Journal of Hydroinformatics

Journal of Hydroinformatics Vol 24 No 2, 295 doi: 10.2166/hydro.2022.070

Characterizing water quality datasets through multi-dimensional knowledge graphs: a case study of the Bogota river basin

Juan D. Rondón Díaz[a] and Luis M. Vilches-Blázquez[b,*]

[a] Universidad Nacional de Colombia – Sede Bogotá, Bogotá, Colombia
[b] Centro de Investigación en Computación, Instituto Politécnico Nacional, Av. Miguel Othón de Mendizabal s/n, UPALM—Zacatenco, Mexico City 07738, Mexico
*Corresponding author. E-mail: lmvilches@cic.ipn.mx

LMV, 0000-0001-5799-469X

ABSTRACT

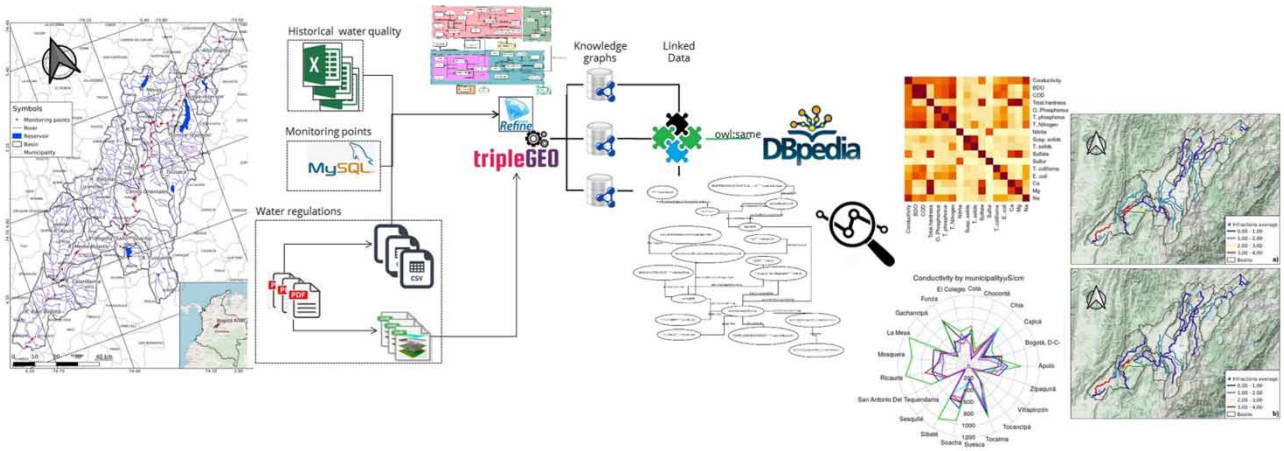
The world is transforming into a predominantly urban space, meaning that cities have to be ready to provide services, for instance, to ensure availability and sustainable management of water and sanitation for all. In this scenario, the water quality evaluation has a crucial role and often needs multiple sources segregated. Our purpose is to build bridges between these data silos to provide an integrated and interoperable view, where different datasets can be provided and combined through knowledge graphs in order to characterize water quality. This work shows the quality of the Bogota river basin's water bodies by analyzing physicochemical and biological properties using spatio-temporal and legal elements. So, our knowledge graphs allow us to discover what, when, and where infractions happened on water quality in a river basin of the most populated cities of Latin America during a critical period (2007–2013), highlighting the presence of high values of suspended solids and nitrites, lower amounts of dissolved oxygen, and the worst water quality during the driest periods (appearing until a maximum of 63 infractions in a year).

Key words: Bogota river basin, knowledge graph, ontology, spatio-temporal characteristics, water quality

HIGHLIGHTS

- A new water quality ontology with three modules composed of diverse international standards.
- Multi-dimensional knowledge graphs about the water quality of the Bogota river basin were developed.
- The water quality characterization using spatio-temporal distribution and legal framework from an integrated and interoperable scenario.

GRAPHICAL ABSTRACT



INTRODUCTION

The world is transforming into a predominantly urban space, and in this sense, cities around the globe have experienced unprecedented growth during the last four decades. According to the United Nations (United Nations 2018a), 55% of the world's population resides in urban places, a proportion that is expected to increase to 70% by 2050. In Latin America and the Caribbean, the tendency toward urbanization has been even more dramatic, going from 50.5% in 1961 to 81% in 2018 (United Nations 2018b).

This growth of urban areas around the world makes cities and their populations more vulnerable to global warming and climate change (Valipour *et al.* 2021) since these spaces are often located in hazardous locations where economic assets and residents increasingly find themselves at elevated risk of climate-related events (Gasper *et al.* 2011). Therefore, this urban context entails that cities need to be ready to confront these *new* scenarios and, among other issues, to ensure availability and sustainable water and sanitation management for all. It is part of the Sustainable Development Goals; concretely, goal 6 is related to the following target: *improving water quality by reducing pollution, eliminating dumping, and minimizing release of hazardous chemicals and materials, halving the proportion of untreated wastewater and substantially increasing recycling and safe reuse globally*.

This scenario points out the well-known vital importance of water because it is essential for life and, consequently, its access to all people must be ensured through an adequate, sufficient, accessible, and innocuous supply concerning what is referred to as its quality (WHO 2017; Pahl-Wostl 2020). In addition, the effects of global warming and climate change affect available water and impact its quality due to factors such as the low dilution of the pollutant load (Radhapyari *et al.* 2021).

The water quality monitoring and analysis is a fertile research area (Resh & Unzicker 1975; Ward *et al.* 1986; Ouyang 2005; Baltacı *et al.* 2008; Behmel *et al.* 2016; Sankary & Ostfeld 2018; Aguilar *et al.* 2019; Alilou *et al.* 2019; Shi *et al.* 2019). This area has a multidisciplinary character that attracts the study of hydrologists, ecologists, environmental scientists, or land/agricultural scientists (Giri & Qiu 2016). So, it often needs multiple and 'third party' data sources to deal with the extent, complexity, and uncertainty of environmental issues (Raymond *et al.* 2010; Liu *et al.* 2013). This involves data unavailable on a single source, entailing diverse models, formats, or vocabularies because they are segregated from the others in most cases (Vilches-Blázquez *et al.* 2014; Delpla *et al.* 2020).

This scenario faces significant technical difficulties in collection methods and analysis strategies due to data remaining as data silos. Therefore, it creates a need for users to deploy and learn a wide diversity of software (Zhao *et al.* 2012) or have some necessary programming abilities (Arribas-Bel 2014). As a result, data analysis converts an opportunity reserved by some users, and many data cannot be examined enough since it entails primary barriers to taking advantage of them (Zhao *et al.* 2012; Arribas-Bel 2014).

In this line, our purpose focuses on building bridges between these data silos to provide an integrated and interoperable view of multi-domain (hydrologic, legal, environmental, and ecological) multi-dimensional (spatio-temporal, multivariate, and multi-valued properties related to physicochemical and biological parameters) water quality data. These bridges are created by means of knowledge graphs that combine multiple and heterogeneous data sources, supporting a holistic understanding of water quality issues (Janowicz *et al.* 2020). We chose to use knowledge graphs because they have appeared as an extension of Semantic Web practices and are embraced by diverse companies such as Google, IBM, Facebook, or Microsoft (Noy *et al.* 2019). Knowledge graphs promote the creation, reuse, and recovery of human and machine-readable structured data about real-world objects using a graph-based representation (Paulheim 2017). Additionally, they entail that water quality data adopt the best practices for exposing, sharing, and integrating data on the Web (Van den Brink *et al.* 2019), so domain experts may convert information into actionable knowledge (Masmoudi *et al.* 2020). In this way, knowledge graphs have become one of the principal ways to integrate diverse data (Cudré-Mauroux 2020) so that multiple heterogeneous datasets can be handled and interlinked in a single system (Krötzsch & Thost 2016; Bellomarini *et al.* 2020).

According to Ehrlinger & Wöß (2016), a knowledge graph is defined as a multi-relational graph formed by entities and relationships between them. These graphs promote the creation, retrieval, and reuse of human- and machine-readable structured data about real-world objects (Paulheim 2017). In this way, knowledge graphs have become one of the principal ways to integrate diverse data (Cudré-Mauroux 2020) so that multiple heterogeneous datasets can be handled and interlinked in a single system (Bellomarini *et al.* 2020). Their adoption entails a new perspective on structuring, integrating, publishing, discovering, accessing, and reusing data (Van den Brink *et al.* 2019) to deal with multiple and fragmented water data. Thus,

knowledge graphs offer experts a semantically connected and interoperable view of data for evidence-based decision-making (Pahl-Wostl 2020) and bring them closer to the expert analysis to end-users to discover issues about water quality (*what*, *when*, and *where*) in a river basin.

Several works highlight the importance of integrating and combining data from multiple and heterogeneous sources using ontologies and Linked Data principles. There are some relevant works in the environmental sciences area (CUAHSI 2010; Hunter *et al.* 2011; Curry *et al.* 2014; DEFRA 2014; Kämpgen *et al.* 2014; Masmoudi *et al.* 2020). However, even though some of these works focused on monitoring water resources, there are no previous contributions where knowledge graphs with spatio-temporal and legal components are developed to characterize water quality in a river basin to the best of our knowledge.

This work makes the following original contributions: (i) it proposes constructing knowledge graphs for integrating various and heterogeneous data sources of the water quality domain. The graphs contain multi-dimensional data associated with physicochemical and biological properties, spatio-temporal elements, and legal characteristics of the Bogota river basin's water bodies; (ii) it presents an ontology-based knowledge representation using diverse international standards; (iii) it assesses our study area's water quality, applying semantic queries in order to exploit the integrated and interoperable view of data; and (iv) it characterizes water quality spatio-temporal distribution using the legal aspects as a cornerstone and sets correlations between various physicochemical and biological properties captured through multiple sampling points. In addition, this work promotes the possibility to create automated machine-to-machine workflows for data discovery and analyses since the developed knowledge graphs utilize several ontologies and standards to automate multi-domain model coupling and data transformation in the context of many hydroinformatics applications. In this sense, this work shows the benefits of using technologies associated with knowledge graphs to integrate diverse data sources and analyze water quality in a case study associated with one of the most populated Latin American cities.

This paper is organized into sections that describe: (1) a brief context of the study area, characteristics of considered data sources, and the applied methodology; (2) related works on water and semantics; (3) the details of the ontology developed by water quality and data transformation and enrichment processes; (4) our results and discussions; and (5) some conclusions and future lines of work.

RELATED WORK

Several works relate hydrological and environmental topics to semantic issues in the literature. So, Yu *et al.* (2015) presented netCDF-LD, a new convention for encoding netCDF based on Linked Data principles. It allowed metadata elements to be presented as Linked Data resources, improving data discovery. CUAHSI (2010) described a system for sharing hydrological information, whereas in DEFRA (2014), the Department for Environment Food & Rural Affairs (DEFRA) published historical and updated information on the quality of bathing waters, as well as information on basins following the Linked Data principles (http://environment.data.gov.uk).

