Enabling OLAP Analyses on the Web of Data
Franck Ravat, Jiefu Song

To cite this version:
Franck Ravat, Jiefu Song. Enabling OLAP Analyses on the Web of Data. 19th IEEE International Conference on Digital Information Management (ICDIM 2016), Sep 2016, Porto, Portugal. pp. 215-224. hal-01466650

HAL Id: hal-01466650
https://hal.archives-ouvertes.fr/hal-01466650
Submitted on 13 Feb 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Open Archive TOULOUSE Archive Ouverte (OATAO)
OATAO is an open access repository that collects the work of Toulouse researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: http://oatao.univ-toulouse.fr/
Eprints ID: 17174

The contribution was presented at ICDIM 2016:


Any correspondence concerning this service should be sent to the repository administrator: staff-oatao@listes-diff.inp-toulouse.fr
Enabling OLAP Analyses on the Web of Data

Franck Ravat
IRIT-Université Toulouse I Capitole
UMR5505, CNRS
Toulouse, France
ravat@irit.fr

Jiefu Song
IRIT-Université Toulouse I Capitole
UMR5505, CNRS
Toulouse, France
song@irit.fr

Abstract—This paper describes a business-oriented analysis environment facilitating analyses of coherent data from Data Warehouses (DWs) and Linked Open Data (LOD) datasets. Specifically, we present a multidimensional modeling solution, named Unified Cube, which provides a single, comprehensive representation of data from multiple sources. Unified Cubes include both concepts close to business terms and user-friendly graphical notations. An implementation framework is proposed to enable unified analyses of warehoused data and LOD. The feasibility of the proposed concepts is illustrated with examples based on real-world datasets.

Keywords—Data Warehouses; Linked Open Data; Unified Conceptual Model; On-line Analytical Processing;

I. INTRODUCTION

For over a decade, Data Warehouses (DWs) have been widely used to support decision-making processes. In a DW, operational data are periodically extracted, transformed and loaded in a multidimensional structure called data cube. A data cube represents warehoused data according to a fact (analysis subject composed of numeric indicators) and dimensions (analysis axes composed of attributes organized according to hierarchies) [1].

Coming from the inside of an organization, warehoused data do not provide all useful information for decision-making. In today's highly dynamic business context, decision-makers need to access several information sources, especially the Web, to obtain comprehensive analytical perspectives [2]. Among the data publically available on the Web, Linked Open Data (LOD) 1, especially analysis-oriented LOD 2, provide promising opportunities to enhance business analyses with semantically interconnected and machine-readable data [3], [4].

Motivating example. In an organization studying the social housing market, a DW contains quarterly and annual submitted and accepted applications for affordable housings by districts and counties (cf. Fig. 1(a)). The DW follows a Relational On-Line Analytical Processing (ROLAP) schema which organizes data according to a fact table (e.g., Affordable Housing) composed of numeric indicators (e.g., Applications and Acceptance) and several dimension tables (e.g., Area and Period) including an ordered set of analytical granularities

http://linkeddata.org

1 Analysis-oriented LOD refer to a type of Linked Open Data describing numeric indicators according to a set of analysis axes. Without specification, all LOD discussed in this paper belong to analysis-oriented LOD.

http://opendatacommunities.org/data

2013 is labeled as Year in the source ROLAP and PeriodRC in the source LOD2; (d) one analytical granularity may group several attributes from heterogeneous sources, e.g., the analytical granularity about district is described by a simple string in ROLAP (e.g., Adur), a classification code in LOD1(e.g., Adur E07000223) and an URI in LOD2 (e.g., http://statistics.data.gov.uk/id/statistical-geography/E07000223).

(e) each source contains different indicators. Some indicators can be analyzed together starting from certain analytical granularities. For instance, Fig. 1 shows independent analysis results revealing the Applications and Acceptance of affordable housings with the supply of affordable housings (i.e. Dwellings) and average rental Charges in the same district named Adur for the same year of 2013.

3
Fig. 1. A ROLAP DW and two LOD datasets about the social housing market.

**Contribution.** Our aim is to provide decision-makers with a business-oriented environment which allows facilitating analyses of coherent data from multiple sources. To this end, we propose a unified view including, in a single schema, all the indicators along with all available analysis axes as well as all the attributes and aggregation paths (coming from the heterogeneous sources). In the previous example, a unified view would enable decision-makers to more easily build the following dashboard (cf. Fig. 1(d)). This dashboard allows better explaining the lower rental charges in the district Adur in 2014 results in more applications for affordable housings, while the greater number of accepted applications is related to the additional supply of affordable housings in 2014.

In this paper, we describe a novel multidimensional model, named **Unified Cube**, which unifies related warehoused data and LOD in a business-oriented way. First, we discuss related work on combining warehoused data with LOD (cf. section II). Second, we present conceptual definitions and graphical notations of **Unified Cubes** (cf. section III). Third, we describe an implementation framework for **Unified Cubes** (cf. section IV). At last, we detail how analyses of data from multiple sources are carried out through an implemented **Unified Cube** (cf. section V).

II. RELATED WORK

In the scientific literature, we can find many state-of-art papers [2], [6] that put forward the idea of including data from DWs and LOD datasets in one analysis. Among these papers, one of the consensuses is that a generic modeling solution should be proposed to unify warehoused data with LOD. The existing work with regards to a unified model can be categorized into three approaches.

