Containment and Complementarity Relationships in Multidimensional Linked Open Data

Marios Meimaris and George Papastefanatos

Institute for the Management of Information Systems, Research Center “Athena”, Greece
{m.meimaris, gpapas}@imis.athena-innovation.gr

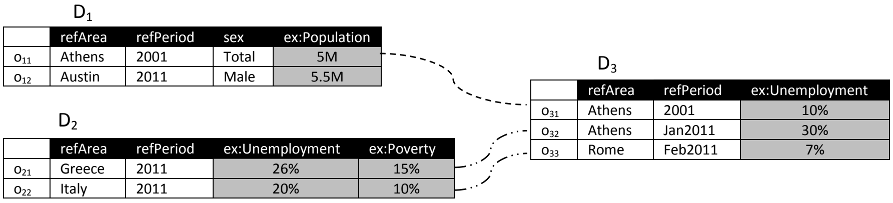
Abstract. The Linked Open Data (LOD) cloud can act as a source of remote multidimensional datasets which are seemingly disparate, but are modeled under common directives and thus often share a common meta-model, dimensions and measures, as well as external codelists. This gives them a latent measure of relatedness that is independent of the publishers' initial intentions, but a derivative of the motivations behind LOD. In this paper we identify the constituents of relatedness between multidimensional LOD data points (observations) modeled with the Data Cube vocabulary, that often exhibit overlapping values both at the schema and at the data level. Treating hierarchies as first-class citizens, we consider observation relatedness in two aspects, namely containment and complementarity, for which we provide formal definitions and representational semantics. Finally, we present a methodology for computing these types of relatedness and we provide an evaluation over real-world datasets.

Keywords: Linked Data, Data Cube, Multidimensional Data, Data Enrichment

1 Introduction

Recently, more and more bodies such as governments, statistical authorities, public and private organizations, research and health centers publish information in the form of multidimensional Linked Open Data (LOD) [2][3] in very different domains, such as census and statistical data, socioeconomic and demographic indicators, clinical trials and health data, environmental and finance data. The abundance of such datasets enable third parties to have access, exploit and combine published data eventually leading to the generation of new quantifiable insights and knowledge, capable of influencing policy-building and decision-making.

Modelling-wise, multidimensional data are traditionally represented as cubes of observations that are instantiated over a fixed set of dimensions and measures. In the LOD paradigm, W3C has proposed the Data Cube Vocabulary [2], a recommendation for modelling and publishing multidimensional datasets. However, one difference between closed-world multidimensional data stores, such as OLAP databases, and LOD datasets is that proper LOD publishing techniques across remote data providers will lead to the existence of common or shared terms between seemingly independent multidimensional datasets, given that reusability and interoperability are prime drivers in the semantic web. Practically, this means that ontologies, codelists and hierarchies that are commonly used in the LOD cloud are likely to appear across different and disparate LOD multidimensional datasets.



D_1

	refArea	refPeriod	sex	ex:Population
o_{11}	Athens	2001	Total	5M
o_{12}	Austin	2011	Male	5.5M

D_2

	refArea	refPeriod	ex:Unemployment	ex:Poverty
o_{21}	Greece	2011	26%	15%
o_{22}	Italy	2011	20%	10%

D_3

	refArea	refPeriod	ex:Unemployment
o_{31}	Athens	2001	10%
o_{32}	Athens	Jan2011	30%
o_{33}	Rome	Feb2011	7%

Figure 1: Candidate relationships between observations

Currently there is little work on definitions and techniques to enable the discovery, representation and exploitation of relationships between individual observations in the web of data [14][2]. An analyst needs to know how observations from different sources are related, e.g., if an observation contains aggregated data with respect to other observations, or if two observations that measure different phenomena can be combined, and exploit such relationships. Consider the motivating example of Fig 1, with 3 multidimensional datasets $D_1 \ldots D_3$ coming from different web sources that measure the population in major global cities, the unemployment and poverty in EU countries and the unemployment in EU cities, respectively. The analyst wishes to explore and annotate the possible interconnection between two observations and specifically detect whether an observation contains an aggregated measurement of another, or whether an observation can be compared with another based on the common dimension values. Current vocabulary and techniques do not allow for users to quickly validate whether or not two observations are hierarchically related, and if so in what level and under what context. In our example, observation o_{11} shares the same dimension values with o_{31}, but they measure two different facts. On the other hand, observation o_{21} that measures unemployment in Greece contains observation o_{32} that measures unemployment in Athens for a sub-period of the same year, although o_{21} measures poverty as well. This knowledge can give insights on how rollups can be performed in order to navigate a data web cube online, or make the two observations comparable. Finding related observations based on one or more similarity factors can help the analyst with exploration, discovery and online browsing of multidimensional datasets