In addition to these works, some initiatives focused on the integral management of water. In Wilson *et al.* (1997), the authors presented the design and development of the Water Resource Management System (WREMS). Besides, different details about the linkages between the data management and optimization components of the system were discussed. Curry *et al.* (2014) it described the usage of Linked Data technologies for the water management domain, presenting their architectural approach and discussing possible water management applications. In Kämpgen *et al.* (2014), the authors formalized the Integrated Water Resources Management (IWRM) domain in an OWL ontology. Also, they used Linked Data and multi-dimensional modeling based on the RDF Data Cube Vocabulary to develop a knowledge base with research data.

On the other hand, there are some works where water quality is dealt with semantic technologies. In Jajaga *et al.* (2016), the authors provided details about a system that handles stream data and detects inadequate water quality statuses. The system also recognized the potential sources of pollution by extending a previously developed ontology (Jajaga *et al.* 2015). In Jajaga *et al.* (2015), it is described as an extension of the SSN ontology with a module for distinguishing the potential pollution sources. It was validated in the context of the InWaterSense project. In Hunter *et al.* (2011), the authors developed a framework and set of services to facilitate streamlined access to real-time, near-real-time, and static datasets related to water resource management in South East Queensland. Moreover, the authors also built ontologies and semantic querying tools to connect some datasets and perform management actions to water quality indicators in particular regions and periods.

Several works have appeared in the literature about our study area. Rodríguez-Jeangros *et al.* (2018) and Díaz-Casallas *et al.* (2019) focused on generating inputs to evaluate the river basin's pollution degree utilizing mathematical models or water

quality indexes. Giraldo & Garzón (2002), Miranda *et al.* (2008), and Rodríguez Forero *et al.* (2009) performed a diagnosis of the Bogota river basin as a result of the loss of water quality. These works considered the water resource state and its effects on soils, agricultural activities, or influence on animals' and plants' conditions.

Despite existing various proposals related to this work, we identified no previous approaches where knowledge graphs were developed from multi-dimensional data sources (water quality, spatio-temporal, and legal components). Hence, these graphs have not been used to assess water quality in a river basin. Moreover, there are no approaches where the spatial dimension of water quality is dealt with from a semantic perspective, more concretely, using GeoSPARQL, and neither appear an ontology-based knowledge representation to model multi-dimensions of water quality in the literature.

MATERIALS AND METHODS

Study area

The Bogota river basin is located in the Department of Cundinamarca (Colombia), comprises 589.46 hectares (around 32% of the Department's surface), and is 336 km in length. The river basin is composed of 19 sub-basins, whose borders are: on the North with the *Boyacá* Department, with the Department of *Tolima* at the South, while to the West the basin limits to the municipalities of *Guayabal de Síquima*, *Sasaima*, *Bituima*, *San Francisco*, *Pacho*, *Albán*, *La Vega*, *Supatá*, and *Pacho*. To the East, the basin limits the municipalities of *Nile*, *Silvania*, *Chipaque*, *Choachí*, *Nilo*, and *Ubaque*. A map of this river basin with its sub-basins and monitoring points is depicted in Figure 1.

This river basin flows through 25 municipalities (Ricaurte, Agua de Dios, Girardot, Tocaima, El Colegio, Apulo, Soacha, Anapoima, San Antonio de Tequendama, Tena, La Mesa, Mosquera, Funza, Bogotá, Cota, Chía, Sopó, Cajicá, Tocancipá, Gachancipá, Sesquilé, Zipaquirá, Suesca, Chocontá, and Villapinzón) where more than 9 million people are living. Hence, the river basin is a critical resource for these urban context and their people.

However, the Bogota river presents several potential pollution problems, which have their origin mainly in domestic wastewater discharges, slaughters of livestock, and industrial wastewater. It occasions an environment composed of local wastewater producers and spills from animals and several activities related to leather tanning, mining, textile, papermaking, and glass, among other activities. Consequently, it is necessary to monitor water properties adequately to formulate control strategies to promote better water quality (Camacho 2020).

Data sources and pre-processing

To generate the knowledge graphs and, later, assess the water quality in the Bogota river basin from an interoperable and holistic perspective, we took into account the following heterogeneous and multi-dimensional data sources.

Historical water quality data

The historical water quality data were gathered from the Regional Autonomous Corporation of Cundinamarca (also known as CAR), the principal environmental agency responsible for collecting and managing the Bogota river basin. These data are associated with laboratory reports of the physicochemical and microbiological analysis belonging to 81 sampling points within the Bogota river basin. The water quality parameters, originally stored in multiple spreadsheets in the Microsoft Excel format with different characteristics, correspond to two monitoring campaigns (dry and wet seasons) per year from 2007 to 2020. The mentioned spreadsheets contain the following parameters: *Sample number*, *flow rate*, *conductivity*, *Biochemical Oxygen Demand*, *Chemical Oxygen Demand* (COD: mg/L), *total hardness*, *orthophosphate*, *Total Phosphorus* (TP: mg/L), *ammonia*, *Total Nitrogen* (TN: mg/L), nitrate, *nitrite*, *dissolved oxygen*, *pH*, *Suspended Solids* (SS: mg/L), *total solids*, *sulfate*, *sulfur*, *turbidity*, *total coliform*, *Escherichia coli*, *cadmium*, *calcium*, *chromium VI*, *chromium*, *magnesium*, *manganese*, *lead*, *sodium*, and *sodium adsorption ratio*. In addition, these datasets also incorporate data about rain, temperature, timestamp, sampling, and water type.

Considering multiple and heterogeneous spreadsheets provided by the CAR, we performed a pre-processing on these data to harmonize their structure and clean up (deleting redundant data and atypical records, setting sampling dates, standardizing measurement units, etc.) using the R language to facilitate the next steps of these datasets.

Water regulations

The official regulations source considered in this work is a PDF format (concretely, scanned documents). This source comprises the range with thresholds of allowed concentration values according to each water body of the Bogota river basin.

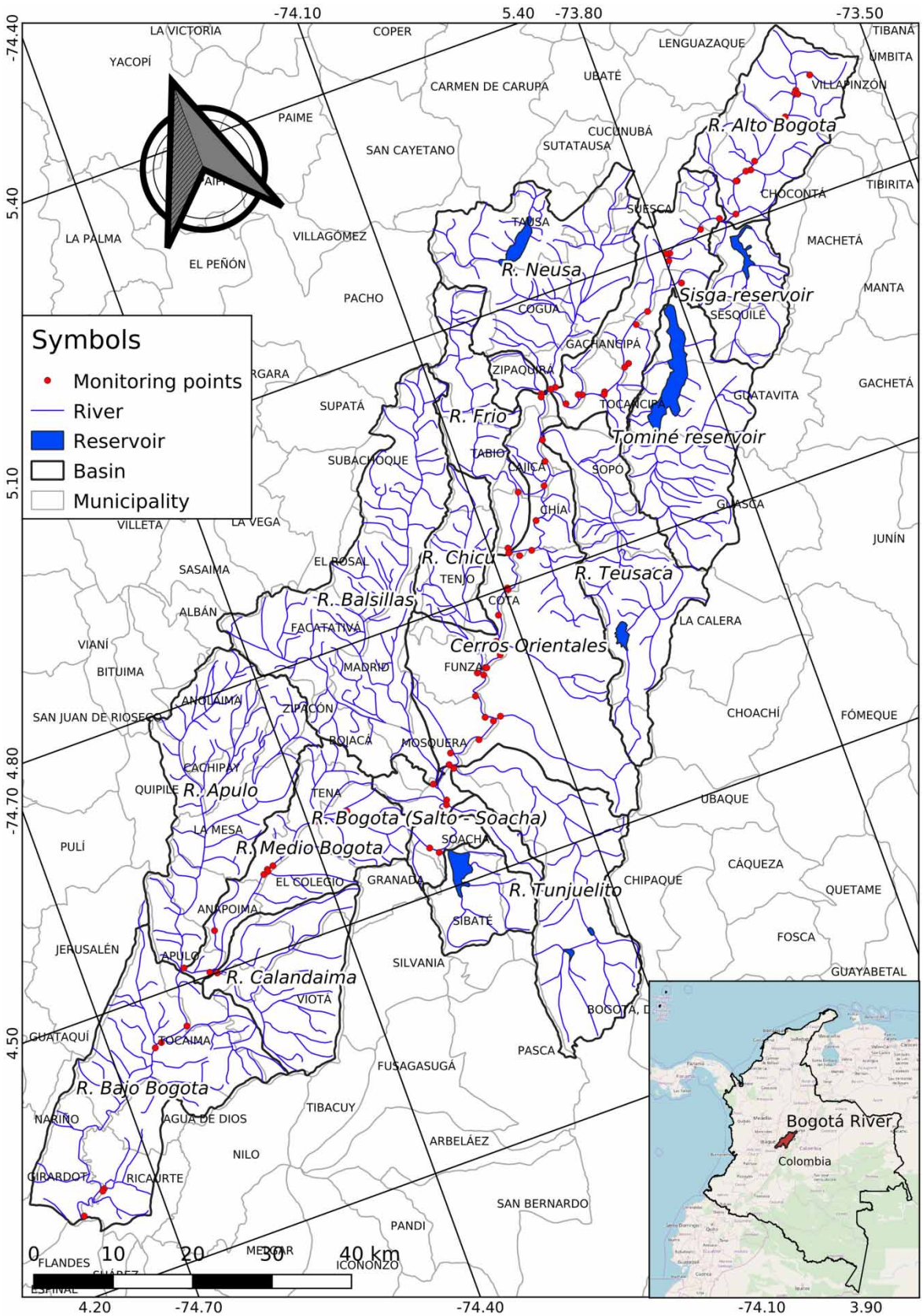


Figure 1 | The Bogota river basin.

Considering the initial situation of these data sources, we developed a comma-separated value (CSV) file with the following information:

- *CLASS I* presents the values of water used for human consumption and domestic with conventional treatment, conservation of flora, fauna, and agricultural and livestock uses.
- *CLASS II* is related to water-use values for human consumption and domestic with conventional treatment, agricultural use with restrictions, and livestock use.
- *CLASS III* describes the values assigned to the quality of the reservoirs, lagoons, wetlands, and other lenticular water bodies located within this river basin.
- *CLASS IV* defines agricultural use values with restrictions and livestock.
- *CLASS V* sets values of uses for power generation and industrial use.

Spatial data

We utilized the coordinates associated with 81 monitoring points of the Bogota river basin, stored in a MySQL database, and information about administrative boundaries (municipalities). Both spatial datasets were in MAGNA-SIRGAS – the conventional reference system for Colombia (EPSG:3116).

Additionally, the hydrological units' data are described in the legal documents mentioned above. It is important to notice that although it is spatial information, there are no official cartographic references related to these units. Therefore, we built the boundaries of water bodies and watersheds of the Bogota river basin's hydrological units contained in the aforementioned legal documents. So, we created a shapefile with these units, adopting the EPSG:3116 to guarantee the accuracy of the data using the national reference system. To these geometries (polygons), we added an attribute with the class (*I*–*V*) defined for each unit by the regulatory documents.

Methodology

In order to integrate diverse mentioned datasets and generate our water quality knowledge graphs of the Bogota river basin, we adopted Linked Data principles (http://linkeddata.org) because they propose best practices for providing, sharing, and integrating data on the Web (Heath & Bizer 2011). The Linked Data principles (Berners-Lee 2006) comprise: (1) to use of URIs as names for things; (2) to use of HTTP URIs so that people can look up those names; (3) when someone looks up a URI, provide useful information, through standards RDF and SPARQL Query Language for RDF (SPARQL); and (4) to include links to other URIs so that they can discover more things. These best practices are being adopted by an increasing number of data providers, pointing to creating a global data space with billions of assertions, that is, the Web of Data (Heath & Bizer 2011).