Firstly, the DW community tries to build a classical OLAP schema based on warehoused data combined with LOD. More specifically, the authors of [7]–[9] treat LOD as other external data which should be stored in a DW through Extract, Transform and Load (ETL) processes. However, due to the resource-consuming ETL processing, it is difficult to guarantee the freshness of LOD materialized in a local, stationary repository. Moreover, since warehoused LOD are queried in an offline way, decision-maker can hardly obtain up-to-date information during analyses.

Secondly, the LOD community aims at transforming warehoused data into an RDF graph through a customizable mapping language such as R2RML. The resulted schema is expressed in an RDF-based modeling vocabulary which describes the multidimensional structure of a LOD dataset. Among the existing vocabularies, we can cite the RDF Data Cube Vocabulary (QB) as the current W3C standard to publish multidimensional statistical data. In [10], the authors propose the QB4OLAP vocabulary by adding more multidimensional characteristics to QB, like multiple analytical granularities and aggregation functions associated with a measure. Yet, datasets published in QB4OLAP and QB vocabularies still become standalone LOD sources on the Web. No solution has been proposed to enable interoperability between independent data sources.

Thirdly, a joint effort between the two communities consists of combining standalone sources together to enable simultaneous analyses of both warehoused data and LOD. The authors of [11] discuss research challenges and application prospects of merging data from multiple sources without any concrete proposal. The first solution was proposed by the LOD community. [12] proposes IGOLAP vocabulary to represent data from both DWs and LOD datasets according to a multidimensional structure. However, the work of [12] requires warehoused data to be transformed into a stationary RDF dataset before being analyzed with LOD, which reminds us of the drawbacks of the work based on ETL processes. Moreover, IGOLAP is an RDF-based vocabulary which represents data at the logical level. A business-oriented modeling solution is still missing to enable non-expert users to exploit coherent data at a higher level, i.e., conceptual level.
To overcome these drawbacks, Unified Cubes should be generic enough to be independent of specific modeling solutions in the domains of DW and LOD. Moreover, a Unified Cube model should support on-the-fly analyses of both warehoused data and LOD. To facilitate the analysis tasks of non-expert users, only concepts close to business terms should be included in a Unified Cube.

III. CONCEPTUAL MODELING OF UNIFIED CUBES

The objective of Unified Cubes is to provide decision-makers with a single, comprehensive representation of all useful data. This representation should be oriented to business analyses, so that non-expert users can easily interact with a Unified Cube without the need for specific knowledge in DW and LOD domains. To this end, a Unified Cube should fix a generic modeling language which

- organizes warehoused data and LOD according to (a) one analysis subject containing numeric indicators and (b) some analysis axe including analytical granularities according to one or several aggregation paths;
- allows gathering several aggregation paths from different sources together in one analysis axis, even these aggregation paths do not start from the same analytical granularity;
- describes an analytical granularity with several attributes from multiple sources while managing the relationships between heterogeneous attribute instances;
- supports on-the-fly and automatic extraction of detailed data from the sources;
- associates a numeric indicator with only a set of summarizable analytical granularities instead of all analytical granularities on all analysis axes;
- provides user-friendly graphical notations to facilitate the data exploitation task of decision-makers.

A. Analysis Axis: Dimension

A dimension corresponds to a unified vision of attributes which are related to one analysis axis. The organization of these attributes consists of the definition of one or several analytical granularities within a dimension.

The analytical granularities of a dimension may form one or several aggregation paths (i.e. hierarchies). Due to the unification of attributes belonging to different analytical granularities, and contrary to classical multidimensional model, a dimension of Unified Cube may contain two hierarchies sharing no common lowest analytical granularities. For instance, in a temporal dimension, the union of analytical granularities week-year and month-quarter-year results in two aggregation paths without a common starting point. The definition of dimension should be generic enough to include this specificity.

Notations. In the remainder of the paper, a superscript represents an element that the base belongs to, e.g., name\textsuperscript{Adur} means the name of Adur. A subscript indicates the index of an element in a set, e.g., district\textsubscript{k} corresponds to the k-th one in a set of districts.

Definition 1. A dimension representing an analysis axis composed of a sequence analytical granularities is denoted as D\textsubscript{i}={n\textsuperscript{D}; l\textsuperscript{D1}; l\textsuperscript{D2}; \ldots; l\textsuperscript{Dn}}, where:

- n\textsuperscript{D} is the dimension name;
- l\textsuperscript{Di} is a set of levels, each level represents a distinct analytical granularity;
- \( \leqslant_{D} \) is a reflexive binary relation which associates a child level \( l_{c} \in l\textsuperscript{D1} \) with a parent level \( l_{p} \in l\textsuperscript{Dn} \), such as \( l_{c} \leqslant_{D} l_{p} \).

Example. We identify a geographical dimension denoted Geography from the three datasets in Fig. 2. The dimension Geography named Geography includes two levels, such as Geography\textsubscript{i}={l\textsubscript{District}; l\textsubscript{County}}. The binary relation \( \leqslant_{Geography} \) connects the level l\textsubscript{District} to its parent level l\textsubscript{County}, such as l\textsubscript{District} \( \leqslant_{Geography} \) l\textsubscript{County}.

Remark. By removing the constraint on the unique lowest level, a hierarchy including month, quarter and year from one source (i.e., \( \epsilon^{21} = \{\text{month}; \text{quarter}; \text{year}\} \)) and another one containing week and year from a different source (i.e., \( \epsilon^{22} = \{\text{week}; \text{year}\} \)) can be unified in a dimension \( D_{\text{S1S2}} \) which includes a set of levels \( L^{D_{\text{S1S2}}} = \{l\text{\_month}; l\text{\_quarter}; l\text{\_year}\} \). The same hierarchies can be found within the new dimension \( D_{\text{S1S2}} \), such as \( l\text{\_month} \leqslant_{D_{\text{S1S2}}} l\text{\_quarter} \leqslant_{D_{\text{S1S2}}} l\text{\_year} \) which do not share a unique lowest level.

