Towards Interoperable Open Statistical Data

Evangelos Kalampokis , Areti Karamanou , and Konstantinos Tarabanis

University of Macedonia, Thessaloniki, Greece

{ekal,akarm,kat}@uom.edu.gr

Abstract. An important part of Open Data is of statistical nature and describes economic and social indicators monitoring population size, inflation, trade, and employment. Combining and analysing Open Data from multiple datasets and sources enable the performance of advanced data analytics scenarios that could result in valuable services and data products. However, it is still difficult to discover and combine open sta- tistical data that reside in different data portals. Although Linked Open Statistical Data (LOSD) provide standards and approaches to facilitate combining statistics on the Web, various interoperability challenges still exist. In this paper, we define interoperability conflicts that hamper com- bining and analysing LOSD from different portals. Towards this end, we start from a thorough literature review on databases and data ware- houses interoperability conflicts. Based on this review, we define inter- operability conflicts that may appear in LOSD. We defined two types of schema-level conflicts namely, naming conflicts and structural conflicts. Naming conflicts include homonyms and synonyms and result from the different URIs used in the data cubes. Structural conflicts result from different practices of modelling the structure of data cubes.

Keywords: Open Data・Linked statistical data・Interoperability

Introduction

During the last years, an increasing number of governments, public authorities, and companies have opened up their data providing a vast amount of Open Data through numerous portals [18]. Today, more than 2600 Open Data portals operate around the globe providing access to Open Data. Open Data promise to offer many benefits to the society including transparency, accountability and economic growth by stimulating the creation of added value data-driven services and products [16].

An important part of Open Data is of statistical nature and describe eco- nomic and social indicators monitoring the population size, inflation, trade, and employment [10]. Statistical data are often described in a multidimensional man- ner. This means that a measure is described based on a number of dimensions, e.g., unemployment rate (measure) for different countries and years (dimensions) [15]. This type of data can be conceptualized as a cube, where the location of a cell is specified by the values of the dimensions, while the value of a cell specifies the measure. We onwards refer to these data as “data cubes” or just “cubes”.

Integrating data from different sources will unleash the full potential of Open Data [26,27,33,35]. This will enable, for example, performing combined analyt- ics on top of data published by different national statistics offices [17]. Linked data has been introduced as a promising paradigm towards this direction, since it facilitates data integration on the Web. In data cubes, linked data has the potential to realize the vision of performing data analytics on top of previously isolated cubes across the Web [19]. An important step towards this direction is the RDF data cube (QB) vocabulary [11], which enables modelling Linked Open Statistical Data (LOSD) in a standardised manner.

Today, many Open Data portals use standard vocabularies, such as QB, to publish LOSD. These include the portals of the Scottish Government, the UK Department for Communities and Local Government (DCLG), the Italian National Institute of Statistics (ISTAT), the Flemish Government, and the Irish Central Statistics Office. However, the flexibility of these vocabularies allows portals to use different practices when applying a vocabulary. As a result, the produced data become non-interoperable and thus isolated in data portals.

The issue of data interoperability is not new but has been raised in the past in the context of traditional databases and data warehouses. In particular, scientific literature (e.g., [8,20,28]) has already investigated

This is a post-print version of the following paper Kalampokis E., Karamanou A., Tarabanis K. (2019) Towards Interoperable Open Statistical Data. In: Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_14

and defined the different types of interoperability conflicts that result from creating different relational models or from having inconsistent data.

The aim of this paper is to define interoperability conflicts that hamper combining and analysing LOSD from different data portals. To this end, we first identify interoperability conflicts of databases and data warehouses using a thorough literature review and, subsequently, map those conflicts to LOSD interoperability conflicts.

The rest of the paper is organized as follows: Sect. 2 presents the approach of this research, Sect. 3 presents the background knowledge required to understand this research, Sect. 4 provides a review of the interoperability conflicts of tradi- tional databases and data warehouses, while Sect. 5 presents the interoperability conflicts of LOSD. Finally, Sect.6 summarizes the results and identifies open research issues.