Approach Overview. These points raise issues that we attempt to approach in the context of this paper. Namely, we provide semantics and techniques to measure containment- and schema-based similarities between multidimensional observations coming from different sources with varying degrees of LOD-imposed term overlaps. More specifically, given an observation o_g, and a set of observations O from different sources, we define two relatedness properties, namely observation *containment* and *complementarity*. Containment captures whether an observation measures an aggregated phenomenon with respect to other observations. It compares and determines whether values from the dimensions of o_g contain *fully* or *partially* (i.e. are hierarchical ancestors of) values of the dimensions of another observation. Complementarity refers to characterizing two observations based on their ability to be combined and provide extended/enriched information. More specifically, we extend the notion of schema complement, defined in [15] and apply it at the instance level of observations in order to annotate whether two observations can complement each other's information. Then, we provide a technique for computing these properties as follows: first all observations are placed in an occurrence matrix that represents them as data vectors in a multidimensional feature space encoding both schema and data information along with dimension hierarchies. The occurrence matrix is used to compute the complementarity matrix, which enables us to derive complementarity between observations. Then, the occurrence matrix is transformed to a set of k $|O|\times|O|$ containment matrices, where k is the distinct number of dimensions. Full and partial containment for all pairs of observations are given by adding all matrices together.

Contributions. In short, the contributions of this work are as follows: (a) We define *new relationships* between individual observations and *extend* the Data Cube terms with properties for representing: *full* and *partial observation containment* between two observations as derivatives of the hierarchical relationships between their dimension values, and *observation complementarity* as a means of comparison and correlation of different measures; (b) we provide an *efficient technique* of computing these properties based on occurrence and similarity matrices; finally (c) we *evaluate* our techniques over real-world multidimensional datasets.

This paper is structured as follows: section 2 provides background knowledge and related work, section 3 defines the new properties, section 4 provides the techniques for computing them, while section 5 proposes possible extensions to Data Cube Vocabulary. Finally, section 6 presents an evaluation of our approach and section 7 concludes the paper and presents future directions of this work.

2 Related Work

Generally, the problem of finding related multidimensional observations has been addressed within the contexts of Linked Data and Online Analytical Processing (OLAP) [10] [11]. Data mining techniques in OLAP are known as Online Analytical Mining (OLAM) [7]. OLAM research works study problems such as clustering/classifying OLAP cubes [9], detecting outliers [8], performing intelligent aggregations [5] and building recommender systems for OLAP sessions based on either query formulation or observation similarity [1][16]. Applications of these approaches aim to enable discovery of latent knowledge, promote exploratory analysis [6], improve OLAP query efficiency [9] and so on. In this context, Aligon et al. [1] study the problem of finding similarities between OLAP sessions, i.e sequences of queries that are applied online. To this end they define similarity functions by conducting user-based analysis and they compute similarity of sessions by decomposing the session queries' features and then using the Levenshtein distance, Dice's coefficient, term frequency-inverse document frequency (tf-idf) and the Smith-Waterman algorithm. They find the latter to be the best performing measure for their purposes.

Baikousi et al. [16] provide distance functions categorized over their relation with the hierarchy space. Similarly to our approach, they consider hierarchies to be of prime importance in the problem and base their distance functions on hierarchies. Finally, they summarize the hierarchical distances with two approaches, simple summation and the Hausdorff distance and they find that both approaches are equally effective. Hsu et al [4] apply multidimensional scaling methods (MDS) and hierarchical clustering (HC) in order to find similarity between OLAP reports of the same cube. They formally define the problem and its constraints, such as identifying when two OLAP reports are comparable, and conclude that a combination of MDS and HC yields the best results.

In the context of Linked Data, similarity or relatedness between entities has been a main component of entity resolution, record linkage and interlinking [17][18][19]. These approaches deal with discovering links between RDF nodes from different datasets in efficient ways by using distance-based techniques. Statistical linked open data have been addressed by [13] in the context of online analysis and exploration, and in [14] as a use case scenario for data source contextualization. To the best of our knowledge, this is the first work that addresses the definition, representation and computation of relationships between individual multidimensional LOD observations.