We define a sub-dimension as a part of dimension containing only a subset of levels.

Definition 2. A sub-dimension of \( D_{i} \), denoted \( D_{k} = \{n_{D}; l_{D1}; \ldots; l_{Dn}\} \), corresponds to the part of the dimension \( D_{i} \) started with the level \( l_{k} \), where

- \( n_{D} \) is the name of the sub-dimension;
- \( l_{Dk} \) is the subset of levels, \( l_{Dk} \subseteq l_{Dn} \), \( \forall l_{i} \in l_{Dk} \), \( l_{i} \leqslant_{D} l_{j} \);
- \( \leqslant_{D} \) is the same binary relation of the one on the dimension \( D_{i} \).

Example. A sub-dimension of the temporal dimension may be \( D_{\text{Time\_Year}} \) named Time-Year with \( L_{\text{Time\_Year}} = \{\text{Year}\} \), which represent the subpart of the dimension \( D_{\text{Time}} \) that measures from the datasets LOD1 and LOD2 can be calculated along.

B. Analytical Granularity: Level

A level indicates a distinct analytical granularity by grouping together a set of attributes from different sources. For each attribute, the level keeps its name and an extraction formula allowing querying the attribute instances on-the-fly. Some mappings are also included within a level to manage the correlation and child-parent relations between heterogeneous attributes instances.

Definition 3. A level represents a distinct analytical granularity on a dimension. A level is denoted as \( l_{k} = \{n_{k}; \mathcal{N}_{k}; \mathcal{C}_{k}; \mathcal{G}_{k}\} \), where:
• *n* is the name of the level;

• \( \mathcal{A} = \{a_1, \ldots, a_n\} \) is a finite set of attributes. Each attribute \( a_i \) is a pair \((n, A)\), where \( n \) is the name of the attribute and \( A \) is an extraction formula through query algebra (i.e., relational algebra [13] and SPARQL algebra) indicating the instances of \( a_i \). The domain of an attribute is denoted as \( \text{dom}(a_i) \);

• \( \mathcal{E} : \text{dom}(a_i) \rightarrow 2^{\text{dom}(a)} \) is a symmetric correlative mapping which associates an attribute \( a_i \) with its related ones at the same level;

• \( \mathcal{R} : 2^{\text{dom}(a)} \rightarrow \text{dom}(a_i) \) is a rollup mapping implementing the binary relation between two levels. It connects the instances of child attributes with the instances of a parent attribute at an adjacent level.

**Remark.** Table I. shows the algebraic form of commonly used SPARQL queries.

<table>
<thead>
<tr>
<th>SPARQL Query</th>
<th>SPARQL Algebra</th>
</tr>
</thead>
<tbody>
<tr>
<td>SELECT *</td>
<td>(BGP (TRIPLE ?s ?p ?o))</td>
</tr>
<tr>
<td>WHERE { ?s ?p ?o}</td>
<td></td>
</tr>
<tr>
<td>SELECT ?s ?p</td>
<td>(PROJECT(?s ?p))</td>
</tr>
<tr>
<td>SELECT ?s1 ?s2</td>
<td>(PROJECT(?s1 ?s2))</td>
</tr>
<tr>
<td>WHERE { ?s1 ?p ?o1}</td>
<td>(FILTER (&lt; ?s1 5))</td>
</tr>
<tr>
<td>FILTER (?s1 &lt; 5)</td>
<td>(LEFTJOIN)</td>
</tr>
<tr>
<td>OPTIONAL {?s2 ?p1 ?o2}</td>
<td>(BGP (TRIPLE ?s2 ?p1 ?o2))</td>
</tr>
<tr>
<td>FILTER (?s2 &gt; 10)</td>
<td>(BGP (TRIPLE ?s2 ?p1 ?o2))</td>
</tr>
<tr>
<td>SELECT ?s (COUNT(?o) as ?nh)</td>
<td>(PROJECT(?s ?nh))</td>
</tr>
<tr>
<td>WHERE { ?s ?p ?o}</td>
<td>(FILTER (&gt; ?s 10))</td>
</tr>
<tr>
<td>GROUP BY ?s</td>
<td>(EXTEND(?nh))</td>
</tr>
<tr>
<td>HAVING (COUNT(?o) &gt; 2)</td>
<td>(BGP (TRIPLE ?s ?p ?o)))</td>
</tr>
</tbody>
</table>

**Example.** The level \( l_{\text{District}} \) of the dimension \( D_{\text{geo}} \) contains a set of attributes (i.e., \( \mathcal{A}_{\text{District}} = \{a_{\text{District}}, a_{\text{Description}}, a_{\text{AreaAS}}, a_{\text{AreaRC}}\} \) from different sources: the attributes \( a_{\text{District}} \) and \( a_{\text{Description}} \) come from the ROLAP DW, while the \( a_{\text{AreaRC}} \) and \( a_{\text{AreaAS}} \) are extracted from the LOD1 and LOD2 datasets respectively (cf. Fig. 2). To indicate how attribute instances can be retrieved from sources, an extraction formula is defined for each attribute within a level. For instance, the attribute \( a_{\text{District}} \) is associated with an extraction formula \( E_{\text{District}} = \Pi_{\text{District}}(\text{Area}) \), while the attribute \( a_{\text{AreaAS}} \) is connected with an extraction formula as follows:

\[
E_{\text{AreaAS}}(\text{Area}) = \text{SELECT ?o \text{ FROM DataAccess/rq23/rq24-algebra.html}}
\]

The correlative mapping \( C_{\text{District}} \) associates the instances of the attribute \( a_{\text{District}} \) with the ones of its descriptive attribute \( a_{\text{Description}} \) coming from the same ROLAP DW and the related attributes \( a_{\text{AreaRC}} \) and \( a_{\text{AreaAS}} \) in LOD datasets, for instance,

\[
\text{prefix g: http://opendatacommunities.org/data/housing-market/2012/prefix/}
\]

Table I. shows the algebraic form of SPARQL queries.

