A Foundation for Spatial Data Warehouses on the Semantic Web

Gur, Nurefsan; Pedersen, Torben Bach; Zimányi, Esteban; Hose, Katja

Published in:
Semantic Web

DOI (link to publication from Publisher):
10.3233/SW-170281

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
A Foundation for Spatial Data Warehouses on the Semantic Web

Gur, Nurefsan; Pedersen, Torben Bach; Zimányi, Esteban; Hose, Katja

Published in:
Semantic Web

DOI (link to publication from Publisher):
10.3233/SW-170281

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
A Foundation for Spatial Data Warehouses on the Semantic Web

Nurefşan Güra,b,*, Torben Bach Pedersen a, Esteban Zimányib and Katja Hosea
a Center for Data Intensive Systems, Aalborg University, Selma Lagerlöfsvej 300, DK-9220 Aalborg Ø, Denmark
E-mail: {nurefsan,tbp,khose}@cs.aau.dk
b Department of Computer and Decision Engineering, Université Libre de Bruxelles, Avenue F. D. Roosevelt 50, B-1050 Brussels, Belgium
E-mail: {nurefsan.gur,ezimanyi}@ulb.ac.be

Abstract. Large volumes of geospatial data are being published on the Semantic Web (SW), yielding a need for advanced analysis of such data. However, existing SW technologies only support advanced analytical concepts such as multidimensional (MD) data warehouses and Online Analytical Processing (OLAP) over non-spatial SW data. To remedy this need, this paper presents the QB4SOLAP vocabulary, which supports spatially enhanced MD data cubes over RDF data. The paper also defines a number of Spatial OLAP (SOLAP) operators over QB4SOLAP cubes and provides algorithms for generating spatially extended SPARQL queries from the SOLAP operators. The proposals are validated by applying them to a realistic use case.

Keywords: Spatial OLAP, Spatial data, Multidimensional data, Data modelling, RDF, SPARQL

1. Introduction

The Semantic Web (SW) has evolved, from focusing mostly on data publishing to also support increasingly complex queries such as interactive analytical queries. Simultaneously, the data available on the SW has evolved from being simple, most alphanumeric data, to also include complex data such as spatial data. Indeed, geospatial data is now common on the SW, but it remains difficult to analyze it.

In a non-SW context, the main tools for interactive data analyses have been Data Warehouses (DWs) and Online Analytical Processing (OLAP) tools and queries. DWs store large volumes of data and are designed with a multidimensional (MD) modeling approach, which has shown itself to be intuitive for interactive data analytics. Concretely, DWs consist of MD data cubes. The cells of the cube represent the topic of analysis, and associate observation facts with numerical measures that can be aggregated. For example, a sales fact cube has measures such as QuantitySold and SalesPrice. Facts are linked to dimensions, which provide contextual information, e.g., sales date, product, and location. Dimensions are perspectives, which are used to analyze data, and are organized into hierarchies with levels, e.g., Store, City, and Region, that allow users to analyze and aggregate measures at different levels of detail. Levels have a set of attributes that describe the characteristics of the level members.

In traditional DWs, the location dimension is widely used, but as a conventional dimension with alphanumeric data and thus only nominal reference to spatial concepts such as areas and places. This does not allow manipulating through spatial location data or deriving topological relations among the hierarchy levels of the location dimension. This yields a demand for truly spatial DWs for better analysis purposes. Including the geometric information of the location data, sig-
significantly improves the analysis process (i.e., proximity analysis of the locations) with additional perspectives by revealing dynamic spatial hierarchy levels and new spatial members.

Similarly, providing deep spatial analytics support for spatial SW data is very valuable. Spatial data requires specific treatment techniques, in particular encoding, special functions and different manipulation methods, which should be considered in the modeling process and querying. The current state of the art for the geospatial Semantic Web focuses on techniques for publishing, linking and querying spatial data, but supports only “plain” spatial SW data (without support for spatial DW concepts such as spatial hierarchies, levels, and measures) and does not consider analytical queries over spatial RDF data (see Section 2 for details).

**Problem Definition.** The proliferation of open geospatial data on the SW creates possibilities for advanced analysis of such data. Many examples exist of spatial Linked Open Data (LOD) published on the SW as RDF. These datasets have observations and measures that are well suited for analytical queries (e.g., water/air quality measurements, immigration rates, EU subsidies in agriculture, crop revenue, etc.). However, such datasets are typically not modeled with spatial dimension levels and hierarchies. Thus, they cannot be queried with interactive spatial analytical queries (a.k.a. SOLAP) on the SW. In the current state of the SW, if a (spatial) DW user would like to query the existing spatial RDF data from the SW with SOLAP operations, the user needs to download the RDF data, map it to a relational data model (i.e., with a snowflake schema), and then import it into a traditional spatial data warehouse in order to query with SOLAP, which is slow, labor-intensive, and stores the data in a non-open format.

**Our Approach.** On the contrary, annotating spatial RDF datasets with QB4SOLAP [17,16] allows users to define spatial multidimensional concepts on top of existing RDF data. Hence, the user can create and publish spatial data warehouses on the Semantic Web, which can be easily queried with SOLAP operations. Fig. 1 depicts the general workflow scenario, where the spatial RDF datasets from endpoints can be annotated with QB4SOLAP. This makes it possible for end users to use SOLAP queries. However, writing a SOLAP query in SPARQL can be very complicated for users inexperienced with SPARQL (e.g., traditional DW users). Due to the lack of MD semantics of spatial RDF data and the lack of translation techniques from high-level SOLAP expressions to SPARQL, there is a considerable entry barrier for advanced spatial data analysis on the SW for data warehouse users.

**Contributions.** In order to address these issues, this paper makes a number of contributions. First, we propose QB4SOLAP, a generic and extensible vocabulary (metamodel) for spatial DWs on the SW. QB4SOLAP extends the most recent stable version of the QB4OLAP vocabulary with spatial concepts. We provide a full formalization of QB4SOLAP. The key concepts of spatial cube members, spatial hierarchies and levels, spatial measures, spatial aggregate functions (e.g., union, buffer, and convex–hull) and topological relations among spatial dimension and hierarchy level members (e.g., within, intersects, and overlaps), are defined. Second, we define a number of analytical Spatial OLAP (SOLAP) operators over the model including giving formal semantics of the operators. The operators support advanced analytical queries over MD geospatial SW data. Third, we provide algorithms for generating spatially extended SPARQL queries for individual and nested SOLAP operators, which allows writing SOLAP queries without knowledge of RDF/SPARQL. Fourth, we validate the vocabulary, operators, and query generation algorithms by applying them to a realistic use case.

**Paper structure.** The remainder of the paper is structured as follows. Section 2 discusses related work. Section 3 defines preliminary spatial and OLAP concepts. Section 4 defines the QB4SOLAP vocabulary, while Section 5 defines the SOLAP operators. Section 6 provides the SPARQL query generation algorithms. Finally, Section 7 concludes the paper and points to future research.

---

2. UK Environmental Data: [http://environment.data.gov.uk](http://environment.data.gov.uk)
3. Danish Agricultural Data: [https://datahub.io/dataset/govagribus-denmark](https://datahub.io/dataset/govagribus-denmark)
2. Related work

DW and OLAP technologies have been successful for analyzing large volumes of data [1]. Combining DW/OLAP technologies with RDF data makes RDF data sources more easily available for interactive analysis. The following work concerns the integration of DW/OLAP with the SW.

**DW/OLAP and Semantic Web.** Using OLAP to analyze SW data is considered in several approaches. Kämpgen et al. propose an extended model [23] on top of the RDF Data Cube Vocabulary (QB) [6] for interacting with statistical linked data via OLAP operations directly in SPARQL. However, it has the inherent limitations of QB and thus cannot support OLAP dimensions with hierarchies and levels, and built-in aggregate functions. Etcheverry et al. introduce QB4OLAP [11] as an extended vocabulary based on QB, with a full MD metamodel, supporting OLAP operations directly over RDF data with SPARQL queries. Nath et al. considers creating an Extract–Transfo rm–Load (ETL) framework for semantic data warehouses [8]. Varga et al. presents a comprehensive methodology for dimensional enrichment of statistical LOD by using QB4OLAP and provide a SW-based OLAP engine for traditional DW users [38]. However, these approaches and vocabularies support neither spatial DWs nor provide SOLAP operators for the SW.

**Spatial DW and OLAP.** The constraint representation of spatial data has been the focus in many fields from databases to AI [32]. Extending OLAP with spatial features has also attracted the attention of the data warehousing community. Bédard et al. first introduced the term SOLAP [4] in 1997. SOLAP systems [33,9] since then, have significantly been improved. Respectively, various papers improve the spatial aggregation functions and techniques [29,26,7,15].

Several conceptual models are proposed for representing spatial data in data warehouses. Stefanovic et al. [19] considers constructing and materializing spatial cubes in their proposed model. The Multi-Dim conceptual model, introduced by Malinowski and Zimányi [27], cope s with spatial features and is extended in [37], to include complex geometric features (continuous fields), with a set of operations and an MD calculus supporting spatial data types. Gómez et al. [14] propose an algebra and a general framework for OLAP cube analysis on discrete and continuous spatial data. Even though spatial data warehousing is thus widely studied, those studies are limited to traditional non-semantic spatial data warehouses and SOLAP techniques. The work above neither considers semantic web data nor spatial analytical querying in SPARQL.

**Geospatial Semantic Web.** The Open Geospatial Consortium (OGC) has proposed GeoSPARQL [3] as a vocabulary to represent and query spatial data in RDF using an extension to SPARQL. Kyzirakos et al. present a comprehensive survey of data models and query languages for linked geospatial data in [25], and propose a semantic geospatial data store called Strabon in [24]. Strabon has an extensive query language called stSPARQL, which is however limited to the specific environment. LinkedGeoData is a significant contribution on interactive transformation of OpenStreetMap data to RDF data [36]. GeoKnow [34] is a more recent project with focus on linking geospatial data from heterogeneous sources. Andersen et al. considers publishing/converting open spatial data as Linked Open Data [2]. However, none of these works consider the MD aspects of geospatial data or allow querying with SOLAP on the SW, unlike QB4SOLAP. The QB4SOLAP vocabulary is validated with both the running example use case, the GeoNorthwind data cube, as well as a substantial real-world use case, the GeoFarmHerdState data cube [16]. GeoFarmHerdState is a spatial data cube about livestock holdings in Denmark, which integrates environmental and geographical open data from several sources, thus enabling a range of interesting SOLAP queries.

In summary, none of the related work, which is surveyed in the fields of “DW/OLAP and the SW”, “Spatial DW and OLAP”, and “Geospatial Semantic Web” provides a substantial foundation for modeling and querying spatial data warehouses on the Semantic Web, unlike the QB4SOLAP vocabulary, SOLAP operators, and SPARQL generation algorithms presented in this paper.

3. Preliminary concepts

In this section, we describe the spatial objects and the spatial operations that manipulate them. Then, we introduce the data cubes and spatial enhancement on them as spatial data cubes. Finally, we show the traditional OLAP operations, which manipulate data cubes, and explain the Spatial OLAP (SOLAP) operators, which manipulate spatial data cubes.

---

5http://www.openstreetmap.org
3.1. Spatial objects

A spatial object represents a real-world object whose geographic features are important for an application. These geographic features are encoded using the geometry data type. Point, Line, and Polygon are the basic instantiable types of the geometry data type. Coordinates for geometry data type are generally given in 2-dimensions with X, Y values. Geometries are associated with a spatial reference system (SRS), which describes the coordinate space in which the geometry is defined. There are several SRSs and each of them are identified with a spatial reference system identifier (SRID). The World Geodetic System (WGS) is the most well-known SRS and the latest version is called WGS84, which is also used in our use case.

3.2. Spatial operations

There is a set of spatial operations that can be applied on spatial data. We grouped these operations into classes, based on the common functionality of the operators. These classes are defined next.

Definition 1. (Spatial aggregation) The operators in the spatial aggregation class \( S_{\text{agg}} \) aggregate two or more spatial objects and return a new spatial object. Union, Intersection, ConvexHull, and Minimum-BoundingRectangle (MBR) are example operators of this class. Some spatial functions such as ConvexHull or MBR can also be interpreted as unary spatial functions with a single parameter, but here we only consider the aggregate versions of the functions. In order to make this clear, the aggregate versions of those functions are given with a prefix “Aggr” in the QB4SOLAP vocabulary (Fig. 6). For our purpose, it is enough to group all spatial aggregate functions into a single group, although more fine-grained classification proposals for spatial aggregate functions exist [7].

Definition 2. (Topological relations) The operators in the topological relation class \( T_{\text{rel}} \) are commonly expressed in the RCC8 and DE-9DIM models [31,10]. Topological relations are Boolean predicates that specify how two spatial objects are related to each other. Examples of topological relations are Intersects, Disjoint, Equals, Overlaps, Contains, Within, Touches, Covers, CoveredBy, and Crosses.

Definition 3. (Numeric operations) The operators in the numeric operation class \( N_{\text{op}} \) take one or more spatial objects and return a numeric value. Perimeter, Area, NoOfInteriorRings, Distance, HaversineDistance, and NoOfGeometries are example operators of this class.

3.3. Data cubes

Data warehouses store large volumes of data for decision support. They are based on the multidimensional model, which views data in an n-dimensional space, usually called a data cube. The cells of the cube represent the observation facts for analysis with a set of attributes called measures (e.g., a sales fact cube with measures product quantity and price). Facts are linked to dimensions, which provide perspectives to analyze

---

6RCC8 (Region Connection Calculus) describes regions in Euclidean space or in a topological space by their possible relations to each other.

7DE-9DIM (Dimensionally Extended Nine-Intersection Model) is a topological model that describes spatial relations of two geometries in two dimensions.
data (e.g., sales date, product, and customer location). Dimensions are organized into hierarchies, which allow users to aggregate measures at various levels of detail. Hierarchies are composed of levels and there is always a unique top level `All` with just one member `all`. Levels have a set of attributes that describe the characteristics of the level members.

An example of a data cube with three dimensions (Customer, Time, and Product) and one measure (Quantity) is given in Fig. 2. Each cell in the cube is an observation fact, which is characterized by dimension and measure values. The hierarchies of this cube are given in Fig. 4a–c. Thus, in the cube shown in Fig. 2, the Product dimension is given at the Category level, the Time dimension at the Quarter level, and the Customer dimension at the City level. Measure values represent the measure Quantity of the sold products.

