Open Statistics: The Rise of a New Era for Open Data?

Evangelos Kalampokis, Efthimios Tambouris, Areti Karamanou, Konstantinos Tarabanis

Open Statistics: The Rise of a new Era for Open Data?

Evangelos Kalampokis[1,2], Efthimios Tambouris[1,2], Areti Karamanou[1,2], and Konstantinos Tarabanis[1,2]

[1] University of Macedonia, Thessaloniki, Greece
[2] Information Technologies Institute, Centre for Research & Technology – Hellas, Thermi, Greece
{ekal,tambouris,akarm,kat}@uom.gr

Abstract. A large part of open data concerns statistics, such as demographic, economic and social data (henceforth referred to as Open Statistical Data, OSD). In this paper we start by introducing *open data fragmentation* as a major obstacle for OSD reuse. We proceed by outlining data cube as a logical model for structuring OSD. We then introduce *Open Statistics* as a new area aiming to systematically study OSD. Open Statistics reuse and extends methods from diverse fields like Open Data, Statistics, Data Warehouses and the Semantic Web. In this paper, we focus on benefits and challenges of Open Statistics. The results suggest that Open Statistics provide benefits not present in any of these fields alone. We conclude that in certain cases OSD can realise the potential of open data.

Keywords: Open data, statistical data, open statistics, linked data, data analytics

1 Introduction

Today an increasing number of public authorities, international organisations and even enterprises publish Open Data [1, 2]. Open Data refers to data that *can be freely used, re-used and redistributed by anyone*[1]. In the public sector, opening up government data aims to increase transparency and boost economic growth. Indeed, estimates suggest that the potential of Open Data is tremendous [3]. For example, a study conducted by the McKinsey Global Institute estimated the global annual economic potential value of Open Data to $3 trillion [4]. Against this general euphoria however, studies reveal that publishing open data does not automatically provide benefits [5, 6]. Thus, we are still far from suggesting that the potential of open data has been realised. On the contrary, further research is needed in promising areas.

In this respect, an obvious route for further research is to understand the nature of open data. Policy documents and research in the area suggest that

[1] http://opendefinition.org

a large part of open data is numerical and, more specifically, concerns statistics [7]. Examples include demographics (e.g. census data), social data (e.g. on unemployment and poverty), economic data (e.g. number of new businesses) etc. In this paper we refer to these as Open Statistical Data (OSD). The fact that OSD is a large part of open data was the main motivation for our research. OSD are numerical hence can be easily processed and visualised while significant knowledge already exists in areas such as statistics and data warehouses.

In this paper, we present Open Statistics as a new field to systematically investigate OSD and the creation of value from them. Open Statistics reuse methods from diverse fields like Open Data, Semantic Web, Statistics, Machine Learning and Data Warehouses. More specifically, Open Statistics use existing knowledge on Open Data (such as processes and formats used to publish open data) as background environment. In this environment, Open Statistics reuse but, more importantly, in many cases redefines and extends existing methods from other areas, e.g. for data integration, analysis and visualisation. As a result, Open Statistics provide benefits that go much beyond what was possible in each separate field. This suggests that Open Statistics can actually constitute a significant field of research that, under certain conditions, could enable realising the full potential of Open Data.

The research work that we present in this paper is exploratory [8] as we aim to scope out the magnitude of Open Statistics and to provide and initial understanding about it. In general, exploratory research is research conducted for a problem that has not been clearly defined. It often occurs before we know enough to make conceptual distinctions or posit an explanatory relationship [9]. As a result, we include the following activities in our approach:

- Study datasets from two open data portals at different administrative levels. In particular we focus on the UK's national open data portal[2] and the European Union's open data portal[3] and we study statistical datasets related to *unemployment*.
- Review literature related to research areas overlapping with Open Statistics. In particular, we reuse existing knowledge from (a) Open Data because open statistical data is a major part of them, (b) Data Warehouse and Online Analytical Processing (OLAP) because data cube model seems appropriate to conceptualise OSD, (c) Statistics as a valuable way to create value out of OSD, and (d) Linked Data as a vital technological enabler to achieve the full potential of Open Statistics.

The rest of this paper is organised as follows. In section 2 we present the existing situation in OSD. In section 3 we present a major challenge for OSD reuse, namely data fragmentation. In section 4 we outline the data cube model.In section 5 we introduce Open Statistics and show how it is related to other fields of study. In section 6 we illustrate the benefits of Open Statistics while in section 7 we present the relevant challenges. In section 8 we discuss the finding while

[2] `http://data.gov.uk`
[3] `http://www.europeandataportal.eu/`

section 9 presents the main conclusions of the work and directions of future research.