**Definition 4.** A Fact corresponds to an analysis subject composed of a set of measures. A fact is denoted as \( F = \{n^i; M^f\} \) where:

• \( n^i \) is the name of the fact;

• \( M^f = \{m_1, \ldots, m_p\} \) is a finite set of numeric indicators called measures. Each measure \( m_i \) is a pair \( \langle n_{me}, E_{me}\rangle \), where \( n_{me} \) is the name of the measure, \( E_{me} \) is an extraction formula pointing to the images of the measure in the source. The set of values of a measure \( m_i \) is denoted as \( \text{val}(m_i) \).

**Example.** The fact named Affordable Housing contains four measures, namely \( m_{\text{Applications}}, m_{\text{Acceptance}}, m_{\text{Dwelling}}, \) and \( m_{\text{Charges}} \). The measure \( m_{\text{Applications}} \) has an extraction formula: \( E_{\text{Applications}} = \Sigma_{\text{Applications}} \) (AffordableHousing). The measure \( m_{\text{Dwellings}} \) has an extraction formula as follows:

\[
\text{prefix m: http://opendatacommunities.org/data/affordable-housing/type/2012/prefix/}
\]

**D. Unified Cube**

A Unified Cube contains an analysis subject (i.e., fact) described by a set of analysis axes (i.e., dimensions). A measure from one source may only be computed with regards to a subset of levels on a dimension (i.e., a sub-dimension). A generalization of the dimension-measure relationship is needed to associate a measure with a sub-dimension starting from any level in Unified Cube.

**Definition 5.** A Unified Cube is a n-dimensional finite space describing a fact with some dimensions. A Unified Cube is denoted as \( UC = \{F; D; LM\} \), where:

• \( F \) is a fact containing a set of measures;

• \( D = \{D_1, \ldots, D_k\} \) is a finite set of dimensions;
• \( \text{LM} \): \( l \mapsto l \) is a level-measure mapping which associates a set of summarizable levels with a measure \( m \) (\( m \in \mathcal{M} \)), such as \( \forall i \in [1..n], l^i \mapsto l^i \) is the set of levels on the sub-dimension \( D^i \) of \( D^j (D^j \in \mathcal{D}) \) starting from \( l^k \).

We propose a set of user-friendly graphical notations for a Unified Cube based on the fact-dimension model [14] with minor modifications (cf. Fig. 2). The graphical notation aims at facilitating data exploitation at the schema level for non-expert users. Measures sharing the same related dimensions at the same levels are grouped together. For readability purposes, the graphical notations do not include concepts involving data instances (e.g., correlative mapping and rollup mapping).

\[ \text{Fig. 2. A Unified Cube about affordable housings.} \]

**Example.** Fig. 2 shows a Unified Cube which is built upon the warehoused data and the two LOD datasets in the motivating example. It contains three dimensions \( \mathcal{D} = \{ \text{Geography}, \text{Time}, \text{Type} \} \). Each measure is associated with its related levels. For the sake of readability, in the graphical notation the level-measure mappings are represented only between the lowest levels of sub-dimension and measures:

\[ \text{LM}: \{ \langle \text{Time} \cdot \text{Year}, \text{Geography} \rangle \mapsto \{ \text{Charges} \}; \langle \text{Time} \cdot \text{Quarter}, \text{Geography} \rangle \mapsto \{ \text{Applications}, \text{Acceptance} \}; \langle \text{Time} \cdot \text{Quarter}, \text{Type} \rangle \mapsto \{ \text{Dwellings} \} \} \]

**IV. IMPLEMENTATION OF A UNIFIED CUBE**

In this section, we describe a generic, reusable implementation framework for Unified Cubes (cf. Fig. 3). The objective of the framework is to enable analyses of both warehoused data and LOD by representing different components of a unified schema and relationships among data from multiple sources. To do so, the implementation framework should include:

- a metamodel for Unified Cubes with a compatible instantiation algorithm. (cf. section IV.A Schema module);
- tables of correspondences allowing keeping related data in a coherent environment (cf. section IV.B Relationships module).

**A. Schema Module**

The Schema module aims at representing the overall structure of data from multiple sources. A metamodel representing the concepts related to the multidimensional schema of a Unified Cube is included within this module. Fig. 4 shows the class diagram representation of the metamodel.

\[ \text{Fig. 4. Metamodel of Unified Cubes' schema.} \]

In this metamodel, each class implements one concept of Unified Cubes. Note that binary relations between level (i.e., \( \equiv \)) and level-measure mappings (i.e., \( \text{LM} \)) are implemented through associations. Moreover, each attribute and each measure are associated with an executable query (i.e., \( \text{queryM} \) or \( \text{queryA} \)) translated from the extraction formulae. By executing a query in a corresponding source whose query endpoint (i.e., \( \text{URLM} \) and \( \text{URLA} \)) is recorded, data can be extracted in real-time, so that decision-makers can always obtain up-to-date information during analyses.

