Modeling Scientometric Indicators Using A SDMX Ontology

Víctor Iván López Rodríguez (victorlrdz26@gmail.com)
Tecnologico de Monterrey: Instituto Tecnologico y de Estudios Superiores de Monterrey
https://orcid.org/0000-0002-7362-7694

Hector G. Ceballos
Tecnologico de Monterrey: Instituto Tecnologico y de Estudios Superiores de Monterrey

Research

Keywords: Scientometric Indicators, SDMX Ontology, Graph Database, CRISP-DM

Posted Date: September 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-842177/v1

RESEARCH

Modeling Scientometric Indicators using a SDMX Ontology

Victor Lopez-Rodriguez* and Hector G. Ceballos

*Correspondence:
a00817161@tec.mx
School of Engineering and Sciences, Tecnologico de Monterrey, Nuevo Leon, MX
Full list of author information is available at the end of the article

Abstract

Scientometrics is the field of study and evaluation of scientific measures such as 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 Satistical 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 in a Graph Database (Neo4j) the linked data produced by our approach.

Keywords: Scientometric Indicators; SDMX Ontology; Graph Database; CRISP-DM

Introduction

Nowadays, the growth of Scientometrics in different contexts has an impact in the way to analyze this information. Every year this impact generates an immense volume of information and it tends to become difficult to analyze. Braun et al. (1987) stated that the definition of Scientometrics focused in the study if scientific information is that Scientometrics analyzes the quantitative aspects of the generation, propagation and utilization of scientific information in order to contribute to a better understanding of the mechanism of scientific research activities [1].

At present, the Research Office from the School of Engineering and Sciences at Tecnologico de Monterrey faces a problem getting statistical information about their current and past research works.

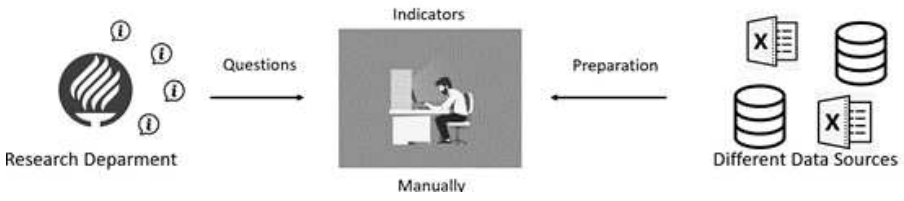


Figure 1 Actual Process

The actual process shown in Figure 1, begins with the Research Office asking questions related to the Scientometrics Indicators that include several measurements about articles, research works and academic papers. The term scientometrics was coined by Vassily V. Nalimov in the 1960s, and refers to the science of measuring and analyzing of science, such as a discipline's structure, growth, change, and interrelations [2]. In [3] Vlinker refers to Scientometric Indicator as the measure of a single scientometric aspect of scientometric systems represented by a single sientrometric set with a single hierachichal level also called gross indicators. The workers receive the questions and they interpret the question with their knowledge and context to know what and where to search. A bad interpretation can lead to several problems in finding the correct result of the indicator. An example can be the following question: 'How many Professors are currently doing research with a master student?' For this example, we are going to take the concept of Professor. For the Research Office, the concept Professor is related to the concept Person who has an active contract of complete-time in a certain period (this can be obtained from the context of the question). For the workers who get the answer to the question, the concept of Professor is related to the concept Person and Researcher and who has an active contract of full and partial time and also consider external professors. 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 questions, and a wrong answer leads to bad decisions.

After receiving the question and making an interpretation of it, the worker starts to perform statistical operations within the concepts related to it. Information is obtained from different sources and it is difficult to assure that it is up to date. Statistical operations are made and the worker answers back to the Research Office with the answer. In this stage of the process we can observe that is done manually and information is obtained from different resources, in most of the cases worksheets.

The main idea is to extract a sample of Scientometric Indicators used in the Research Office from Tecnologico of Monterrey. This sample will contain also historic data and will be chosen according the necessity of the requirement and experimentation of distinct types of indictators. Data will be transformed manually to the Resource Description Framework (RDF) and a tool will be constructed for automation (specifically for the sample of indicators extracted).RDF is a standard model for data interchange on the Web and it has features that facilitate data merging even if the underlying schemas differ, and it specifically supports the evolution of schemas over time without requiring all the data consumers to be changed [4]. The generation of an Ontology will be perform and this serie of indicators would be load to the graph database Neo4j and experimentation of queries and visualizations will be perform there.

The main units of the ontology are concepts, relations between them, and their properties. Relations and properties are represented in the form of triplets (subject, predicate, object or property value). Each concept has a universal resource identifier (URI) assigned to it [5].In our research work our triplets represent the Scientometric Indicators with their according relations and the properties of the value.