3 Problem definition

We consider that the problem space is composed by *n* datasets modeled and validated by the integrity constraints imposed by the Data Cube Vocabulary. A dataset is composed of its schema (i.e. dimensions, measures and attribute definitions), and its data (i.e. observations). The values in dimensions are provided by a fixed set of coded lists (code-value pairs) that are hierarchically structured in levels. Flat coded lists, i.e., simple enumerations, are considered to be hierarchies with exactly one level. These are presented formally in the following:

Definition 1 (Cube Structure): Let D={$D_1, ..., D_n$} be the set of all input datasets. A dataset $D_i \in D$ is composed by the set of observations, O_i, and the set of schema definitions, S_i, and O={$O_1, ..., O_n$} and S={$S_1, ..., S_n$} are the sets of all observations and schema definitions in D. Furthermore, a schema S_i consists of the sets of dimension P_i and measure properties M_i defined in D_i, i.e., $S_i=\{P_i,M_i\}$. Let $P=\bigcup_1^n P_i = \{p_1, p_2, \ldots, p_k\}$ and $M=\bigcup_1^n M_i = \{m_1, m_2, \ldots, m_l\}$ be the set of all *k* distinct dimensions and *l* measure properties in D. Any $p_j \in P$, $m_j \in M$ can belong to more than one S_i, as dimension and measure properties are reused among sources. An observation $o \in O_i$ is an entity that instantiates all dimension and measure properties defined in S_i. The value that observation o_i has for dimension p_j ish_i^j.

Definition 2 (Coded list terms): Each dimension property $p_j \in P$ takes values from a fixed coded list, i.e. a set of code – value pairs, $C(p_j)=\{c(p_j)_1, \ldots c(p_j)_m\}$, $j=1..k$, (for simplicity we write c_{ji} instead of $c(p_j)_i$). The coded list defines a hierarchy such that when $c_{ji} \succ c_{jm,}$ then c_{ji} is an ancestor of c_{jm}. Furthermore, we define c_{jroot} as the top concept in the code list of p_j, i.e., an ancestor of all other terms in the coded list, such that $\forall c_{ji}: c_{jroot} \succ c_{ji}$. Every coded list term is an ancestor of itself, i.e. $\forall c_{ji}: c_{ji} \succ c_{ji}$.

In Figure 2, we present sample coded lists for the dimensions present in motivating example of Fig. 1.

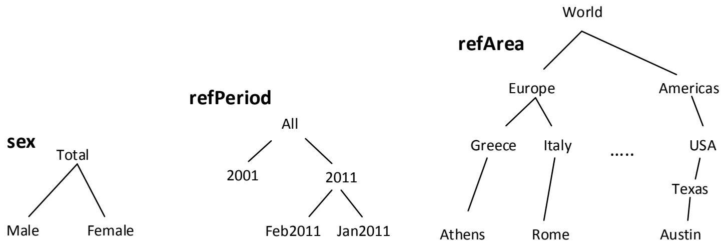


Fig. 2. Hierarchical coded list for the dimensions in Fig 1

Next we provide the definitions for containment and complementarity properties. Similarly to [14] [15], we apply the notion of complementarity between two observations for denoting whether these are comparable, i.e., they have the same dimension values but measure different phenomena. This is represented by the following.

Definition 3 (Observation Complement): Given two observations o_a and o_b and their dimensions P_a and P_b, o_a is observation complement to o_b when:

$$(P_a \subseteq P_b) \wedge \left(\forall p_i \in P_a \cap P_b : h_a^i = h_b^i\right) \wedge \left(\forall p_j \in P_b \setminus P_a : h_b^j = c_{jroot}\right)$$

We denote this relationship with $\mathsf{Compl}(o_a, o_b)$ or equivalently o_a Compl o_b. Def. 1 states that the dimensions in o_b must be a superset of the dimensions in o_a and the common dimensions must have the same values. All other dimension values of o_b must be equal to the root of the dimension hierarchy, thus providing no further specialization. For example, an observation measuring poverty in Greece per year is observation complement with an observation measuring the population in Greece per year for all genders.

Furthermore, a containment relationship captures whether an observation measure is an aggregation of the measures of the contained observations. For example, an observation measuring the population of Greece implicitly contains all observations measuring the population of sub-regions of Greece. We distinguish between full and partial containment. The former denotes that all contained observations can be combined in a roll-up operation for being observation complement with the containing one, while the latter denotes that both contained and containing observation must be rolled-up on their disjoint dimensions for being observation complement. These are presented in the following definition.