2 Existing Situation

Today, opening government data is a political priority in many countries worldwide including the USA and the EU. As a result, an exponentially increasing amount of government data is rapidly opening. International organizations (such as the World Bank) also open up their data. A five-star model has been proposed by Tim Berners-Lee to evaluate the maturity of open data[4].

More specifically, OSD is currently provided by governments and organisations through data portals at the international, European, national or regional level. At the international level, organisations provide OSD related to countries in data portals such as the World Bank data portal[5], the Organisation for Economic Cooperation and Development (OECD) data portal[6] and the United Nations Educational, Scientific and Cultural Organisation (UNESCO) data portal[7]. At the European level, OSD are provided through the official European Data Portal[8] and the data portal of Eurostat[9]. At the national level, OSD are provided by the national open data portals (e.g. the data.gov.uk in the UK) but also by the National Statistical Offices (e.g. the Office for National Statistics[10] in the UK). Finally, at the regional level, OSD are published by local agencies, cities or even boroughs of cities in local data portals such as the data portal of the city of Brussels[11] and the data portal of the Camden borough of London[12]. Finally, data portals also serve as single points of access and, apart from providing data regarding their administrative level, they also provide links to datasets that are published at data portals of lower levels.

3 Open Statistical Data Fragmentation

As already stated, a large part of open data are numerical thus potentially easy to process and visualise. In reality however studies suggest that open data reuse is limited. In this section, we investigate practical obstacles for OSD reuse. We do not consider obstacles related to legal and organisational issues at the side of the publishers. Instead, we concentrate on the side of the end user, who is interested to reuse open data.

[4] https://www.w3.org/DesignIssues/LinkedData.html
[5] http://data.worldbank.org/
[6] http://stats.oecd.org/
[7] http://opendata.unesco.org/
[8] http://europeandataportal.eu/
[9] http://ec.europa.eu/eurostat/data/database
[10] http://ons.gov.uk
[11] http://opendata.brussel.be/
[12] http://opendata.camden.gov.uk/

For the purposes of this research, we searched two major open government data portals, namely the UK data portal and the European data portal. In both case, according to our scenario, we were interested to reuse open data about *unemployment*.

We first searched the UK data portal for datasets using the keyword *unemployment*. This resulted in 122 results, which provided access to 56 files and 610 links to other portals (e.g. to the UK's Office for National Statistic) and thus to other files. We opened and examined these files one by one and find out that only 13 out of 56 are relevant to unemployment and that 7 out of 13 provide structured numeric values in a machine readable format.

Most importantly, however, those datasets measure unemployment based on different characteristics (also called dimensions). For example, we found datasets for unemployment in different geospatial levels (e.g. in the city of London, in the Camden borough, or in the different wards of Camden), age groups, gender or time duration (e.g. annual, quarterly or monthly unemployment). Relative datasets also measure unemployment using different units of measure (e.g. unemployment rate or thousands of unemployed people). Finally, different datasets may employ different methods for measuring unemployment e.g. based on the UK's Office for National Statistics (OSN) estimations, based on the number of people that claim Job Seekers Allowance (JSA) or based on the International Labour Organization's (ILO) model.

We then searched European data portal using *unemployment* keyword. This search returned 120 datasets mainly from Eurostat. Again, those datasets describe unemployment using different dimensions and units of measures. Different datasets also measure unemployment in different context (e.g. in the context of education and training or in regional statistics). This also means that these datasets are located in different parts of the portals.

In summary, our research revealed that searching the two open data portals for useful data on unemployment results in large numbers of datasets and links.

We call *open data fragmentation* the situation where collections of relevant open data are broken down into many pieces that are not close together. This definition is actually an adaptation of the definition of data fragmentation in computing.

Unemployment is not the only case where relevant data are fragmented. In the case of OSD, fragmentation is actually the rule rather than the exception. Therefore, we suggest that in order for OSD (and therefore Open Data in general) to be useful the problem of open data fragmentation has to be sufficiently addressed. We acknowledge that other obstacles already mentioned in the literature are also important. However, in this paper we concentrate on overcoming the obstacle of open data fragmentation.