In our previous work [15], we describe the designing of a conceptual Unified Cube through a schema definition process. In the remainder of this section, we complete our previous work by detailing how a Unified Cube is implemented. To automate implementations of Unified Cubes’ schemas, we propose an algorithm which instantiates the metamodel with a conceptual Unified Cube (cf. Algorithm 1).

We apply the algorithm to the conceptual Unified Cube in Fig. 2. The obtained instantiated metamodel is composed of (a) 1 instance of the Fact class, (b) 4 instances of the Measure class, (c) 3 instances of the Dimension class, (d) 5 instances of the Level class with 12 associations implementing the level-measure mapping and 4 associations representing the binary relation between levels, and (e) 10 instances of the Attribute class. Fig. 3 shows an extract of the instantiated metamodel related to the measure \( m_{\text{charges}} \).

**Algorithm 1. Unified Cube Schema Implementation**

**Input:** A Unified Cube \( \equiv \{ \mathcal{F}, \mathcal{D}, \mathcal{LM} \} \).

**Output:** An instantiated metamodel.
Begin
1. For each $D_i \in D$
2. For each $l_j \in L_{D_i}$
3. Instantiate the Level class: $l_{\text{meta}} = \text{new Level}(n^d)$;
   For each $a_i \in A^d$
4. Translate the query algebra of the extraction formula $E_{\text{ex}}$ into a query $Q_{a_i}$;
5. Get the attribute's query endpoint $URL_{a_i}$;
6. Instantiate the Attribute class: $a_{\text{meta}} = \text{new Attribute}(n^x, Q_{a_i}, URL_{a_i})$;
7. Instantiate the association between $a_{\text{meta}}$ and $l_{\text{meta}}$
8. End for
9. For each $l_j \leq l_i (l_i, l_j \in L_{D_i})$
10. Instantiate the association between child level $l_{\text{meta}}$ and the parent level $l_{\text{meta}}$
11. End for
12. End for
13. Instantiate the Fact class $F_{\text{meta}} = \text{new Fact}(n^f)$;
14. For each $m_k \in M^k$
15. Translate the query algebra of the extraction formula $E_{\text{ex}}$ into a query $Q_{m_k}$;
16. Get the measure's query endpoint $URL_{m_k}$;
17. Instantiate the Measure class: $m_{\text{meta}} = \text{new Measure}(n^m, Q_{m_k}, URL_{m_k})$;
18. Instantiate the association between $m_{\text{meta}}$ and $F_{\text{meta}}$
19. For each level $l_j \subseteq l_k$, $l_k \subseteq L$ associated with $m_k$, $F$:
   $L_{l_k} \times \ldots \times L_{l_k} \Rightarrow m_k$
20. For each level $l_j \subseteq L_{l_k} \times \ldots \times L_{l_k}$
21. Instantiate the association between $l_{\text{meta}}$ and $m_{\text{meta}}$
22. End for
23. End for
End

B. Relationships Module

The Relationships module handles related data from different sources. It includes (a) a toolkit identifying related attribute instances and (b) tables of correspondences keeping the obtained relationships between attribute instances.

a) Step I: Identification of Related Data.

In the context of Unified Cubes, identifying the relationships between data consists of assessing the correlation and the child-parent relations between attribute instances from different sources. Numerous techniques have been proposed in the scientific literature to identify the relationships between warehoused data and LOD. Two comprehensive surveys of automatic matching between related schemas and data instances can be found in [16], [17]. The authors of [18] focus on some semi-automatic techniques which take users' needs into account.

Among all above-mentioned work, one consensus is that there is no "one-size-fits-all" method for aligning data of all types [19], and such a universal method does not fall within the scope of our work. One objective of the Relationships module is to provide schema designers with a toolkit facilitating the task of discovering related data. A designer can simply specify an appropriate method implemented within the toolkit. Then the implementation framework automatically searches for related data in corresponding sources according to the chosen method.

In the context of the Unified Cube of our running example, we implement some of the most effective methods allowing identifying the correlation and child-parent relationships between data. These methods can be categorized into three groups.

- The first group is based on an intermediate ontology with a comprehensive coverage of the related concepts in several datasets (i.e., containing enough matches between related data).
- The second group is applicable to data instances sharing related labels by calculating the string-based similarity. Normalization techniques and external thesaurus are often used to improve the obtained result.
- The third group includes methods use reasoning techniques to deduce the relations between data. This group of methods is especially useful to obtain rollup mappings between attributes at two adjacent levels. By referring to existing correlative and/or rollup mappings within a Unified Cube, the reasoning process is as follows:

$$\text{let } a_{i1}, a_{j1} \in A^d, a_{j2} \in A^d (l_i \leq l_j);$$

- Correlative mappings:
  $$\forall l_i: \text{dom}(a_{i1}) \rightarrow \text{dom}(a_{j2}) \wedge l_j: \text{dom}(a_{j1}) \rightarrow \text{dom}(a_{j2})$$
  $$\Rightarrow \exists l_k: \text{dom}(a_{i1}) \rightarrow \text{dom}(a_{j1})$$

- Rollup mappings:
  $$\forall l_i: \text{dom}(a_{i2}) \rightarrow \text{dom}(a_{j2}) \wedge l_j: \text{dom}(a_{j1}) \rightarrow \text{dom}(a_{j1})$$
  $$\Rightarrow \exists l_k: \text{dom}(a_{i2}) \rightarrow \text{dom}(a_{j1})$$