Definition 4 (Partial and full containment): Given two observations o_a and o_b, their dimensions P_a and P_b and their measures M_a and M_b, partial containment between o_a and o_b exists when:

$$(\exists M_i \in M_a \cap M_b) \wedge (P_a \subseteq P_b) \wedge (\exists p_i \in P_a \cap P_b : h_a^i \succ h_b^i)$$

An observation o_a partially contains o_b when (i) there is one M_i shared between o_a and o_b, (ii) the dimensions of o_a are a subset of the dimensions of o_b and (iii) there exists at least one dimension whose value for o_a is a hierarchical ancestor of the *respective* dimension value for o_b. We denote this as $\mathsf{Cont}_{\mathsf{partial}}(o_a, o_b)$ or equivalently o_a $\mathsf{Cont}_{\mathsf{partial}}$ o_b. Similarly, full containment between o_a and o_b exists when:

$$(\exists M_i \in M_a \cap M_b) \wedge (P_a \subseteq P_b) \wedge (\forall p_i \in P_a \cap P_b : h_a^i \vdash h_b^i)$$

That is, an observation o_a fully contains o_b when (i) there is one M_i shared between o_a and o_b, (ii) the dimensions of o_a are a subset of the dimensions of o_b and (iii) values of all dimensions for o_a are hierarchical ancestors of the *respective* dimension values for o_b. We denote this with $\text{Cont}_{\text{full}}(o_a, o_b)$ or equivalently o_a $\text{Cont}_{\text{full}}$ o_b. Observe by definition that the containment property is not symmetric and that given o_a $\text{Cont}_{\text{full}}$ o_b we cannot derive that o_b $\text{Cont}_{\text{full}}$ o_a. Based on the above definitions, our problem can be outlined as follows.

Problem: Given a set of source datasets *D*, and *O* the set of observations in *D*, for each pair of observations o_i, $o_j \in O$, $i \neq j$, assess whether a) o_i $\text{Cont}_{\text{full}}$ o_j, b) o_i $\text{Cont}_{\text{partial}}$ o_j and c) o_i OC o_j. In the following section, we provide our techniques for computing these properties.

4 Computing containment and complementarity properties

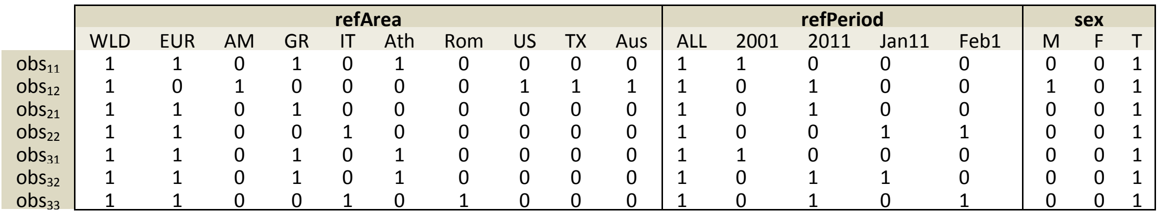
Our technique for computing containment and complementarity properties considers that observations are data vectors in a multidimensional feature space composed of all schema definitions and coded list values in *D*. In addition, the feature set is enriched with all ancestor values in the hierarchy of each coded list, up to the higher common ancestor of all values in *D*. This is represented by an occurrence matrix that captures occurrence (1 or 0) of a dimension, measure definition and coded list value in the set of observations. The occurrence matrix is used for calculating complementarity properties between two observations. It is, then, used for constructing containment matrices that are used for calculating containment properties.

4.1 Constructing the Occurrence Matrix

Each o_i defines a bit vector $\mathbf{o_i}$ of |C|+|P|+|M| dimensions and all $o_i \in O$ yield a |O|x|C+P+M| occurrence matrix **OM** that consists of the following sub-matrices:

- $\mathbf{OM_C}$ is the |O|x|C| matrix defined by the occurrences of coded list values in the respective dimension values of all observations. Each value $c_{ji} \in C_j$ corresponding to dimension p_j is treated as a feature, i.e., a column in $\mathbf{OM_C}$. Hierarchical containment is encoded into $\mathbf{OM_C}$ using a bottom-up algorithm that places a value of 1 in column c_{ji} if the value h_a^j of the dimension p_j of o_a is equal to the feature c_{ji} and then gives the value of 1 to all columns corresponding to the parents of c_{ji}. Finally, we fill with 1's the c_{jroot} of all observations that do not contain p_j in their schema.
- $\mathbf{OM_P}$ and $\mathbf{OM_M}$ are the |O|x|P| and |O|x|M| matrices defined by the occurrences of dimension and measure properties in each observation. Each dimension and measure definition is considered as a feature and is marked with 1 if o_a contains it and 0 if not. The measure values are not taken into account. Therefore, **OM**= [$\mathbf{OM_C}$, $\mathbf{OM_P}$, $\mathbf{OM_M}$].