### 3.4. Spatial data cubes

A spatial data cube contains both conventional and spatial dimensions. A spatial dimension is a dimension, which includes at least one spatial level in which the application should store the spatial characteristics of the members. Similarly, a hierarchy is a spatial hierarchy if it has at least one spatial level. Spatial characteristics of the levels are captured by their geometries and can be recorded in the spatial attributes of the level. A spatial fact is a fact that relates several dimensions in which, two or more are spatial.

For example, consider a Sales spatial fact, which has spatial dimensions Customer and Supplier, each with a spatial hierarchy Geography composed of spatial levels City → State → Country → All (Fig. 4c). These spatial levels record the spatial characteristics of its members with spatial attributes: Customer, Supplier, and City using a point spatial data type, whereas State and Country with a multi-polygon spatial data type.

Following the philosophy of spatially extended MultiDim conceptual model [37], MD concepts such as levels are considered to be spatial, only if they record the spatial characteristics of the concepts as geometries. For instance, “continent” might be considered as a spatial object, in theory or in other vocabularies. However, if there is no information about the geometry of the continents in the schema and in the instance data, continent does not become a spatial level (Def. 7), although continent might still be a traditional level of the spatial hierarchy Geography with alphanumeric attributes (i.e., continent name, code, and etc.).

Spatial data cubes typically have spatial measures, which are also represented by a spatial data type. An example is a SalesPoint measure that stores the location of sales. Fig. 7 shows the multidimensional schema of the GeoNorthwind data warehouse, which is used as running example in the paper.

### 3.5. OLAP operators

OLAP operators are used for expressing queries over data cubes. The traditional OLAP operators are given next.
The slice operator removes a dimension from a cube by selecting one instance in a dimension level. An example is “slice on City is equal to Odense”.

The dice operator selects the cells in a cube that satisfy a Boolean condition. An example is “dice on the first and last quarter of the year”.

The roll-up operator aggregates measures along a hierarchy to obtain data at a coarser granularity. An example is “roll-up to the Country level” (Fig. 3).

Finally, the drill-down operator disaggregates measures along a hierarchy to obtain data at a finer granularity. It is the inverse operation of roll-up. Starting from the cube in Fig. 3, an example is “drill-down to the City level”.

3.6. Spatial OLAP operators

Spatial OLAP (SOLAP) operates on spatial data cubes. SOLAP increases the analytical capabilities of OLAP by taking into account the spatial information in the cube. SOLAP operators involve spatial conditions or spatial functions by using the spatial operators defined in Sect. 3.1. Spatial conditions specify constraints on the geometries associated to cube members or measures, while spatial functions derive new data from the cube, which can be used, e.g., to derive dynamic spatial hierarchies or levels, as explained in the following example. Spatial extensions of the common OLAP operators are formally defined in Sect. 5.

Table 1
Sample (instance) data for the Sales cube

<table>
<thead>
<tr>
<th>Customer City</th>
<th>Supplier</th>
<th>Total Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Düsseldorf</td>
<td>c1</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>c2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>c3</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>c4</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>c5</td>
<td></td>
</tr>
</tbody>
</table>

Table 2
Roll-up of the Sales cube

<table>
<thead>
<tr>
<th>City</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Düesseldorf</td>
<td>21</td>
</tr>
<tr>
<td>Dortmund</td>
<td>34</td>
</tr>
<tr>
<td>Münster</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 3
S-Roll-up of the Sales cube

<table>
<thead>
<tr>
<th>CityClosest Supplier</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>Düesseldorf</td>
<td>25</td>
</tr>
<tr>
<td>Dortmund</td>
<td>20</td>
</tr>
<tr>
<td>Münster</td>
<td>33</td>
</tr>
</tbody>
</table>

Example 1. Consider the summarized data for the Sales cube given in Table 1, where a ‘–’ is used if there are no sales to customers from the corresponding suppliers. The data in Table 1 is shown on the map in Fig. 5, where the arrows on the map between the supplier and customer locations represent the distance. The quantities of sold products are shown along these arrows.

The hierarchies in Fig. 4a–c can be used to perform classical roll-up operations, where measures are aggregated from a child to a parent level. An example of such a roll-up operator is expressed by the query “total sales to customers by city”, whose results is given in Table 2.

On the other hand, as shown in Table 4 and Fig. 5, some customers may be closer to suppliers from other cities. For example, customer c3 is related to its city Dortmund by using traditional Geography hierarchy, but the customer is closer to the city Düesseldorf of supplier s1. Similarly, customer c4 in city Dortmund is closer to the city Münster of supplier s3. Fig. 4d shows...
a new dynamic spatial hierarchy that can be obtained
with a spatial roll-up (s-roll-up) operator that expresses
the query “total sales to customers by city of the closest
supplier”. Such queries are not possible to express
on conventional hierarchies with traditional OLAP.

The hierarchy in Fig. 4d is created on the fly with the help
of a spatial function computing the distance between
customer and supplier locations. Therefore, using the s-roll-up operator, sales to customers are aggre-
gated by city of the closest suppliers, where Dortmund
has a significant drop off in the quantity of the sales
from 34 (Table 2) to 20 (Table 3).

4. The QB4SOLAP vocabulary

In this section, we formally define how to represent
(spatial) data cubes in RDF. We use as running example
the GeoNorthwind data warehouse whose concep-
tual schema is given in Fig. 7.

The QB4OLAP [11] vocabulary allows to define
cube schemas and cube instances as RDF triples.
QB4OLAP is an extension of the RDF Data Cube
Vocabulary (QB) [6] with multidimensional concepts
in order to be able to support OLAP operations di-
rectly over RDF data with SPARQL queries. We ex-
 tended QB4OLAP (v1.2)8 with spatial concepts to
give QB4SOLAP [17]. We based our extension on
GeoSPARQL [30], a standard from the Open Geospa-
tial Consortium (OGC) for representing and querying
geospatial linked data for the Semantic Web. Since our
base vocabulary QB4OLAP uses the MultiDim con-
ceptual model to describe the multidimensional con-
cepts, we base our definitions on a spatially extended
version of MultiDim conceptual model [37] for spa-
tial extension of the MD concepts. Fig. 6 shows the
QB4SOLAP vocabulary for representing a spatial cube
schema and spatial cube members as RDF triples. A
cube schema defines the structure of the cube in terms
of dimension levels, measures, aggregation functions
(e.g., SUM, AVG, COUNT) on measures, spatial ag-
gregation functions (S_agg in Def. 1) on spatial mea-
ures, dimensions hierarchies, and parent–child relations-
ships between levels (including their cardinality and
topological relationships for spatial levels). These
schema level metadata are used to define multidimen-
sional datasets in RDF. Cube members are the in-
stances of a cube schema that represent level mem-
bors, facts, and measure values. As we will show in
Sect. 6, we use the schema level metadata to produce
SPARQL queries that implement SOLAP operators on
cube members.

Terms with capitalized initials and non-italic font
in Fig. 6 represent RDF classes, terms with capital-
ized initials and italic font represent RDF instances,
and terms with non-capitalized initials represent RDF
properties. Classes in external vocabularies are de-
picted in light gray background and font. RDF Cube
(QB), QB4OLAP, and QB4SOLAP classes are shown,
respectively, with white, light gray, dark gray back-
gounds. Original QB terms are prefixed with qb:9.
QB4OLAP and QB4SOLAP terms are prefixed, re-
spectively, with qb4o:10 and qb4so:11. Spatial
classes and properties are prefixed with geo:12.

In what follows, we first define formally RDF
triples, and then discuss how to describe (spatial) mul-
tidimensional data using QB4OLAP and QB4SOLAP.

Definition 4. (RDF triple) An RDF triple \( t = (s, p, o) \)
consists of three components: \( s \) is the subject, \( p \) is the predicate, and \( o \) is the object. RDF triples are defined
over
\[
T = (\mathcal{I} \cup \mathcal{B}) \times \mathcal{I} \times (\mathcal{I} \cup \mathcal{B} \cup \mathcal{L})
\]

where \( \mathcal{I} \) is the set of IRIs, \( \mathcal{B} \) is the set of blank nodes,
and \( \mathcal{L} \) is the set of literals.

A set of RDF triples is referred to as a graph. We
denote a QB4SOLAP graph by \( \mathcal{G} \), where \( \mathcal{G} \subset T \).The
cube schema and cube instances are subsets of this
graph and are denoted, respectively, by \( G^S \) and \( G^I \),
where \( G^S \subset \mathcal{G} \) and \( G^I \subset \mathcal{G} \).

Given an MD element \( x \in (\mathcal{I} \cup \mathcal{B}) \) in a schema
graph or instance graph \( \mathcal{G} \), we define by \( \mathcal{G}_x \) the sub-
graph of \( \mathcal{G} \) for \( x \), where \( \mathcal{G}_x \subset \mathcal{G} \). We define the func-
tion \( id(x) : \mathcal{G} \rightarrow \mathcal{I} \), that given a MD element \( x \) returns
its identifier \( \mathcal{I} \) from the graph \( \mathcal{G} \). We use superscript
notation to indicate the type of the identifier from the
cube schema graph \( (G^S) \) and cube instance graph \( (G^I) \),
e.g., \( id^S(x) \) for a cube schema identifier and \( id^I(x) \)
for a cube instance identifier.

---
8QB4OLAP v1.2: https://github.com/lorenae/qb4olap/blob/master/rdf/qb4olap.1.2.ttl
9RDF cube: http://purl.org/linked-data/cube#
10QB4OLAP: http://purl.org/qb4olap/cubes#
11QB4SOLAP: http://w3id.org/qb4solap#
12GeoSPARQL: http://www.opengis.net/ont/geosparql#
4.1. Defining spatial data cube schemas with QB4SOLAP

An n-dimensional cube schema $CS$ is a tuple $CS = (D, M, F)$, with a set of dimensions $D$, a set of measures $M$, and a fact $F$. A dimension $d \in D$ has a set of hierarchies $H(d)$. Each hierarchy $h \in H(d)$ is organized into a set of levels $L(h)$. Each level $l \in L(h)$ has a set of attributes $A(l)$. Each attribute $a \in A(l)$ is defined over a domain. Each measure $m \in M$ is also defined over a domain.

We define next how to represent a cube schema $CS$ in RDF using QB4SOLAP. We denote the RDF graph of the cube schema $G^S$. In the examples we prefix the elements of $G^S$ with gnw:. We follow a similar naming convention for schema elements as in QB4OLAP. If there is a possibility of confusion for different MD concepts with same schema name, i.e., customer dimension and customer level, we suffix the dimensions with Dim (e.g., gnw:customerDim for dimension, and gnw:customer for level). The subgraph of $G^S$ that refers to a specific schema element $x$ is denoted by $G^S_x$ and the unique identifier of $x$ is denoted by $id^S(x)$.

Definition 5. (Dimensions) An n-dimensional cube schema $CS$ has a set of dimensions $D = \{d_1, \ldots, d_n\}$ and each dimension $d_i$ has a set of hierarchies $H(d_i)$ (Def. 6). Each dimension $d_i \in D$ is defined in the cube schema graph $G^S$ with qb:DimensionProperty. Each hierarchy $h \in H(d_i)$ is linked to its dimension $d_i$ with the qb4o:hasHierarchy property. The RDF graph formulation of the dimensions $D$ is represented as

$$G^S_D = \bigcup_{i=1}^{n} G^S_{d_i}$$
where
\[ G^S_{H(d)} = \bigcup_{i=1}^{m} G^S_{h_i} \]

\[ G^S_{h_i} = \{(id^S(h_i) \text{ rdf:type qb:Hierarchy})\} \cup \{(id^S(h_i) \text{ qb4o:hasHierarchy } id^S(d))\} \cup \bigcup_{l \in L(h_i)} \{(id^S(h_i) \text{ qb4o:hasLevel } id^S(l))\} \]

**Extension 5.** (Spatial dimensions) A dimension is spatial if it has at least one spatial level. A spatial dimension \(d_s\) belongs to the set of spatial dimensions \(D_s\), which is a subset of the set of dimensions \(D\), such that \(d_s \in D_s \subseteq D\).

**Example 2.** The triples below show how some of the dimensions of the GeoNorthwind DW (Fig. 7) are represented in RDF using Def. 5 and Ext. 5. As we will see below, the Customer and Supplier dimensions are spatial as they both have a spatial hierarchy Geography.

\[
\begin{align*}
\text{# Dimensions} & \\
\text{gnw:customerDim rdf:type qb:DimensionProperty} ; \\
\text{gnw:hasHierarchy gnw:customerGeography} & . \\
\text{gnw:supplierDim rdf:type qb:DimensionProperty} ; \\
\text{gnw:hasHierarchy gnw:supplierGeography} & . \\
\text{gnw:employeeDim rdf:type qb:DimensionProperty} ; \\
\text{gnw:hasHierarchy gnw:categories} & . \\
\text{gnw:productDim rdf:type qb:DimensionProperty} ; \\
\text{gnw:hasHierarchy gnw:productDim} & .
\end{align*}
\]

**Definition 6.** (Hierarchies) A dimension \(d\) has a set of hierarchies \(H(d) = \{h_1, \ldots, h_m\}\), where each hierarchy \(h_i\) has a set of levels \(L(h_i)\) (Def. 7). Each hierarchy \(h_i \in H(d)\) is defined in the cube schema graph \(G^S\) with the \text{qb4o:hasHierarchy} predicate and is linked with its dimension \(d\) by the \text{qb4o:inDimension} property. Each level \(l \in L(h_i)\) that belongs to a hierarchy \(h_i\) is defined with the \text{qb4o:hasLevel} property. The RDF graph formulation of the hierarchies \(H(d)\) is represented as

\[ G^S_{H(d)} = \bigcup_{i=1}^{m} G^S_{h_i} \]

**Example 3.** The triples below show how some of the hierarchies of the GeoNorthwind DW (Fig. 7) are represented in RDF using Def. 6 and Ext. 6. As we will see below, the Geography hierarchies in the Customer and Supplier dimensions are spatial since they have spatial levels (City, State, etc.)

\[
\begin{align*}
\text{# Hierarchies} & \\
\text{gnw:customerGeography rdf:type qb4o:Hierarchy} ; \\
\text{qb4o:inDimension gnw:customerDim} ; \\
\text{qb4o:hasLevel gnw:customer} & , \text{gnw:city}, \text{gnw:state}, \text{gnw:country} . \\
\text{gnw:supplierGeography rdf:type qb4o:Hierarchy} ; \\
\text{qb4o:inDimension gnw:supplierDim} ; \\
\text{qb4o:hasLevel gnw:supplier} & , \text{gnw:city}, \text{gnw:state}, \text{gnw:country} . \\
\text{gnw:categories rdf:type qb4o:Hierarchy} ; \\
\text{qb4o:inDimension gnw:productDim} ; \\
\text{qb4o:hasLevel gnw:product} & , \text{gnw:category} .
\end{align*}
\]