4 The Data Cube Model

The study of datasets in both the UK and the European open data portals reveals that (a) OSD can be conceptualised using the traditional data cube (or

just cube) model that was initially introduced in Data Warehouses and that (b) different datasets provide fragmented views of a cube.

Although research in data warehouses is active for more than two decades, concepts and systems lack a uniform theoretical basis with regards to models that define data cubes and operations that are performed on data cubes [10–14]. In general, however, a data cube is specified by a set of dimensions and a set of measures. The dimensions create a structure that comprises a number of cells, while each cell includes a numeric value for each measure of the cube. Let us consider as an example a cube from Eurostat with three dimensions, namely time in years, geography in countries, and age group, that measures the employment rate. An example of a cell in this cube would define that the percentage of unemployed people between 25 and 49 years old in France in 1999 is 10.2 % (Figure 1). This conceptual cube could have been created using numeric values from multiple datasets.

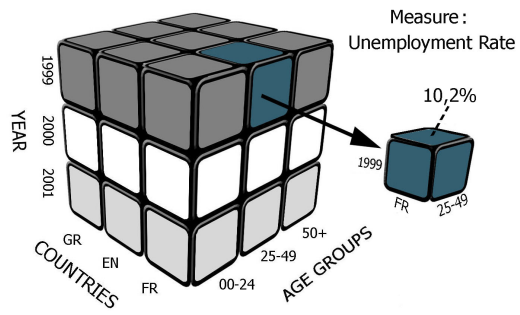

Fig. 1. Open Statistical Data modelled as a cube

5 Open Statistics

Open Statistics is a field aiming to systematically study Open Statistical Data (OSD). The main idea behind Open Statistics is that we concentrate only on Open Data that are actually statistics. This is a large part of all Open Data but clearly not all Open Data are OSD.

Open Statistics mainly capitalise on existing knowledge on Open Data and mainly Open Government Data. The majority of the existing body of knowledge in those areas is applicable in Open Statistics although in some cases some precaution is necessary. In this environment, methods from three other areas are reused and in some cases redefined. These are Data Warehouses and OLAP, Statistics and the Semantic Web (mostly Linked Open Data).

Open Statistics reuse the concept of data cubes for OSD logical organisation. It shows how data are logically connected and not necessarily how they are physically connected.

Open Statistics also reuse Online Analytical Processing (OLAP) methods, such as slicing and dicing. In some cases however those methods are redefined since OLAP was initially introduced in a close environment. In Open Statistics we have the possibility of performing operations not needed before. For example, searching for similar data cubes in the same or different open data portal is an essential operation to overcome open data fragmentation.

Open Statistics involve analysing OSD with statistical methods, such as Pearson's correlation, linear regression, and logistic regression, and techniques such as panel data analysis and even statistical learning analysis in order to explain or predict phenomena. In the context of Open Statistics the exploitation of these methods and techniques will be redefined.

Finally, Open Statistics capitalise on the Linked Open Data technology (LOD) and more specifically on the LOD implementation of the data cube model, termed RDF Data Cube (QB) vocabulary. This provides the necessary technological infrastructure for Open Statistics.

6 Open Statistics Potential

Some of the most valuable methods that are used to exploit data include Online Analytical Processing (OLAP), correlation of cross-sectional data, time-series correlation, panel data analysis and creation of predictive models. These methods can also be used to analyse OSD.

OLAP refers to the technique of performing complex analysis over the information stored in a DW. The multidimensional nature of OSD allows performing OLAP on top of them in order to explore and get different views of the data. For example, OLAP can be used to view only selected part of data (slice or dice), to view a reoriented view of the data (pivot) or to navigate among different levels of the data along a specific dimension (drill-down or roll-up).

Cross-sectional data [15] provide observations of phenomena at a single point of time. Correlation of cross-sectional data can be, hence, used in statistics to measure and interpret the extent to which two measured variables are related to each other within a single point of time. Linear regression is the mostly used method to explore the correlation between two measured variables. Cross-sectional data correlation can be used to assess possible associations between different phenomena described by OSD e.g. unemployment rate and poverty rate in the UK in 2015.

Correlation can be also used to measure and interpret the relationship between measured variables over time. In this case, correlation is applied in data that are modeled as time-series. OSD can be easily formulated as time series data as they usually measure the same phenomenon at successive time intervals. Time-series correlation can then be applied in order to explore the relationship between different phenomena over time e.g. unemployment rate and poverty rate in the UK over the last ten years.