TABLE II. shows five correlative and six rollup mappings identified after applying suitable methods implemented in the Relationships module. Note that no method is needed to identify mappings already existing in the sources. For instance, mappings 1, 6 and 9 are obtained by directly referring to the ROLAP DW, while mapping 4 is directly identified because the instances of the attributes $a_{\text{PeriodAS}}$ and $a_{\text{PeriodRC}}$ are denoted by using the same identifier (e.g., the year of 2013 is denoted as <http://reference.data.gov.uk/id/government-year/2013-2014> for both $a_{\text{PeriodAS}}$ and $a_{\text{PeriodRC}}$). The relations between attribute instances involved in the above-mentioned mappings are called direct, since it can be identified by directly querying the corresponding sources. Relations embedded in other mappings are called deductive, because it is identified by using additional processing methods. For instance, the correlative mapping between the attributes $a_{\text{District}}$ and $a_{\text{AreaAS}}$ (cf. mapping 2 in TABLE II. ) is obtained by calculating the substring test similarity $\tilde{\sigma}(s_1, s_2)$, such as

$$\tilde{\sigma}(s_1, s_2) = \frac{2|s_1 \cap s_2|}{|s_1| + |s_2|},$$

where $s_1$ and $s_2$ are two attribute instances, $s$ is the longest common substring between $s_1$ and $s_2$. 

A correlative mapping associates one attribute instance with at most one instance of another attribute. Therefore, for a pair of attributes \( a_i \) and \( a_j \), attributes instances that can possibly be materialized corresponds to the smaller set of attribute instances (i.e., \( \min|\{\text{dom}(a_i), \text{dom}(a_j)\}| \)). Within a level \( l_k \), there exist at most \( \frac{|\mathcal{D}|(x)\cdot(|\mathcal{D}|(x)-1)}{2} \) non-redundant pairs of attributes that can possibly be associated together through correlative mappings. Let \( \mathcal{D}_{x}^{t_i} : \text{dom}(a_i) \rightarrow \text{dom}(a_j) \) be a correlative mapping between attributes \( a_i \) and \( a_j \), for a n-dimensional Unified Cube, the Relationships module materializes at most the following number of attributes instances involved in correlative mappings.

$$\sum_{i=1}^{n} \sum_{d=1}^{\left|\mathcal{D}(x)\right|} \sum_{k=1}^{\left|\mathcal{D}(y)\right|} \min(|\text{dom}(a_i)|, |\text{dom}(a_j)|)$$

A rollup mapping associates one or several instances of a child attribute with at most one instance of a parent attribute. The maximum number of attributes instances possibly being materialized for a rollup mapping between a pair of attributes \( a_i \) and \( a_j \), \( a_i \in \{l^i\}, a_j \in \{l^j\}, l^i < l^j \), \( \text{dom}(a_i) \). Therefore, for a n-dimensional Unified Cube, the Relationships module materializes no more than the following number of attributes instances for the implementation of rollup mappings.

$$\sum_{i=1}^{n} \sum_{d=1}^{\left|\mathcal{D}(x)\right|} \sum_{k=1}^{\left|\mathcal{D}(y)\right|} \text{dom}(a_i)$$

In both formulae, only dimensions are involved. For a multidimensional dataset, the size of dimension is much smaller than the size of fact. For instance, in a test set of SSB benchmark [20], dimensions represent basically 1% to 6% of the total data volume, and the larger the data scale becomes, the lower the proportion of dimensions accounts for [21].

From the above discussions, it is apparent that the materialization of deducive mappings allows managing related data from different sources with minimum cost. Within the implementation framework, two types of table of correspondences are used to materialize correlative mappings and rollup mappings. As shown in Fig. 6, the one implementing correlative mappings associates the instances (i.e., INSTANCE) of an attribute (i.e., ATTRIBUTE) with the related ones (i.e., COR_INS) of a correlative attribute (i.e., COR_ATT) within the same level, while the other one managing rollup mappings connects a set of instances (i.e., CHILD_INS) of a child attribute (i.e., CHILD_ATT) with an instance (i.e., PARENT_INS) of a parent attribute (i.e., PARENT_ATT). The content of a table of correspondence is preprocessed and updated at the beginning of an analysis process to keep the materialized data up-to-date.

### Table II. List of Mappings and Identification Methods

<table>
<thead>
<tr>
<th>Mapping</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ( \mathcal{D}<em>{x}^{t_1} : \text{dom}(a</em>{x_1}) \rightarrow \text{dom}(a_{x_2}) )</td>
<td>substring test similarity</td>
</tr>
<tr>
<td>2. ( \mathcal{D}<em>{x}^{t_1} : \text{dom}(a</em>{x_1}) \rightarrow \text{dom}(a_{x_2}) )</td>
<td>intermediate ontology</td>
</tr>
<tr>
<td>3. ( \mathcal{D}<em>{x}^{t_1} : \text{dom}(a</em>{x_1}) \rightarrow \text{dom}(a_{x_2}) )</td>
<td>substring test similarity</td>
</tr>
<tr>
<td>4. ( \mathcal{D}<em>{x}^{t_1} : \text{dom}(a</em>{x_1}) \rightarrow \text{dom}(a_{x_2}) )</td>
<td>reasoning (cf. mappings 2 and 6)</td>
</tr>
<tr>
<td>5. ( \mathcal{D}<em>{x}^{t_1} : \text{dom}(a</em>{x_1}) \rightarrow \text{dom}(a_{x_2}) )</td>
<td>reasoning (cf. mappings 3 and 7)</td>
</tr>
<tr>
<td>6. ( \mathcal{D}<em>{x}^{t_1} : \text{dom}(a</em>{x_1}) \rightarrow \text{dom}(a_{x_2}) )</td>
<td>reasoning (cf. mappings 4 and 10)</td>
</tr>
</tbody>
</table>

Fig. 5. Extract of metamodel instantiated with the ROLAP DW and the two LOD datasets.
After implementing the Unified Cube of our running example, 1474 attribute instances are materialized in the Relationships module, which corresponds to about 5.1% of data in the sources. Selective materialization allows reducing significantly the amount of data stored in the implementation framework, which minimizes the cost of updating and storing data from the sources.