The $\mathbf{OC_C}$ of the example of Fig. 1, given the hierarchies shown in Figure 2, is shown in Table 2. The sub-matrix $\mathbf{OM_C}$ can be further broken down in separate sub matrices for each coded list, i.e., $\mathbf{OM_C}$ = [$\mathbf{OM_{C1}}$, …, $\mathbf{OM_{Ck}}$] where $\mathbf{OM_{Ci}}$ is a sub-matrix that represents occurrences for all values of dimension p_i and k=|C|. Therefore OM becomes [$\mathbf{OM_{C1}}$,…, $\mathbf{OM_{Ck}}$, $\mathbf{OM_P}$, $\mathbf{OM_M}$].



	refArea										refPeriod					sex		
	WLD	EUR	AM	GR	IT	Ath	Rom	US	TX	Aus	ALL	2001	2011	Jan11	Feb1	M	F	T
obs_{11}	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	1
obs_{12}	1	0	1	0	0	0	0	1	1	1	1	0	1	0	0	1	0	1
obs_{21}	1	1	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	1
obs_{22}	1	1	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	1
obs_{31}	1	1	0	1	0	1	0	0	0	0	1	1	0	0	0	0	0	1
obs_{32}	1	1	0	1	0	1	0	0	0	0	1	0	1	1	0	0	0	1
obs_{33}	1	1	0	0	1	0	1	0	0	0	1	0	1	0	1	0	0	1

Table 1: Sub-matrix OM_C for the example of Fig 1.

4.2 Pair-wise observation containment

Containment matrices. A containment matrix $\mathbf{CM_i}$ is a |O|x|O| bit-vector matrix that captures pair-wise containment relationships between observations, for each dimension p_i. If cell $CM_i[o_a,o_b] > 0$ then observations o_a and o_b are hierarchically related for dimension $p_i \in P$ (e.g. o_a refers to Greece and o_b refers to Europe), while $CM_i[o_a,o_b]= 0$ holds otherwise.

Computation of containment matrices. To compute a containment matrix $\mathbf{CM_i}$, we first take each bit array $\mathbf{OM_{Ci}}$ in $\mathbf{OM_C}$ and we apply a containment function *sf* for all pairs of observations o_a and o_b by considering the rows of o_a and o_b in $\mathbf{OM_{Ci}}$ as two bit vectors, *a* and *b*, respectively. Then, two observations are hierarchically related if the bit-wise AND operation between their corresponding bit vectors yields one of the two bit vectors as shown in [12]. Given this, we define *sf* for a pair of observations o_a, o_b and their bit-vectors a,b resp. in $\mathbf{OM_{Ci}}$ as the following conditional function:

$$sf(o_a,o_b)|_{C_i} = \begin{cases} 1, & (a\ AND\ b) = b \\ 0, & otherwise \end{cases}$$

$sf(o_a,o_b)|_{C_i}$ means that we apply *sf* for o_a and o_b in $\mathbf{OM_{Ci}}$ and if the AND between a and b gives the bit vector b, then o_a is contained by o_b. If a=b then the relationship still holds. By applying this function for each sub matrix in $\mathbf{OC_C}$ we acquire a set of *k* containment matrices of |O|x|O| dimensions, $\mathbf{CM_1}$, …, $\mathbf{CM_k}$, each capturing containment information for a given dimension property. Then, addition of all $\mathbf{CM_i}$ yields the *Overall Containment Matrix* **OCM**, which holds full and partial relationships in the form of normalized similarities between pairs of observations as follows:

$$\boldsymbol{OCM} = \frac{\sum_{i=1}^{k} u_i CM_i}{\sum_{i=1}^{k} u_i}$$

OCM retains values in the range of [0,1]. Full containment is given when two observations have similarity 1, i.e., o_a $Cont_{full}$ o_b iff cell $OCM[o_a, o_b]$=1. Partial containment is given when two observations have similarity between 0 and 1, i.e., o_a $Cont_{partial}$ o_b

iff OCM[o_a, o_b]>0. Finally, no containment between observations exists when OCM[o_a, o_b]=0.

4.3 Pair-wise observation complementarity

Following the definition of observation complementarity, we use **OM** to assess whether the dimension values between pairs of observations are the same, given that $P_a \subseteq P_b$ holds. To check if $P_a \subseteq P_b$ is true between two observations o_a and o_b we apply a bit-wise AND to the bit vectors a, b of o_a and o_b the same way as when trying to compute **CM** matrices, by using function *sf*(o_a, o_b). This is because values in $\mathbf{OM_P}$ capture whether an observation has a dimension in its schema. Given this, we want to assess whether two bit vectors in $\mathbf{OM_P}$ are related via a containment property, which justifies the use of function *sf*, this time on the matrix of the dimension properties. Then, we check if the dimension values of the two observations are equal:

$$cf(o_a, o_b) = \begin{cases} 1, & (sf(o_a, o_b)|_P = 1) \text{ AND } (\mathrm{a} = b) \\ 0, & otherwise \end{cases}$$

$sf(o_a, o_b)|_P$ means that we apply *sf* for o_a and o_b in $\mathbf{OM_P}$, while a and b are the bit vectors in $\mathbf{OM_C}$. This results in a |O|x|O| Complementarity matrix that gives the value of 1 for complementary observations and 0 otherwise, i.e., o_a Compl o_b iff Complementarity [o_a, o_b]>0.