Panel data [16] (or longitudinal data) can be used to model multi-dimensional data over time. Panel data are able to contain observations of multiple phenom-

ena over multiple time periods. Panel data can be used on top of OSD to explore how a measured variable changes over time. For example, panel data can be used to explore the relationship between unemployment rate in all countries of Europe and the poverty rate in all countries of Europe the last ten years.

Predictive models are created and assessed in the context of predictive analytics in order to make empirical predictions using data and statistical or data mining methods [17]. In general, the goal of predictive models is to predict the output of a variable value (Y) for new observations given their input values (X) based on historical data. OSD can be used as historical data for the creation as well as the assessment of predictive models.

According to our view Open Statistics introduce two types of OSD exploitation: the *problem-driven* approach and the *data-driven* approach.

The problem-driven approach follows the traditional data exploitation paradigm that aims to solve a well-defined problem. In this case, one of the main challenges is to discover the appropriate data and Open Statistics can support this task. For example, a problem-driven type of scenario could be the following: *"I would like to explore a phenomenon"*. In this scenario, if we consider as an example the phenomenon of unemployment in European countries, the first requirement towards exploring unemployment would be to discover all relevant OSD. These can be datasets that measure unemployment from different European countries and in various time periods, provided by a single or various data portals. Relevant OSD can be then combined to provide a single view of unemployment and then analysed using different methods of analysis to produce interesting results. OLAP analysis, for example, could be used to view unemployment in Italy, Greece and Spain in years 2014 and 2015.

Another problem-driven type of scenarios could be the following: *"I would like to explore the relationship between two or more phenomena"*. For example, we would like to explore the relationship between unemployment and poverty in the countries of Europe. Towards this end, we need again to discover relevant OSD. Once we have the datasets, we can combine them and then apply on them methods such as cross-sectional correlation, time-series correlation or panel data analysis.

As a result, the most important task in problem-driven scenarios is the discovery of relevant data. In the new reality of Open Statistics the vision is to facilitate the discovery of this relevant data. Hence, the main benefit of Open Statistics in this approach is that it will allow the easy and effective discovery of relative OSD that can be then analysed using the methods of analysis described above in order to solve specific problems.

The data-driven approach is a bottom-up approach compelled by OSD. Specifically, this approach aims at identifying unexplored results starting from a dataset at hand. A data-driven type of scenarios could be the following: *"I would like to explore phenomena out of OSD"*. In order to solve this problem, we could start from a single dataset and then search for relevant datasets, combine and analyse them in order to discover possible relationships or other interesting conclusions.

In data-driven scenarios the benefits of Open Statistics can be even greater since the different methods of analysis that can be applied on OSD need to be redefined. Specifically, starting from a specific dataset, OLAP can be used to enhance this dataset by finding relevant datasets (e.g. that measure the same variable in a different year). This will facilitate the inspiration of innovative solutions or unexpected results that were not known before. Moreover, correlation (cross-sectional or time series) could be used to identify unexpected relationships with other datasets. Starting again from a specific dataset, data-driven correlation and panel data in Open Statistics will allow to go bottom-up and identify and create new and, maybe, unexpected relationships with other datasets. Finally, OSD can be used as the basis for the creation and assessment of predictive models. These predictive models could then be reused by different applications, in the same way that open data is reused.

7 Open Statistics Challenges

This section presents a preliminary analysis of the main challenges towards the vision of Open Statistics.

7.1 Data Integration

A big challenge in Open Statistics is related to overcoming the data fragmentation problem. OSD integration is required in order to be able to achieve the vision of Open Statistics. Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data [18]. Because, however, OSD can be conceptualised as cubes, the data integration problem in Open Statistics can be though as the problem of combining cubes.

Although cubes integration has being studied in data warehouses literature for more than a decade [13, 19, 20], OSD have introduced new requirements in the area. Traditionally, an organisation had a collection of measures that were important to its operation. These measures were organised in a data warehouse. In Open Statistics, however, data providers make available for reuse in an ad-hoc manner multiple datasets that can actually comprise parts of a bigger cube with multiple measures, dimensions, and hierarchies. On the other hand, however, users may need data that require the integration of these datasets or even the data cubes that can be created by integrating the datasets. Moreover, in most of the traditional theoretical frameworks cubes integration was only presented as part of a generic framework aiming at conceptualise cubes and thus they do not describe in detail cubes integration. As a result, cube integration has to be studied under this new perspective.