Scientometric Indicators are used in Tecnologico de Monterrey by the Research Office for decision making. This kind of indicators are also compared with a target value in which can result in a completeness or failure for certain kind of activities. Braun et al. mentioned that one of the most challenging aims of scientometric research is to build indicator systems characterizing the research of scientific communities [6]. In our case of study it applies different topics and suborganizations of our institution. As we mentioned before, this indicators are calculated from different kind of sources and the workers in charge of obtaining this indicators may vary the sources or the operations to get the final result. This Scientometric Indicators are calculated each year and historical data is also available and stored in an excel worksheet. In order for the society to tract and learn from its own vast knowledge about events and things, it needs to be able to gather statistical information from heterogeneous and distributed sources to uncover insights, make predictions, or build smarter systems that the society needs to progress [7].

One example of a Scientometric Indicator is Citation Count. It is the sum of citations received to date by institutional outputs and answers the question of how many total citations an institution's output has received [8]. This metric is available by the Common European Research Information Formation (CERIF) that is dedicated to the development of Current Research Information Systems (CRIS Systems) and a template both for data exchange between CRISs and for mediating access to multiple heterogeneous distributed CRISs [9].

The main goal of our research work is to construct a model of Scientometric Indicators using RDF for the description of the resources and the Satistical Data and Metadata Exchange (SDMX) as an extension of the vocabulary for dimensions and measures. Another goal is to implement this model in a graph database and evaluate queries and visualizations.

The SDMX initialized by seven international institutions (UN, World Bank, IMF, Eurostat, BIS, and European Central Bank) with the purpose of improving efficiency and enhance quality using advance technology for share and exchange statistical data and metadata among institutions for interoperability. The SDMX covers how to represent statistical data in flat files and as XML to the definition of fact, dimension, and measure [10]. It provides flexibility to structure its data according to their specific informational needs and usage scenarios [11].

The motivation of this research work is to have a reliable model of Scientometric Indicators that ensures data integration and interoperability. We look that the model is flexible in terms of modifications (updatable), easy to maintain, quick access and reusable for several applications.

The research questions stated for this work are the following: Which Scientometric Indicators could be represented on the model? Which dimensions need to be defined for modeling? How would be the volume of information using the Scopus database? Which benefits do we obtain by using a semantic representation? The

last research question also focuses if the model can be load in a graph database (Neo4j) and which types of queries and visualizations can be done in Neo4j.

Literature Review

In [12] the modeling from a city indicator with a semantic approach was performed. In this research work an ontology development is model by the representation of city indicator definitions. It includes a model that represents key aspects such as membership extent, temporal extent, spatial extent and measurement of populations. The topic of population in an ontology let the researchers a different perspective on how it needed to be represented as a model. RDF Data Cube Vocabulary is also used in the specification of dimensions but SDMX standard was implemented as part of the solution. The evaluation of the ontology was divided in how representing population as definition of indicators, consistency of indicator definitions against the interpretation of a city and how it can be use 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 [13] Thiry et al. presented an interactive tool for a question answering system that access statistical databases and it follows the SDMX standard. In this research work they take a look in 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 measure 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.

In [14] they converted data in a relational database form collected from the Semantic Web Journal to RDF and published them as RDF. It contains also an entire time line for each paper together with metadata from the Semantic Web Journal (SWJ) unique open and transparent review process. This give insights of scientific networks and new trends. BIBO ontology was extended for capturing information about paper's time line.

In [15] Osborne et al. presented a novel approach for clustering authors according their citation distribution. They introduced the Bibliometric Data Ontology (BiDo) which allows an accurate representation of such clusters. BiDO is a modular OWL 2 ontology that allow the description of bibliometric data of people, articles, journals and other entities described by SPAR Ontologies in RDF. BiDO has kinds of biblimetric 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 describien the research career of authors. The difference with our research work is the use of SDMX. SDMX allows to have multi-dimensions in the scientometric indicators available such as citation count. Multi-dimension helps to go deep in information and answer specific questions through queries.

Data Methods

Data for this research work is taken from an official worksheet 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 100 indicators and each of them belong to a list of categories. In the list of categories we can find the following: Publications and Cites, Intellectual Vitality, Patents, Students, Researches, Science Divulgation, and Rankings.

The worksheet list each Scientometric Indicator with the person responsible of calculating it, historical data per years, if the indicator is evaluated by quinquennium it stores also the range of years evaluated and the actual value if it is yet available. If the Scientometric Indicator also has a dimension of school, level of education, researcher level, among others, the values of the indicator are shown including this dimensions. The dimension school reference to VIVO, a Semantic Web-based network of institutional ontology-driven databases to enable national discovery, networking, and collaboration via information sharing about researchers and their activities [16].

For experimentation aspects, we took a sample of 10 Scientometric Indicators shown in Table 1. These indicators were selected because we have two kinds of data.