V. ANALYSIS PROCESSING

Besides the two modules presented in section IV, the implementation framework also includes a third module, named Mediator, whose aim is to enable decision-makers to analyze both warehoused data and LOD. It (a) deals with an analytical need expressed through the graphical notations of a Unified Cube (cf. arrow 1 in Fig. 7) and (b) yields an analysis result based on data extracted from multiple sources (cf. arrow 5 in Fig. 7).

![Fig. 7. Analysis processing within the implementation framework](image)

Since most decision-makers are not expert in neither DW nor LOD domain, the Mediator module should facilitate analysis tasks by including functionality to:

- automatically extract data related to the analysis from the sources. To do so, we propose an algorithm which generates a set of executable queries based on the measures and attributes chosen by a decision-maker (cf. arrows 2 and 3);
- include all data in one unique analysis result to offer comprehensive analytical perspectives. To this end, we present an algorithm which combines all partial results together by referring to the related attribute instances materialized in the implementation framework (cf. arrows 4.1, 4.2 and 5).

A. Partial Query Generation

By referring to the extraction formulae implemented in the metamodel, the Mediator module can directly extract the instances of one measure or one attribute from one source. In the case of a unified analysis including several measures and attributes, a simple concatenation of the extraction formulae from the metamodel is not enough. The Mediator module must enrich and/or rewrite the queries hosted in the metamodel to adapt to the specificities of data sources.

- When an analysis involves attributes at different levels, the Mediator module complete queries by adding joins and grouping predicates.
- When an analysis includes measures and attributes from multiple sources, the Mediator module generates queries containing the corresponding attributes and measures from the same source.

To automate the generation of queries within the Mediator module, we propose the following algorithm.

**Algorithm 2. Partial Query Generation**

**Input:** A set of chosen attributes $A_{sub}$ ($A_{sub} \subseteq U_{D_{i \in D}} U_{k \in E_{i}} A^{k}$), a set of chosen measures $M_{sub}$ ($M_{sub} \subseteq M$)

**Output:** A set of partial queries $Q$

**Begin**

1. For each $m_{k} \in M_{sub}$
   1. Get associated levels of $m_{k}$: $L_{m_{k}}=L_{m_{k}}.getLvls()$;
   2. Get the query of the measure $q_{m_{k}}=m_{k}.getQueryM()$;
   3. Remove irrelevant attributes from $q_{m_{k}}$
   4. For each $a_{i} \in A_{sub}$
      1. Get the level of $a_{i}$: $P_{a_{i}}=a_{i}.getLevel()$;
      2. If $m_{k}.getURLM() \neq a_{i}.getURLA()$
         1. Find an attribute $a_{j}$, such as $a_{j}.getLvl()=P_{a_{i}}$
            1. Get the query $q_{a_{j}}$ of the attribute $a_{j}$;
            2. Add $q_{a_{j}}$ in $q_{m_{k}}$;
      3. Find a set of level $L_{sub}$ ($L_{sub} \subseteq L_{m_{k}}$) such as $\forall l_{i} \in L_{sub}$
         1. Get the dimension $l_{i}.getDim()=P_{a_{i}}.getDim() \land l_{i} \notin P_{a_{i}}$;
         2. If $|L_{sub}|>1$
            1. Add join predicates between $m_{k}$ and $a_{i}$ in $q_{m_{k}}$;
            1. Add a grouping predicate at the end of $q_{m_{k}}$;
            2. End if
   4. End if
5. End if
6. $Q=Q \cup q_{m_{k}}$;
7. End for
8. End for

**End**

To better explain how the algorithm works, we illustrate its execution process with an example. An analysis need about the number of applications (i.e., $m_{app}$) for social housings with the average rental charges (i.e., $m_{charge}$) by year

$\forall D_{i \in D}, U_{k \in E_{i}} A^{k} = A^{1} \cup \ldots \cup A^{|E_{i}|}$ is the set of attributes within the dimension $D_{i}$. Thus, $U_{D_{i \in D}} U_{k \in E_{i}} A^{k}$ represents all attributes of a Unified Cube.
(i.e., \(a_{\text{Year}}\))^r consists of a cross-source analysis. It calculates two measures from ROLAP DW and LOD2 dataset according to an attribute from the ROLAP DW. Two queries are generated for this analysis after the execution of the algorithm (cf. Fig. 8). The first query \(Q_1\) is sent to the ROLAP DW. It contains the measure \(m_{\text{Applications}}\) and the attribute \(a_{\text{Year}}\), with corresponding join and grouping predicates. For the second query \(Q_2\), the algorithm firstly refers to the metamodel to find a correlative attribute (i.e., \(a_{\text{PeriodRC}}\)) in the LOD2 dataset which is at the same level of \(a_{\text{Year}}\). Then a query is directly generated from the extraction formulae of \(m_{\text{Charges}}\) and \(a_{\text{PeriodRC}}\), by simply adding a grouping predicate.