Our work follows the methodological guidelines proposed by Vilches-Blázquez *et al.* (2014) to generate the knowledge graphs. This methodology suggests an iterative incremental life cycle model characterized by a continuous process of improvement and extension of data and presents the subsequent steps: (1) specification, (2) modeling, (3) RDF generation, (4) links generation, (5) publication, and (6) exploitation. Further details about the guidelines and their steps can be found in Vilches-Blázquez *et al.* (2014). Next, we provide some aspects of different phases with several examples related to our work.

CONSTRUCTING A WATER QUALITY ONTOLOGY NETWORK

Ontologies have been proposed as formal models of how a domain is observed and present an accurate, logical description of the intended meaning of terms, data structures, and other components describing the real world (Hakimi *et al.* 2020). In addition, ontologies have been applied to the explication of hidden and implicit knowledge to overcome semantic heterogeneity problems (Wache *et al.* 2001) since they allow modeling semantic relationships between distinct structures and forming an integrated and coherent view of multiple and heterogeneous datasets (Krötzsch & Thost 2016).

Consequently, one of the main elements of this work is to develop an ontology-based knowledge representation to model water quality semantically, incorporating its spatio-temporal and legal components. For achieving this goal, we applied the NeOn methodology (Suarez-Figueroa *et al.* 2012), which defines nine scenarios that cover commonly occurring circumstances in the ontological development process, e.g., when available ontologies require to be aligned, re-engineered, modularized, and integrated with non-ontological resources, putting a particular emphasis on reusing and re-engineering knowledge resources (ontological and non-ontological). Further details of the NeOn methodology and its related scenarios

are provided in Suarez-Figueroa *et al.* (2012). Next, we show details associated with different modules that define our ontology.

Creating modules of the ontology

The water quality ontology comprises three modules associated with water observations, regulations, and spatio-temporal components to represent multiple parameters and dimensions of diverse data sources:

- *Water observations module*: The development of the water observation module was built by applying Scenario 2 (reusing and re-engineering non-ontological resources), Scenario 3 (reusing ontological resources), and Scenario 4 (reusing and re-engineering ontological resources) of the NeOn methodology. This module contains physicochemical and microbiological parameters, sampling points, hydrological units, and associated methods.

  Among non-ontological resources (Scenario 2), we considered the aforementioned spreadsheets, where parameters and observations of the study area appear; dictionaries, which were included to harmonize the list of physicochemical and microbiological parameters according to the CUAHSI Hydrologic Information System (HIS) (http://hiscentral.cuahsi.org); and WaterML (https://www.opengeospatial.org/standards/waterml), an Open Geospatial Consortium (OGC) standard for the representation of water observations data.

  Regarding ontological resources, we selected and reused (Scenario 3) three ontological resources to develop this module: the standard RDF Data Cube (https://www.w3.org/TR/vocab-data-cube/), which provides a means to publish multi-dimensional data on the Web using the W3C RDF (Resource Description Framework) standard; the Observations & Measurements (O&M) ontology, which is associated with an OGC standard (https://www.ogc.org/standards/om) and is taken as a starting point for creating an ontology according to the OGC WaterML 2.0 standard; and the INWATERSENSE system (INWS) ontology (Ahmedi *et al.* 2013), a Semantic Sensor Network-based ontology for water quality management developed to promote water quality classification based on different regulatory authorities.

  Although the mentioned ontologies are reused, they do not constantly adapt appropriately to the necessities of this work. Therefore, we carried out a re-engineering of ontological resources (Scenario 4). For instance, we performed a re-engineering process of the INWS ontology since this proposal models water quality focusing on sensors data, and our work handles data provided by manual sampling.
- *Water regulations module*: This module addresses water parameter thresholds regulated by the Colombian authorities. We used national water regulations to develop this module, which define the degree of compliance, permitted uses of water bodies, and the set of rules that apply to each sample over time. To incorporate these resources into our ontology network, we employed Scenario 2 since they are available as PDF documents and, therefore, are non-ontological resources.
- *Spatio-temporal module*: This module is developed following Scenario 3 of the mentioned methodology. On the one hand, we reused the GeoSPARQL standard (Perry & Herring 2012) for spatial elements, a vocabulary for describing geospatial data in RDF, and an extension to the SPARQL query language for processing geospatial data. This proposal allows distinct geometries (e.g., points, lines, polygons, and multipoints), adds multiple coordinate reference systems, and expresses spatial relations for querying geographic datasets (e.g., intersects, touches, and overlaps). Then, geometries are defined by the class *Geometry*, and the coordinates can be encoded using Well-Known Text (WKT) or Geography Markup Language (GML). On the other hand, we incorporated the DBpedia ontology (http://wiki.dbpedia.org/services-resources/ontology), a cross-domain ontology generated based on the frequently used infoboxes within Wikipedia. Concerning time issues, they are modeled with the Time Ontology (http://www.w3.org/TR/owl-time/), an ontology for temporal notions developed by W3C that implements a vocabulary about durations and date-time information.

Creating modules of the ontology

Herein, we aim to show how described modules are connected in order to build the water quality ontology. For that, we defined some steps between each developed module:

- *Specification of taxonomy mappings*: Various semantic equivalences between elements of developed modules were set. For instance, the *qb:Observation* class of RDF Data Cube vocabulary is defined as an *owl:EquivalentTo sosa:Observation* class, part of the Semantic Sensor Network ontology. It is formalized in the axiom definition (1).

$$qb{:}Observation \equiv sosa{:}Observation \tag{1}$$

- *Re-engineering*: The *o&m:SamplingPoint* class, an element of the Observation & Measurement standard presented in the water observation module, was defined as a *rdfs:subClassOf geosparql:Point*. This latter class is a *rdfs:subClassOf geosparql:Geometry*. The axiom (2) formalizes this example of the mentioned re-engineering process.

$$o\&m{:}SamplingPoint \sqsubseteq geosparql{:}Point \sqsubseteq geosparql{:}Geometry \tag{2}$$

- *Setting relations*: Multiple relations were established between the mentioned modules. For example, a hydrological unit has associated geometry and a type of legal parameters through water classes. It is formalized in our ontology network according to the axiom (3), where the *wq:HydrologicalUnit* class, which is in the water observation module, has a *geosparql:asWKT* to represent its polygonal information.

$$wq{:}HydrologicalUnit \exists geosparql{:}asWKT.geometry \tag{3}$$

Moreover, elements associated with the water observation module were connected to the water regulation module. For instance, the *wq:HydrologicalUnit* (water observation module) and *wr:WaterClass* (water regulation module) are connected using the *wr:hasWaterClass* relation. The axiom (4) shows the formalization of this connection between elements of different modules.

$$wr{:}hasWaterClass \exists wr{:}hasWaterClass.wr{:}WaterClass \tag{4}$$

In this way, we developed our ontology, defining relations between different modules to provide a spatio-temporal and legal context for the water quality assessment in the study area. Figure 2 shows an overview of the developed ontology network supported by Protégé (https://protege.stanford.edu/) and expressed in OWL2 (https://www.w3.org/TR/owl2-overview/).

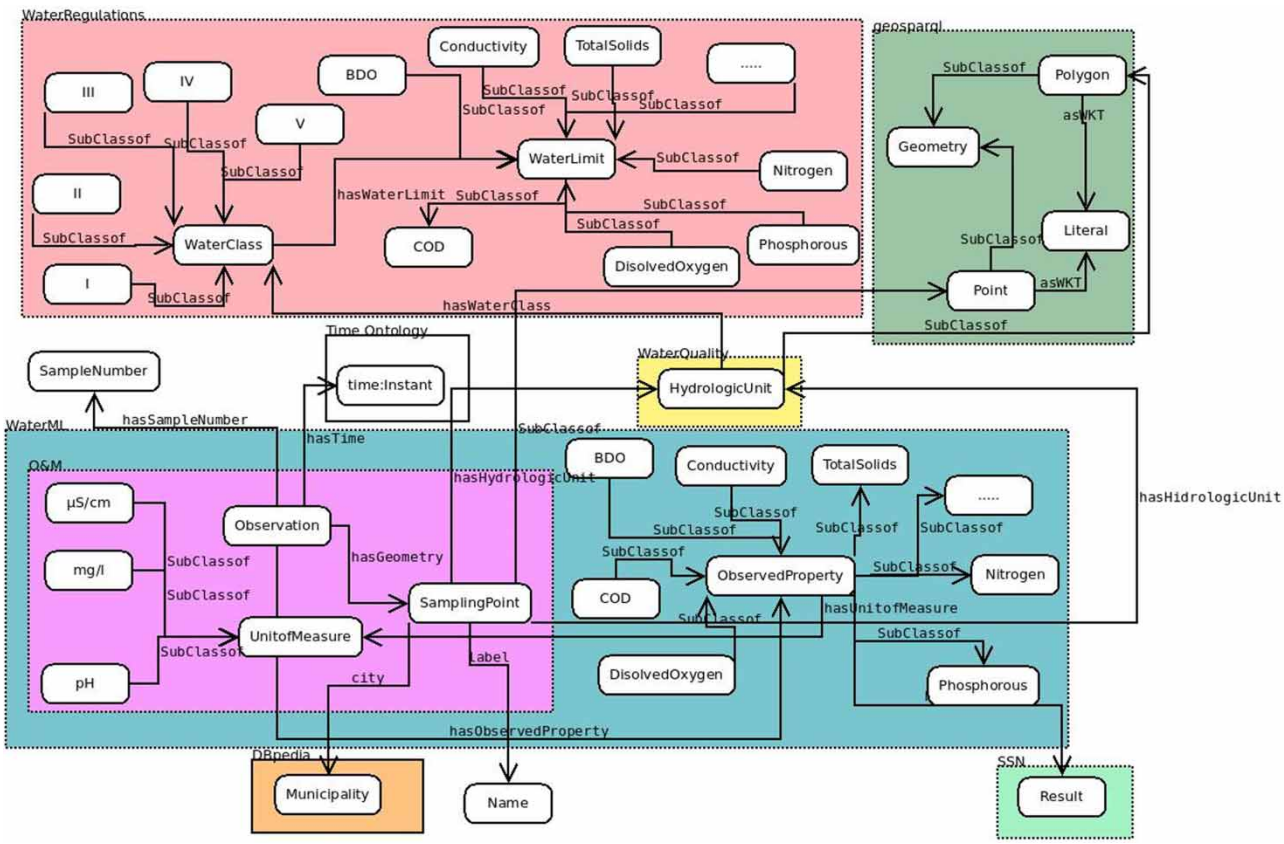


Figure 2 | An overview of the developed ontology.