- The first kind of indicators store their values only per year.
- The second kind of indicators store their values per year and the dimension according the indicator.

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

After selecting the sample of Scientometric Indicators to model for answering the research questions, we will prepare the information as we need it to able to automate the conversion.

The first step is to make a manual conversion of the 10 Scientometric Indicators listed in the original worksheet and obtain as an outpout 10 different csv files with the values of each indicator stored in a tabular way. In Figure 2 we can observe an example of extracting 3 Scientometric Indicators and store them in 3 different csv files.

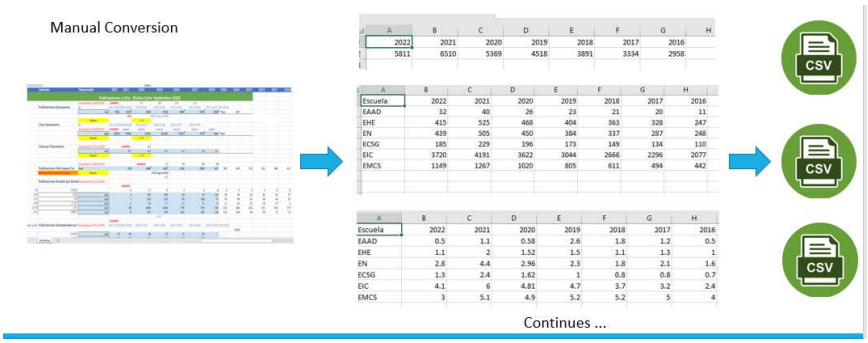


Figure 2 Extraction of Scientometric Indicators

The next step is to write manually the head of the RDF files. We are building a 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 to load. A Turtle file allows writing down an RDF graph in a compact textual form. An RDF graph is made up of triples consisting of a subject, predicate and object [17]. We divided the RDF file format in Vocabularies, Dataset, Data Structure Definition (DSD), Dimension Properties, Measure and Dimension Properties, Concept Scheme and Observations.

Vocabularies

The vocabularies used in our RDF files were the following: owl, qb, rdf, rdfs, sdmx-attribute, sdmx-dimension, skos and xsd.

Dataset

In this section of the RDF file, we created an Universal Resource Identifier (URI) for each Scientometric Indicator. The identifier for an object is assigned when the object is created and remains constant throughout the lifetime of the object [18]. An example is shown in Figure 3. We can observe that we use the rdf property label in order to be able to identify it quicker for future tasks like queries. Data Structure Definition (DSD) is also defined by the cube property.

```
# DataSet

<https://www.tec.mx/ontos/indicators/DSSchoolPub>   rdf:type qb:DataSet ;
 qb:structure <https://www.tec.mx/ontos/dsd/dsd_schoolpub> ;
 rdfs:label "school citation dataset"@en ;
.
```

Figure 3 URI example for the Scientometric Indicator School Publications

Data Structure Definition (DSD)

The Data Structure Definition describes the structure or metamodel of one or more statistical datasets and it defines attributes, measures and dimensions called components [19]. 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 unit measure, some examples could rely on number of publications, number of cites, number of researchers, etc. We also define the component of sdmx dimension and it includes schools. In Figure 4 we can observe how both are defined.

```
#  Data Structure Definition (DSD)

<https://www.tec.mx/ontos/dsd/dsd_schoolpub>   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/DSSchoolPub#school> ;     ] ;
 qb:component [       rdf:type qb:ComponentSpecification ;
  qb:measure <https://www.tec.mx/ontos/dsd/cs/DSSchoolPub#numberofpublications> ;     ] ;
 rdfs:label "dsd for datacube quinquennial school publication"@en ;
.
```

Figure 4 Data Structure Definition

Measure and Dimension Properties

In this section we defined with an specific URI both measure and dimension of the Scientometric Indicator and a label is assigned. As we mentioned before, we have two kinds of Indicators, general (only year dimension) and dimensional (year and school). All the indicators have the measure property define and the dimension property only of applicable.

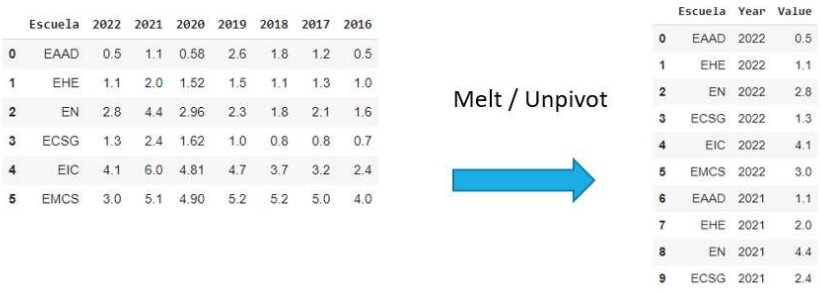
Concept Scheme

In the Concept Scheme section, if a sdmx dimension exist for that Scientometric Indicator, then Concept Scheme is written. We use SKOS and it consists of a set of RDF properties and RDFS clauses that can used to express the content and structure of a concept scheme as an RDF graph [20]. The term concept scheme can be define as a set of interrelated concepts without modeling those concepts as formal classes [21]. In this case for the applicable indicators we same concept scheme repeats by reusing them and obtain the benefits of the framework.