Research Approach

The research approach of this paper includes the following steps:

- – Step 1: Understand interoperability conflicts in databases and data ware- houses. Towards this end, we conduct a systematic literature review on inter- operability conflicts of database and data warehouses based on the state-of- the-art analysis method proposed by Webster and Watson [37]. According to this method, we initially perform a systematic search to accumulate a set of relevant scientific papers. Then, we perform a concept-centric analysis on the papers to extract a list of interoperability conflicts and their definitions.
- – Step 2: Define LOSD interoperability conflicts based on the results of Step 1.

3 Background

This section briefly presents the background knowledge required to understand the contents of this paper. In particular, we describe (1) the Data Cube model, that is used to describe multidimensional data, (2) the main concepts of Linked Statistical Data, and (3) an overview of the official data portals that host LOSD.

3.1 The Data Cube Model

The data cube model was introduced to cover the needs of the Online Analytical Processing (OLAP) and data warehouse systems. A data cube has been defined in various ways. However, according to all definitions a data cube comprises: [3,6,12,35] (1) measures, which represent numerical values (e.g., unemployment), and (2) dimensions, which provide contextual information for the measures (e.g., geospatial or temporal dimension). In addition, each dimension has a set of distinct values e.g., a temporal dimension may have values like 2000, 2001 etc. Finally, the dimensions may be hierarchically organized into levels representing different granularities. For instance, the geospatial dimension may have levels like country, region, city etc.

An example of a data cube has one measure (i.e. unemployment) and three dimensions (i.e. year, countries, age group). The distinct values of the dimension “year” are 1999, 2000, and 2001, of the dimension “countries” are GR, EN, and FR, and of dimension age group are 00–24, 25–49, and 50+. All the dimensions have one hierarchical level, however there could be more e.g., the geospatial dimension may have both countries and regions.

3.2 Linked Statistical Data

Linked data are based on the Semantic Web philosophy and technologies and are mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inferencing.

This is a post-print version of the following paper Kalampokis E., Karamanou A., Tarabanis K. (2019) Towards Interoperable Open Statistical Data. In: Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_14

The QB vocabulary [11] is a W3C standard for publishing statistical data on the Web using the linked data principles. The core class of the vocabulary is the qb:DataSet that represents a data cube, which comprises a set of dimen- sions (qb:DimensionProperty), measures (qb:MeasureProperty), and attributes (qb:AttributeProperty). Attributes are used to represent structural metadata such as the unit of measurement. Finally a data cube has multiple qb:Observation that describe the cells of the data cube.

It is a common practice to re-use predefined code lists to populate the dimen- sion values. For example, the values of a geospatial dimension can be populated by a code list defining the geographical or administrative divisions of a country. The code lists can be specified using either the QB vocabulary or the W3C stan- dard Simple Knowledge Organization System (SKOS) [24] vocabulary. The val- ues of the code list may also include hierarchical relations which can be expressed using the SKOS vocabulary (e.g., using the skos:narrower property), the QB vocabulary (e.g., using the qb:parentChildProperty) or the XKOS[2] vocabulary (e.g., using the xkos:isPartOf property).

Finally, the UK Government Linked Data Working Group[3] has developed a set of common concepts (e.g., dimensions, measures, attributes, and code lists) that can be reused. The definitions of these concepts are based on the SDMX guidelines[4]. For example, dimensions like sdmx:timePeriod, sdmx:refArea, and sdmx:sex, and measures like sdmx:obsValue have been proposed. Although these resources are not part of the QB vocabulary, they are currently widely used.

3.3 Portals with Linked Statistical Data

Today, a large volume of Linked Statistical Data is provided on the Web through dedicated data portals. For example, the Scottish Government provides official data on "Neighborhood Statistics" as Linked Statistical Data. In particular, they provide access to 238 data cubes categorized to 18 themes such as housing and transport. In addition, the UK's Department for Communities and Local Government (DCLG) provides statistical data that describe various indicators including local government finance and housing and homelessness. In particular, they provide access to 167 data cubes categorized to 14 themes (e.g., homeless- ness, societal well-being). The environmental department of the Flemish govern- ment also provides nine data cubes that describe environmental data as Linked Statistical Data, while the portal site of the Official Statistics in Japan (e-Stat) provides 78 data cubes from seven sources of statistics such as a population census, an economic census, and a labor force survey [1]. Finally, the Italian National Institute of Statistics (ISTAT) and the Irish Central Statistics Office have published as Linked Statistical Data the Italian Census 2011 (8 cubes) and the Irish Census 2011 (682 cubes) respectively.