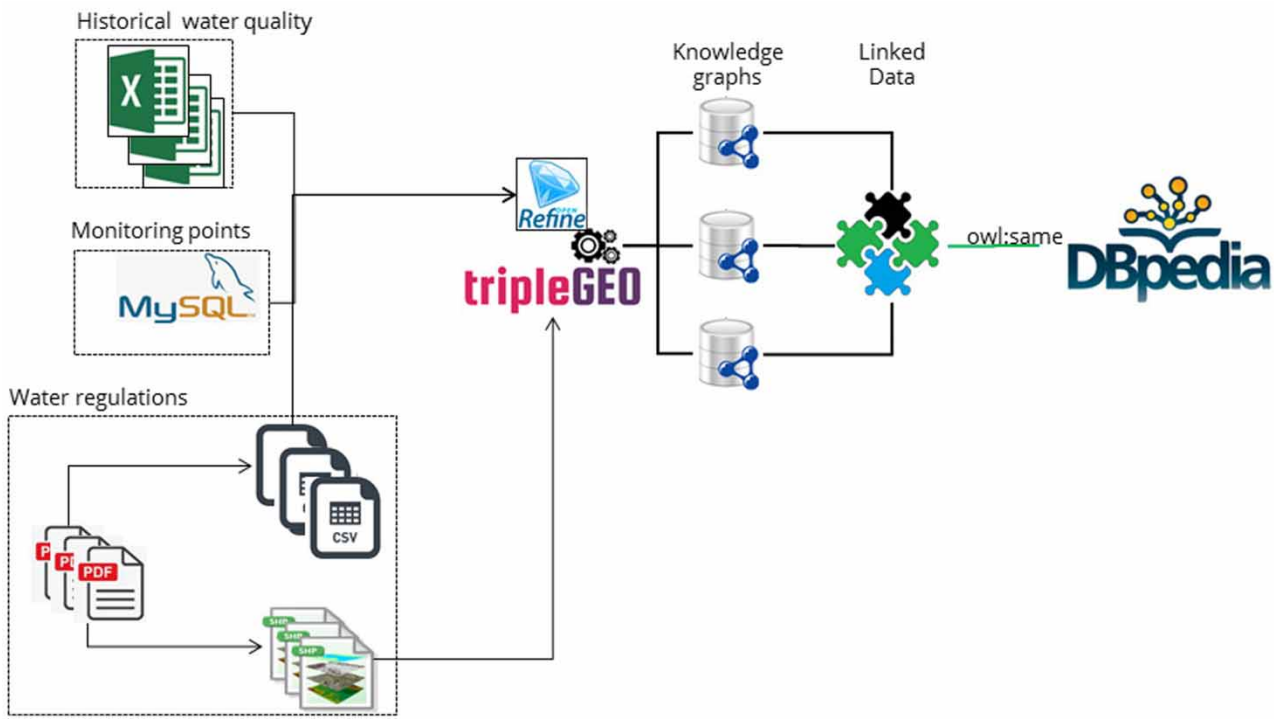


Figure 3 | Our workflow for transforming and enriching water quality data.

DATA TRANSFORMATION AND ENRICHMENT

The deployed workflow to transform and link multiple datasets described previously is shown in Figure 3. These datasets are converted into RDF according to Linked Data principles using our developed ontology. It allows generating three integrated knowledge graphs that provide (semantic) interoperability to multi-dimensional water data. Next, we provide details of this workflow's steps.

Data transformation

We selected RDF as the conventional way to generate the knowledge graphs since we required harmonizing different formats of our datasets (databases, shapefiles, CSV, and Microsoft Excels files) to evade employing proprietary formats and seek a (semantically) interoperable approach. Furthermore, RDF (https://www.w3.org/TR/rdf11-concepts/) is the standard knowledge representation language for the Semantic Web, providing a framework for sharing, publishing, and querying structured data on the Semantic Web (McDonald & Levine-Clark 2018). It is important to note that we utilized diverse URIs defined in our ontology network to transform original data into RDF. These URIs were set in each of the tools used in the proposed workflow, allowing our knowledge graph to be generated using our developed ontology.

- *Historical water quality data*: As we mentioned previously, historical water quality data are collected in laboratory reports in the Microsoft Excel format, where its parameters have associated some spatio-temporal characteristics. The conversion process to RDF was carried out with LODRefine, an OpenRefine (http://openrefine.org/) distribution with integrated extensions to achieve the transition from tabular data to Linked Data.
- *Water regulations data*: These data were initially described in a PDF document and transformed into a CSV file after a preprocessing task. Herein, each parameter's limits (maximum values) and water uses are defined according to each water body in the study area. Considering the input file format (CSV), we also executed the transformation process with LODRefine.
- *Spatial data*: The spatial data correspond to the monitoring points and the hydrological units collected in the considered data. Each of these datasets has a different provenance and characteristics. In the case of monitoring points, they were stored in a MySQL database using the MAGNA-SIRGAS coordinate reference system (EPSG:3116), whereas hydrological

units were created as a shapefile with boundaries for water bodies, watersheds, and their defined classes for each case in the Bogota river basin.

In both cases (monitoring points and hydrological units), we used TripleGeo (Patroumpas *et al.* 2014) to reproject them to WGS84 (EPSG:4326) and transform these files into RDF afterward. We reprojected WGS84 (EPSG:4326) to manipulate these data using standards since some technologies do not support projected coordinate systems, as in the MAGNA-SIRGAS case. Besides, it is worth noting that these datasets were transformed according to GeoSPARQL and serialized as WKT.

Connecting and enriching data

After knowledge graphs generation, we also connected some elements of these (RDF) graphs using the spatial component and enriched them with other external datasets published on the Web of Data (https://lod-cloud.net/) following the fourth principle of Linked Data.

With regard to spatial linking, we validated the connection between each water quality sample and its spatial locations. For this purpose, several spatial joins were executed using the QGIS tool (https://qgis.org/). In this process, we took the monitoring point's coordinates as a reference and applied methods for performing overlay analysis with administrative boundaries to check data accuracy. This process allowed setting spatial links between diverse used datasets.

On the other hand, we enriched our data with the Web of Data setting *owl:sameAs* links with DBpedia (https://wiki.dbpedia.org/), which is a cross-domain knowledge base as a result of a community effort to gather structured information from Wikipedia and to put this information available on the Web (Auer *et al.* 2007). In this way, we utilized LODRefine to connect municipalities from our knowledge graphs with DBpedia, enriching our data with information from this knowledge base.

All different outputs of each transformation, connection, and enrichment process (RDF dump) feed our graph database, allowing us to deploy our knowledge graphs. These graphs permitted the integration of all diverse and multi-dimensional data sources considered in this work, achieving an integrated and interoperable view of water quality for the study area. An example of this integration is shown in Figure 4, where the sample '1343' has a conductivity value, geometry, hydrological unit, use, and belongs to a municipality. The related RDF graph is presented in Figure 5.

RESULTS AND DISCUSSION

This section addresses the results of exploiting generated knowledge graphs. This exploitation allows multi-dimensional graphs to assess water quality in the Bogota river basin and identify water quality issues around *what*, *when*, and *where*. Next, we provide details of the knowledge graphs' query process, describe how they are used to exploit spatio-temporal and legal characteristics, and set correlations among water quality parameters.

Querying geospatial dimension

The knowledge graphs are stored and deployed by a graph database, called RDF store or knowledge bases. According to Battle & Kolas (2012), they can better manage different kinds of restrictions that relational databases struggle with or are not expected to complete: queries with multiple joins across entities, with mutable properties (Weiss *et al.* 2008), and ontological inference on graphs.

Our work deployed Parliament as our graph database, where we can ask our knowledge graphs through semantic queries using SPARQL. Listing 1 presents a query example to recover all those monitoring points (*SamplingPoint*) with samples (*?Sample*) that exceeded thresholds (*?BODLimit*) according to its specific use (*?class*) established by the regulatory framework in the case of the BOD parameter, as well as their associated coordinates (*?lat*, *?long*), hydrological units (*HydrologicUnit*), and date (*?Date*).

Listing 1: Semantic query with spatial and legal elements

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX observation: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#>
PREFIX wml: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX units: <http://www.opengis.net/def/uom/OGC/1.0/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