Observe that the time complexity of computing containment and complementarity matrices for N observations is O(N^2), further optimization is left as future work. Two approaches to improve running time are (i) parallelism and (ii) reducing the search space by taking into account characteristics of the incoming schemata. For example, if we can conclude that there can be no containment and/or complementarity between observations of D_i and D_j just by examining their schema, we do not need to perform any computations on observation pairs between D_i and D_j.

5 Complementarity and containment properties in Data Cube

We propose simple extensions to the Data Cube Vocabulary such that complementarity and containment, full and partial, between observations can be represented. We define three properties, containment, partialContainment and fullContainment, where partialContainment is a sub-property of the generic containment property, and fullContainment is a sub-property of partialContainment, which reflects the fact that full containment is a specialization of partial containment. As an example, a relationship $Cont_{partial}(o_a, o_b)$ is then modelled as shown in the top part of Figure 3. The containment relationship becomes a blank node of the appropriate type and is reified to include information on o_b and other possible metadata on the relationship. Similarly, complementarity is denoted with the property imis:complement, as shown in the bottom part of Figure 3.

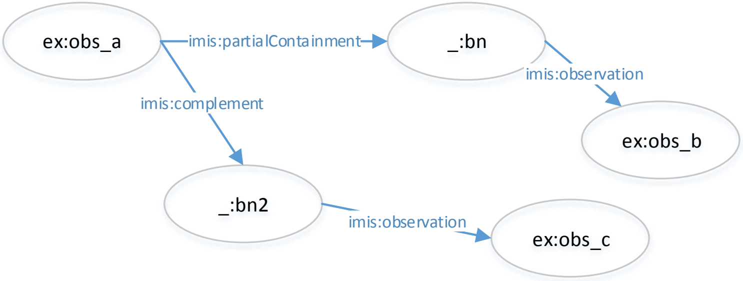


Fig. 3. o_a partially contains o_b. Also, o_a complements o_c.

6 Experimental Evaluation and Discussion

In this section we present the evaluation of our approach over real-world statistical datasets. Our experiments were performed using Java and Apache Jena for handling the RDF models and creating the matrices, and R for computations on the matrices, application of the functions *sf* and *cf* and so on.

Datasets. Four datasets have been used. D_1 and D_2 measure poverty in two different sets of EU sub-regions and periods, D_3 measure population in three EU countries and their first-level sub-regions and D_4 measure households with internet access in seven EU countries and their sub-regions. They exhibit an overlap of 5 dimensions (location, time, sex, unit and age) and 3 measures (poverty, population and households with internet access). The mappings between the coded lists for the common dimension values have been manually created. Observe that creating the mappings for different dimension values is an orthogonal work to our approach; many approaches from the fields of entity resolution and LOD interlinking can be applied. The datasets are either downloaded as RDF or converted using a conversion script written in Java. Eurostat Linked Data Wrapper[1], the Eurostat database[2] and World Bank[3] were used as sources. The datasets where pre-processed to include data about EU countries, as well as selected sub-regions based on official EU geo classifications (NUTS[4]), for various time periods, taken from the Gregorian calendar classification of data.gov.uk[5]. The dataset structures are summarized in Table 2.

#of obs.	*refArea*	*refPeriod*	*sex*	*unit*	*age*	*poverty*	*internet*	*population*
D_1 (539)	85 regions,20 countries	2004-2011	N/A	Yes	Yes	Yes	N/A	N/A
D_2 (1693)	293 regions, 33 countries	2003-2010	N/A	Yes	N/A	Yes	N/A	N/A
D_3 (629)	42 regions, 3 countries	2009-2013	M, F, Total	Yes	N/A	N/A	N/A	Yes

[1] http://estatwrap.ontologycentral.com/
[2] http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database
[3] http://data.worldbank.org/
[4] http://nuts.geovocab.org/
[5] http://datahub.io/dataset/data-gov-uk-time-intervals

D_4 (316)	65 regions,7 countries	2009-2013	N/A	N/A	N/A	N/A	Yes	N/A

Table 2: Dimensions and measures of the input datasets. Measures are marked in grey

Computing containment and complementarity properties over the observations of the four datasets resulted in the creation of multiple relationships summed up in Table 3. The results do not include self-containing or self-complementing observations. We have defined three metrics, **full**, **partial** and **compl**. For a pair of datasets D_i and D_j, they measure the total number of pairs that exhibit full containment, partial containment and observation complementarity respectively, as a percentage of the total possible number of pairs in the given two datasets, minus the diagonal. As can be seen, most new relationships are partial containments, which is a reasonable result given that it is the most weakly defined relationship in terms of its prerequisites. The strictest relationship, observation complementarity, resulted in linking 0.03% of the total possible observation pairs. Sample observations participating in the newly created relationships can be seen in fig 4. The created links are modeled after our proposed vocabulary and uploaded in RDF form in an Openlink Virtuoso store[6].