**Definition 7.** (Levels) A hierarchy \(h\) has a set of levels \(L(h) = \{l_1, \ldots, l_k\}\) and each level \(l_i\) has a set of attributes \(A(l_i)\) (Def. 8). Each level \(l_i \in L(h)\) is defined in the cube schema graph \(G^S\) with the \text{qb4o:LevelProperty} predicate. Each attribute \(a \in A(l_i)\) is linked to its level \(l_i\) with the \text{qb4o:hasAttribute} property. The RDF graph formulation of the levels \(L(h)\) is represented as

\[ G^S_{L(h)} = \bigcup_{i=1}^{k} G^S_{l_i} \]

where
\[ G^S_{l_i} = \{(id^S(l_i) \text{ rdf:type qb4o:LevelProperty})\} \cup \bigcup_{A(l_i)} \{(id^S(l_i) \text{ qb4o:hasAttribute } id^S(a))\} \]

**Extension 7.** (Spatial levels) A level is spatial if it has an associated geometry. A spatial level \(l_{s_i}\) belongs to the set of spatial levels \(L_s(h)\), which is a subset of the set of levels \(L(h)\), such that \(l_{s_i} \in L_s(h) \subseteq L(h)\). The geometry of a spatial level is defined in the cube schema graph \(G^S\) with the \text{geo:hasGeometry} property. The RDF graph formulation of the spatial levels \(L_s(h)\) is represented as

\[ G^S_{L_s(h)} = \bigcup_{i=1}^{k} G^S_{l_{s_i}} \]
where

\[ G^S_{l_1} = \{ (id^S(l_1) \text{ rdf:type qb4o:LevelProperty}) \} \cup \{ (id^S(l_1) \text{ geo:hasGeometry geo:Geometry}) \} \cup \bigcup_{a \in A(l_1)} \{ (id^S(l_1) \text{ qb4o:hasAttribute id^S(a)}) \} \]

Example 4. The triples below show how some of the levels of the GeoNorthwind DW (Fig. 7) are represented in RDF using Def. 7 and Ext. 7. Note that the Customer and City levels are spatial as they have a geometry that is specified at the level definition.

Example 4. The triples below show how some of the levels of the GeoNorthwind DW (Fig. 7) are represented in RDF using Def. 7 and Ext. 7. Note that the Customer and City levels are spatial as they have a geometry that is specified at the level definition.

\[ \text{# Levels} \]
\[ \text{gnw:customer rdf:type qb4o:LevelProperty ;} \]
\[ \text{qb4o:hasAttribute gnw:customerID ;} \]
\[ \text{qb4o:hasAttribute gnw:customerName ;} \]
\[ \text{qb4o:hasAttribute gnw:address ;} \]
\[ \text{qb4o:hasAttribute gnw:postalCode ;} \]
\[ \text{geo:hasGeometry gnw:customerGeometry .} \]

\[ \text{Definition 8. (Attributes) A level } l \text{ has a set of attributes } A(l) = \{ a_1, \ldots, a_p \}, \text{ which defines the characteristics of the level members. One among these attributes, denoted as } a_{ID}, \text{ specifies a surrogate key for the level, i.e., the value of } a_{ID} \text{ uniquely identifies the members of the level. For simplicity, we assume that it is the first attribute in the set of attributes } A(l), \text{ i.e., } a_1 = a_{ID}. \text{ Each attribute } a_i \in A(l) \text{ is defined in the cube schema graph } G^S \text{ with the } \text{qb4o:LevelAttribute predicate} \text{ and is linked to its level } l \text{ with the } \text{qb4o:inLevel property}. \text{ Each attribute } a_i \text{ is defined as ranging over XSD literals } \mathcal{L} \text{ using the } \text{rdfs:range property}. \text{ The RDF graph formulation of the attributes } A(l) \text{ is represented as} \]

\[ G^S_{A(l)} = \bigcup_{i=1}^{p} G^S_{a_i} \]
Extension 8. (Spatial attributes) An attribute is spatial if it is defined over a spatial domain. A spatial attribute $a_i$ belongs to the set of spatial attributes $A(l)$, which is a subset of the set of attributes $A(l)$, such that $a_i \in A(l) \subseteq A(l)$. The RDF graph formulation of the spatial attributes is similar as in Def. 8. However, the attribute must range over spatial literals $L_s$ i.e., a well-known text literal (WKT) from OGC schemas. Further, the domain of the attribute should be specified with the rdfs:domain property, which must be a geometry. Finally, the attribute must be specified as spatial object with the rdfs:subClassOf property. The RDF graph formulation of the spatial attributes $A_i(l)$ is represented as

$$G^{S}_{A_{i}(l)} = \bigcup_{i=1}^{p} G^{S}_{A_{i}}$$

where

$$G^{S}_{A_{i}} = \\
\{\text{id}^{S}(a_i) \text{ rdf:type qb4o:LevelAttribute})\bigcup \\
\{\text{id}^{S}(a_i) \text{ qb4o:inLevel id}^{S}(l))\bigcup \\
\{\text{id}^{S}(a_i) \text{ rdfs:range } L_s\}\bigcup \\
\{\text{id}^{S}(a_i) \text{ rdfs:subPropertyOf geo:Geometry})\bigcup \\
\{\text{id}^{S}(a_i) \text{ rdfs:subClassOf geo: SpatialObject}\}$$

Example 5. The triples below show how some of the attributes of the GeoNorthwind DW (Fig. 7) are represented in RDF using Def. 8 and Ext. 8. Note that the Customer level has a spatial attribute (Customer geometry). It is represented as a WKT literal that defines a Point type from the Geometry class, which is a subclass of Spatial Object.

```
# Attributes
gnw:customerID rdf:type qb4o:LevelAttribute ; 
qb4o:inLevel gnw:customer; 
rdfs:range xsd:Integer .
```

Definition 9. (Hierarchy steps) A hierarchy $h$ has a set of hierarchy steps $HS(h) = \{hs_1, \ldots, hs_q\}$, which define the structure of the hierarchy in relation with its corresponding levels. A hierarchy step $hs_i = (l_c, l_p, card) \in HS(h)$ entails a roll-up relation between a lower (child) level $l_c$ to an upper (parent) level $l_p$ with a cardinality $card$. The cardinality $card \in \{1, 1-n, n-1, n-n\}$ describes the number of members in one level that can be related to a member in the other level for both the child and the parent levels.

Each hierarchy step $hs_i$ is defined in the cube schema graph $G^S$ as a blank node _:hs_i $\in B$ with the qb4o:HierarchyStep predicate. Each hierarchy step is linked to its hierarchy with the qb4o:inHierarchy property. The child and parent levels are linked in a hierarchy step with the qb4o:childLevel and qb4o:parentLevel properties, respectively. The cardinality $card$ of a hierarchy step is defined by the qb4o:pcCardinality property. The RDF graph formulation of the hierarchy steps $HS(h)$ is represented as

$$G^{S}_{HS(h)} = \bigcup_{i=1}^{q} G^{S}_{hs_i}$$

where

$$G^{S}_{hs_i} = \\
\{\text{_:hs}_i \text{ rdf:type qb4o:HierarchyStep})\bigcup \\
\{\text{_:hs}_i \text{ qb4o:inHierarchy id}^{S}(h))\bigcup \\
\{\text{_:hs}_i \text{ qb4o:parentLevel id}^{S}(l_p))\bigcup \\
\{\text{_:hs}_i \text{ qb4o:childLevel id}^{S}(l_c))\bigcup \\
\{\text{_:hs}_i \text{ qb4o:pcCardinality id}^{S}(card))\}$$
**Extension 9.** (Spatial hierarchy steps) A hierarchy step is spatial if it relates a spatial child level \( l_c \) and a spatial parent level \( l_p \), in which case it entails a topological relationship between these spatial levels. A spatial hierarchy step is then a tuple \( h_{s_i} = (l_c, l_p, \text{card}, \text{topoRel}) \) where the topological relation \( \text{topoRel} \) belongs to the \( \mathcal{T}_{rel} \) class (Def. 2). The topological relation between parent-child levels of a spatial hierarchy step is defined by the \( \text{qb4so:pcTopoRel} \) property. The RDF graph formulation of the spatial hierarchy steps \( HS(h) \) (w.r.t. Def. 9) is represented as

\[
G_{HS(h)}^S = \bigcup_{i=1}^{q} G_{h_{s_i}}^S
\]

where

\[
G_{h_{s_i}}^S = G_{h_{s_i}}^S \cup \{\text{_:hs}_{i} \text{qb4so:pcTopoRel} \text{id}^S(\text{topoRel})\}
\]

**Example 6.** The triples below show how the hierarchy steps of the GeoNorthwind schema in the Customer dimension are represented in RDF using Def. 9 and Ext. 9. Note that all hierarchy steps are spatial and have an associated topological relation.

```
# Hierarchy steps
_:customerGeography_hs1 a qb4o:HierarchyStep ;
  qb4o:inHierarchy gwn:customerGeography ;
  qb4o:childLevel gwn:customer ;
  qb4o:parentLevel gwn:city ;
  qb4o:pcCardinality qb4o:ManyToOne ;
  qb4so:pcTopoRel qb4so:Within .
_:customerGeography_hs2 a qb4o:HierarchyStep ;
  qb4o:inHierarchy gwn:customerGeography ;
  qb4o:childLevel gwn:city ;
  qb4o:parentLevel gwn:state ;
  qb4o:pcCardinality qb4o:ManyToOne ;
  qb4so:pcTopoRel qb4so:Within .
_:customerGeography_hs3 a qb4o:HierarchyStep ;
  qb4o:inHierarchy gwn:customerGeography ;
  qb4o:childLevel gwn:state ;
  qb4o:parentLevel gwn:country ;
  qb4o:pcCardinality qb4o:ManyToOne ;
  qb4so:pcTopoRel qb4so:Within .
```

**Definition 10.** (Partial order on levels) The hierarchy steps \( HS(h) \) of a hierarchy \( h \) define a partial order on the levels \( l \in L(h) \). The reflexive and transitive closure of the partial order is denoted as \( \sqsubseteq \), with a unique base level \( l_b \) and a unique top level \( All \), where all levels \( l \) are such that \( l_b \sqsubseteq l \) and \( l \sqsubseteq All \).

**Definition 11.** (Measures) An \( n \)-dimensional cube schema has a set of measures \( M = \{m_1, \ldots, m_r\} \), which record the values of a phenomena being observed. Each measure \( m_i \) is defined in the cube schema graph \( G^S \) with the \( \text{qb:MeasureProperty} \) predicate. Similarly to attributes, each measure \( m_i \) is defined as ranging over XSD literals \( \mathcal{L} \) with the \( \text{rdfs:range} \) property. The RDF graph formulation of the measures \( M \) is represented as

\[
G_{M}^S = \bigcup_{i=1}^{r} G_{m_i}^S
\]

where

\[
G_{m_i}^S = \{\text{id}^S(m_i) \text{rdfs:range} \mathcal{L} \} \cup \{\text{id}^S(m_i) \text{rdfs:range} \mathcal{L}_s \}
\]

**Extension 11.** (Spatial measures) A measure is spatial if it is defined over a spatial domain as in spatial attributes (Ext. 8). A spatial measure \( m_i \) belongs to the set of spatial measures \( M_s \), which is a subset of the set of measures \( M \), such that \( m_i \in M_s \subseteq M \). The RDF formulation of the spatial measures is similar as in Def. 11. However, the domain should range over spatial literals \( \mathcal{L}_s \). The RDF graph formulation of the spatial measures \( M_s \) (w.r.t. Def. 11) is represented as

\[
G_{M_s}^S = \bigcup_{i=1}^{r} G_{m_i}^S
\]

where

\[
G_{m_i}^S = \{\text{id}^S(m_i) \text{rdfs:range} \mathcal{L}_s \} \cup \{\text{id}^S(m_i) \text{rdfs:range} \mathcal{L}_s \text{subClassOf geo: SpatialObject} \}
\]

**Example 7.** The triples below show how the measures of the GeoNorthwind DW (Fig. 7) are represented in RDF using Def. 11 and Ext. 11. Note that SalesPoint is a spatial measure, which records the location of the
stores in which the sales occurred. It is defined over Geometry domain as a Point type with WKT literal.

# Measures
gnw:quantity rdf:type qb:MeasureProperty ;
rdfs:range xsd:integer .
gnw:unitPrice rdf:type qb:MeasureProperty ;
rdfs:range xsd:decimal .
gnw:discount rdf:type qb:MeasureProperty ;
rdfs:range xsd:decimal .
gnw:salesAmount rdf:type qb:MeasureProperty ;
rdfs:range xsd:decimal .
gnw:freight rdf:type qb:MeasureProperty ;
rdfs:range xsd:decimal .
gnw:salesPoint rdf:type qb:MeasureProperty ;
rdfs:domain geo:Point;
rdfs:range geo:wktLiteral ;
rdfs:subClassOf geo:SpatialObject .

Definition 12. (Fact) In an $n$-dimensional cube schema $\mathcal{CS} = (D, M, F)$, the fact $F$ defines the structure of a cube with the qb:DataStructureDefinition property. The dimensions are given as components of the fact and are defined with the qb4o:level property. We assume that the fact $F$ links the dimensions at the lowest granularity level, therefore $\text{qb4o:level}$ links the lowest (base) level $i_b$ of each dimension $d_i$, which is denoted as $i_b(d_i)$. The cardinality $\text{card}$ of the relationship between a dimension level and a fact is represented with the $\text{qb4o:cardinality}$ property. Similarly, the measures are given as components of the fact and are defined with the $\text{qb:measure}$ property. The aggregate function $\text{aggr}$ associated to each measure is represented with the $\text{qb4o:aggregateFunction}$ property. The RDF graph formulation of the fact $F$ is given in the following equation.

\[
\mathcal{G}_F = \{(id^S(F))
\text{rdf:type qb:DataStructureDefinition}) \cup
\bigcup_{d_i \in D}
\{(id^S(F)) \text{qb:component}
\text{qb4o:level id}^S(i_b(d_i)) ;
\text{qb4o:cardinality id}^S(\text{card})\}\cup
\bigcup_{m_i \in M}
\{(id^S(F)) \text{qb:component}
\text{qb:measure id}^S(m_i) ;
\text{qb4o:aggregateFunction id}^S(\text{aggr})\}\}
\]

Extension 12. (Spatial fact) A fact is spatial if it relates several levels, where two or more are spatial.