Observations

In this section, automation is performed to generate the observations of each Scientometric Indicator. Multidimensional data is represented as observation that are instantiated over predefined dimensions and measures [22]. We generated two interactive python notebooks, one for general indicators and other one for indicators

with a 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 read it. The next step is to melt or unpivot the data in order to be able to loop on it and generate the observations. In Figure 5 we can observe an example of this procedure.



	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

	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

Figure 5 Melt/Unpivot of Data

Before it enters to the loop, it receives 3 parameters to build the observations. The parameters are the indicator name, measure name and the measure label. Observations are built automatically and manually pass them to each file. An example of observations of the same Scientometric Indicator are shown in Figure 6.

```
<https://www.tec.mx/ontos/observation/DSDocumentSchoolCites/2022/EAAD/count>  rdf:type qb:Observation ;
 qb:dataSet <https://www.tec.mx/ontos/indicators/DSDocumentSchoolCites> ;
 sdmx-attribute:unitMeasure <http://qudt.org/vocab/unit#Number> ;
 sdmx-dimension:refPeriod <http://reference.data.gov.uk/id/year/2022> ;
 rdfs:label "number of cites per document made to EAAD in 2022" ;
 <https://www.tec.mx/ontos/dsd/cs/DSDocumentSchoolCites#school> <https://www.tec.mx/ontos/conceptscheme/schools/concept/EAAD> ;
 <https://www.tec.mx/ontos/dsd/cs/DSDocumentSchoolCites#numberofcites> 0.5
 .


<https://www.tec.mx/ontos/observation/DSDocumentSchoolCites/2022/EHE/count>  rdf:type qb:Observation ;
 qb:dataSet <https://www.tec.mx/ontos/indicators/DSDocumentSchoolCites> ;
 sdmx-attribute:unitMeasure <http://qudt.org/vocab/unit#Number> ;
 sdmx-dimension:refPeriod <http://reference.data.gov.uk/id/year/2022> ;
 rdfs:label "number of cites per document made to EHE in 2022" ;
 <https://www.tec.mx/ontos/dsd/cs/DSDocumentSchoolCites#school> <https://www.tec.mx/ontos/conceptscheme/schools/concept/EHE> ;
 <https://www.tec.mx/ontos/dsd/cs/DSDocumentSchoolCites#numberofcites> 1.1
 .
```

Figure 6 Example of observations

After passing through all this steps for each Scientometric Indicator, we will have 10 RDF files ready for the modeling in a graph database.

Methodology

In this research work we will guide us with a methodology commonly use in data science projects. The name of the methodology is called CRISP-DM and it it stands for Cross Industry Standard Process for Data Mining. It is a process model with six phases that naturally describe the data science life cycle [23].

Modeling

In this phase of the Methodology we describe the load of this RDF files into the graph database Neo4j. It is a native graph data store built from the ground up, to leverage not only data but also data relationships and it connects data as it's stored, enabling queries never before imagined, at speeds never thought possible [24]. Neo4J uses natives graph storage which provides the freedom to manage and store data in a highly disciplined manner. and it is considered the most popular and used graph database worldwide, used in areas such as health, government, automotive production, military area, among others [25].

Lal defines in [26] some advantages of using a graph database:

- Query performance of a graph database is a few orders of magnitude better than a relational data base or other NoSQL alternatives.
- Flexibility and agility in adding relationships, nodes types, properties without making changes to existing queries.
- The schema is dictated by the application and leads to lesser ambiguity.
- The design to delivery time is reduced.

In order to load RDF triplets to the graph database we needed to install a plugin called n10semantics from the Neo4j labs. This plugin enables the use of RDF and its associated vocabularies like (OWL,RDFS,SKOS and others) for data interchange. This plugin is also use to build integration with RDF generation and consuming components [27]. 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

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 had the default value. After the graph initialization we will proceed to use a store procedure from the n10s plugin for importing the RDF file containing the Scientometric Indicator. This store procedure is called import.fecht and we call it from the browser terminal of the graph database and it received as parameters the location of the RDF file and the RDF format, in our case Turtle. An example of the load is shown in Figure 7.