A large volume of Linked Statistical Data has also been published by third parties activities (unofficial). For example, a linked data transformation of Euro- stat's data[5], which was created in the course of a research project, includes more than 5,000 cubes. Moreover, few statistical datasets from the European Central Bank, World Bank, UNESCO and other international organizations have been also transformed using the QB vocabulary in a third party activity [7].

4 A Systematic Literature Review on Interoperability Conflicts of Databases and Data-Warehouses

Based on the method described in Sect. 2 and after limiting out irrelevant papers, we resulted in 18 papers. We analysed these papers following a concept-centric approach i.e., we synthesized the literature by grouping and studying the main identified interoperability conflicts. After excluding conflicts that are not appli- cable to Linked Statistical Data, we resulted in two types of conflicts namely schema conflicts and data conflicts [4,20,28]. Schema conflicts are further clas- sified into naming and structural conflicts. Data conflicts are further classified into scaling, precision, representation, and data value conflicts. Table 1 presents the scientific papers that are related to each type of conflict.

Table 1. Concept-centric analysis of the literature

This is a post-print version of the following paper Kalampokis E., Karamanou A., Tarabanis K. (2019) Towards Interoperable Open Statistical Data. In: Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_14

	Schema conflicts			Data conflicts			
	Naming	Structural		Scaling	Precision	Representation	Data value
		Schematic discrep.	Isomorph.				
Tseng [35]	✓	-	-	✓	-	-	✓
Kim [20]	✓	✓	✓	✓	✓	✓	✓
Berger [4]	✓	✓	✓	✓	✓	-	✓
Ram [28]	✓	-	✓	✓	✓	✓	✓
Reddy [29]	✓	✓	-	✓	✓	-	-
Batini [2]	✓	✓	-	-	-	-	-
Sheth [31]	✓	✓	✓	✓	✓	✓	✓
Channah [8]	✓	-	-	-	✓	-	-
Doan [14]	✓	✓	-	-	-	-	✓
Bruckner [5]	✓	-	-	-	-	-	-
Spaccapietra [32]	✓	✓	-	✓	-	-	-
Lee [21]	-	✓	✓	✓	✓	✓	-
Lee [22]	✓	✓	✓	✓	✓	✓	-
Sboui [30]	✓	-	-	✓	-	-	-
Mangisengi [23]	✓	-	-	-	-	-	-
Diamantini [13]	✓	-	✓	✓	-	-	✓
Neumayr [25]	-	-	-	✓	-	-	-
Torlone [34]	✓	-	-	✓	-	-	✓

Based on the Entity Relationship (ER) model [9], entities, relations and attributes are the main components that can be used to model data. Schema-level conflicts result from using the components of data in different ways. Data-level conflicts result from incompatible or inconsistent data.

In the rest of the section we describe the above type of conflicts. To this end, we need to define the similarity relationship between two terms. In particular, two terms are “semantically similar” if they refer to the same concept, while they are “semantically unrelated” if they refer to different concepts. In order to support the description of the conflicts, we use the example presented in Fig. 1. The example presents two datasets (both schema and data) that describe a com- pany’s sales. The schema of dataset 1 contains three entities: “product”, “sales” and “date”, while the schema of dataset 2 contains two entities: “product” and “sales”. Accordingly, dataset 1 includes three tables, while dataset 2 includes two tables.