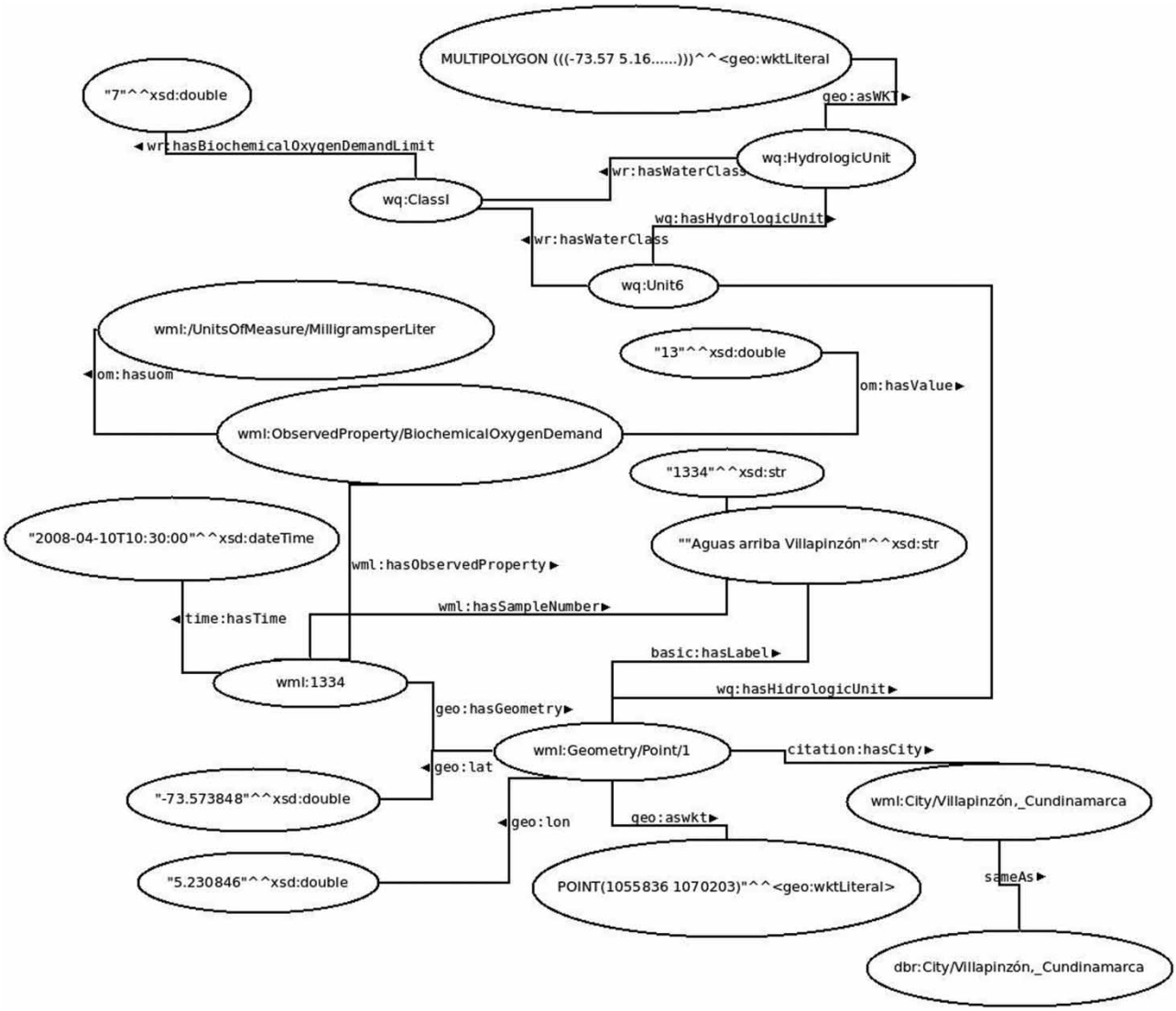


Figure 4 | RDF graph generated for observations and their sampling points.

```
PREFIX sampling: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#>
PREFIX wr: <http://www.waterregulations.gov.co/>
PREFIX citation: <http://def.seegrid.csiro.au/isotc211/iso19115/2003/citation#>

SELECT DISTINCT ?Sample ?Date ?class ?BOD ?BODLimit ?lat ?long
WHERE {
  ?unit a wr:HydrologicUnit.
  ?unit wr:hasWaterClass ?class.
  ?class <http://www.waterregulations.gov.co#hasBiochemicalOxygenDemandLimit>
  ?BODLimit.
  ?unit geo:asWKT ?geounit.
  ?point a sampling:SamplingPoint ;
  <http://www.w3.org/2003/01/geo/wgs84_pos#lat> ?lat ;
  <http://www.w3.org/2003/01/geo/wgs84_pos#long> ?long.
  ?Sample citation:date ?Date.
  ?Sample geo:hasGeometry ?point.
```

```
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix wr: <http://www.waterregulations.gov.co> .
@prefix wq: <http://www.waterquality.gov.co> .
@prefix dbr: <http://dbpedia.org/resource/> .
@prefix time: <https://www.w3.org/TR/owl-time/#> .
@prefix wml: <http://WaterML.org/> .
@prefix iso19156: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/observation#> .
@prefix iso19103: <http://def.seegrid.csiro.au/isotc211/iso19103/2005/basic#> .
@prefix sfm: <http://def.seegrid.csiro.au/isotc211/iso19156/2011/sampling#> .
@prefix iso19115: <http://def.seegrid.csiro.au/isotc211/iso19115/2003/citation#> .
@prefix class: <http://www.waterregulations.gov.co/WaterClass/> .

# Observation (Sample)
wml:1334 a iso19156:Observation ;
    geo:hasGeometry <http://WaterML.org/Geometry/Point/1> ;
    wml:SampleNumber "1334" .

#Observed parameter (Biochemical Oxygen Demand - BOD)
_:nodelat4pofs9x2 iso19156:observedProperty wml:BiochemicalOxygenDemand ;
    iso19156:value "2" ;
    iso19103:uom  <http://www.WaterML.org/UnitsOfMeasure/MilligramsPerLitre> .

#Location of observation (Sample)
<http://WaterML.org/Geometry/Point/1> a sfm:SamplingPoint ;
    rdfs:label "Aguas arriba Villapinzon" ;
    geo:asWKT "POINT(1055836 1070203)"^^<geo:wktLiteral> ;
    iso19115:city dbr:Villapinzon,_Cundinamarca ;

# Date of observation
wml:1334 wml:WaterTemperature _:nodelat4pofs9x29 ;
    time:Interval "2008-04-10T10:30:00"^^xsd:dateTime .

# Related Hydrologic Unit and class with its coordinates
wq:Unidad6 a wq:HydrologicUnit ;
    wr:hasWaterClass class:ClassI ;
    geo:asWKT "<http://www.opengis.net/def/crs/OGC/1.3/CRS84>
    MULTIPOLYGON (((-73.5737087682013 5.16128464047719.....)))
    "^^<http://www.opengis.net/ont/sf#wktLiteral> .

# Class with value of parameter
class:ClassI wr:hasBiochemicalOxygenDemandLimit "7.0" .
```

Figure 5 | A depiction of the obtained RDF graphs.

?Sample <http://WaterML.org/BOD> ?bodnode.
?bodnode observation:value ?BOD.
?point rdfs:label ?name.
?point geo:asWKT ?geopoint.
FILTER (geof:sfContains(?geounit, ?geopoint) && ?BOD>?BODLimit)
}

The coordinates of each *SamplingPoint* are located at each *HydrologicUnit*. So, we included spatial analysis for making it explicit using (*sf:contains*) operator of GeoSPARQL. These units have a table of allowed maximum values (*?BODLimit*) that are a reference to detect infractions in the water quality of the Bogota river basin by means of a logical sentence (*?BOD* > *? BODLimit*).

The outputs of this query present the integration results of considered datasets in this work. They can be handy for performing analysis tasks of an expert responsible for evaluating where the regulatory framework was transgressed according to specific uses and destinies for each water resource. In this way, our approach helps to answer questions about *where* (place) and *when* (time) water quality parameters were exceeded.

This query is an example of question types used as inputs in the following results of our work, which were generated using the R language and various packages such as SPARQL, Shiny, and Leafet.

Temporal analysis and legal framework

Our approach allows contextualizing the monitoring points in their temporal and legal context since we generated knowledge graphs using historical series with physical–chemical parameters, location, hydrological units, and hydrological classes.

Thereby, we can identify *what* (parameters) and *when* (date) limits were exceeded according to the Bogota river basin's legal framework.

As an example, Figure 6 depicts the temporal distribution of diverse parameters (*Total coliforms*, *BOD*, *Suspended solids*, *Nitrites*, Nitrates, *Dissolved oxygen*, and *pH*) in *Socotá* during 2007–2013, a critical period for the water quality in the river basin that caused a subsequent World Bank project for clean up and flood control (https://bit.ly/3ylCxic). Additionally, the mentioned figure includes thresholds established by law, such as its upper (red line) and lower (blue line) limits.

In this case, we observed that diverse parameters surpass setting limits, such as *Suspended solids* and *Nitrites*, which were exceeded during the different years of the mentioned period. These cases are of particular interest since high values of both parameters may affect the health of citizens who drink this water. For example, high values of *Nitrites* affect oxygen fixation and its subsequent transport through the body tissue, especially in childhood (Basulto *et al.* 2014). In contrast, *Dissolved oxygen* presented lower values in 2009, 2011, and part of 2012 without complying with the law's provisions for more than half of the evaluated time. It is essential to remark that *Dissolved oxygen* is fundamental to guarantee favorable conditions for the aerobic organism and degrade organic material. On the opposite side, *Nitrates* and *pH* reported permissible values at this monitoring point.

We think these parameter variations are associated with anthropic activities carried out within this section of the river basin. These (domestic, agricultural, and industrial) activities, together with climate sensitivity to oscillation (Valipour *et al.* 2021) (dry and wet periods) and a lack of an adequate treatment of discharges, affect the water quality, entailing many parameter variations, and that they constantly exceed the maximum allowed by current legislation, as shown in Figure 6.

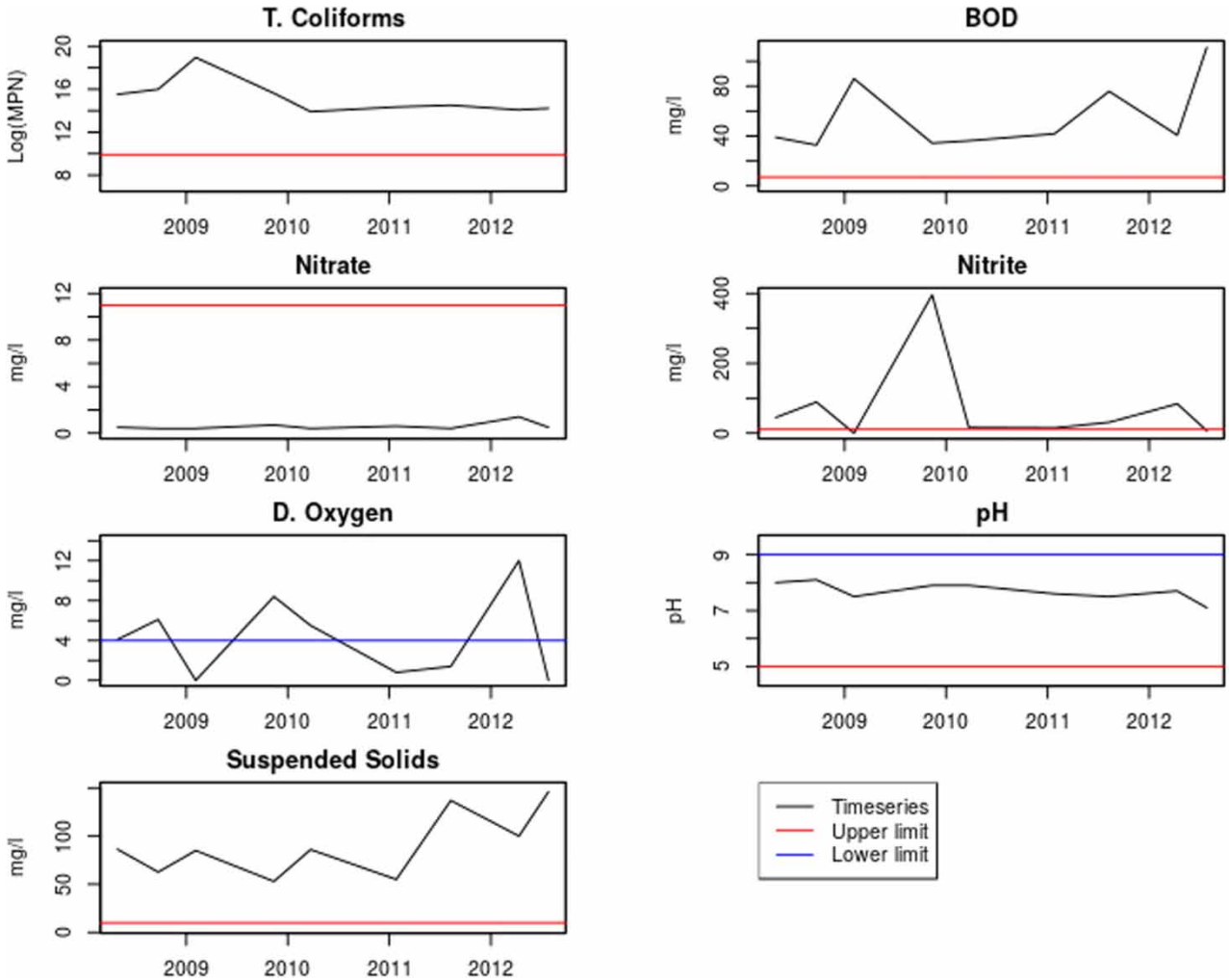


Figure 6 | Time series of the Socotá municipality. Please refer to the online version of this paper to see this figure in color: http://dx.doi.org/10.2166/hydro.2022.070.

Correlating parameters

We also evaluated the correspondence between physicochemical and microbiological parameters through the Pearson correlation. This type of correlation indicates the extent to which two variables are linearly related. Figure 7(a) shows the matrix with correlations for multiple parameters considered in this work. Thus, we identified that the strongest correlation appears on *Conductivity*, *BOD*, *COD*, and *Nutrients* (*Phosphorus* and *Nitrogen* with each other). Moreover, it is worth noticing a lower correlation between parameters as *Sulfate*, *Calcium*, *Magnesium*, and *Total hardness*, or the relationship between *Suspended* and *Total solids*.