This procedure is repeated for all Scientometric Indicators. We did it manually to make sure all the triplets were loading correctly. After completing all the list of Scientometric Indicators our graph database is ready to be evaluated with some queries. In Figure 8 we show how all the nodes and relationships are stored in our graph database of Scientometric Indicators use in Tecnologico de Monterrey.

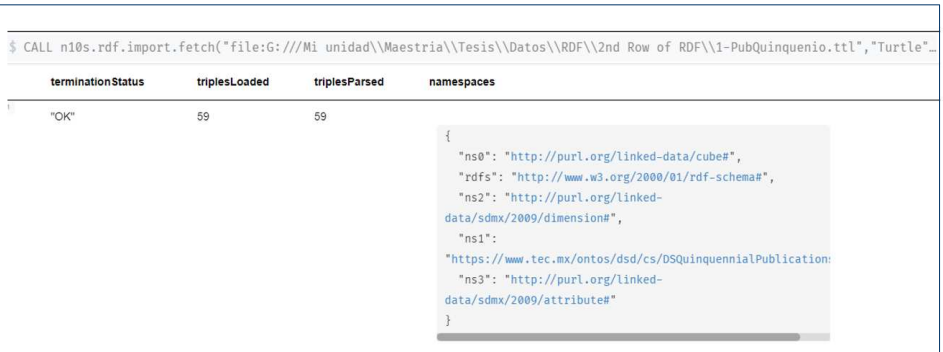


Figure 7 Example of RDF file load to Neo4j

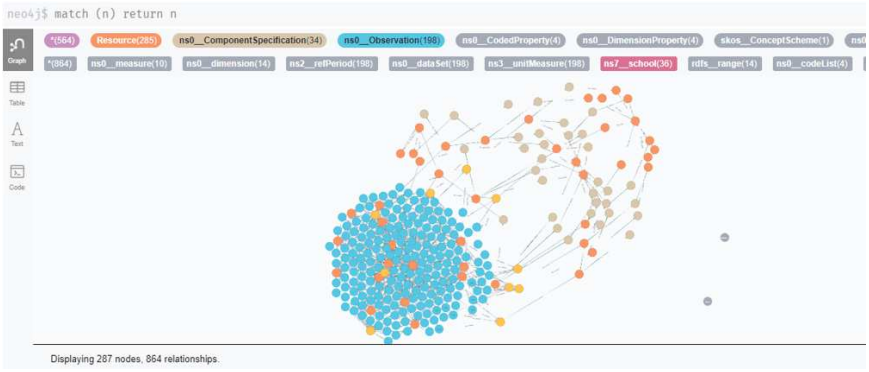


Figure 8 Scientometric Indicators graph database

Results

In this section we will proceed to test some queries using Cypher that comes along with Neo4j. Cypher is a powerful, graph-optimized query language that 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 [28]. We will 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 Apache Jena Fuseki server.

The first test is to query all the nodes that have a relationship with the year 2020. In this case, we query 2 types of node: Observation and Resource. In the node of type Observation we will find all the values of measures found in the properties of each node. We also query a relationship between this nodes call refPeriod that indicates that this node are related to the year we are querying. We filter the years by filtering the property URI with the URI from 2020 that we are working on in all the Scientometric Indicators. In Figure 9 we observe the result of the query.

The second test is a query to obtain all the nodes of type observation that have a relationship with a period of time. We add another relationship of type school from

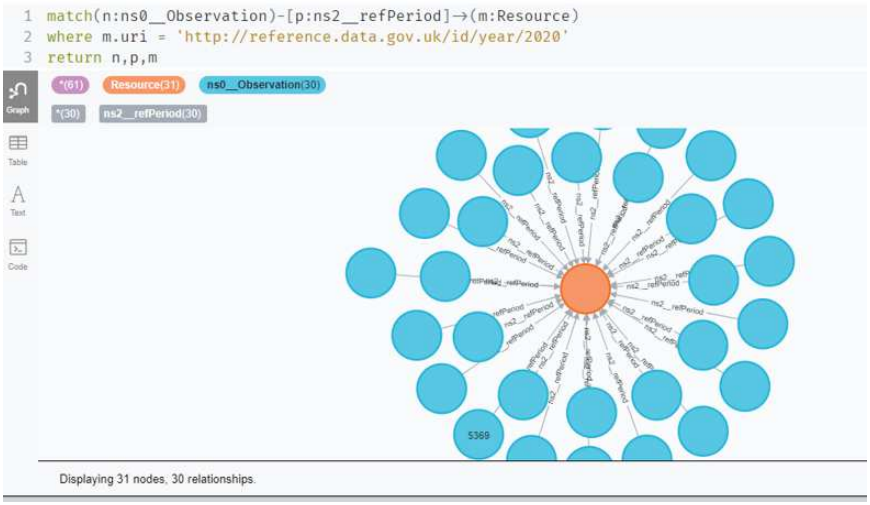


Figure 9 1st Query Test

the observation to the node of type resource. We filter the year by only querying by the URI of 2019, that is the property of the node of type resource (year). We can see in the extraction that every node of type school and type year are connected through relationships to the node of type observation and this node of type observation has the numerical value of the Scientometric Indicator. It is important to clarify that the node of type resource (year and school) can be filtered through their URIs but in this case we are showing all the values. In Figure 10 we observe the result of this query and the selection of a node of type resource (school dimension) and the properties values (label and URI) of this node.