An interesting case of OSD integration involve the expansion of an initial cube by using data from other cubes. For example, in terms of our first scenario, we can expand a dataset about unemployment in European countries in different years by reusing cubes with unemployment data in lower geographical level. This

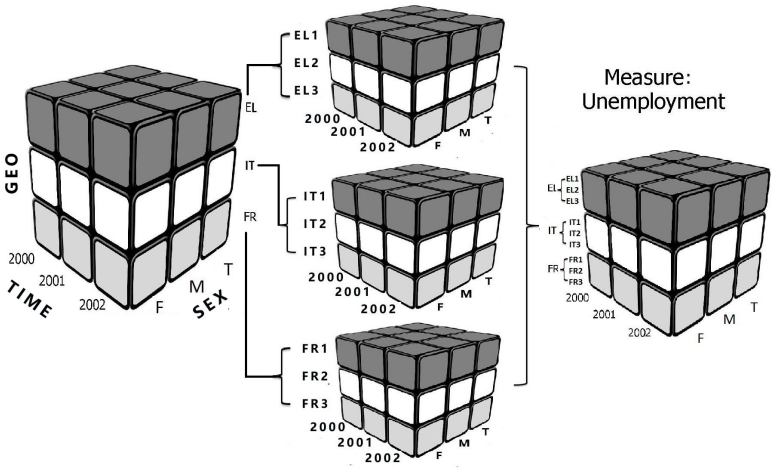

Fig. 2. Cube integration that enables the expansion of an initial cube

will result in a new cube with data of unemployment in two levels of geography (Figure 2).

A second interesting case of OSD integration involves the creation of a cube from the intersection of two other cubes. For example, in terms of the same scenario, we integrate two cubes with the same dimensions but with different values of dimensions. The resulted cube contains only the intersection of these values as presented in Figure 3.

These types of cube integration pose some interesting requirements that need to be further analysed and formally defined.

7.2 Data Analysis

Data analysis challenges are mainly related to data-driven scenarios where automatic processing of data is required. All different statistical analysis methods and techniques should be studied in the context of cubes and specific requirements that would enable automated and massive analyses should be defined.

Moreover, different analyses could present controversial results for the same phenomenon depending on the statistical methods and/or the data that have been employed. For example, [21] reviewed 68 studies about the relationship between crime and the unemployment rate and he found that only less than half of these studies have found positive significant effects of the unemployment on crime rates. So, it is important statistical analyses and models to also open up and connect to OSD [22].

7.3 Technologies

Linked data technologies has been early proposed as the most effective way for opening up data on the Web [23]. In the case of OSD this is particular true

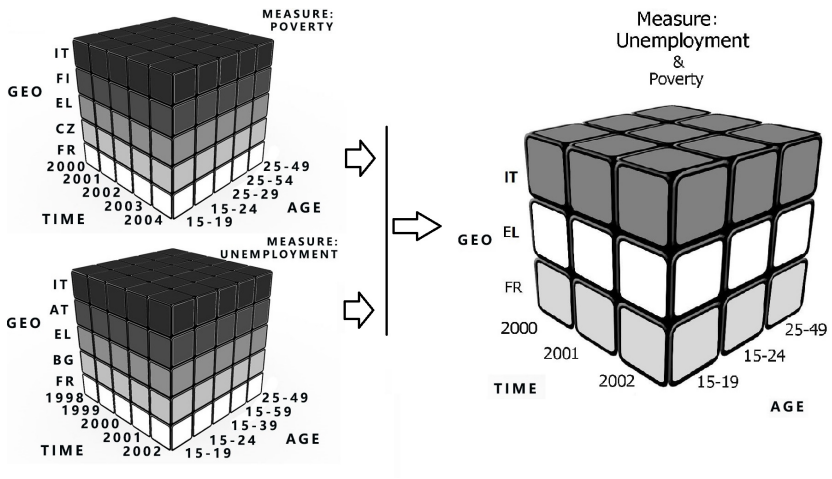

Fig. 3. Cube integration that enables the intersection of two cubes

as it will not only facilitate data integration but also enable the realisation of data-driven scenarios that require automatic data processing [24].