As shown in Fig. 8, generated partial queries are sent to the Mediator module. The Mediator module returns a result that is displayed in a unified way. We propose the following algorithm to generate the Multidimensional Table based on the partial query results received from the Mediator module and the mappings of related attributes instances obtained from the Relationships module.

**Algorithm 3. Generation of Analysis Result**

**Input:** A set of partial results \(\mathcal{R}\), each result \(r_i (r_i \in \mathcal{R})\) contains a set of measures \(\mathcal{M}\) and a set of attributes \(\mathcal{A}\) from one data source; A set of mappings \(\mathcal{Map}\), each mapping, \(f_{pq}: \text{dom}(a_p) \rightarrow \text{dom}(a_q)\) represents either the correlative mapping or the rollup mapping between the attributes \(a_p\) and \(a_q\).

**Output:** A Multidimensional Table \(\mathcal{MT} = \{S_{\mathcal{MT}}, \mathcal{AX}_{\mathcal{MT}}\}\)

**Begin**

1. For each \(r_i \in \mathcal{R}\)
2. For each \(a_{pq} \in \mathcal{A}\)
3. Get the dimension of \(a_{pq}\): \(\mathcal{D}_{pq} = a_{pq}.\text{getLvl}().\text{getDim}()\).
4. Add \(a_{pq}\) in the set of attributes \(\mathcal{A}\) on the corresponding analysis axis of \(\mathcal{MT}\), such as \(3a_{\mathcal{MT}} = \mathcal{AX}_{\mathcal{MT}} \cup \{a_{pq}\}\). For each attribute instance \(i_{\mathcal{MT}}\) in \(\mathcal{AX}_{\mathcal{MT}}\):
5. If \(\exists f_{pq}, \text{dom}(a_p) \rightarrow \text{dom}(a_q) (f_{pq} \in \mathcal{Map})\) and \(a_p \in \mathcal{A}_{\mathcal{MT}}\) Associate \(i_{\mathcal{MT}}\) with the related instances of the attribute \(a_q\).
6. Else
7. Create a new header \(i_{\mathcal{MT}}\) in the \(\mathcal{MT}\)
8. End if
9. End for
10. End For
11. For each \(m_j \in \mathcal{M}\)
12. Add \(m_j\) in \(\mathcal{M}_{\mathcal{MT}}\);
13. For each measure value \(v_{m_j}\), \(v_{m_j} \in \text{val}(m_j)\)
14. If there exists a measure value \(v_{m_k}\) (\(m_k \in \mathcal{M}_{\mathcal{MT}}\)) described by the same related attributes instances as \(v_{m_j}\) in the \(\mathcal{MT}\)
15. Display \(v_{m_j}\) in the same cell as \(v_{m_k}\);
16. Else
17. Display \(v_{m_j}\) in a new cell
18. End if
19. End for
20. End For
21. End for
22. End For
23. End for

**End.**

During the execution of the algorithm, each analysis axis in a Multidimensional Table is built by (a) combining different attributes referring to the same dimension (cf. lines 3 and 4) and (b) organizing related attribute instances into the same headers (cf. lines 5 to 11). The analysis subject in a
**Multidimensional Table** is obtained by (a) unifying the measures from all partial results of query (cf. line 14) and (b) displaying multiple measure values together if they are described by related attribute instances involved in *correlative* mappings or *rollup* mappings (cf. lines 15–21).

VI. CONCLUSION AND DISCUSSIONS

Under today's highly competitive business environment, warehoused data alone do not provide enough information during analyses. External data, especially *analysis-oriented* LOD, should also be included in a decision-making context to offer multiple perspectives to decision-makers. In this paper, we describe a novel modeling solution, named *Unified Cube*, which represents data from DWs and LOD datasets in a generic and business-oriented way.

As an extension of classical multidimensional models, a *Unified Cube* organizes warehoused data and LOD according to one analysis subject (i.e., fact) and a set of analysis axes (i.e., dimensions). Due to the unification of analytical granularities (i.e., levels) from different sources, we generalize the definition of dimension to allow including several hierarchies sharing no common lowest level. The concept of level is also extended to regroup multiple attributes if they refer to the same analytical granularity. Heterogeneity among attribute instances from multiple sources is managed by *correlative* mappings, while the child-parent relations between attribute instances at adjacent levels are implemented through *rollup* mappings. The concept of *level-measure* mapping is also included in *Unified Cubes* to associate a numeric indicator (i.e., measure) with a set of summarizable levels.

To enable non-expert users to analyze warehoused data with LOD, we propose an implementation framework compatible with *Unified Cubes*. Three modules are included within the framework. The *Schema* module hosts a metamodel implementing the multidimensional structure of a *Unified Cube*. The instantiation of the metamodel yields a non-materialized *Unified Cube* supporting on-the-fly analyses among multiple data sources. The *Relationships* module provides a toolkit allowing identifying related attributes instances from heterogeneous sources. A set of tables of correspondences materializes the mappings between related data to avoid repetitive relationship processing during analyses. The third module named *Mediator* automatically translates a unified analysis need into a set of executable queries. At the end of an analysis, the *Mediator* module provides one unique result including all useful data for an analysis.

In the future, we intend to include a graphical querying language within our implementation framework to further facilitate the analysis tasks of non-expert users. We also intend to extend *correlative* and *rollup* mappings to more generic ones, e.g., *n*-ary mappings between three or more attributes. A more long-term objective consists of maximizing the analysis efficiency by finding an optimized selective materialization strategy according to different data types.

**References**