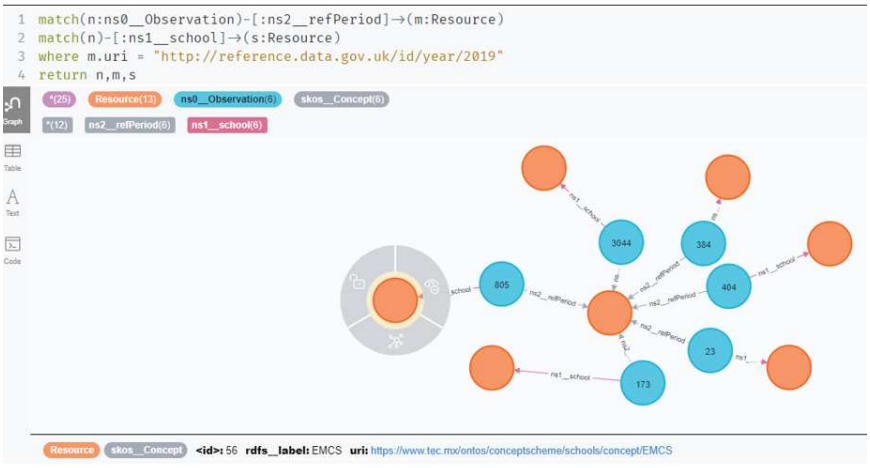


Figure 10 2nd Query Test

The third test is a query to obtain an average value in a period of time of a Scientometric Indicator. We continue with the same guide of matching a node of

type observation with a relationship of refPeriod to the node of type resource (year dimension). We filter data by filtering the resource node with the URI of the year 2019. Instead of returning a set of nodes and relationships, we return the average of the property called number of publication of the node of type observation. In this case we are returning the average of number of publications of all the school from Tecnologico de Monterrey in 2020. In Figure 11 we show the numerical result of the query.

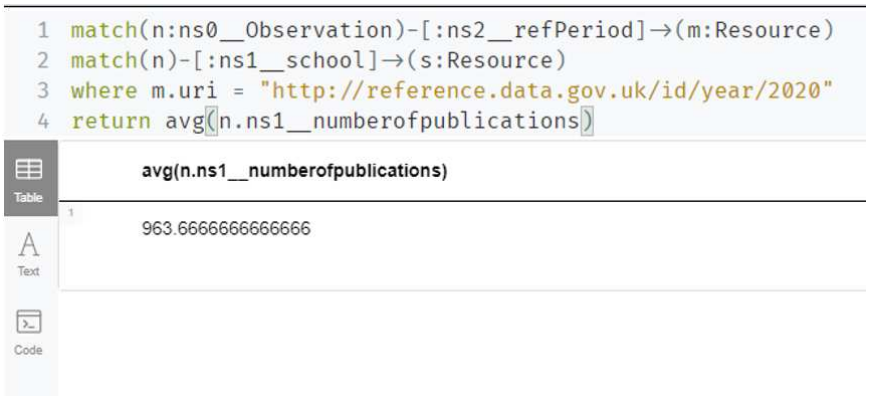


Figure 11 3th Query Test

The last test is about a query calculating new Scientometric Indicators with information available in our graph database. As we already have the Scientometric Indicators of Quinquennial Publications and Quinquennial Cites, it is possible to calculate a new Scientometric Indicator called Cites per Document by dividing the number of cites by the number of publications. It is possible to make this type of calculation in Neo4j. In Figure 12 we can observe that we are matching 2 different nodes of type observation (x and y) using the relationship of type period to the node of type resource (year dimension). We filter data by using the property URI of the node of type resource and only have data from 2020. For knowing the type of Scientometric Indicator in the nodes of type observation (x and y) we filter data by using the property URI of each node and therefore have those 2 values available. In this query we only show the nodes and relationships.

In Figure 13 we can observe that by using the properties number of publications and number of cites of the nodes of type observation (x and y), we can use these values and proceed with the operation with these values. By dividing them we obtain a numerical value that represent the Cites per Document in 2020 of Tecnologico de Monterrey.