This is a post-print version of the following paper Kalampokis E., Karamanou A., Tarabanis K. (2019) Towards Interoperable Open Statistical Data. In: Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_14

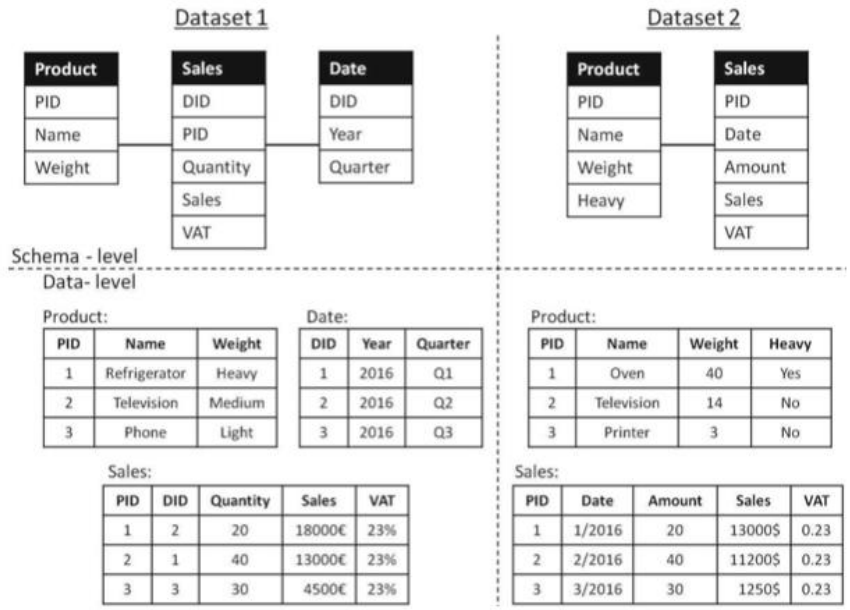

Fig. 1. Example of two database schema and data

4.1 Schema-Level Conflicts

Schema-level conflicts are classified into naming and structural conflicts. Naming conflicts. Various names are used for the components of a dataset's schema [29] resulting in a proliferation of names as well as possible conflicts among them. There are two types of schema-level conflicts [2,8,14,20,29,31]:

– Homonyms. This type of conflict results from two semantically unrelated components that have the same name. In our example, a homonym conflict results from the fact that "Weight" of entity "Product" refers to total weight in dataset 1 and in net weight in dataset 2.

– Synonyms. This type of conflict results from two semantically similar com- ponents that have different names. For example the attribute "Quantity" of entity "Sales" in dataset 1 and the attribute "Amount" of the same entity in dataset 2 have different names although they refer to the same term. Multiple languages can also cause synonym problems, e.g., week (English) and woche (German) refer to the same concept using different language [5].

Structural conflicts. This type of conflict occurs when two semantically similar components use different modeling approaches [2]. In particular, there are two types of structural conflicts [20–22,29,31,32]:

– Schematic discrepancies. This type of conflict occurs when the logical structure of a set of attributes and their values belonging to an entity class in one schema are organized to form a different structure in another schema [28,31]. For example, in dataset 1, "date" is an entity while in dataset 2 an attribute of the "Sales" entity. A specific case of this conflict is defined in [31], where the value of an attribute in one case corresponds to an attribute in another case. For example, in dataset 1, the attribute "weight" of the entity "Product" has the value "heavy" in the first record of the table that corresponds to an attribute of the second dataset's "Product" entity.

– Schema Isomorphism. This type of conflict occurs when two semanti- cally similar entities have different number of attributes [20,21,28,31], e.g., in dataset 2 the "Product" entity has one extra attribute (i.e. "Heavy") related to the same entity at dataset 1.

4.2 Data-Level Conflicts

This is a post-print version of the following paper Kalampokis E., Karamanou A., Tarabanis K. (2019) Towards Interoperable Open Statistical Data. In: Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_14

Data-level conflicts are classified into data scaling, data precision, data repre- sentation, and data value conflicts.
Data scaling conflicts. They result from data that are stored in semantically similar attributes and use different units of measure [20,22,29–31]. For instance, “sales” in dataset 1 are measured in euros and in dataset 2 in dollars.