A spatial fact may also have a topological relation $\text{topoRel}$ that must be satisfied by the related spatial levels, which is represented with $\text{qb4so:topologicalRelation}$. This object property allows to specify a topological relation in fact-level relationship of spatial facts. The RDF graph formulation of such a fact is simply by adding the property of fact-level topological relation consecutively to the cardinality property as given in the following equation.

\[
\mathcal{G}^S_{F_s} = \{(id^S(F_s))
\text{rdf:type qb:DataStructureDefinition}) \cup
\bigcup_{d_i \in D}
\{(id^S(F_s)) \text{qb:component}
\text{qb4o:level id}^S(i_b(d_i)) ;
\text{qb4so:cardinality id}^S(\text{card}) ;
\text{qb4so:topologicalRelation id}^S(\text{topoRel})\}\cup
\bigcup_{m_i \in M}
\{(id^S(F_s)) \text{qb:component}
\text{qb:measure id}^S(m_i) ;
\text{qb4o:aggregateFunction id}^S(\text{aggr})\}\}
\]

Example 8. The triples below show how the fact of the GeoNorthwind DW (Fig. 7) is represented in RDF using Def. 12. Sales fact does not impose any topological relation between its spatial dimensions Supplier and Customer. SalesPoint is a spatial measure, which has a spatial aggregate function (AggrConvexHull).

# Cube definition
gnw:GeoNorthwind rdf:type
qb:DataStructureDefinition ;
  # Lowest level for each dimension
  qb:component [qb4o:level gnw:customer ;
  qb4o:cardinality qb4o:ManyToOne] ;
  qb:component [qb4o:level gnw:supplier ;
  qb4o:cardinality qb4o:ManyToOne] ;
  qb:component [qb4o:level gnw:product ;
  qb4o:cardinality qb4o:ManyToOne] ;
  ...;
  # Cube measures
  qb:component [qb:measure gnw:quantity ;
  qb4o:aggregateFunction qb4o:Sum] ;
  qb:component [qb:measure gnw:unitPrice ;
  qb4o:aggregateFunction qb4o:Sum] ;
  qb:component [qb:measure gnw:discount ;
  qb4o:aggregateFunction qb4o:Avg] ;
  ...;
  qb:component [qb:measure gnw:salesPoint ;
4.2. Defining spatial data cube members with QB4SOLAP

We have explained in Sect. 4.1 how a data cube schema can be represented in RDF with QB4SOLAP. We show next how to use this schema to represent the instances of the GeoNorthwind DW (Fig. 7) in RDF. We denote by $G^l$ the RDF graph of the data cube instances. In the examples, we prefix the elements of $G^l$ with $\text{gnwi:}$. The subgraph of $G^l$ that refers to a specific cube instance $x$ is denoted by $G^l_x$ and the unique identifier of $x$ is denoted by $id(x)$.

**Definition 13.** (Level members) A level $l$ has a set of level members $LM(l) = \{lm_1, \ldots, lm_n\}$. Each level member $lm_i$ has a unique IRI $id(lm_i) \in I$, which is linked in the cube instance graph $G^l$ with the $\text{qb4o:LevelMember}$ predicate. A level member is related to its level by the $\text{qb4o:memberOf}$ property. The RDF graph formulation of the level members $LM(l)$ is represented as

$$G^l_{LM(l)} = \bigcup_{i=1}^{n} G^l_{lm_i}$$

where

$$G^l_{lm_i} = \{ (id(lm_i) \text{ rdf:type } \text{qb4o:LevelMember}) \cup \{ (id(lm_i) \text{ qb4o:memberOf } id(l)) \} \}$$

**Definition 14.** (Attributes of level members) A level member $lm$ has a set of attributes $A(lm) = \{a_1, \ldots, a_p\}$, which are used to describe the characteristics of the level member (Def. 8). Each attribute $a_i$ is linked to the level member with the identifier $id^S(a_i)$. We denote by $lm \leadsto v_{a_i}$ the value $v_{a_i}$ that a level member $lm$ associates to attribute $a_i$. This value is given as a literal $L$ such that $v_{a_i} \in L$. The RDF graph formulation of the attributes $A(lm)$ is represented as

$$G^l_{A(lm)} = \bigcup_{i=1}^{p} G^l_{a_i}$$

where

$$G^l_{a_i} = \{ (id(lm) \text{ id}^S(a_i) v_{a_i}) | lm \leadsto v_{a_i} \}$$

**Definition 15.** (Partial order on level members) A hierarchy step $hs = (l, l_p, \text{card})$ between a child level $l$ and a parent level $l_p$ defines a set of roll-up relations $RU(hs) = \{r_1, \ldots, r_k\}$ where each $r_i = \text{lm}_{c_i} \subseteq \text{lm}_{p_i}$ relates a child level member $\text{lm}_{c_i} \in LM(l)$ to a parent level member $\text{lm}_{p_i} \in LM(l_p)$. These roll-up relations define a partial order between level members with regard to Def. 10 and are expressed using the property $\text{skos:broader}$. The RDF graph formulation of the roll-up relations $RU(hs)$ is represented as

$$G^l_{RU(hs)} = \bigcup_{i=1}^{k} G^l_{r_i}$$

where

$$G^l_{r_i} = \{ (id(lm_c) \text{ skos:broader } id(lm_p)) | r_i = \text{lm}_{c_i} \subseteq \text{lm}_{p_i} \}$$

**Example 9.** The triples below show how some level members of the GeoNorthwind DW (Fig. 7) are represented in RDF usingDefs. 13–14.

```
gnwi:customer_1 rdf:type qb4o:LevelMember ;
  qb4o:memberOf gnw:customer ;
  gnw:customerID 1 ;
  gnw:customerName "Alfreds Futterkiste" ;
  gnw:address "Obere Str. 57" ;
  gnw:postalCode "12209" ;
  gnw:customerGeo
    "POINT(13.099 52.401)""geo:wktLiteral ;

gnwi:city_6 rdf:type qb4o:LevelMember ;
  qb4o:memberOf gnw:city ;
  gnw:name "Berlin" ;
  gnw:cityGeo
    "POINT(13.4060 52.519)""geo:wktLiteral ;
    skos:broader gnwi:state_224 .
```

**Definition 16.** (Fact members) A fact $F$ has a set of fact members $FM(F) = \{f_1, \ldots, f_t\}$, which are the instances of the data cube. Each fact $f_i \in FM$ has a unique IRI $id^l(f_i) \in I$, which is linked in the cube instance graph $G^l$ with the $\text{qb:Observation}$ predicate.

A fact member $f_i$ is related to a set of dimension levels $l(f_i) = \{l_1, \ldots, l_s\}$ and has a set of measures $M(f_i) = \{m_1, \ldots, m_k\}$. Each dimension level $l_j$ is linked to the level member with the identifier $id^S(l_j)$ and each measure $m_k$ is linked to the level member with the identifier $id^S(m_k)$. We denote by $f \leadsto v_{l_j}$ and $f \leadsto v_{m_k}$, respectively, the dimension values and
measures values associated with a fact \( f \). The value \( v_i \in I \) is the identifier of a level member in \( LM(l_j) \). Further, the value \( v_{mk} \) for every measure \( m_k \) is a literal such that \( v_{mk} \in L \). The RDF graph formulation of the fact members \( FM(F) \) is represented as

\[
G^f_{FM(F)} = \bigcup_{i=1}^{t} G^f_i
\]

where

\[
G^f_i = \{(id(f_i) \text{ rdf:type qb:Observation}) \cup \bigcup_{l_j \in LM(f_i)} \{(id(f_i) \text{ id}(l_j) \text{ id}(v_{ij}) \mid f_i \rightarrow v_{ij})\} \cup \bigcup_{m_k \in M(f_i)} \{(id(f_i) \text{ id}(m_k) \text{ id}(v_{mk}) \mid f_i \rightarrow v_{mk})\}
\]

**Example 10.** The triples below show how a fact member of the GeoNorthwind DW (Fig. 7) is represented in RDF using Defs. 12–16. Note that the fact member and corresponding level members relating to dimensions are given with the prefix \( \text{gnwi:} \). The complete SPARQL query exam-

\begin{verbatim}
gnwi:sale_10613_1 rdf:type qb:Observation ;
gnwi:customer gnwi:customer_1 ;
gnwi:supplier gnwi:supplier_6 ;
gnwi:product gnwi:product_13 ;
... gnwi:quantity 8 ;
gnwi:unitPrice "6,00"ˆˆxsd:decimal ;
gnwi:discount "0,10"ˆˆxsd:decimal ;
gnwi:salesPoint 
"POINT(23.08 42.34)"ˆˆgeo:wktLiteral.
\end{verbatim}

5. Semantics of SOLAP operators

This section defines a formal algebra for SOLAP operators. Examples of the operators are provided after their definitions. The complete SPARQL query examples are given at our website\(^{13}\) and can be tested at our public endpoint\(^{14}\). The query runtimes for each SOLAP operator are given in Appendix A1, Table 5 for the use case dataset GeoNorthwind (Sect. 4.1, Fig. 7). These operators can be applied on spatially enhanced multidimensional data cubes (Sect. 3.3). The presentation defines the semantics of a SOLAP operator by logically specifying the typical OLAP operators with spatial functions and conditions. Spatial functions and conditions can be selected from a range of operation classes, which can be applied on spatial data types (Sect. 3.1). Let \( S \) be the set of any spatial operators where \( S = (S_{agg} \cup T_{rel} \cup N_{op}) \), used to represent a spatial predicate \( \phi^S \in S \) or a spatial function \( f^S \in S \), which is in a SOLAP operator. The following SOLAP operators are defined with a spatial extension to the well-known OLAP operators, which are given in the remarks.

**Remark 17.** (Slice) The slice operator removes a dimension from a cube \( C \) by selecting one instance in a dimension level. For example, the query “slice on customers in the city of Odense” is a slice operation. (Cube is the sales, dimension is the customer, level in dimension is the city and the value is Odense, which is sliced out from the cube).

**Definition 17.** (S-Slice) The s-slice operator removes a dimension from a cube \( C \) by choosing a single spatial attribute value \( v_s \in L_s \) (Ext. 8) in a spatial level \( l_s \) (Ext. 7).

As for the semantics, the s-slice takes an \( n \)-dimensional cube \( C \) as an argument. We assume that the cube has the cube schema \( CS = (D, M, F) \), with the fact members \( f \in FM \) as given in Def. 16. As parameters, s-slice takes a spatial literal value \( v_s \), the base level \( l_b \) and the target (spatial) level \( l_s \) of a dimension \( d_i \). The base level \( l_b \) specifies the dimension \( d_i \) (Def. 16). The target spatial level \( l_s \) is the level, that the spatial literal value \( v_s \) is related.

The operator is defined as: \( SS(C)[l_b, l_s, v_s] = C' \), which returns a cube \( C' \) with \( n - 1 \) dimensions and the schema \( CS' = (D', M', F') \), where \( D' = D \setminus \{d_i\} \), \( M' = M \), and \( F' = F \). The measures \( M \) and the fact type \( F \) remains the same though the new cube \( C' \) has one dimension less.

The s-slice operator selects a subset \( FM' \) from the set of fact members \( FM \) (\( FM' \subseteq FM \)), with respect to the given parameter \( v_s \). Assuming that the granularity of the fact members are at the (lowest) base level of the dimension \( l_b \in L(d_i) \) in the given cube, a partial order exists among the levels, from bottom level to the target spatial level \( l_s \) such that \( l_b \sqsubseteq l_s \). The given parameter \( v_s \) is related to a level member of the level \( l_s \). We say that the fact members are characterized by dimension values, which is written as \( f \rightarrow v_{d_i} \) where \( v_{d_i} \equiv v_{l(b(d_i))} \) (Def. 16). In other words, dimensions

\(^{13}\)http://extbi.cs.aau.dk/SOLAP4SW/queries
\(^{14}\)http://extbi.lab.aau.dk/sparql
are associated to the fact members by the values of the dimensions’ base level members $v_{b(d_i)}$. When the dimension $d_i$ is clear in the context, we will use base level $v_b$ for simplicity reasons.

To sum up, the subset $FM'$ of facts is selected with regards to the partial order on levels from base level $l_b$ to the target level $l_s$. The value $v_b$ in the target level $l_s$ is specified with respect to the given spatial literal value $v_s$. The value of $v_b$ might be equal to a spatial attribute value in the target level $l_s$, thus $v_s$ is characterized by the attribute value $v_s$ and written as $v_b \rightarrow v_s$ (Ex. 11). Or, $v_s$ is an arbitrary spatial literal that entails a topological relation $T_{rel}$ (i.e., within) in a value of the target spatial level $v_s$, which is written as $\exists v_s : \phi^S(v_s)$ where $\phi^S$ is a spatial Boolean predicate that represents a topological relation (Ex. 12).

After applying the s-slicer operator on cube $C$, the new (sub)set of fact members is defined for both cases respectively as follows: $FM' = \{ f \in FM \mid \exists v_b \in LM(l_b), v_s \in LM(l_s) : f \leftarrow v_b \wedge v_s \subseteq v_b \wedge v_s \rightarrow v_s \}$, $FM'' = \{ f \in FM \mid \exists v_b \in LM(l_b), v_s \in LM(l_s) : f \leftarrow v_b \wedge v_s \subseteq v_b, \phi^S(v_s) \}$.

Example 11. With regards to the traditional slice query “slice on customers in the city of Odense”, in s-slice, the user could specify a geometry extent (e.g., polygon coordinates of the city of Odense) as spatial literal for slicing instead of giving a text literal (e.g., “Odense”). So the s-slice query would be; “slice on customers of the city, which has the geometry "POLYGON((10.43951 55.47006, 10.4394729 55.4700361, 10.4392403 (...))”"). More intuitively, instead of the specified spatial literal $v_s \in L_s$, the user could pass a function call as parameter to s-slice, e.g., by querying “slice on customers in the largest city of (southern) Denmark by land area”. The function call should calculate the area of the cities by their geometries where the largest city is selected as a requirement of the s-slicer operator. Both cases are given in the following.