In order to observe the behavior of our graph in Neo4j, we decided to make a comparison of the nodes and relationships against the number of triplets by uploading all the mentioned Scientometric Indicators in the Neo4j graph database and a default graph created in Apache Jeana Fuseki server. Both graphs received as input the 10 RDF files of Scientometric Indicators. In Table 3 we can observe that the

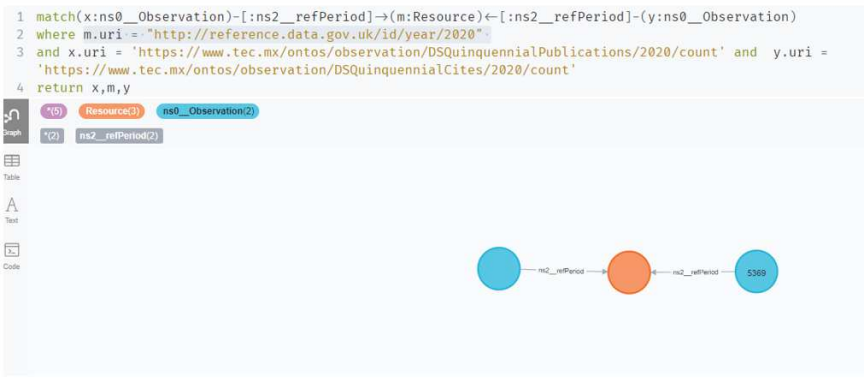


Figure 12 4th Query Test

Figure 13 5th Query Test

Neo4j graph database consists of 287 nodes and 864 relationships and in Table 4 we observe that a RDF graph has 1578 triplets. The Neo4j graph database achieves this quantity of nodes and relationships because all the sdmx-dimensions (time and school) are certainly reused nodes with a relationship with all the scientometric indicator node type.

Table 2 Neo4j Graph Database

Scientometric Indicators	Nodes	Relationships
10	287	864

Table 3 RDF Graph

Scientometric Indicators	Triplets
10	1578

Conclusion and Future Work

Scientometric Indicators are important for universities in terms of decision making. These indicators also are used for rankings and universities try to always improve them. Modeling an Ontology of Scientometric Indicators was a reliable approach to make an standardization on how these indicators are obtained and calculated. In the model, 10 Scientometric Indicators were include such as Publications, Cites, Annual

Publications, Number of Researchers among others. For this indicators we needed to define two dimensions (time and school). A semantic approach gives us benefits such as providing metadata of resources, in our case we are making reference to the site VIVO in which we found triplets about the school dimension and they are suborganizations of the university and this site has also additional information for definition of the indicators. Another benefit of the semantic approach if that we have a standard syntax and making queries get less complex in terms of lines of code and will allow us to use them in applications for the exchange of information. Using a Neo4j graph database based in nodes and relationships allow us to have an understandable and less complex model. Using SDMX improve the accessibility to dimensions and measures.

Future work of this semantic approach is to develop a chatbot application that answer natural language questions about Scientometric Indicators. A chatbot is a machine conversation system that interacts with human users via natural conversational language [29]. Computer based chatbots are getting to be distinctly famous as an intuitive and successful open framework between human and machines [30]. Chatbots represent a potential means for automating customer service because it is provided through online chat and the fact of recent advances in artificial intelligence and machine learning including the general adoption of messaging platforms makes the chatbot application interesting to explore [31]. As customers become more demanding with quick response, accurate answers and professional services, some enterprises are now turning to chatbots to overcome these challenges [32]. The application will perform the translation from natural language to cypher langauge and perform the extraction of information to answer correctly the question. We also want to include metadata for describing the Scientometric Indicators in our model, this will allow us to differentiate the indicators into categories and make them easy to query. In this research work we use a sample of 10 Scientometric Indicators that enable the definition of 2 dimensions. We see the Scientometric Indicators in a general way and possible approach to be more specific is to lower the information to researchers and publications as classes, this can allow us to have several types of queries.

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/`

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

Ethics approval and consent to participate
Not applicable.

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

Consent for publication
Not applicable.

Authors' contributions

V. Lopez: Software, Formal Analysis, Data Curation, Visualization, Writing - Original Draft. H. Ceballos: Conceptualization, Methodology, Validation, Writing - Review & Editing.

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.

Author details

School of Engineering and Sciences, Tecnologico de Monterrey, Nuevo Leon, MX.