Data precision conflicts. They result from data stored in semantically sim- ilar attributes and use different precisions [20,21,29,31]. For example, the “weight” attribute of the Product entity (dataset 1) includes values like “heavy”, “medium”, and “light” while in dataset 2 the “weight” of the Product is mea- sured in kilograms. Moreover, different levels of accuracy may be used e.g., weight can be measured up in milligrams or in grams.

Data representation conflicts. In some cases, although data stored in seman- tically similar attributes have the same unit of measure and precision, they have different formats [20–22,28,31] resulting in data representation conflicts. For example, attribute “VAT” of Sales (dataset 1) is a percentage (e.g., 23%) while in dataset 2 it is a decimal value (e.g., 0.23). Although different formats are used, both values are equivalent. Another example could be the date attributes that may use different formats, e.g., “dd/mm/yy” vs “mm/dd/yyyy’.

Data value conflicts. They result from data have measurements with con- flicting values [14,28,31,34,35]. For example, the Television sales for Q1 2016 in dataset 1 are 13000e while in dataset 2 $11200 (the values are conflicting even after converting euros to dollars). Such conflicts occur due to wrong or obsolete data or when different statistical methods are employed [20].

5 Interoperability Conflicts of Linked Open Statistical Data

In this section we define the conflicts of the literature of traditional databases and data warehouses in the context of LOSD. We define only schema-level conflicts as data-level conflicts depend on the specific values of the data cubes.

5.1 Naming Conflicts

One of the principles of linked data is to name things using URIs. In the case of LOSD, naming conflicts may result from the URIs used for the dimensions, measures, measure units, dimension levels, and dimension values of the data cubes. A common practice in linked data is to reuse standardized vocabularies and create new vocabularies only when required [36]. Naming conflicts mainly occur because some linked data portals reuse standardized vocabularies, while other create their own vocabularies. For example, the SDMX vocabulary is com- monly used by most linked data portals but not by all of them. The two types of naming conflicts in LOSD are Homonyms and Synonyms.

Homonym conflicts result from using the same URI to represent semanti- cally unrelated elements (e.g., the measure property) of data cubes. In partic- ular, LOSD publishers may use the same URI for semantically different mea- sure properties. For example, sdmx-measure:obsValue is often used to represent semantically unrelated measures (e.g., unemployment, poverty etc.).

Synonym conflicts result from using different URIs for semantically similar elements. In particular, synonym conflicts in LOSD result from:

- – Using different URIs for semantically similar measure properties. For exam- ple, some data portals define a new measure property to measure unemploy- ment (e.g., test:unemployment), which is rdfs:subPropertyOf sdmx:obsValue. Other portals, however, define a new measure property that is not related to the sdmx:obsValue.
- – Using different URIs for semantically similar dimension properties. For example, a common practice for the common dimensions (e.g., temporal, geospatial, gender, and age) is to re-use the dimension defined by SDMX (e.g., sdmx:refArea for the geospatial dimension). However, other practices may define a new dimension property (e.g., eg:geo) instead of re-using SDMX.

This is a post-print version of the following paper Kalampokis E., Karamanou A., Tarabanis K. (2019) Towards Interoperable Open Statistical Data. In: Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_14

- – Using different URIs for semantically similar hierarchical data (i.e., relations between data, and levels of hierarchies). For example, some data portals may use the dcterms:isPartOf and dcterms:hasPart to represent hierarchical rela- tions while others may define new URIs.

– Using different URIs for the code lists (i.e, the code list for the unit of measure, the temporal dimension, and of the gender dimension). For example, two alternative practice could be to use the QUDT vocabulary or the DBpedia vocabulary for the unit of the measure.