1. The following SPARQL query shows an s-slice operator, which filters with the given spatial literal by the user.

```
SELECT ?obs WHERE {
  ?obs rdf:type qb:Observation ;
  gwn:customerID ?cust .
  ?cust qb4o:memberOf gwn:customer ;
  skos:broader ?city .
  FILTER (?cityGeo = "POLYGON((10.43951 55.470064, 10.4394729 55.4700361, 10.4392403 (...))") )
}
```

2. The following SPARQL query shows the s-slice operator, which filters with the function call (largest city) returned from inner select. Given the current limitations of SPARQL, there is not an area calculation function from the geometries of the spatial objects during query run time, however we give the query with a notional built-in bif:st_area function.

```
SELECT ?obs WHERE {
  ?obs rdf:type qb:Observation ;
  gwn:customerID ?cust .
  ?cust qb4o:memberOf gwn:customer ;
  skos:broader ?city .
  # Inner select for finding the largest city
  SELECT ?x (MAX(?area) as ?maxArea)
  WHERE {
    ?obs rdf:type qb:Observation ;
    gwn:customerID ?cust .
    ?cust qb4o:memberOf gwn:customer ;
    skos:broader ?city .
    BIND(bif:st_area(?x) as ?area)
  }
  FILTER (?cityGeo = ?x )
}
```

Example 12. With regards to the traditional slice query “slice on customers in the city of Odense”, in this example of s-slice, the user gives a point geometry (i.e., $X,Y$ coordinates of a point as spatial literal) and filter at the given level (i.e., City level) that the given point is within. So the s-slice query would be; “slice on customers of the city, in which the given "POINT(10.43951 55.47006)" is within”.

The following SPARQL query shows an s-slice operator, which filters at the specified level with the given spatial literal by the user.

```
SELECT ?obs WHERE {
  ?obs rdf:type qb:Observation ;
  gwn:customerID ?cust .
  ?cust qb4o:memberOf gwn:customer ;
  skos:broader ?city .
  FILTER (bif:st_within("POINT(10.43951 55.47006)", ?cityGeo))
}
```

Note that the s-slice can be operated in different ways based on the geometry given to the query. In both Ex.s 11 and 12, slice level is given as City, however in Ex. 12 a random $X,Y$ point is given that is falling into the target city. Therefore we need to use within from topological relationships ($T_{rel}$) class in order to verify and filter that city.
Remark 18. (Dice) The traditional dice operator takes a cube and a Boolean condition $\phi$, which returns a new cube containing only the cells that satisfy the Boolean condition $\phi$. Dice operation is analogous to relational algebra, $R$ selection; $\sigma_{\phi}(R)$, but the argument is a cube not a relation. For example, the query “sales to customers of type LLC (Limited Liability Company)” is a dice operation. (Cube is the sales, dimension is the customer, and Boolean condition is the customer type if they are LLC).

Definition 18. (S-Dice) Similarly, the s-dice operator takes an n-dimensional cube $C$ as an argument, which has the cube schema $CS = (D, M, F)$ with the fact members $f \in FM$ as given in Def. 16. As a parameter s-dice takes a spatial Boolean predicate, which is denoted by $\phi^{S}$. The s-dice operator keeps the cells of the cube $C$ that satisfies the spatial predicate over spatial dimension levels $l_d$, attributes $a_s$, and measures $m$.

The semantics of the operator is defined as:
$$SD(C)[\phi^{S}] = C'$$
where spatial predicate $\phi^{S}$ can be applied on spatial level member values $\phi^{S}(v_{l_d})$, spatial attribute values $\phi^{S}(v_{a_s})$, measure values $\phi^{S}(v_{m})$ and/or a combination of these.

$SD$ operator returns a sub cube $C' \subseteq C$, which has the schema $CS = (D', M', F')$ where $D' = D$, $M' = M$, and $F' = F$. Unlike the s-slice operator, s-dice keeps all the dimensions $D$ in the output cube $C'$. The set of measures $M$ and the fact type $F$ also remains the same, though the new cube $C'$ is a subset of the original cube $C$ with filtered fact members $f \in FM'$, which is explained in the following.

The s-dice operator selects a subset $FM'$ of the fact members’ set $FM \subseteq FM'$ with respect to the spatial predicate $\phi^{S}$ on level members as follows;

1. Spatial predicate on level values: $FM' = \{ f \in FM \mid \exists v_{l_d} \in LM(l_d), v_{a_s} \in LM(l_s) : f \rightsideset{\triangleright}{\wedge} v_{l_d} \wedge v_{a_s} \subseteq v_{l_s} \wedge \phi^{S}(v_{m}) \}$.

2. Spatial predicate on level attribute values:
$$FM' = \{ f \in FM \mid \exists v_{l_d} \in LM(l_d), v_{a_s} \in LM(l_s) \wedge v_{a_s} \rightsideset{\triangleright}{\wedge} v_{a_s} \wedge v_{l_d} \rightsideset{\triangleright}{\wedge} v_{l_d} \wedge \phi^{S}(v_{m}) \}.$$ Note that the filtering the facts through level members can be done by $v_{l_d}$ (level values) or attribute values $v_{a_s}$ by applying the spatial predicate $\phi^{S}$. Finally filtering of the facts is on associated measure values is defined in the following;

3. Spatial predicate on measure values of $m_a$: $FM' = \{ f \in FM \mid \exists v_{m_a} \in Codomain(m_a) : f \rightsideset{\triangleright}{\wedge} v_{m_a} \wedge \phi^{S}(v_{m_a}) \}$. For complex cases, i.e., combining these three types; the result set is also followed by combining the basic result sets.

Example 13. The s-dice operator can be implemented on level and attribute values by filtering level members in the cube or on measures by filtering the facts in the cube. In both cases the spatial predicate $\phi^{S}$ is used. The query for the s-dice operator could be “sales to customers, which are located within 5 km distance from their city center” where the s-dice is on level members by filtering the customer level. The spatial predicate $\phi^{S}$ can be interpreted in two different ways (See Appendix A1 for comparison of their query run times).

1. First method is assuming a buffer area of 5 km from the coordinates of city center and checking customers’ locations by within operator from topological relations $\phi^{S} \in T_{rel}$ if it meets the condition. The following SPARQL query shows the implementation of this method on level members.

```sparql
SELECT ?obs WHERE {
  ?obs rdf:type qb:Observation ;
  gnw:customerID ?cust .
  ?cust qb4o:memberOf gnw:customer ;
  skos:broader ?city ;
  FILTER (bif:st_within (?custGeo, ?cityCentGeo, 5))
}
```

2. Second method is checking if the distance from a customer location to the corresponding city center is less than 5 km, by using distance function from numeric operations $\phi^{S} \in N_{op}$. In this case the spatial predicate $\phi^{S}$ is a combination of a spatial function $\phi^{S}$ and a regular Boolean predicate $\phi$. Spatial function is distance from numeric operations and the predicate is less than ($<$). The following SPARQL query shows the implementation of this method for s-dice on level members.

```sparql
SELECT ?obs WHERE {
  ?obs rdf:type qb:Observation ;
  gnw:customerID ?cust .
  ?cust qb4o:memberOf gnw:customer ;
  skos:broader ?city ;
  BIND (bif:st_distance (?custGeo, ?cityCentGeo) AS ?distance) FILTER (?distance < 5 )
}
```

Remark 19. (Roll-up) The traditional roll-up operator aggregates measures according to a dimension hi-
erarchy (by using an aggregate function), in order to obtain measures at a coarser granularity for a given dimension. For example, the query “total amount of sales to customers by city” is a classical roll-up operation. (Cube is the sales, dimension is the customer, level in dimension to roll-up is the city such that customer ⊆ city, measure is the sales amount and aggregate function is the sum in order to calculate the total sales.)

**Definition 19. (S-Roll-up)** Similarly to roll-up operator, s-roll-up aggregates measures \( m \in M \) of a given cube \( C \), by using an aggregate function and a spatial function \( f^S \in S \) (Sect. 3.1) along a spatial dimension’s hierarchy \( h_s \) (Ext. 6), which should have spatial levels \( l_s \) (Ext. 7). However, in s-roll-up the dimension hierarchy is created dynamically on levels by the spatial function \( f^S \). We call this hierarchy a dynamic spatial hierarchy, conceptually from a base level \( l_b \) to the dynamically created target level \( l'_s \) such that \( l_b \sqsubseteq d l'_s \). The instances of the target level \( l'_s \) are obtained by the spatial function \( f^S(l_s) \) that is applied on spatial dimension levels.

As for the semantics, s-roll-up takes an \( n \)-dimensional cube \( C \) as an argument, which has the cube schema \( CS = (D, M, F) \) with the fact members \( f \in FM \) as given in Def. 16. As a parameter s-roll-up takes a spatial function \( f^S \in S \) to operate on levels \( L(d_i) \) and an aggregate function \( agg \) to calculate a measure \( m \) at the higher target level. For simplicity of explanation and without loss of generality, we initially assume that there is only one measure \( m \). The extension of the operator on several measures \( m \in M \) is explained in the last paragraph. S-Roll-up operator is formulated as: \( SRU(C)(f^S(L(d_i)), agg(m)) = C' \), which returns a cube \( C' \) with \( n \)-dimensions and has the same schema \( CS = (D', M', F') \) where \( F' = F, M' = M \), and \( D' = \{ d_i \in D \mid \{d_1, \ldots, d'_i, \ldots, d_n \} \land L(d'_i) = L(d_i) \setminus \{ l_b \sqsubseteq d_l \} \} \cup \{ l'_s \} \). After the s-rollup operation, number of dimensions in \( D \) remains the same, although the base levels and levels below the target level \( l_b \sqsubseteq d l \) of the corresponding dimension \( d_i \) are left out and a new target level \( l'_s \) is added to the set of dimension levels \( L'(d'_i) \) of \( d'_i \).

The set of level members of the level \( l'_s \) is selected with respect to the spatial function on base level members of a spatial dimension such that \( LM(l'_s) = b \{ f^S(v_{l_b}) \mid v_{l_b} \in LM(l_b) \} \) where \( l_b \sqsubseteq d l'_s \iff f^S(v_{l_b}) = v_{l'_s} \), which means that the base level \( l_b \) rolls up along the spatial dynamic hierarchy \( \sqsubseteq d \) to the target new spatial level \( l'_s \) if and only if spatial function on base level \( f^S(v_{l_b}) = v_{l'_s} \) produces the new spatial level members \( v_{l'_s} \). Even though the set of measures \( M \) remains the same, the s-roll-up operator obtains the measure values associated with fact members \( f' \) at a coarser granularity \( l'_s \), which alters the set of facts \( FM' \not\subseteq FM \). In order to create the new set of facts \( FM' \) at the new granularity level \( l'_s \), the Group operator [28] is used to group the facts characterized by the same level members \( v_{l'_s} \in LM(l'_s) \) such that \( Group(v_{l'_s}) = \{ f \in FM \mid \exists v_{l_b} \in LM(l_b) : f \rightsquigarrow v_{l_b} \land v_{l_b} \sqsubseteq d l'_s \} \). The output of the Group operator on level members is a new fact instance \( f' \). In order to aggregate the measure values \( m \), which are associated with the fact members \( f \) we use an aggregate function \( agg \) such that \( agg(f_1, \ldots, f_k) = agg(v_{m_1}, \ldots, v_{m_k}) \) where \( f_i \rightsquigarrow v_{m_i}, i = 1, \ldots, k \).

Finally, the set of the new facts \( FM' \) is constructed, that is given with the associated new level members and aggregated measure values as; \( FM' = \{ f' = Group(v_{l'_s}) \mid \exists v_{l'_s} \in LM(l'_s) : f' \rightsquigarrow v_{l'_s} \land f' \rightsquigarrow agg(Group(v_{l'_s})) \} \).

The extension to multiple measures is similar, which is done by providing and using a separate aggregate function for each measure \( m \in M \).

**Example 14.** The following SPARQL query shows the s-roll-up operator, which is exemplified in Sect. 3.6. The query is “total amount of sales to customers by city of the closest suppliers”. Note that the measures are aggregated up to a new city from customer level of the customer dimension, which is specified as the Closest City. The hierarchy step from customer to city is defined dynamically by a spatial function \( f^S \) (distance from numeric operations \( N_{op} \subseteq S \)), which is then used in a wrapper function to find the closest distance of the suppliers and customers. The levels and level members of (customer), which are below the newly defined level (Closest City) are left out in the result.

```
SELECT ?city (SUM(?sales) AS ?totalSales)
WHERE { ?obs rdf:type qb:Observation ;
  gnw:customerID ?cust ;
?sup np:supN:amount ?totalSales .}
```

```sql
# Inner Select for the total sales to
# the closest supplier of the customer
{ SELECT ?cust (MIN(?distance) AS ?minDistance) WHERE }
```
Remark 20. (Drill-down) Drill-down is the inverse operator of roll-up, which disaggregates previously summarized data to a child level in order to obtain measures at a finer granularity of a given dimension. For example, the roll-up query given in Remark 19 ("total amount of sales to customers by city") aggregates sales by summing up the sales amount, from customer level to city level along a hierarchy. As drill-down operator performs the operation opposite to the roll-up an example would be: “average amount of sales of each supplier, drilled down from the city level to the supplier level”. (Cube is the same as sales, and the hierarchy is the same but the dimension is the supplier, so child level in dimension to drill-down from city level is the supplier such that \( \text{City} \supseteq \text{Supplier} \). Conceptually, a drill-down to level \( l_s \) in a cube \( C \) corresponds to a roll-up to the same level \( l_s \) on the base cube \( \text{BaseCube}(C) \), that is denoted as \( \text{BaseCube}(C) \).

Definition 20. (S-Drill-down) Analogously to drill-down operator, s-drill-down disaggregates measures \( m \in M \) of a given cube \( C \), by using an aggregate function \( \mathbf{f}^S \) (Sect. 3.1) along a spatial dimension’s hierarchy \( h_s \) (Ext. 6), which should have spatial levels (i.e., \( l_s \)) (Ext. 7).

Conceptually, in s-drill-down, the dimension hierarchy is created dynamically on levels by the spatial function \( \mathbf{f}^S \) as in s-roll-up. This is similar to the dynamic spatial hierarchy defined in Def. 19, that is from a spatial parent level \( l_p \) to a dynamically created spatial child level \( l_c \) such that \( l_p \supseteq d l_c \). The target spatial child level \( l_c \) is the output of the spatial function \( \mathbf{f}^S \) on spatial levels \( l_s \in L(d_i) \) of the spatial dimension. Applying s-drill-down to child level \( l_c \) from a parent level \( l_p \) on a cube \( C \) corresponds to applying s-roll-up to the same level \( l_c \) from the base level \( l_b \) on the base cube of \( C \).