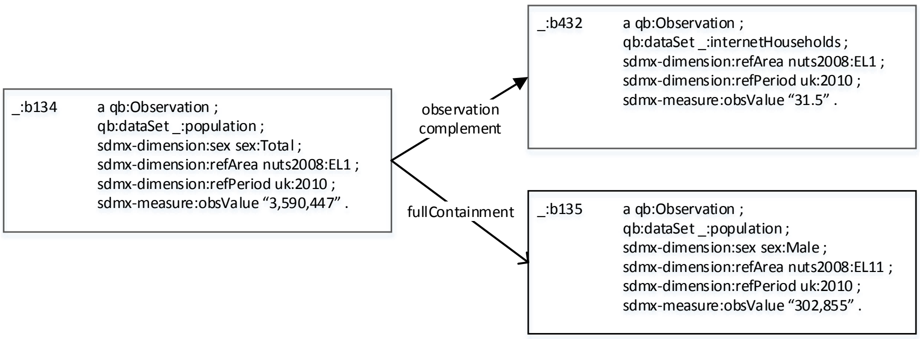
	D_1	**D_2**	**D_3**	**D_4**
D_1	**647 (0.31%) full** **34.3k (16.32%) partial** N/A compl	N/A full N/A partial N/A compl	N/A full N/A partial N/A compl	N/A full N/A partial N/A compl
D_2	**605 (0.02%) full** **605k (14.83%) partial** **1238 (0.04%) compl**	**3370 (0.14%) full** **378k (14.83%) partial** N/A (complement	N/A full N/A partial **204 (0.004%) compl**	N/A full N/A partial N/A compl
D_3	N/A full N/A partial N/A compl	N/A full N/A partial N/A compl	**1k (0.26%) full** **261k (65.9%) partial** N/A compl	N/A full N/A partial N/A compl
D_4	N/A full N/A partial **328 (0.05%) compl**	N/A full N/A partial **218 (0.005%) compl**	N/A full N/A partial **592 (0.07%) compl**	**437 (0.17%) full** **22.2k (22.3%) partial** N/A compl

Table 3: Results of new relationships for test datasets D_1...D_4. Each cell [i,j] contains information on the total number of pairs exhibiting each relationship, as well as the percentage over all possible pair-wise combinations of observations for the combination

Discussion. The relatedness properties that we have defined yield interesting information on how existing observations can be combined across datasets as well as in the same source dataset. The advantages of this work are two-fold, first it creates new information by combining existing facts (complementarity) and second it creates a containment graph among observations that helps exploration, aggregation and discovery of nearby multidimensional observations. The definitions that we have provided can also be studied in terms of their impact at the dataset level, for example, D_3 has 65.9% partial containment in itself, with only 0.26% full containment. This hints that the structure of the dataset's hierarchies is such that there are a few top level concepts

[6] http://83.212.121.252/sparql

in comparison to lower-level concepts, and that also the depth of the hierarchies is not very large. As a matter of fact, this is true for D_3 as it addresses 3 countries and 42 sub-regions, all of them lying on the same level. Figure 4 shows an example of an observation *_:b134* that exhibits both complementarity and containment with *_:b432* and *_:b135* resp. Complementarity shows that we can combine the measures of two observations set on the same dimension values in order to complement their information. Two complementary observations can be joined on their common dimension values and combined into a new observation, whose schema contains the union of their dimension and measure sets. Figure 4 shows how we can combine population and households with internet access because they are both set on region EL1 in 2010 although *_:b432* does not involve gender dimension.



```
_:b134      a qb:Observation ;
            qb:dataSet _:population ;
            sdmx-dimension:sex sex:Total ;
            sdmx-dimension:refArea nuts2008:EL1 ;
            sdmx-dimension:refPeriod uk:2010 ;
            sdmx-measure:obsValue “3,590,447” .
```

```
_:b432      a qb:Observation ;
            qb:dataSet _:internetHouseholds ;
            sdmx-dimension:refArea nuts2008:EL1 ;
            sdmx-dimension:refPeriod uk:2010 ;
            sdmx-measure:obsValue “31.5” .
```

```
_:b135      a qb:Observation ;
            qb:dataSet _:population ;
            sdmx-dimension:sex sex:Male ;
            sdmx-dimension:refArea nuts2008:EL11 ;
            sdmx-dimension:refPeriod uk:2010 ;
            sdmx-measure:obsValue “302,855” .
```

Fig. 4. Containment and complementarity relationship samples of the test results. Both relationships originate from the same observation _:b134.