5.2 Structural Conflicts

Structural conflicts are directly related to the structure of the data cubes. The two types of structural conflicts in LOSD include Schema isomorphism and Schematic discrepancies.
Schema isomorphism conflicts result from defining different number of compo- nents (i.e. dimension, measure, attribute) in semantically similar data cubes. In particular, schema isomorphism conflicts result from:

- – Using different practices to model the measure and its parameters in seman- tically similar data cubes. For example, some data portals may define just the measure of the data cube, while others may define both the measure and the measure type (qb:measureType).
- – Defining different number of measures in semantically similar data cubes. For example, different practices could be to define a single or multiple measures per data cube.
- – Using different practices to model the unit of measure and its parameters in semantically similar data cubes. For example, some data portals may define only the unit of the measure while other may define also the unit multi- plier, which is used to indicate the magnitude in the units of measurements (e.g., hundreds, thousands, tens of thousands etc.).
- – Defining different number of units per measure in semantically similar data cubes. For example, data portals may publish several data cubes with a single or one data cube with multiple units per measure.
- – Using different practices to define hierarchical levels. For example, some data portals may define one data cube to measure all hierarchical levels, while others may define one cube per hierarchical level.

  Schematic discrepancies conflicts result from using different logical constructs to represent the same set of data cube components (e.g., when different practices are used to define more than one measure in a data cube). In particular, schematic discrepancies result from:

– Modelling semantically similar data cubes with multiple measures in differ- ent ways. For example, one practice is to define multiple qb:MeasureProperty components in the qb:DataStructureDefinition of the data cube (one for each measure), an instance of a single measure component in each observation, and an extra qb:measureType dimension that denotes the measure used in the observation. An alternative practice is to define multiple qb:MeasureProperty components in the qb:DataStructureDefinition of the data cube and also an instance of each defined measure component in each observation.

- – Defining the unit of the measure in different levels of semantically similar data cubes. For example, a data portal practice defines the unit of measure in the qb:Observation level, while another practice defines the unit of measure in the qb:MeasureProperty level.
- – Defining the single value dimensions (i.e., dimensions with the same value) of semantically similar data cubes in a different way. For example, a practice could be to define the single value dimension in the qb:Dataset level, and an alternative to define the dimension in the qb:Slice level.
- – Defining aggregated values in a different way. For example, a practice could be to use a hierarchy and define total values on the top of the hierarchy, while another practice to define a unique total URI that could be used in every case (e.g., sdmx:total).
- – Defining the values of the temporal dimension of semantically similar data cubes in a different way. For example, some data portals define the value of the temporal dimension along with its data type, i.e., “2011”^^xsd:date while others define only the value of the temporal dimension (e.g., 2011).

This is a post-print version of the following paper Kalampokis E., Karamanou A., Tarabanis K. (2019) Towards Interoperable Open Statistical Data. In: Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_14

- – Using different ways to associate dimensions with potential values. For exam- ple, some data portals use the qb:codeList property, while others define the rdfs:range of the qb:DimensionProperty as a skos:Concept.

6 Conclusion and Future Work

Interoperability among data cubes is crucial to unleash the full potential of linked statistical data. For example, it will enable performing combined analytics and visualizations on data published by different national statistics offices and other organisations. Currently, all official portals that publish linked statistical data are using the QB vocabulary, however they adopt different practices thus hampering the interoperability among their data.

In this paper we define the interoperability conflicts that hamper combining and analysing LOSD from different data portals. Our study was based on a thorough literature review on databases and data warehouses interoperability conflicts. We defined two types of schema-level conflicts namely, naming conflicts and structural conflicts. Naming conflicts include homonyms and synonyms and are mostly result from the URIs that are used in the data cubes. Structural conflicts result from different practices of modelling the structure of data cubes.

This work used a top-down approach to identify interoperability conflicts of LOSD. In the future, we plan to follow a bottom-up approach and study data cubes of LOSD data portals in order to understand the different practices they use to publish LOSD and, hence, result in interpretability conflicts.

Acknowledgments. This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Program "Human Resources Development, Education and Lifelong Learning 2014–2020" in the context of the project "Integrating open statistical data using semantic technologies" (MIS 5007306).