Example 15. In order to exemplify an s-drill-down, starting from the result cube graph of Ex. 14 ("total amount of sales to customers by city of the closest supplier"), which is at the granularity of City level, we drill down to child level Supplier with the query “average amount of sales of furthest suppliers to their city center, drilled down the from City level to Supplier level”. The following SPARQL query shows the given example.

```
SELECT ?sup (AVG(?sales) AS ?averageSales) 
WHERE { ?obs rdf:type qb:Observation ;
  gwn:supplierID ?sup ;
  gwn:salesAmount ?sales .
?sup qb4c:memberOf gwn:supplier ;
  gwn:supplierName ?supName ;
  skos:broader ?city .
?city qb4c:memberOf gwn:city ;
  gwn:cityGeo ?cityCentGeo .
BIND (bif:st_distance(?supGeo, ?cityCentGeo) AS ?distance)
FILTER (?distance = ?maxDistance)
GROUP BY ?sup
```

In this paper, we focus on direct querying of single data cubes with main SOLAP operators in SPARQL. The integration of several cubes through s-drill-across or set-oriented operations such as \( \text{union, intersection, and difference} \) [5] is out of scope and remained as future work.

6. Generating SOLAP queries in SPARQL via QB4SOLAP

After having defined the high-level SOLAP operators in Sect. 5, this section first describes how to generate SPARQL queries for each of these operators by using the QB4SOLAP metamodel (Sect. 4). Afterwards, this section describes how to create more complex SPARQL queries for nested SOLAP operations.

6.1. Generation algorithms

The generated SPARQL queries \( Q \) are of the form  \( Q = \text{SELECT } R \text{ WHERE } GP \) \), where \( GP \) is a graph pattern containing triple patterns and \( R \) is the (set of) variable(s) that are returned in the result of the query. Triple patterns are based on triples of the form \( (s, p, o) \) (Def. 4), where triple components are replaced by variables. A set of triple patterns defines a graph pattern \( GP \). Given an RDF graph \( G \), a graph pattern \( GP \) is used to search for subgraphs \( \mathcal{G}_H \subseteq G \) matching the pattern. In our algorithm, the graph pattern is...
initially empty, $GP = \emptyset$, and the triple patterns are added incrementally to the body of the WHERE clause: $GP = GP \cup (s p o)$.

RDF datasets published with the QB4SOLAP vocabulary use the skos:broad property to define the roll-up relation from child level to parent level (Defs. 13 and 15). As this is the case for all hierarchy levels in a dimension, every SOLAP query contains such roll-up paths that we need to consider as part of $GP$ in the WHERE clause.

Thus, we define a helper function $RUPath$ (Algorithm 1) that we can use in the SOLAP query generation algorithms.

Algorithm 1: $RUPath(G^S_s(l), l_b, l_s, a_{ID}, ?a_s, ?f)$

Input: $G^S_s(l), l_b, l_s, a_{ID}, ?a_s, ?f$
Output: $GP$

1 begin
2 $GP = (\text{if rdf:type qb:Observation})$
3 $GP = GP \cup (\text{if id}^S(a_{ID}) \land ?l_b \land ?l_b$qb4o:memberOf$id^S(l_b))$
4 foreach $(id^S(l_c), id^S(l_p)) \in G^S_s(l) \mid l_p \subseteq l_s$ do
5     $GP = GP \cup (\exists l_p \text{skos:broader} ?l_p \land$
6     $\exists l_p \text{skos:broader} ?a_s)$
7 return $GP$

Build roll-up path ($RUPath$). The helper function $RUPath$ returns a graph pattern that we can use in the body of the WHERE clause. The roll-up path pattern is created as a path-shaped join of triples with partial order ($\subseteq$, skos:broad) (Defs. 10 and 15). The triple pattern is of the form

$\{(s_1 p_1 o_1), (o_2 p_2 o_2), (o_3 p_3 o_3), \ldots (o_{n-1} p_{n-1} o_n)\}$

where $s_1$ is the root of the graph pattern and corresponds to fact members $f$ from the QB4SOLAP schema (Def. 16), $p_1$ is the predicate $id^S(a_{ID})$ that associates facts with level members $v_{l_1}$ ($f \rightsquigarrow v_{l_1}$, Def. 16), $o_1$ is the variable for the first level member that rolls up to its parent level $o_2$ such that $o_1 \subseteq o_2$ and so on, and the $p_2$ predicate corresponds to the skos:broad property. The last variable in the path $o_n$ corresponds to the target level $l_s$ in order to represent the level member attributes at the target level. The roll-up path starts at the fact instances $f$ (Def. 16). Afterwards, the partial order on level members (Def. 15) from base level $l_b$ to target level $l_s$ is applied. Algorithm 1 sketches the helper function for building the roll-up path for dimensions; from facts to dimension levels with predicates and cube member IRIs defined in the cube schema.

In order to represent such varying parameters at the instance level such as fact members, level members, or parameter values given by the user, and to distinguish these parameters from other parameters in the algorithm, we represent such parameters using variable names with question marks.

We use a FILTER expression to restrict the output data by using a (spatial) Boolean predicate $\phi^S$. A FILTER expression is part of the body of a graph pattern in a SPARQL query. Therefore, it is added to the body of the WHERE clause in the graph pattern $GP$ as $GP = GP \cup (\text{FILTER } \phi^S)$. In the cases where there is a spatial function $f^S(x)$ in the SOLAP operator, it is given in the BIND clause, which is technically a part of the WHERE clause and therefore added to the body of the WHERE clause in a graph pattern $GP$ as $GP = GP \cup (\text{BIND } f^S(x))$. SPARQL 1.1 defines aggregate expressions\(^{15}\), such as SUM, MIN, MAX, AVG, etc.

We apply them on measure values or use them as wrappers in spatial functions. In the following, we often write AGG to represent them.

In the following, we present the SPARQL query generation algorithms for the SOLAP operators defined in Sect. 5. The algorithms take the input parameters and arguments of the SOLAP operator and return a SPARQL query $Q$ that can be executed.

S-Slice generator. To generate a SPARQL query for the s-slice operator $SS_s(l_b, l_s, v_s)$ (Def. 17), we use Algorithm 2. Parameter $v_s$ is a spatial literal value $v_s \in \mathcal{L}_s$ (i.e., POINT or POLYGON) that should be related to a spatial level $l_s$ (Ext. 7). This means that $v_s$ is defined as a polygon geometry that corresponds to a spatial attribute value in the target level $l_s$ (Ex. 11) or $v_s$ is defined as a point geometry that is spatially contained in a spatial attribute value of the target level $l_s$ (Ex. 12). Note that in Ex. 11.1, the given spatial literal has the geometry data type polygon, which corresponds to a spatial level attribute $a_s$ (Ext. 8) at a spatial level $l_s$. Similarly, the spatial function call $f^S(x)$ in Ex. 11.2 returns a polygon that corresponds to a spatial level attribute $a_s$.

\(^{15}\)https://www.w3.org/TR/sparql11-query/#aggregates
Algorithm 2: S-SliceGenerator \( (G^I, v_s, l_b, l_s) \)

\[ Algorithm 2: S-SliceGenerator \( (G^I, v_s, l_b, l_s) \) \]

1. **Input**: \( G^I \), \( v_s \), \( l_b \), \( l_s \)
2. **Output**: \( Q \)

```plaintext
begin
    Q = ∅; GP = RUPath \( (G^I, l_b, l_s, a_{ID}, \text{?as}, \text{?f}) \)
    if \( v_s \) is a POINT then
        GP = GP \( ∪ \) (FILTER \( (\text{st\_within} v_s, \text{?as}) \))
    else if \( v_s = f^S(x) \) then
        Q′ = ∅;
        GP′ = RUPath \( (G^S, l_b, l_s, a_{ID}, \text{?as}, \text{?f}) \)
        GP′ = GP \( ∪ \) (BIND \( f^S(x) \) AS \( ?v_s \))
        Q′ = SELECT ?x AGG(?y) WHERE GP′
        GP = GP \( ∪ \) Q′ \( ∪ \) (FILTER \( ?x = ?as \))
    else
        GP = GP \( ∪ \) (FILTER \( v_s = ?as \))
return Q = SELECT ?f WHERE GP
```

On the other hand, given the spatial literal in Ex. 12, the geometry data type point, which corresponds to the spatial level \( l_s \) via topological relations \( (T_{rel}) \). We consider all these possibilities in the s-slice generator algorithm. We explained these in the following, where the steps are referencing the line numbers in Algorithm 2.

**Line 2.** Get the path for dimension \( d_s \) (e.g., Customer) from the observation facts \( f \) to the base level \( l_b \), and build path-shaped triple pattern paths from the dimension’s base level \( l_b \) to the target spatial level \( l_s \) (e.g., City level). Finally, get level attributes IRIs and variables for the spatial attribute \( a_s \) (e.g., City geometry). All this is done by the RUPath function (Algorithm 1) that is used by the s-slice generator. The following shows an example result of this step that is added to \( GP \):

```sql
{?obs rdf:type qb:Observation ;
  gwn:customerID ?cust .
?cust qb4o:memberOf gwn:customer ;
skos:broader ?city .
}
```

**Line 3.** Check if the spatial literal \( v_s \) is a point geometry type. If true, create a FILTER statement with a spatial Boolean predicate (Line 4) and go to the result (Line 12).

**Line 4.** Build the FILTER statement based on the spatial literal \( v_s \) and the spatial attribute \( a_s \) (Ex. 12). As a result the following lines might be added to the \( GP \):

```sql
FILTER (bif:st\_within("POINT(10.43951 55.47006)"), ?cityGeo))
```

**Line 5.** Check if \( v_s \) is a function call \( f^S(x) \). If true (Ex. 11.2), construct an inner select query to compute the spatial function \( f^S(x) \), then go to the result (Line 12).

**Line 6.** Call the RUPath function in order to link \( a_s \) variables with the fact instances (this time for inner select query \( Q' \)). This step creates a graph pattern \( GP' \) for inner select query \( Q' \), for example:

```sql
{?obs rdf:type qb:Observation ;
gwn:customerID ?cust .
?cust qb4o:memberOf gwn:customer ;
skos:broader ?city .
}
```

**Line 7.** Build a bind statement on \( a_s \) variables for calculating spatial functions (e.g., compute areas). For example, the following lines might be added to graph pattern \( GP' \):

```sql
BIND (bif:st\_area(?x) as ?area) 
```

**Line 8.** Generate the inner select query \( Q' \) based on \( GP' \) generated in Lines 6 and 7. For example \( (Q' \) finds the geometry of the largest city):

```sql
Q' = (SELECT ?x (MAX(?area) as ?maxArea)
WHERE (?obs rdf:type qb:Observation ;
  gwn:customerID ?cust .
?cust qb4o:memberOf gwn:customer ;
skos:broader ?city .
BIF (bif:st\_area(?x) as ?area) )
```

**Line 9.** Build the filter statement with the output of the spatial function \( f^S(x) \), construct \( GP \) (includes \( Q' \)) for the outer query, and go to the result (Line 12). At this stage \( GP \) is constructed in Lines 2 and 8. The following for the filter statement is added to the \( GP \):

```sql
FILTER (?cityGeo = ?x )
```

**Line 11.** If a spatial literal \( v_s \in L \) is given as the parameter instead, build a filter statement that checks if \( v_s \) is equal to the spatial attribute \( a_s \) values, and go to the result (Line 12). For example, the following filter condition might be added to graph pattern \( GP \):

```sql
FILTER (?cityGeo = ?x )
```
FILTER (?cityGeo = "POLYGON((10.43951 55.47006, 10.43947 55.47003, 10.43924 (...))")

Line 12. Finally, the algorithm generates query \( Q \), which can be executed over the fact members \( FM' \). In our running examples we obtain the following cases for the generated s-slice query \( Q \).

**S-Slice operator with a given spatial value as point data type:** The following listing corresponds to the SPARQL output of the running example where the spatial value is given as \( \text{POINT} \) data type (Ex. 12) and filters the level attributes with a spatial predicate within a given level. The graph pattern \( GP \) for the query is created in Lines 2 to 7.

1. \( Q = \text{SELECT} \ ?\text{obs WHERE} \)
2. \{ ?obs rdf:type qb:Observation ;
4. ?cust qb4o:memberOf gwn:customer ;
5. skos:broader ?city .
7. FILTER (bif:st_within("POINT((10.43951 55.47006, 10.43947 55.47003, 10.43924 (...))")})

**S-Slice operator with a spatial function call:** The following listing corresponds to the SPARQL output of the running example where the spatial value is returned from a function call (Ex. 11.2). The graph pattern \( GP' \) for the spatial function call is created in Lines 8 to 13. The graph pattern \( GP \) for the whole query is created in Lines 2 to 12.

1. \( Q = \text{SELECT} \ ?\text{obs WHERE} \)
2. \{ ?obs rdf:type qb:Observation ;
4. ?cust qb4o:memberOf gwn:customer ;
5. skos:broader ?city .
7. FILTER (bif:st_within("POINT((10.43951 55.47006)*, ?cityGeo))")

**S-Dice generator.** To generate a SPARQL query for the s-dice operator, \( SD(C)[\phi^S] \) (Def. 18 – parameter \( \phi^S \) represents a spatial predicate), we follow the steps sketched in Algorithm 3. The algorithm takes parameter \( \phi^S \) as input, which corresponds to a spatial predicate that could represent a topological relation from the \( T_{rel} \) set or a combination of a spatial function (a numeric operation from the \( N_{op} \) set) and a regular predicate \( \phi \). For illustration, we use the example query that we have introduced in Sect. 5 for s-dice (Ex. 13):

"sales to customers, which are located 5 km distance from their city center". In the following, we discuss the main steps of Algorithm 3 with the running example, where the steps are referencing the line numbers in Algorithm 3.

Line 3. The algorithm runs through the levels from base level \( l_b \) to the target spatial level \( l_s \), which are both given in the spatial Boolean predicate \( \phi^S \).

Line 4. Build the roll-up path for those levels using the helper function \( RUP\text{Path} \). Note that, when we apply the roll-up path to the target level, we can also link the level attributes for the target (spatial) level – as, for example, in the last line of the following listing. The output of function \( RUP\text{Path} \) is added to graph pattern \( GP \):

1. \( Q = \text{SELECT} \ ?\text{obs WHERE} \)
2. \{ ?obs rdf:type qb:Observation ;
4. ?cust qb4o:memberOf gwn:customer ;
5. skos:broader ?city .
7. FILTER (bif:st_within("POLYGON((10.43951 55.47006, 10.43947 55.47003, 10.43924 (...))")})

Line 5. Check if \( \phi^S \) is to be implemented as a spatial predicate from topological relations \( T_{rel} \) as interpreted in Ex. 13.1.