7 Conclusions and Future Work

In this paper, we have presented a novel approach for identifying and modeling relationships between observations of multidimensional linked open data. We have defined three new properties, namely full and partial observation containment and observation complementarity between two observations as derivatives of the hierarchical relationships between their dimension values, and as a means of comparison and correlation of their different measures. We have proposed a possible extension on the Data Cube terms for representing these properties and we have provided an evaluation of our approach over real-world statistical datasets.

A future direction concerns techniques for grouping and combining observations into new datasets based on their containment and complementarity properties. Another direction is the incorporation of the Data Cube attributes in our algorithms as these often represent valuable information like units of measure, observation status or folded dimensions and can be used for creating dimension value mappings at the pre-processing stage. Finally, we will explore techniques for optimizing the computation time and scale up the overall performance of our approach.

Acknowledgement. This work has been co-financed by the EU project DIACHRON and by EU and Greek national funds through the Operational Program "Competitiveness and Entrepreneurship" (OPCE II) of the National Strategic Reference Framework (NSRF) - Research Funding Program: KRIPIS.

8 References

1. Aligon, J. et al. 2014. "Similarity measures for OLAP sessions." *Knowledge and information systems* 39, no. 2: 463-489.
2. Cyganiak, R., et al. 2013. "The rdf data cube vocabulary." W3C recommendation.
3. Villazón-Terrazas, B. et al. 2011. "Methodological guidelines for publishing government linked data." In *Linking Government Data*, pp. 27-49. Springer New York.
4. Hsu, K.C., and Li, M.Z. 2011. "Techniques for finding similarity knowledge in OLAP reports." *Expert Systems with Applications* 38, no. 4: 3743-3756.
5. Sathe, G., and Sarawagi, S. 2001. "Intelligent rollups in multidimensional OLAP data." In *VLDB*, vol. 1, pp. 531-540.
6. Giacometti, A., et al. 2009. "Query recommendations for OLAP discovery driven analysis." In *Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP*, pp. 81-88. ACM.
7. Han, J. et al. 1999. "Constraint-based, multidimensional data mining." *Computer* 32, no. 8: 46-50.
8. Aggarwal, C. C., and Yu, P.S. 2001. "Outlier detection for high dimensional data." In *ACM Sigmod Record*, vol. 30, no. 2, pp. 37-46. ACM.
9. Markl, V. et al. 1999. "Improving OLAP performance by multidimensional hierarchical clustering." In *Database Engineering and Applications, 1999. IDEAS'99. International Symposium Proceedings*, pp. 165-177. IEEE.
10. Chaudhuri, S., and Dayal, U. 1997. "An overview of data warehousing and OLAP technology." *ACM Sigmod record* 26, no. 1: 65-74.
11. Vassiliadis, P., and Sellis, T. 1999. "A survey of logical models for OLAP databases." *ACM Sigmod Record* 28, no. 4: 64-69.
12. Aït-Kaci et al. 1989. "Efficient implementation of lattice operations." *ACM Transactions on Programming Languages and Systems (TOPLAS)* 11, no. 1: 115-146.
13. Capadisli, S. et al. 2013. "Linked statistical data analysis." *Semantic Web Challenge*.
14. Wagner, A. et al. 2013. "Discovering related data sources in data-portals." In Proceedings of the First International Workshop on Semantic Statistics, co-located with the the International Semantic Web Conference.
15. Das Sarma, A. et al. 2012. "Finding related tables." In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 817-828. ACM.
16. Baikousi, E. et al. 2011. "Similarity measures for multidimensional data." In *Data Engineering (ICDE), 2011 IEEE 27th International Conference on*, pp. 171-182. IEEE.
17. Ngomo, N., and Auer, S. 2011. "LIMES: a time-efficient approach for large-scale link discovery on the web of data." In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three*, pp. 2312-2317. AAAI Press.
18. Volz, J. et al. 2009. "Silk-A Link Discovery Framework for the Web of Data." *LDOW* 538.
19. Mendes, P.N. et al. 2011 "DBpedia spotlight: shedding light on the web of documents." In *Proceedings of the 7th International Conference on Semantic Systems*, pp. 1-8. ACM.