References

1. Asano, Y., Takeyoshi, Y., Matsuda, J., Nishimura, S.: Publication of statistical linked open data in Japan. In: Proceedings of the 4th International Workshop on Semantic Statistics Co-Located with 15th International Semantic Web Conference (ISWC 2016). CEUR Workshop Proceedings (2016)
2. Batini, C., Lenzerini, M., Navathe, S.B.: A comparative analysis of methodologies for database schema integration. ACM Comput. Surv. 18(4), 323–364 (1986)
3. Berger, S., Schrefl, M.: From federated databases to a federated data warehouse system. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences, pp. 394–394. IEEE (2008)
4. Berger, S., Schrefl, M.: FedDW global schema architect: UML-based design tool for the integration of data mart schemas. In: Song, I.Y., Golfarelli, M. (eds.) DOLAP, Maui, Hawaii, USA, pp. 33–40. ACM, November 2012
5. Bruckner, R.M., Ling, T.W., Mangisengi, O., et al.: A framework for a multidi- mensional OLAP model using topic maps. In: Proceedings of the 2nd International Conference on Web Information Systems Engineering 2001, vol. 2, pp. 109–118. IEEE (2001)
6. Cabibbo, L., Torlone, R.: A logical approach to multidimensional databases. In: Schek, H.-J., Alonso, G., Saltor, F., Ramos, I. (eds.) EDBT 1998. LNCS, vol. 1377, pp. 183–197. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0100985
7. Capadisli, S., Auer, S., Ngonga Ngomo, A.C.: Linked SDMX data. Semant. Web 6(2), 105–112 (2015)
8. Channah, N., Aris, O.: A classification of semantic conflicts in heterogeneous database systems. J. Organ. Comput. 5(2), 167–193 (1995)
9. Chen, P.P.S.: The entity-relationship model—toward a unified view of data. ACM Trans. Database Syst. (TODS) 1(1), 9–36 (1976)
10. Cyganiak, R., Hausenblas, M., McCuirc, E.: Official statistics and the practice of data fidelity, pp. 135–151 (2011). https://doi.org/10.1007/978-1-4614-1767-5 7
11. Cyganiak, R., Reynolds, D.: The RDF data cube vocabulary: W3C recommenda- tion, January 2014
12. Datta, A., Thomas, H.: The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses. Decis. Support. Syst. 27(3), 289–301 (1999)
13. Diamantini, C., Potena, D., Storti, E.: Data mart reconciliation in virtual innova- tion factories. In: Iliadis, L., Papazoglou, M., Pohl, K. (eds.) CAiSE 2014. LNBIP, vol. 178, pp. 274–285. Springer, Cham (2014). https://doi.org/10.1007/978-3-319- 07869-4 26
14. Doan, A., Halevy, A.Y.: Semantic integration research in the database community: a brief survey. AI Mag. 26(1), 83–94 (2005)

This is a post-print version of the following paper Kalampokis E., Karamanou A., Tarabanis K. (2019) Towards Interoperable Open Statistical Data. In: Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_14