Line 6. Create a filter statement with a spatial predicate and the spatial level attribute \( a_s \), which is referenced in the roll-up path (Line 4). For our running example, the filter statement is applied on customers that are located within a buffer area of 5 km from their city centers. The spatial predicate \( st\_within \) is used from the topological relations. The following lines are added to graph pattern \( GP \):
FILTER (bif:st_within(?custGeo, ?cityCentGeo, 5))

Line 7. Check if $\phi^S$ is to be implemented as a combination of a spatial function $f^S(x)$ and a regular predicate $\phi$ as interpreted in Ex. 13.2.

Algorithm 3: $S$–DiceGenerator $(G^I_G, \phi^S) : Q$

Input: $G^I_G, \phi^S$
Output: $Q$

1 begin
2 \[ Q = \emptyset; GP = \emptyset \]
3 for $l_b, l_s \in \phi^S$ do
4 \[ GP = GP \cup RUPath (G^I_G, l_b, l_s, a_{ID}, ?a_x, ?f) \]
5 if $\phi^S$ is a spatial predicate then
6 \[ GP = GP \cup (\text{FILTER } \phi^S(?a_x)) \]
7 else if $\phi^S$ uses a spatial function $f^S(x)$ and a regular Boolean predicate $\phi$ then
8 \[ GP = GP \cup (\text{BIND } f^S(x) \text{ AS } ?v_x) \]
9 \[ GP = GP \cup (\text{FILTER } \phi(?v_x)) \]
10 return $Q = \text{SELECT } ?f \text{ WHERE } GP$

Lines 8, 9. Create a bind statement based on a spatial function (i.e., calculate st_distance between customers and city center) and a filter statement based on the assigned values with a regular predicate (i.e., less than 5 km). The following lines are added to graph pattern $GP$:

\[ \text{BIND (bif:st_distance (?custGeo, ?cityCentGeo) AS ?distance) FILTER ( ?distance < 5 )} \]

Line 10. Generate query $Q$ for selecting the facts $f \in FM'$ matching the incrementally created graph pattern $GP$ in the previous steps. In our running examples we obtain the following cases for the generated s-dice query $Q$.

$S$-Dice operator with $\phi^S$: The following listing is the SPARQL query generated for the running example (Ex. 13.1), where the spatial predicate is interpreted as a topological relation. The graph pattern $GP$ for the query is created in Lines 2 to 8.

\[ \text{FILTER (bif:st_within(?cityCentGeo, ?custGeo, 5))} \]

$S$-Dice operator with $f^S(x)$ and a regular Boolean predicate $\phi$: The following listing is the SPARQL query generated for the running example (Ex. 13.2), where the spatial predicate is interpreted as a combination of a spatial function and a regular predicate. The graph pattern $GP$ for the query is created in Lines 2 to 9.

\[ \text{FILTER (bif:st_within(?cityCentGeo, ?custGeo, 5))} \]

S-Roll-up Generator. To generate a SPARQL query for the s-roll-up operator from a high-level SOLAP expression, $SRU(C)[f^S(L(d_i)), agg(m)]$ (Def. 19), where parameter $f^S(L(d_i))$ denotes a spatial function on spatial level members and $agg(m)$ is an aggregate function on measures. For illustration, we use the query example for s-roll-up given in Sect. 5 for s-roll-up (Ex. 14): “total amount of sales to customers by city of the closest suppliers”. We follow the main steps sketched in Algorithm 4 in the following.

Lines 2, 3. Build the roll-up path using helper function $RUPath$. In addition to the variables given in the $RUPath$ function, we also need to consider measures and measure value variables (Line 3) since we aggregate the measures. A measure is specified in the following listing of the running example as gnw:salesAmount. The following lines are added to the graph pattern $GP$:

\[ \text{FILTER ( ?distance < 5 )} \]

Line 4. Build inner select subquery to apply the spatial function $f^S$ on the spatial level members $L(d_i)$ (i.e., Customer, Supplier). In the example,
we will use this information to create a dynamic spatial hierarchy from the Customer to the City level.

**Algorithm 4: SRUGenerator**(G_C^(d)\(, f^S(L(d)), \text{agg}(m))): Q

Input: G_C^(d), f^S(L(d)), \text{agg}(m)

Output: Q

begin

1. \(Q = \emptyset; GP = \emptyset\)
2. \(GP = \emptyset\)
3. \(GP = GP \cup (\text{if id}^S(m) ?m)\)
4. for \(f^S(L(d_i))\) do
   5. \(GP' = \text{RUPath}(G_C^(d), l_b, l_s, a_{ID}, ?a_s, ?f)\)
   6. \(Q' = \emptyset\)
   7. \(Q' = \text{SELECT} ?x (\text{AGG}(?y_s) AS ?y_g)\)
      \(\text{WHERE} GP' \text{ GROUP BY} ?x\)
   8. \(GP = GP \cup Q' \cup (\text{FILTER} ?x = ?a_s \&\& f^S(x) = ?v_y)\)
   9. let \(l_s = l_s'\)
10. return \(Q = \text{SELECT} ?f ?l' _s \text{agg}(?m) \text{WHERE} GP\)
      \(\text{GROUP BY} ?f ?l'_s\)

Line 5. Call RUPath for the inner select subquery to link the geometry attributes of base level members with different variables and create a graph pattern \(GP'\) for the inner select. The following lines are added to the graph pattern \(GP'\):

\[
\{?obs \text{ rdf:type qb:Observation} ;
\text{gnw:customerID} ?cust1 ;
\text{gnw:supplierID} ?sup1 .
?sup1 \text{ gnw:supplierGeo} ?sup1Geo .
?cust1 \text{ gnw:customerGeo} ?cust1Geo .
\]

Line 6. Build the bind statement in order to calculate the spatial function \(f^S(L(d_i))\) on spatial level members. For the running example the spatial function is \(\text{st\_distance}\). The following lines are added to the graph pattern \(GP'\):

\[
\text{BIND (bif:st\_distance(?cust1Geo, ?sup1Geo) AS ?distance)}
\]

Line 7. Generate the inner select query \(Q'\) using graph pattern \(GP'\) (Lines 5 and 6). Select the corresponding level members (Customer level for the running example) and group them in a group by statement on the selected level members. Note that this is where the spatial function \(f^S(L(d_i))\) is called with a wrapper expression (e.g., \(\text{MIN}, \text{MAX}, \text{etc.}\)) to find the closest distance. The following lines illustrate the inner select query \(Q'\):

\[
Q' = (\text{SELECT} ?cust1 (\text{MIN}(?distance) AS ?minDistance)) \text{WHERE} GP'
\]

Lines 8, 9. Build the filter statement for the whole query based on the output of the spatial function, which is calculated in the inner select subquery. Then, add the filter and inner select subquery to the main graph pattern \(GP'\) (Line 8). The filter statement for the running example is:

\[
\text{FILTER} (?cust = ?cust1 \&\& \text{bif:st\_distance} (?custGeo, ?supGeo) = ?minDistance)
\]

Note that in Line 9, the spatial target level \(l_s\) (City) is altered to a dynamic spatial level \(l'_s\) since applying the spatial function creates a dynamic hierarchy.

Line 10. Generate query \(Q\) for computing the facts \(f \in FM'\) based on graph pattern \(GP\) created in the previous steps. The measures are aggregated at the spatial target level (closest City, which is dynamically selected). The group by statement is applied on the fact members and target level members. In our running example we obtain the following case for the generated s-roll-up query \(Q\).

**S-Roll-up operator**: The following listing shows the generated SPARQL query. Graph pattern \(GP'\) for the inner select subquery is created in Lines 15 to 22 and the graph pattern \(GP\) for the whole query is created in Lines 3 to 24.

\[
Q = \text{SELECT} ?city \text{ (SUM(?sales) AS ?totalSales) WHERE}
\]
\[
\{ ?obs \text{ rdf:type qb:Observation} ;
\text{gnw:customerID} ?cust ;
\text{gnw:supplierID} ?sup ;
\text{gnw:salesAmount} ?sales .
?sup \text{ gnw:supplierGeo} ?supGeo .
?cust \text{ gnw:customerGeo} ?custGeo .
\}
\]

\[
\text{GROUP BY} ?city \text{ WHERE}
\]
\[
\{ ?obs \text{ rdf:type qb:Observation} ;
\text{gnw:customerID} ?cust1 ;
\} \text{ GROUP BY} ?cust1 \text{ (MIN(?distance) AS ?minDistance) WHERE}
\]
\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]

\[
\}
\]
GROUP BY ?city ?obs

(?custGeo, ?supGeo) = ?minDistance)}

FILTER (?cust = ?cust1 && bif:st_distance

AS ?distance)}

19

20

21

22

23

24

25

GROUP BY ?city ?obs

S-Drill-down Generator. The semantics of the s-drill-
down operator are defined in the same way as for
the s-roll-up operator with the condition that the input
cube C for s-roll-up is obtained using a function
BaseCube such that SDD(C)[f^S(L(d_i)), agg(m)] =
SRU(BaseCube(C))[f^S(L(d_i)), agg(m)] (Def. 20).
Therefore, no generator algorithm and steps are spe-
cified since an s-drill-down operator corresponds to a
rewriting of an s-roll-up operator, which is ob-
tained with a Base function that calls the base cube
graph in SRUGenerator as; SDDGenerator =
SRUGenerator (Base(G^c_{(C)}), f^S(L(d_i)), agg(m)).

6.2. Nested SOLAP operations to SPARQL

We now show how a SPARQL query can be gen-
erated for a nested SOLAP expression. In general, a
nested set of SOLAP operators can be rewritten into
an expression with an additional s-dice, on top of a
series of s-roll-ups, on top of one or more s-slices,
on top of an s-dice, i.e., (s-dice1_2(s-roll-up1, . . . s-roll-
upk(s-slice1, . . . s-slice3(s-dice1(C)))))

Let us begin with a simpler nested form that shows
the most typical pattern, namely (s-roll-up (s-slice
(s-dice(C)))))), where initially a subcube graph is selected
by s-dice. Afterwards, an s-slice is performed on a
higher level of a dimension. Then, an s-roll-up is ap-
plied, which aggregates the measures in the sliced cube
from a lower level to a higher level. Finally, we could
also perform another s-dice for filtering the measures.
There may be several s-slices and s-roll-ups in be-
tween.

We formulate the nested SOLAP query as
\(3(s\text{-}roll-up \ 2(s\text{-}slice \ 1(s\text{-}dice(C))))\) and apply our running ex-
amples such that the enumeration of operators can be
interpreted as follows: \(^1\) Get the subcube graph of cus-
tomers that are located within a 5 km distance from
their city center, \(^2\) slice on the customers of the largest
country, (which drops the dimension and leaves out
all the other countries) and \(^3\) get the total amount of
sales for customers by the city of their closest suppliers
( aggregates the measure Sales amount from Customer
to Closest City level). Finally, we may also perform
another (s-)dice on measures, e.g., filtering the total
amount of sales greater than 10500. To perform nested
SOLAP operators, we identify a set of principles to be
considered by the algorithm.

**Principle 1:** Perform s-dice in the beginning or at
the end.

**Principle 2:** If there are several s-roll-up or s-slice
operations call their generator algorithms repeat-
edly.

**Principle 3:** Always separate FILTER clauses when a
SOLAP generator algorithm is used. Enu-
erate separated FILTER clauses. If a SOLAP
operator is the final function added to the graph
pattern, do not separate the FILTER clause.

**Principle 4:** Build the final graph pattern with the
separated and enumerated FILTER clauses with
respect to Principle 3.

**Principle 5:** Drop the main SELECT clause from
each SOLAP generator algorithms and build only
one SELECT that is added to the query at the end.

**Principle 6:** Separate the GROUP BY clause and
AGG functions from the s-roll-up generator algo-
rithms (and enumerate them), and build add them
to the main (outer) SELECT clause at the end.

---

**Algorithm 5:** WriteSPARQL((SRU(C)[f^S(L(d_i)),

agg(m)] (SS(C)[b, l_1, v_a,(SD(C)[\phi^S]])]:Q

Input: (SRU(C)[f^S(L(d_i)), agg(m)](SS(C)[b, l_1, v_a,(SD(C)[\phi^S]])

Output: Q

begin

1 begin

2 \(Q = \emptyset ; GP = \)

RUPath \((G^S_{(C)}, l_1, l_s, a_{ID}, ?a_s, ?f)\)

3 \(GP = GP \cup (f \text{id}^S(m) \ ?m)\)

4 \(GP^1 = S\text{-}DiceGenerator(G^I_{(C)}, \phi^S)\)

\(<\text{FILTER}^{1} \ \text{\{SELECT}}\)

5 \(GP = GP \cup GP^1\)

6 \(GP^2 = S\text{-}SliceGenerator(G^I_{(C)}, v_s, l_1, l_s)\)

\(<\text{FILTER}^{2} \ \text{\{SELECT}}\)

7 \(GP = GP \cup GP^2\)

8 \(GP = GP \cup \text{FILTER}^{1} \cup \text{FILTER}^{2} \cup\)

9 \(SRUGenerator(G^I_{(C)}, f^S(L(d_i)), agg(m))\ \)\)

SELECT \ GROUP BY \ AGG^1 \ \text{\{WHERE}} GP

10 return \(Q = SELECT ?m \ AGG^1(?m) \ \text{\{WHERE}} GP

GROUP BY ?l_s \ \text{\{WHERE}} GP\)
To separate the FILTER clauses, we call SOLAP generator algorithms without their FILTER clause and enumerate each FILTER clause for each SOLAP generator algorithm that is used, i.e., S-SliceGenerator \((G^1 \cup \phi^S) \setminus \text{FILTER}^1\) (Algorithm 5, Line 4). Then, we build the final graph pattern with these separated FILTER clauses i.e., \(GP = GP \cup \text{FILTER}^1 \cup \text{FILTER}^2\) (Line 8). When the last SOLAP generator algorithm is called, the output is directly added to the graph pattern without separating its FILTER clause (Line 9). Throughout the algorithm, all the SELECT clauses are omitted and combined into one SELECT in the output on Line 10. According to Principle 6, if there are any GROUP BY clauses and AGG functions (on measures) in inner selects, we eliminate them with "\(\setminus\)" from the inner selects (Line 9) and finally build the main (outer) select query with (Line 10). Note that in the algorithm, the general graph pattern \(GP\) is initially created by the RUPATH function (Line 2) and incremented with triple patterns for selected measures (Line 3).