From correlations shown in Figure 7(a), we focused on parameters such as *Conductivity*, *BOD*, *COD*, and some *Nutrients*, concretely *Phosphorus* (*Total* and *Ortho*) and Nitrogen, and straightforwardly evaluated their relations (see Figure 8). This additional analysis confirmed the positive association in all mentioned parameters. A special mention should be given to the correlation between *BOD* versus *COD* ($R = 0.9$) and *Ortho Phosphorus* versus *Nitrogen* (R=0.82), showing that a linear function could describe the variables.

Additionally, we formalized a mechanism to estimate parameter values when there is no information in our knowledge graphs. For that, we define multi-parameter linear models (Chambers 1991) that have been adjusted for *BOD*, *COD*, *Conductivity*, *Phosphorus*, *Ortho Phosphorous*, and *Nitrogen*. Considering these parameters, we applied a probabilistic test where the mentioned parameters in each one of the following models were the only ones whose coefficients move away from zero and, therefore, we included them in each proposed model (see Table 1).

Thereby, the associated parameters are *Conductivity* and, mostly, *COD* in the *BOD* model. The *Conductivity* (*C*) model only included *Nitrogen* and *Phosphorus*, whereas the *COD* model was calculated considering *BOD*, *Phosphorus* (*P*), and *Nitrogen*. The *Phosphorus* model is composed of *COD*, *Conductivity*, and *Nitrogen*. The *Ortho Phosphorous* (*OP*) model only contemplates *BOD*, *Phosphorus*, and *Nitrogen* (*N*). Finally, the proposed model for *N* considers *COD*, *C*, *P*, and *OP*. Checking the results of these models, we identified that the best scores of multi-parameter linear regression were associated with *BOD* and *COD* models, with $R^2 = 0.81$ in both cases followed by *N* proposal with $R^2 = 0.8$. At the same time, *Conductivity* presented a lower score with an $R^2 = 0.51$.

As we can see above, the best performance models allow an estimation of *BOD* and *COD*, although the latter case showed that *Conductivity* was not a good predictor. Concerning *Nutrients*, the best model applied to *Nitrogen* over *Phosphorus* (*Total* and *Ortho*). Although all the model parameters presented the least adjustment, we identified that *Conductivity* only considers the *Nutrients* as adequate predictors.

Spatial distribution of correlation

Taking into account the previous correlations, we adopted *Conductivity* as an indicator of several physicochemical and microbiological parameters to provide a spatial distribution. So, we considered the average records of *Conductivity* by

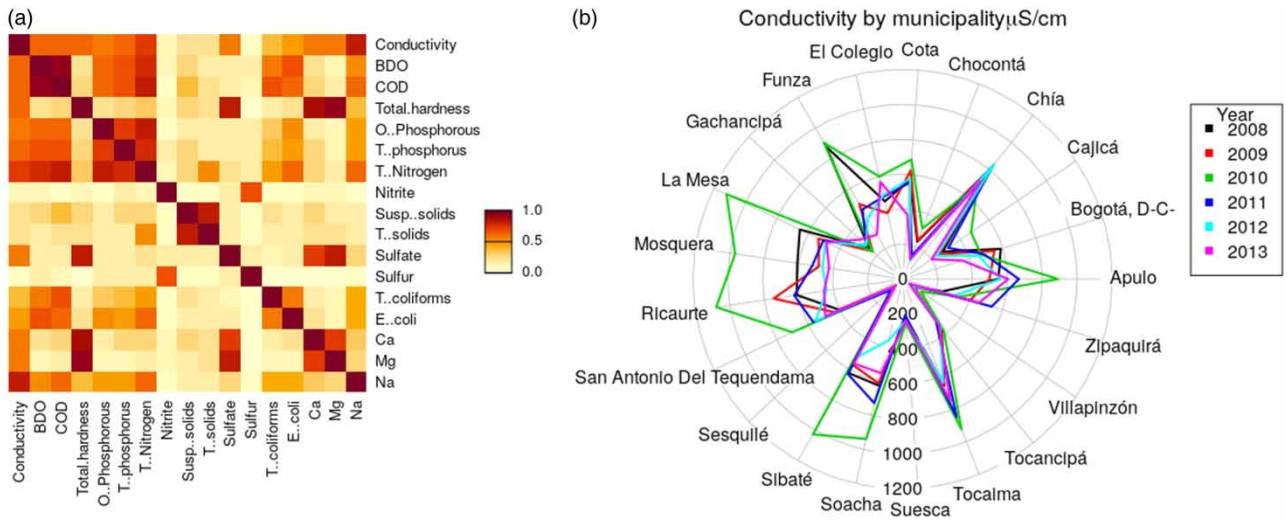


Figure 7 | (a) Pearson's correlation matrix and (b) conductivity by each municipality.

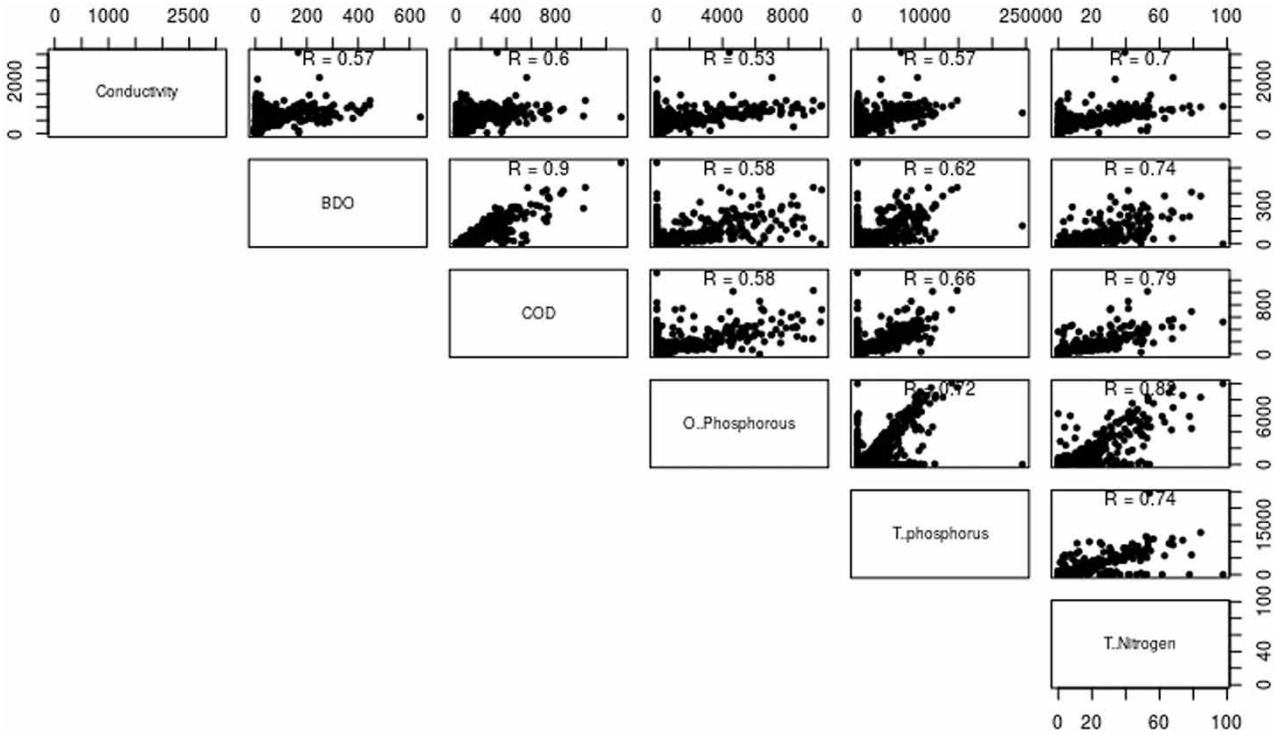


Figure 8 | Scatterplot matrix.

Table 1 | Results of multiple regression models

Model	Explanatory variables	R^2	Var	*t*-values	*p*-values	Res. std. error
BOD	$-11.68 + (\text{COD} * 0.42) + (C * 13.17\text{E}{-}03)$	0.81	COD	4.40E + 01	<2E − 016	33.89
			C	2.90E + 00	3.87E − 03	
C	$183.4 + (P * 2.305\text{E} - 02) + (N * 1.127\text{e} + 01)$	0.51	*P*	3.99E + 00	7.49E − 05	241.3
			N	1.24E + 01	<2E − 16	
COD	$19.05 + (BOD * 1.21) + (P * 9.92\text{E} - 03) + (N * 2.45)$	0.81	BOD	1.83E + 01	<2E − 16	61.63
			P	5.16E + 00	5.16E + 00	
			N	8.29E + 00	1.17E − 15	
P	$-30.32 + (\text{COD} * 6.08) + (C * 0.99) + (N * 42.78) + (\text{OP} * 0.17)$	0.66	COD	7.89E + 00	2.18E − 14	1,402
			C	3.61E + 00	3.41E − 03	
			N	4.90E + 00	1.33E − 06	
			OP	2.72E + 00	6.77E − 03	
OP	$-14.34 + (\text{BOD} * 3.87) + (N * 76.47)$	0.68	BOD	3.86E + 00	1.28E − 04	1,072
			N	1.86E + 01	<2E − 016	
N	$-0.51 + (\text{COD} * 3.5\text{E} - 02) + (C * 8.7\text{E} - 04) + (P * 1.1\text{E} - 04) + (\text{OP} * 3.3\text{E} - 04)$	0.8	COD	8.95E + 00	<2E − 016	7.28
			C	6.26E + 00	8.66E − 16	
			P	4.90E + 00	1.33E − 06	
			OP	1.11E + 01	<2E − 016	

municipality and year (within the considered period) to build a radar chart (see Figure 7(b)) and assess the quality status provided by the monitoring points in each municipality of the Bogota river basin. Thus, we distinguished that *Sesquilé*, *Villapinzón*, *Chocontá*, and *Suesca* reported the lowest *Conductivity* values. These municipalities are characterized by being located at the upper level of the basin. On the other hand, *Ricaurte*, *La Mesa*, *Mosquera*, *San Antonio del Tequendama*, *Sibaté*, *Soacha*, and *Apulo* are located mainly in the lower and middle parts of the basin, registering the highest values. This

pattern is bright to light an evident deterioration in water quality from the upper to lower level of the river basin, from *Villapinzón* (born place) to the mouth into the Magdalena River in the municipality of *Girardot* and *San Antonio del Tequendama*. Moreover, we identified that most cases with the highest *Conductivity* records were given in 2010, while the best conditions for this parameter were reported in 2013.

Regarding the spatial component of our knowledge graphs, we analyzed water quality considering the monitoring periods on the Bogota river basin during dry and wet seasons. These sets of samples are combined with established thresholds, allowing us to check the legal framework's degree of compliance. Figure 9 shows the spatial distribution of infractions averaged during two seasons (dry (see Figure 9(a)) and wet (see Figure 9(b)) seasons) according to associated classes.