15. Gnanadesikan, R.: Methods for Statistical data Analysis of Multivariate Observa- tions, vol. 321. Wiley, Hoboken (2011)
16. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, adoption barriers and myths of open data and open government. Inf. Syst. Manag. 29(4), 258–268 (2012). https://doi.org/10.1080/10580530.2012.716740
17. Kalampokis,E.,Tambouris,E.,Tarabanis,K.:Linkedopencubeanalyticssystems: potential and challenges. IEEE Intell. Syst. 31(5), 89–92 (2016)
18. Kalampokis, E., Tambouris, E., Tarabanis, K.: A classification scheme for open government data: towards linking decentralised data. Int. J. Web Eng. Technol. 6(3), 266–285 (2011)
19. Kalampokis, E., Tambouris, E., Tarabanis, K.: Linked open government data ana- lytics. In: Wimmer, M.A., Janssen, M., Scholl, H.J. (eds.) EGOV 2013. LNCS, vol. 8074, pp. 99–110. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642- 40358-3 9
20. Kim, W., Seo, J.: Classifying schematic and data heterogeneity in multidatabase systems. Computer 24(12), 12–18 (1991). https://doi.org/10.1109/2.116884
21. Lee, C., Chen, C.J., Lu, H.: An aspect of query optimization in multidatabase sys- tems. SIGMOD Rec. 24(3), 28–33 (1995). https://doi.org/10.1145/211990.212011
22. Lee, K.H., Kim, M.H., Lee, K.C., Kim, B.S., Lee, M.Y.: Conflict classification and resolution in heterogeneous information integration based on XML schema. In: Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, TENCON 2002, vol. 1, pp. 93–96. IEEE (2002)
23. Mangisengi, O., Huber, J., Hawel, C., Essmayr, W.: A framework for supporting interoperability of data warehouse islands using XML. Data Warehous. Knowl. Discov. 2114, 328–338 (2001). https://doi.org/10.1007/3-540-44801-2 32
24. Miles, A., Bechhofer, S.: SKOS simple knowledge organization system reference: W3C recommendation, August 2009
25. Neumayr, B., Schrefl, M., Thalheim, B.: Hetero-homogeneous hierarchies in data warehouses. In: Song, I.Y., Golfarelli, M. (eds.) Proceedings 7th Asia-Pacific Conference on Conceptual Modelling, Brisbane, Australia, January 2010
26. Pedersen, T., Pedersen, D., Riis, K.: On-demand multidimensional data integra- tion: toward a semantic foundation for cloud intelligence. J. Supercomput. 65(1), 217–257 (2013). https://doi.org/10.1007/s11227-011-0712-3
27. Perez, J., Berlanga, R., Aramburu, M., Pedersen, T.: Integrating data warehouses with web data: a survey. IEEE Trans. Knowl. Data Eng. 20(7), 940–955 (2008). https://doi.org/10.1109/TKDE.2007.190746
28. Ram, S., Park, J.: Semantic conflict resolution ontology (SCROL): an ontology for detecting and resolving data and schema-level semantic conflicts. IEEE Trans. Knowl. Data Eng. 16(2), 189–202 (2004)
29. Reddy, M., Prasad, B.E., Reddy, P., Gupta, A.: A methodology for integration of heterogeneous databases. IEEE Trans. Knowl. Data Eng. 6(6), 920–933 (1994)
30. Sboui, T., B´edard, Y., Brodeur, J., Badard, T.: A conceptual framework to support semantic interoperability of geospatial datacubes. In: Hainaut, J.L., et al. (eds.) ER 2007. LNCS, vol. 4802, pp. 378–387. Springer, Heidelberg (2007). https://doi. org/10.1007/978-3-540-76292-8 44
31. Sheth, A.P., Kashyap, V.: So far (schematically) yet so near (semantically). In: Proceedings of the IFIP WG2: Conference on Semantics of Interoperable Database Systems, Lorne, Victoria, Australia, pp. 283–312, November 1992
32. Spaccapietra, S., Parent, C., Dupont, Y.: Model independent assertions for inte- gration of heterogeneous schemas. VLDB J. 1(1), 81–126 (1992)
33. Torlone, R.: Two approaches to the integration of heterogeneous data warehouses. Distrib. Parallel Databases 23, 69–97 (2008)
34. Torlone, R.: Interoperability in data warehouses. In: Liu, L., O¨zsu, M.T. (eds.) Encyclopedia of Database Systems, pp. 1560–1564. Springer, Boston (2009). https://doi.org/10.1007/978- 0- 387- 39940- 9
35. Tseng, F.S., Chen, C.W.: Integrating heterogeneous data warehouses using XML technologies. J. Inf. Sci. 31(3), 209–229 (2005). https://doi.org/10.1177/ 0165551505052467
36. W3C: Best practices for publishing linked data. W3C Working Group Note (2014)
37. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: writing a literature review. Manag. Inf. Syst. Q. 26(2), 3 (2002)

This is a post-print version of the following paper Kalampokis E., Karamanou A., Tarabanis K. (2019) Towards Interoperable Open Statistical Data. In: Lindgren I. et al. (eds) Electronic Government. EGOV 2019. Lecture Notes in Computer Science, vol 11685. Springer, Cham. https://doi.org/10.1007/978-3-030-27325-5_14