**Example 16.** ((\(3^s\)-roll-up (\(2^s\)-slice (\(1^s\)-dice(C)))))

1. Get the subcube graph of customer that are located within a 5 km distance from their city center, slice on the customers of the largest country, (which drops the dimension and leave out all the other countries) and get the total amount of sales of customers by the city of their closest suppliers (aggregates the measure Sales amount from Customer to Closest City level).

2. The query is written starting from the innermost operator S-dice to the outermost operator S-roll-up.

```sparql
SELECT ?city (SUM(?sales) AS ?totalSales)
WHERE {
  ?obs rdf:type qb:Observation ;
  gwn:customerID ?cust ;
  gwn:supplierID ?sup ;
  gwn:salesAmount ?sales .
  ?cust g4o:memberOf gwn:customer ;
  gwn:customerGeo ?custGeo ;
  skos:broader ?city .
  ?sup g4o:memberOf gwn:supplier ;
  gwn:supplierGeo ?supGeo ;
  skos:broader ?city .
  ?city g4o:memberOf gwn:city ;
  gwn:cityGeo ?cityCentGeo ;
  skos:broader ?country .
  ?country g4o:memberOf gwn:country ;
  gwn:countryGeo ?countryGeo .
  WHERE {
    ?obs rdf:type qb:Observation ;
    gwn:customerID ?cust .
    ?cust g4o:memberOf gwn:customer ;
    skos:broader ?city .
    ?country gwn:countryGeo ?x .
    FILTER (bif:st_area(?x) as ?area) ;
    BIND(?obs rdf:type qb:Observation ;
    gwn:customerID ?cust1 ;
    gwn:supplierID ?sup1 .
    ?sup1 gwn:supplierGeo ?sup1Geo .
    BIND(bif:st_distance(?cust1Geo, ?sup1Geo)
    AS ?distance) ;
    GROUP BY ?city1 }
  }
  FILTER (?cust = ?cust1 && bif:st_distance
  (?custGeo, ?supGeo) = ?minDistance)}
  GROUP BY ?city
```

The graph pattern \(GP\) is initially created with the RUPATH function for the corresponding levels and level attributes (Lines 3 to 18 of the generated SPARQL query in the listing above). The first operator is called by function S-DiceGenerator, where the first FILTER clause of the outer selected is added to the query (Line 37). The second operator is called by the S-SliceGenerator function excluding its FILTER clause (Lines 19 to 27), which is followed by the SRUAGGR function without GROUP BY and AGG statements (Lines 28 to 36). Note that in Line 36, GROUP BY is applied on the lower level Customer, and the actual GROUP BY for the target City level is applied in the last line (Line 40). Separated FILTER clauses for the S-DiceGenerator and S-SliceGenerator functions are later added to the graph pattern (Algorithm 5, Line 8) corresponding to Lines 37 and 38 in the above example. The main outer select query is defined in the first line by specifying the target level (City) and aggregate function on measures (sum of the total Sales).

7. Conclusions and future work

Motivated by the need for a formal foundation for spatial data warehouses on the Semantic Web, this pa-
per made a number of contributions. First, it proposed the QB4SOLAP vocabulary (metamodel), which supports spatially enhanced multidimensional (MD) data cubes over RDF data. This allows users to publish MD spatial data in RDF format. Second, the paper defines a number of spatial OLAP (SOLAP) operators over the defined QB4SOLAP cubes, allowing spatial analytical queries over RDF data, and gives their formal semantics. Third, the paper provides algorithms for generating spatially extended SPARQL queries from individual and nested SOLAP operators, allowing users to write their spatial analytical queries in our high-level SOLAP language instead of the lower-level and more complex SPARQL. Fourth, the vocabulary, operators, and query generation algorithms are validated by applying them to a realistic use case.

Fig. 8 presents our future vision of SOLAP on the Semantic Web with regards to the current state of the art and ongoing work. In order to verify the algorithms, which are defined in this paper, we developed GeoSemOLAP [18], a SPARQL query generation tool. GeoSemOLAP allows users to perform SOLAP operations and generate SPARQL queries interactively through a GUI. We refer to our screencast\(^{16}\) for a detailed demonstration of GeoSemOLAP.

Publishing spatial DWs on the SW allows users to also exploit the existing external geo-vocabularies (e.g., GeoNames, etc.)\(^{17}\) by defining spatial levels and hierarchies from external open data sources. Our main ongoing work thus focuses on developing an RDF2SOLAP enrichment module that performs the multidimensional annotation of existing spatial RDF datasets with QB4SOLAP in a (semi-)automatic fashion.

Additional interesting aspects of future work would be, for instance, extending the formal techniques and algorithms for generating SOLAP queries in SPARQL to work over multiple RDF cubes, i.e., to support s-drill-across, and supporting spatial aggregation (s-aggregation) with user-defined functions over spatial measures. It would be also interesting to increase efficiency by extending our spatial data warehouse with techniques that been developed in the context of RDF data cubes and SPARQL analytical queries in general, e.g., materialization and optimizing the physical layout [13,21,22], and to enable efficient support of a broad range of external sources by considering aspects such as federated processing of analytical queries [20] and schema heterogeneity [35]. We will also consider more efficient representations of the data, e.g., by removing redundancies. Furthermore, it would be interesting to extend QB4SOLAP and GeoSemOLAP [18] to handle highly dynamic spatio-temporal data and queries, as for instance, found in large-scale transport analytics [12].

Acknowledgment. This research is partially funded by the European Commission through the Erasmus Mundus Joint Doctorate Information Technologies for Business Intelligence (EM IT4BI-DC) and the Danish Council for Independent Research (DFF) under grant agreement no. DFF-4093-00301.

References


[28] T. B. Pedersen, C. S. Jensen, and C. E. Dyreson. A Foundation...


Appendix

A1. Query Run Times

Table 5 presents the query runtimes for the SOLAP operator examples (Ex. 12, Ex. 13.1, Ex. 13.2, Ex. 14, and Ex. 15) given in Sect. 5 and a nested SOLAP operator example (Ex. 16) given in Sect. 6.2. The SOLAP queries are tested against an instance of the use case dataset (GeoNorthwind) that we have discussed in Sect. 4. In total, 48677 triples are obtained from the GeoNorthwind dataset. The published RDF graph instance is denoted as C in Table 5. We used Virtuoso Open Source Edition (Column Store and multi threaded) Version 7.2.5 on an Ubuntu 14.04 server with 2.30GHz CPU and 16 GB RAM to publish the triples at the SPARQL endpoint http://lod.cs.aau.dk:8890/sparql. The actual queries used in the paper are available at http://extbi.cs.aau.dk/SOLAP4SW/queries.

<table>
<thead>
<tr>
<th>SOLAP Operators</th>
<th>Query Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ex. 12 (s-slice(C))</td>
<td>0.07</td>
</tr>
<tr>
<td>Ex. 13.1 (s-dice(C))</td>
<td>0.09</td>
</tr>
<tr>
<td>Ex. 13.2 (s-dice(C))</td>
<td>1.01</td>
</tr>
<tr>
<td>Ex. 14 (s-roll-up(C))</td>
<td>2.03</td>
</tr>
<tr>
<td>Ex. 15 (s-drill-down(C))</td>
<td>1.86</td>
</tr>
<tr>
<td>Ex. 16 (s-roll-up (s-slice (s-dice(C))))</td>
<td>3.04</td>
</tr>
</tbody>
</table>

The query runtime of the s-slice query in Ex 12 is measured as 0.07 seconds, which is an efficient SOLAP operator to execute, since filtering the given spatial literal is done only for the records of the specified level instances.

The query runtimes of the s-dice queries in Ex. 13.1 and Ex. 13.2 are measured as 0.09 and 1.01 seconds respectively, for the two alternative ways of implementing s-dice in SPARQL. It is clear that the first use of the s-dice operator is more efficient compared to the second, even though both return the same results. The first one performs s-dice by filtering the instances of the customers through their geometries that are within a buffer area of a circle with 5 km radius and the city center geometry as the center point of the circle. The second one performs s-dice by measuring the distances between each customer’s location and their city center, and then applies a filter of those measurements to find the ones that are less than 5 km. This is a more expensive approach due to measuring the distances between each customer and city instances with a spatial distance function.

The query runtime of the s-roll-up query in Ex. 14 is measured as 2.03 seconds. The s-roll-up operator has a longer response time than the other operators as it combines a spatial distance function with aggregate operators. Due to its complexity, s-roll-up requires an inner select where it binds the distances of specified spatial level attributes (e.g., customer and supplier geometry), which is calculated with a spatial distance function. Those distance measurements are then
filtered in the outer select with respect to the aggregate function.

The query runtime of the s-drill-down query in Ex. 15 is measured as 1.86 seconds. S-drill-down operates in a very similar manner as s-roll-up. Theoretically, s-drill-down operates on an already spatially rolled up data cube, whilst it is implemented in practice as an s-roll-up on the base cube. The RDF data is at the lowest granularity, which is equal to the definition of the base cube. The SPARQL query of the s-drill-down operator is very similar to the s-roll-up operator. The difference of the query runtime between Ex. 14 and Ex. 15 is directly related to the number of instances of the specified levels and the complexity of the spatial functions.

The query runtime for the nested SOLAP query in Ex. 16 is measured as 3.04 seconds. The nesting in the query is the main reason why it has a longer response time. Moreover, the nested query has more triple patterns and clauses that need to be evaluated during query execution. Hence, it takes longer than evaluating a single SOLAP operator.

A2. Table of Contents

Table 6 summarizes a list of all the definitions, extensions, remarks, and algorithms. Definitions are enumerated starting from "1". Extensions are enumerated starting from "5" to be consistent with the base definition of the extension, e.g., Def. 5 (Dimensions) is followed by Ext. 5 (Spatial dimensions). Remarks are enumerated starting from "17" for OLAP operators, which are given before the definitions of the corresponding SOLAP operator, e.g., Remark 17 (Slice) is followed by Def. 17 (S-Slice). Algorithms are enumerated starting from "1".

<table>
<thead>
<tr>
<th>Type</th>
<th>No.</th>
<th>Topic</th>
<th>Sect.</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>1</td>
<td>Spatial aggregation</td>
<td>3.2</td>
<td>4</td>
</tr>
<tr>
<td>Definition</td>
<td>2</td>
<td>Topological relations</td>
<td>3.2</td>
<td>4</td>
</tr>
<tr>
<td>Definition</td>
<td>3</td>
<td>Numeric operations</td>
<td>3.2</td>
<td>4</td>
</tr>
<tr>
<td>Definition</td>
<td>4</td>
<td>RDF triple</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Definition</td>
<td>5</td>
<td>Dimensions</td>
<td>4.1</td>
<td>8-9</td>
</tr>
<tr>
<td>Extension</td>
<td>5</td>
<td>Spatial dimensions</td>
<td>4.1</td>
<td>9</td>
</tr>
<tr>
<td>Definition</td>
<td>6</td>
<td>Hierarchies</td>
<td>4.1</td>
<td>9</td>
</tr>
<tr>
<td>Extension</td>
<td>6</td>
<td>Spatial hierarchies</td>
<td>4.1</td>
<td>9</td>
</tr>
<tr>
<td>Definition</td>
<td>7</td>
<td>Levels</td>
<td>4.1</td>
<td>9</td>
</tr>
<tr>
<td>Extension</td>
<td>7</td>
<td>Spatial levels</td>
<td>4.1</td>
<td>9-10</td>
</tr>
<tr>
<td>Definition</td>
<td>8</td>
<td>Attributes</td>
<td>4.1</td>
<td>10-11</td>
</tr>
<tr>
<td>Extension</td>
<td>8</td>
<td>Spatial attributes</td>
<td>4.1</td>
<td>11</td>
</tr>
<tr>
<td>Definition</td>
<td>9</td>
<td>Hierarchy steps</td>
<td>4.1</td>
<td>11</td>
</tr>
<tr>
<td>Extension</td>
<td>9</td>
<td>Spatial hierarchy steps</td>
<td>4.1</td>
<td>12</td>
</tr>
<tr>
<td>Definition</td>
<td>10</td>
<td>Partial order on levels</td>
<td>4.1</td>
<td>12</td>
</tr>
<tr>
<td>Definition</td>
<td>11</td>
<td>Measures</td>
<td>4.1</td>
<td>12</td>
</tr>
<tr>
<td>Extension</td>
<td>11</td>
<td>Spatial measures</td>
<td>4.1</td>
<td>12</td>
</tr>
<tr>
<td>Definition</td>
<td>12</td>
<td>Fact</td>
<td>4.1</td>
<td>13</td>
</tr>
<tr>
<td>Extension</td>
<td>12</td>
<td>Spatial fact</td>
<td>4.1</td>
<td>13</td>
</tr>
<tr>
<td>Definition</td>
<td>13</td>
<td>Level members</td>
<td>4.2</td>
<td>14</td>
</tr>
<tr>
<td>Definition</td>
<td>14</td>
<td>Attributes of level members</td>
<td>4.2</td>
<td>14</td>
</tr>
<tr>
<td>Definition</td>
<td>15</td>
<td>Partial order on level members</td>
<td>4.2</td>
<td>14</td>
</tr>
<tr>
<td>Definition</td>
<td>16</td>
<td>Fact members</td>
<td>4.2</td>
<td>14-15</td>
</tr>
<tr>
<td>Remark</td>
<td>17</td>
<td>Slice</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>Definition</td>
<td>17</td>
<td>S-Slice</td>
<td>5</td>
<td>15-16</td>
</tr>
<tr>
<td>Remark</td>
<td>18</td>
<td>Dice</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>Definition</td>
<td>18</td>
<td>S-Dice</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>Remark</td>
<td>19</td>
<td>Roll-up</td>
<td>5</td>
<td>17-18</td>
</tr>
<tr>
<td>Definition</td>
<td>19</td>
<td>S-Roll-up</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>Remark</td>
<td>20</td>
<td>Drill-down</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>Definition</td>
<td>20</td>
<td>S-Drill-down</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>Algorithm</td>
<td>1</td>
<td>RUPath</td>
<td>6.1</td>
<td>20</td>
</tr>
<tr>
<td>Algorithm</td>
<td>2</td>
<td>S-SliceGenerator</td>
<td>6.1</td>
<td>21</td>
</tr>
<tr>
<td>Algorithm</td>
<td>3</td>
<td>S-DiceGenerator</td>
<td>6.1</td>
<td>23</td>
</tr>
<tr>
<td>Algorithm</td>
<td>4</td>
<td>SRUGenerator</td>
<td>6.1</td>
<td>24</td>
</tr>
<tr>
<td>Algorithm</td>
<td>5</td>
<td>WriteSPARQL</td>
<td>6.2</td>
<td>25</td>
</tr>
</tbody>
</table>