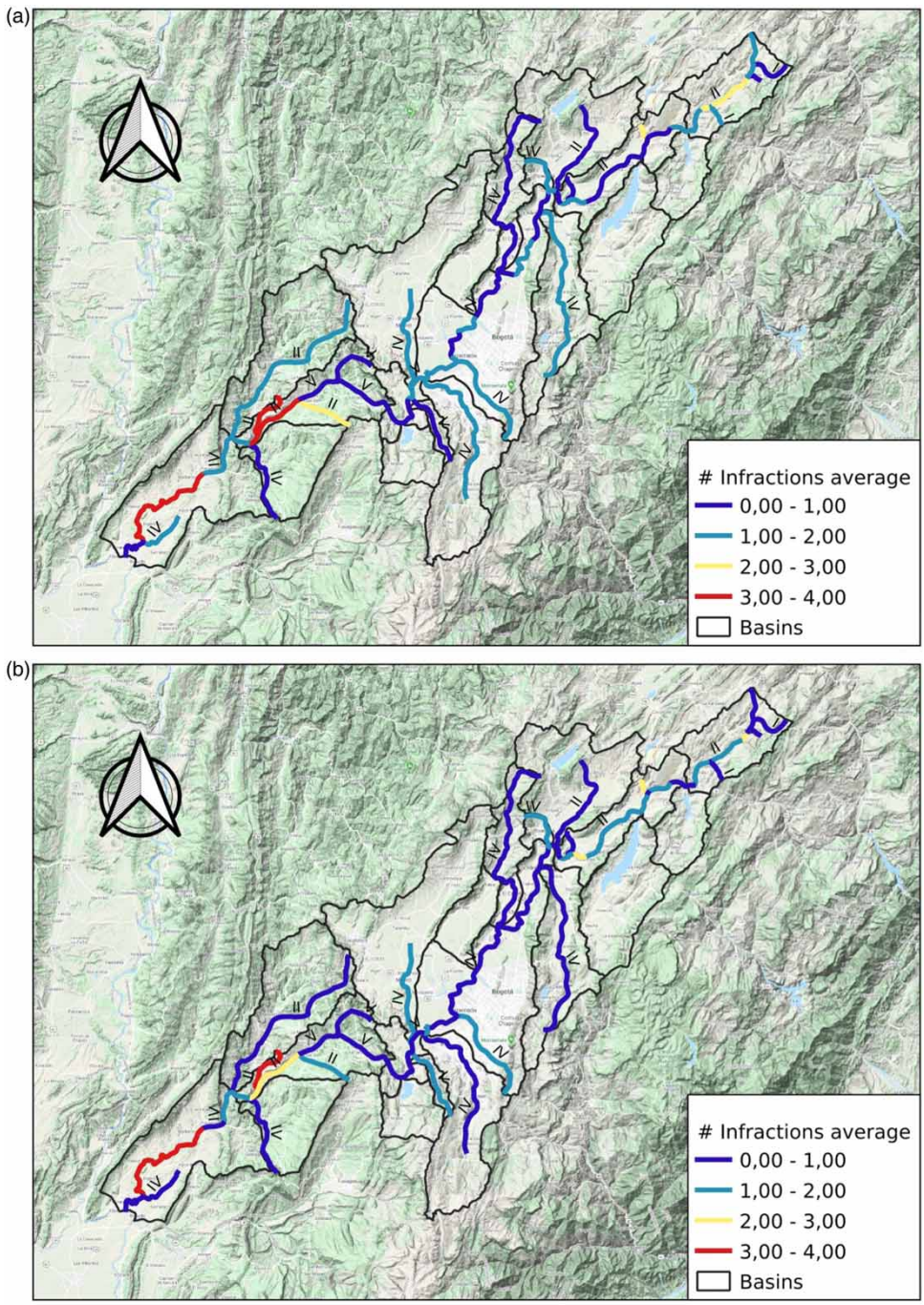


Figure 9 | Average infractions during (a) dry season (Campaign 1) and (b) wet season (Campaign 2).

This analysis of seasons permitted us to identify some patterns. As we mentioned before, the water quality deteriorates when caudal flows toward the mouth in the Southern zone. However, we also discovered some differences between seasons since more infractions are performed during the driest periods. This is possibly associated with less runoff in the basin rivers, which does not help to dilute the pollutants. Likewise, according to the regulatory framework, this river basin's primary channel is the only one that presents the least infraction cases. Nevertheless, it does not entail better water quality conditions, but results are related to areas with fewer restrictions on water quality.

On the other hand, we identified that infractions were more recurrent for the lower and middle parts of the river basin due to uses located in this area, destined for agricultural, industrial, and power generation activities. However, we also discovered exceptions related to some tributary rivers defined as sources of domestic use and present stricter water quality thresholds. In this sense, special attention has to be paid to violations in more restrictive areas since, although they offer a minor number of cases, these infractions affect water bodies destined for the population. This can contribute to the progressive decline of water quality and harm public health (Shi *et al.* 2019).

The number of infractions compiled in Table 2 verifies those behaviors described previously during dry and wet seasons. In the mentioned table, we identified that the most recurrent violations correspond to II and III, whereas Class I presents the most compliance due to its location (in the upper parts of the river basin), where the water bodies have less human intervention.

This work was developed considering a case study of the Bogota river basin in Colombia. Nevertheless, we think our contributions can be applied to other regions with similar/different climates and water quality circumstances. Thereby, for instance, we built an ontology network using several standards (e.g., WaterML, GeoSPARQL, and O&M) that could be integrated into or interoperable with other hydrologic information/decision support/sensor systems. Additionally, the proposed workflow to transform and connect several and heterogeneous datasets, building multi-dimensional water quality knowledge graphs, can be applied to other river basin environments to accomplish integrated and interoperable systems that allow discovering what, when, and where infractions happened on water quality through the exploitation of the spatio-temporal and legal characteristics. However, we are aware that each river basin has particular features that can require adapting or extending some elements of our approach, but this work can be considered a starting point in these specific cases.

CONCLUSIONS AND FUTURE WORK

This paper provided details about the generation of ontology-based knowledge graphs related to the Bogota river basin's water quality. For that, we integrated heterogeneous and multi-dimensional data sources that contain water quality parameters, spatio-temporal, and regulatory data. Furthermore, we developed a new ontology that reuses numerous standards from

Table 2 | Number of infractions by campaign, class, and year

Campaign	Year	Class I	Class II	Class IV	Class V
1	2008	8.00	55.00	10.00	6
2		3.00	55.00	38.00	9
1	2009	4.00	49.00	63.00	13
2		3.00	43.00	16.00	3
1	2010	6.00	44.00	22.00	5
2		0.00	34.00	4.00	0
1	2011	0.00	44.00	44.00	14
2		0.00	49.00	27.00	4
1	2012	1.00	55.00	30.00	4
2		4.00	52.00	29.00	5
1	2013	4.00	49.00	28.00	7
2		0.00	27.00	18.00	0
	Avg	2.75	42.77	27.42	7

different domains such as GeoSPARQL (spatial), O&M (Observations and Measurements), SSN (sensors), or RDF Data Cube (statistics), among others, as well as non-ontological resources associated with standards, for instance, the case of WaterML.

The developed process allowed us to provide (semantic) interoperability to original data to break down with existing data silos, producing our water quality knowledge graphs. Moreover, these graphs offer experts a connected and interoperable view of data and bring the expert analysis closer to end-users.

With respect to the results, it is worth noticing that our approach supports answering questions. Therefore, we discovered what, when, and where infractions happened on water quality in the Bogota river basin, exploiting spatio-temporal and legal characteristics of the knowledge graphs. In addition, we performed correlations between diverse parameters and defined multi-parameter linear models for those cases where information does not exist.

Future work will concentrate on applying hydrological modeling to our knowledge graphs to generate accurate data about water resources where there is a lack of available data related to spatial or time dimensions. In this way, we will incorporate a context associated with spatio-temporal hydrological modeling for diverse parameters. Likewise, we plan to work with current WaterML extensions and associated proposals related to instrumentation and groundwater in order to extend our developed ontology.

COMPETING INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Aguilar, J. A. P., Campo, J., Meneu, S. N., Gimeno-García, E. & Andreu, V. 2019 Analysis of existing water information for the applicability of water quality indices in the fluvial-littoral area of Turia and Jucar Rivers, Valencia, Spain. *Applied Geography* **111**, 102062. doi:10.1016/j.apgeog.2019.102062.

Ahmedi, L., Jajaga, E. & Ahmedi, F. 2013 An ontology framework for water quality management. In: *Semantic Sensor Networks 2013 (SSN2013) – ISWC* (O. Corcho, C. Henson, P. Barnaghi, eds.). CEUR, Sydney, pp. 35–50.

Alilou, H., Nia, A. M., Saravi, M. M., Salajegheh, A., Han, D. & Enayat, B. B. 2019 A novel approach for selecting sampling points locations to river water quality monitoring in data-scarce regions. *Journal of Hydrology* **573**, 109–122. doi:10.1016/j.jhydrol.2019.03.068.

Arribas-Bel, D. 2014 Accidental, open and everywhere: emerging data sources for the understanding of cities. *Applied Geography* **49**, 45–53. doi:10.1016/j.apgeog.2013.09.012.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. & Ives, Z. 2007 DBpedia: a nucleus for a web of open data. In: *The Semantic Web* (K. Aberer, K-S. Choi, N. Noy, D. Allemang, K-I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber & P. Cudré-Mauroux, eds.). Springer, Berlin, Heidelberg, pp. 722–735.

Baltacı, F., Onur, A. K. & Tahmiscioğlu, S. 2008 Water quality monitoring studies of Turkey with present and probable future constraints and opportunities. *Desalination* **226** (1–3), 321–327. doi:10.1016/j.desal.2007.02.114.

Basulto, J., Manera, M. & Baladia, E. 2014 Dietary intake of nitrate in Spanish infants and children and risk of methemoglobinemia. *Pediatría Atención Primaria* **16** (61), 65–69. http://dx.doi.org/10.4321/S1139-76322014000100013.

Battle, R. & Kolas, D. 2012 Enabling the geospatial semantic web with parliament and GeoSPARQL. *Semantic Web* **3** (4), 355–370. doi:10.3233/SW-2012-0065.

Behmel, S., Damour, M., Ludwig, R. & Rodriguez, M. J. 2016 Water quality monitoring strategies – a review and future perspectives. *Science of the Total Environment* **571**, 1312–1329. doi:10.1016/j.scitotenv.2016.06.235.

Bellomarini, L., Sallinger, E. & Vahdati, S. 2020 Knowledge graphs: the layered perspective. In: *Knowledge Graphs and Big Data Processing, LNCS 12072* (V. Janev, D. Graux, H. Jabeen & E. Sallinger, eds), pp. 20–34. https://doi. org/10.1007/978-3-030-53199-7_2.

Berners-Lee, T. 2006 *Linked Data Design Issues*. Available from: https://www.w3.org/DesignIssues/LinkedData.html (accessed 17 May 2021).

Camacho, L. A. 2020 The paradox of the availability of poor water quality in the Colombian rural sector. *Revista de Ingeniería*. Universidad de Los Andes. https://doi.org/10.16924/revinge.49.6.

Chambers, J. M. 1991 *Statistical Models in S*. CRC Press, Inc, New York.

CUAHSI – CUAHSI's Hydrologic Information System (CUAHSI-HIS) 2010 *CUAHSI-HIS*. Available from: http://his.cuahsi.org/ (accessed 17 May 2021).

Cudré-Mauroux, P. 2020 Leveraging knowledge graphs for big data integration: the XI pipeline. *Semantic Web* **11** (1), 13–17. doi:10.3233/SW-190371.

Curry, E., Degeler, V., Clifford, E., Coakley, D., Costa, A., Van Andel, S. J., van de Giesen, N. & Kouroupetroglou, C. 2014 Linked water data for water information management. In: *11th International Conference on Hydroinformatics*.