References

1. Vitanov, N.K., Vitanov, Vitanov, N.: Science Dynamics and Research Production. Springer, ??? (2016)
2. Hood, W.W., Wilson, C.S.: The literature of bibliometrics, scientometrics, and informetrics. Scientometrics **52**(2), 291–314 (2001)
3. Vinkler, P.: The Evaluation of Research by Scientometric Indicators. Elsevier, ??? (2010)
4. Roussey, C., Pinet, F., Kang, M.A., Corcho, O.: An Introduction to Ontologies and Ontology Engineering. In: Ontologies in Urban Development Projects vol. 1, pp. 9–38. Springer, London (2011). doi:10.1007/978-0-85729-724-2_2.
5. Shachnev, D., Karpenko, D.: Using subject area ontology for automating processes in sphere of scientific investigation and education. Programming and Computer Software **44**(1), 15–22 (2018)
6. Braun, T., s Schubert, A., *et al.*: Scientometric Indicators: a 32 Country Comparative Evaluation of Publishing Performance and Citation Impact. World Scientific, ??? (1985)
7. Capadisli, S., Auer, S., Riedl, R.: Towards linked statistical data analysis. In: SemStats@ ISWC (2013)
8. Colledge, L.: Snowball Metrics Recipe Book. Elsevier, ??? (2017)
9. Jörg, B.: The common european research information format model (cerif). CRISs for the European e-Infrastructure. Data Science Journal (in Print, 2009) (2009)
10. WEB, M.D.O.S.: Modeling and querying spatiotemporal multidimensional data on semantic web: A survey. Journal of Theoretical and Applied Information Technology **97**(23) (2019)
11. Karaiskos, D.C., Xinidis, D., Bonis, V.: R&d statistics information system: An interoperability tail between cerif and sdmx. Procedia Computer Science **106**, 87–94 (2017)
12. Fox, M.S.: The semantics of populations: A city indicator perspective. Journal of Web Semantics **48**, 48–65 (2018)
13. 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) (2020)
14. Hu, Y., Janowicz, K., McKenzie, G., Sengupta, K., Hitzler, P.: A linked-data-driven and semantically-enabled journal portal for scientometrics. In: International Semantic Web Conference, pp. 114–129 (2013). Springer
15. Osborne, F., Peroni, S., Motta, E.: Clustering citation distributions for semantic categorization and citation prediction (2014)
16. Krafft, D.B., Cappadona, N.A., Caruso, B., Corson-Rikert, J., Devare, M., Lowe, B.J., Collaboration, V., et al.: Vivo: Enabling national networking of scientists (2010)
17. Beckett, D., Berners-Lee, T., Prud'hommeaux, E., Carothers, G.: Rdf 1.1 turtle. World Wide Web Consortium, 18–31 (2014)
18. Sayers, C., Eshgi, K.: The case for generating uris by hashing rdf content. Aug **22**, 1–13 (2002)
19. Cyganiak, R., Field, S., Gregory, A., Halb, W., Tennison, J.: Semantic statistics: Bringing together sdmx and scovo. In: LDOW (2010)
20. Miles, A., Matthews, B., Wilson, M., Brickley, D.: Skos core: simple knowledge organisation for the web. In: International Conference on Dublin Core and Metadata Applications, pp. 3–10 (2005)
21. Baker, T., Bechhofer, S., Isaac, A., Miles, A., Schreiber, G., Summers, E.: Key choices in the design of simple knowledge organization system (skos). Journal of Web Semantics **20**, 35–49 (2013)
22. Chaudhuri, S., Dayal, U.: An overview of data warehousing and olap technology. ACM Sigmod record **26**(1), 65–74 (1997)
23. Shearer, C.: The CRISP-DM model: the new blueprint for data mining. J Data Warehousing **5**, 13–22 (2000)
24. Neo4j: Neo4j Graph Database (2021). `https://neo4j.com/product/#graph-database` Accessed Accessed 20 Jun 2021
25. Fernandes, D., Bernardino, J.: Graph databases comparison: Allegrograph, arangodb, infinitegraph, neo4j, and orientdb. In: Data, pp. 373–380 (2018)
26. Lal, M.: Neo4j Graph Data Modeling. Packt Publishing Ltd, ??? (2015)
27. Neo4j Labs: neosemantics (n10s): Neo4j RDF & Semantics toolkit (2021). `https://neo4j.com/labs/neosemantics/` Accessed Accessed 20 Jun 2021
28. Neo4j: Cypher Query Language (2021). `https://neo4j.com/product/#cypher` Accessed Accessed 20 Jun 2021

29. Shawar, B.A., Atwell, E.S.: Using corpora in machine-learning chatbot systems. International journal of corpus linguistics **10**(4), 489–516 (2005)
30. Ranoliya, B.R., Raghuwanshi, N., Singh, S.: Chatbot for university related faqs. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1525–1530 (2017). IEEE
31. Følstad, A., Nordheim, C.B., Bjørkli, C.A.: What makes users trust a chatbot for customer service? an exploratory interview study. In: International Conference on Internet Science, pp. 194–208 (2018). Springer
32. Diederich, S., Brendel, A.B., Lichtenberg, S., Kolbe, L.: Design for fast request fulfillment or natural interaction? insights from an experiment with a conversational agent (2019)