2019;4:1182.
18. Neo4j. Neo4j graph database; 2021. https://neo4j.com/product/#graph-database.
19. Fernandes D, Bernardino J. Graph databases comparison: AllegroGraph, ArangoDB, InfiniteGraph, Neo4J, and OrientDB. In: Proceedings of the 7th international conference on data science, technology and applications; 2018. p. 373-80.
20. Stothers JA, Nguyen A. Can Neo4j replace PostgreSQL in healthcare? AMIA Summits Transl Sci Proceed. 2020;2020:646–53.
21. Neo4j Labs. Neosemantics (n10s): neo4j RDF & semantics toolkit; 2021. https://neo4j.com/labs/neosemantics/.
22. Neo4j. Cypher query language; 2021. https://neo4j.com/product/#cypher.
23. Wiemer H, Drowatzky L, Ihlenfeldt S. Data mining methodology for engineering applications (DMME)—a holistic extension to the CRISP-DM model. Appl Sci. 2019;9:403–8.
24. Zouaghi I, Mesmoudi A, Galicia J, Bellatreche L, Aguili T. Query optimization for large scale clustered RDF data. In: DOLAP; 2020. p. 56–65.
25. Vlachou A, Doulkeridis C, Glenis A, Santipantakis GM, Vouros GA. Efficient spatio-temporal RDF query processing in large dynamic knowledge bases. In: For Computing Machinery A, editor. Proceedings of the 34th ACM/SIGAPP symposium on applied computing; 2019. p. 439–47.
26. Escobar P, Candela G, Trujillo J, Marco-Such M, Peral J. Adding value to linked open data using a multidimensional model approach based on the RDF data cube vocabulary. Comput Stand Interfaces. 2020;68:103378.
27. Grévisse C, Rothkugel S. An SKOS-based vocabulary on the swift programming language. In: Springer, editor. International semantic web conference; 2020. p. 244–58.
28. Biagetti MT. Ontologies (as knowledge organization systems); 2020. https://www.isko.org/cyclo/ontologies.
29. W3C. Introduction to SKOS; 2021. https://www.w3.org/2004/02/skos/intro.
30. Brandi G, Matteo TD. Predicting multidimensional data via tensor learning; 2021. arXiv:abs/2002.04328.
31. Følstad A, Araujo T, Papadopoulos S, Law ELC, Granmo OC, Luger E, et al. Chatbot research and design. Amsterdam: Springer; 2020.
32. Ahmadvand H, Goudarzi M, Foroutan F. Gapprox: using Gallup approach for approximation in Big Data processing. J Big Data. 2019;6:1–24.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.