DEFRA – Department for Environment, Food and Rural Affairs 2014 *Defra Open Data Strategy*. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/267934/pb14109-defra-open-data-strategy-131219.pdf (accessed 17 May 2021).

Delpla, I., Proulx, F. & Rodríguez, M. J. 2020 A methodology to prioritize spatio-temporal monitoring of drinking water quality considering population vulnerability. *Journal of Environmental Management* **255**, 109869. doi:10.1016/j.jenvman.2019.109869.

Díaz-Casallas, D. M., Castro-Fernández, M. F., Bocos, E., Montenegro-Marin, C. E. & González Crespo, R. 2019 2008–2017 Bogota river water quality assessment based on the water quality index. *Sustainability* **11** (6), 1668. doi:10.3390/su11061668.

Ehrlinger, L. & Wöß, W. 2016 *Towards a Definition of Knowledge Graphs*. SEMANTiCS (Posters, Demos, SuCCESS), p. 48.

Gasper, R., Blohm, A. & Ruth, M. 2011 Social and economic impacts of climate change on the urban environment. *Current Opinion in Environmental Sustainability* **3** (3), 150–157. https://doi.org/10.1016/j.cosust.2010.12.009.

Giraldo, E. & Garzón, A. 2002 The potential for water hyacinth to improve the quality of Bogota River water in the Muña Reservoir: comparison with the performance of waste stabilization ponds. *Water Science and Technology* **45** (1), 103–110. doi:10.2166/wst.2002.0014.

Giri, S. & Qiu, Z. 2016 Understanding the relationship of land uses and water quality in Twenty First Century: a review. *Journal of Environmental Management* **173**, 41–48.

Hakimi, O., Gelpi, J. L., Krallinger, M., Curi, F., Repchevsky, D. & Ginebra, M. P. 2020 The devices, experimental scaffolds, and biomaterials ontology (DEB): a tool for mapping, annotation, and analysis of biomaterials data. *Advanced Functional Materials* **30** (16), 1909910. doi:10.1002/adfm.201909910.

Heath, T. & Bizer, C. 2011 Linked data: evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology* **1** (1), 1–136. doi:10.4018/978-1-60960-593-3.ch008.

Hunter, J., Becker, P., Alabri, A., Van Ingen, C. & Abal, E. 2011 Using ontologies to relate resource management actions to environmental monitoring data in south east Queensland. *International Journal of Agricultural and Environmental Information Systems (IJAEIS)* **2** (1), 1–19. doi:10.4018/jaeis.2011010101.

Jajaga, E., Ahmedi, L. & Ahmedi, F. 2015 An expert system for water quality monitoring based on ontology. In *Research Conference on Metadata and Semantics Research*. Springer, Cham, pp. 89–100. https://doi.org/10.1007/978-3-319-24129-6_8.

Jajaga, E., Ahmedi, L. & Ahmedi, F. 2016 StreamJess: a stream reasoning framework for water quality monitoring. *IJMSO* **11** (4), 207–220.

Janowicz, K., Gao, S., McKenzie, G., Hu, Y. & Bhaduri, B. 2020 GeoAI: spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science* **34** (4), 625–636. doi:10.1080/13658816.2019.1684500.

Kämpgen, B., Riepl, D. & Klinger, J. 2014 SMART research using linked data-sharing research data for integrated water resources management in the Lower Jordan Valley. In *SePublica*.

Krötzsch, M. & Thost, V. 2016 Ontologies for knowledge graphs: breaking the rules. In: *International Semantic Web Conference*. Springer, pp. 376–392. https://doi.org/10.1007/978-3-319-46523-4_23.

Liu, Q., Bai, Q., Kloppers, C., Fitch, P., Bai, Q., Taylor, K., Fox, P., Zednik, S., Ding, L., Terhorst, A. & McGuinness, D. 2013 An ontology-based knowledge management framework for a distributed water information system. *Journal of Hydroinformatics* **15** (4), 1169–1188.

Masmoudi, M., Karray, M. H., Lamine, S. B. A. B., Zghal, H. B. & Archimede, B. 2020 MEMOn: modular environmental monitoring ontology to link heterogeneous earth observed data. *Environmental Modelling & Software* **124**, 104581. doi:10.1016/j.envsoft.2019.104581.

McDonald, J. D. & Levine-Clark, M. 2018 *Encyclopedia of Library and Information Sciences*. CRC Press, Boca Raton.

Miranda, D., Carranza, C., Rojas, C. A., Jerez, C. M., Fischer, G. & Zurita, J. 2008 Accumulation of heavy metals in soil and plants of four vegetable crops irrigated with water of Bogota river. *Colombian Journal of Horticultural Science* **2** (2), 180–191.

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A. & Taylor, J. 2019 Industry-scale knowledge graphs: lessons and challenges. *Queue* **17** (2), 48–75. https://doi.org/10.1145/3331166.

Ouyang, Y. 2005 Evaluation of river water quality monitoring stations by principal component analysis. *Water Research* **39** (12), 2621–2635. doi:10.1016/j.watres.2005.04.024.

Pahl-Wostl, C. 2020 Adaptive and sustainable water management: from improved conceptual foundations to transformative change. *International Journal of Water Resources Development* **36** (2–3), 397–415. doi:10.1080/07900627.2020.1721268.

Patroumpas, K., Alexakis, M., Giannopoulos, G. & Athanasiou, S. 2014 TripleGeo: an ETL tool for transforming geospatial data into RDF triples. In: *Edbt/Icdt Workshops*. pp. 275–278.

Paulheim, H. 2017 Knowledge graph refinement: a survey of approaches and evaluation methods. *Semantic Web* **8** (3), 489–508. doi:10.3233/SW-160218.

Perry, M. & Herring, J. 2012 *GeoSPARQL: A Geographic Query Language for RDF Data*. Available from: https://portal.opengeospatial.org/files/?artifact_id=47664 (accessed 17 May 2021).

Radhapyari, K., Datta, S., Dutta, S. & Barman, R. 2021 Impacts of global climate change on water quality and its assessment. In: *Water Conservation in the Era of Global Climate Change* (Thokchom, B., Qiu, P., Singh, P. & Iyer, P. K., eds). Elsevier, pp. 229–275. https://doi.org/10.1016/B978-0-12-820200-5.00011-7.

Raymond, C. M., Fazey, I., Reed, M. S., Stringer, L. C., Robinson, G. M. & Evely, A. C. 2010 Integrating local and scientific knowledge for environmental management. *Journal of Environmental Management* **91** (8), 1766–1777. https://doi.org/10.1016/j.jenvman.2010.03.023.

Resh, V. H. & Unzicker, J. D. 1975 Water quality monitoring and aquatic organisms: the importance of species identification. *Journal (Water Pollution Control Federation)*, 9–19. doi:10.2307/25038592.

Rodríguez Forero, A., Mantilla, J. F. G. & Martínez, R. S. 2009 Accumulation of lead, chromium, and cadmium in muscle of capitan (*Eremophilus mutisii*), a catfish from the Bogota River basin. *Archives of Environmental Contamination and Toxicology* **57** (2), 359–365. doi:10.1007/s00244-008-9279-2.

Rodríguez-Jeangros, N., Camacho, L. A., Rodríguez, J. P. & McCray, J. E. 2018 Integrated urban water resources model to improve water quality management in data-limited cities with application to Bogotá, Colombia. *Journal of Sustainable Water in the Built Environment* **4** (2), 04017019. doi:10.1061/JSWBAY.0000846.

Sankary, N. & Ostfeld, A. 2018 Analyzing multi-variate water quality signals for water quality monitoring station placement in water distribution systems. *Journal of Hydroinformatics* **20** (6), 1323–1342.

Shi, B., Bach, P. M., Lintern, A., Zhang, K., Coleman, R. A., Metzeling, L., McCarthy, D. T. & Deletic, A. 2019 Understanding spatiotemporal variability of in-stream water quality in urban environments – a case study of Melbourne, Australia. *Journal of Environmental Management* **246**, 203–213. doi:10.1016/j.jenvman.2019.06.006.

Suarez-Figueroa, M. C., Gomez-Perez, A., Motta, E. & Gangemi, A. 2012 *Ontology Engineering in a Networked World*. Springer, Berlin, Heidelberg. doi:10.1007/978-3-642-24794-1_1.

United Nations, Department of Economic and Social Affairs, Population Division 2018a *World Urbanization Prospects 2018*. Available from: https://population.un.org/wup/ (accessed May 17 2021).

United Nations, Department of Economic and Social Affairs, Population Division 2018b *World Urbanization Prospects: The 2018 Revision*. Available from: https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf (accessed 17 May 2021).

Valipour, M., Bateni, S. M. & Jun, C. 2021 Global surface temperature: a new insight. *Climate* **9**, 81. https://doi.org/10.3390/cli9050081.

Van den Brink, L., Barnaghi, P., Tandy, J., Atemezing, G., Atkinson, R., Cochrane, B., Fathy, Y., García Castro, R., Haller, A., Harth, A., Janowicz, K., Kolozali, Ş., van Leeuwen, B., Lefrançois, M., Lieberman, J., Perego, A., Le-Phuoc, D., Roberts, B., Taylor, K. & Troncy, R. 2019 Best practices for publishing, retrieving, and using spatial data on the web. *Semantic Web* **10** (1), 95–114. doi:10.3233/SW-180305.

Vilches-Blázquez, L. M., Villazón-Terrazas, B., Corcho, O. & Gómez-Pérez, A. 2014 Integrating geographical information in the Linked Digital Earth. *International Journal of Digital Earth* **7** (7), 554–575. doi:10.1080/17538947.2013.783127.

Wache, H., Voegele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H. & Huebner, S. 2001 Ontology-based integration of information – a survey of existing approaches. In *Proceedings of the Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 108–117. Available from: http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-47/wache.pdf (accessed 17 May 2021).

Ward, R. C., Loftis, J. C. & McBride, G. B. 1986 The 'data-rich but information-poor' syndrome in water quality monitoring. *Environmental Management* **10** (3), 291–297. doi:10.1007/BF01867251.

Weiss, C., Karras, P. & Bernstein, A. 2008 Hexastore: sextuple indexing for semantic web data management. *Proceedings of the VLDB Endowment* **1** (1), 1008–1019.

WHO – World Health Organization 2017 *Guidelines for Drinking-Water Quality*. Available from: https://www.who.int/publications/i/item/9789241549950 (accessed 17 May 2021).

Wilson, R. L., Reely, B. T. & Cox, M. 1997 The water resource management system (WREMS): linking data management and operational optimization. *Annals of Operations Research* **72**, 105–124. doi:10.1023/A:1018940103278.

Yu, J., Car, N. J., Leadbetter, A., Simons, B. A. & Cox, S. J. 2015 Towards linked data conventions for delivery of environmental data using netCDF. In: *International Symposium on Environmental Software Systems*. Springer, Cham, pp. 102–112. https://doi.org/10.1007/978-3-319-15994-2_9.

Zhao, P., Foerster, T. & Yue, P. 2012 The geoprocessing web. *Computers & Geosciences* **47**, 3–12. doi:10.1016/j.cageo.2012.04.021.

First received 21 May 2021; accepted in revised form 9 February 2022. Available online 23 February 2022