The RDF Data Cube (QB) vocabulary [25] is a *W3C* standard for publishing data cubes on the Web using the RDF (Resource Description Framework) and the linked data principles. The core class of the vocabulary is the *qb:DataSet* that represents a cube. A cube is connected to a *qb:DataStructureDefinition* which in turn contains a set of components that can be a *qb:DimensionProperty*, a *qb:MeasureProperty* or a *qb:AttributeProperty*. The first one defines the dimensions of the cube, the second the measures, while the third structural metadata such as the unit of measurement. Finally a cube has multiple *qb:Observation* that describe the cells of the cube.

At the moment, a number of statistical datasets are freely available on the Web as linked data cubes. For example, the European Commission's Digital Agenda provides its Scoreboard as linked data cubes. An unofficial linked data transformation of Eurostat's data[13], created in the course of a research project, includes more than 5,000 linked data cubes. Few statistical datasets from the European Central Bank, World Bank, UNESCO and other international organisations have been also transformed to linked data in a third party activity[14]. Census data of 2011 from Ireland and Greece and historical censuses from the Netherlands have been also published as linked data cubes [26, 27]. Moreover, many official efforts launched by governmental organisations (owning the data) are using the QB vocabulary to publish their data as linked data cubes. For example, the Scottish Government, the UK Department for Communities and Local Government, the Italian National Institute of Statistics, the Flemish Govern-

[13] `http://eurostat.linked-statistics.org`
[14] `http://270a.info`

ment, the Irish Central Statistics Office and the European Commission's Digital Agenda have published their data using the QB vocabulary.

Although all the above efforts use the same vocabulary, they often adopt different practices, thus hampering the data integration. The result is the creation of cubes that cannot be integrated despite the use of linked data technologies [28]. Interoperability conflicts that hamper data integration have been extensively studied in the context of relational databases and data warehouses. Examples of such conflicts include naming, structural, data scaling, data precision, and data representation conflicts [11, 29–31]. It is essential, however, to identify all the types of conflicts that may hamper data cube integration in the context of Open Statistics and linked data. Moreover, it is important to come up with and agree on best practices to be followed by statistical data publisher in order to overcome these types of conflicts.

Finally, software tools that support important functionalities related to linked data cubes creation and exploitation have been recently developed [32, 33]. However, we need to overcome challenges related to performance especially in the case of exploiting cubes from multiple data stores [34].

8 Conclusion

An increasing number of public authorities and international organisations publish Open Data. Despite the great expectations of open data movement, studies reveal that publishing open data does not automatically provide benefits. At the same time, policy documents and research in the area suggest that a large part of open data is numerical and, more specifically, concerns statistics.

In this paper, we introduced Open Statistics as a new field to systematically investigate Open Statistical Data. Open Statistics reuse methods from diverse fields like Open Data, Semantic Web, Statistics, Data Warehouses, and OLAP.

Towards this end, we initially studied datasets in both the UK and the European open data portals. We concluded that that (a) OSD can be conceptualised using the traditional data cube (or just cube) model that was initially introduced in Data Warehouses and that (b) different datasets provide fragmented views of a cube.

Thereafter we presented the potential of Open Statistics and we described how OSD redefines traditional statistical analysis methods such as OLAP, panel data, and statistical learning. We also presented challenges related to the achievement of Open Statistics. The challenges were categorised in three categories, namely data integration, data analysis, and technology.

In summary, the results suggest that Open Statistics provide benefits not present in any of these fields alone. We conclude that in certain cases OSD can realise the potential of open data.

Acknowledgments. This work is funded by the European Commission within the H2020 Programme in the context of the project OpenGovIntelligence (`http://OpenGovIntelligence.eu`) under grand agreement No. 693849.

References

1. Kalampokis, E., Tambouris, E., Tarabanis, K.: A classification scheme for open government data: towards linking decentralized data. Int. J. Web Eng. Technol. 6(3), 266-285 (2011)
2. Attard, J., Orlandi, F., Scerri, S., Auer, S.: A systematic review of open government data initiatives. Government Information Quarterly 32(4), 399-418 (2015)
3. Susha, I., Zuiderwijk, A., Janssen, M., Gronlund, A.: Benchmarks for Evaluating the Progress of Open Data Adoption: Usage, Limitations, and Lessons Learned. Social Science Computer Review 33(5), 613-630 (2014)
4. Manyika, J., Chui, M., Bughin, J., Dobbs, R., Bisson, P., Marrs, A. McKinsey Global Institute D (2013)
5. Janssen, M., Charalabidis, Y., Zuiderwijk, A.: Benefits, Adoption Barriers and Myths of Open Data and Open Government. Information Systems Management 29(4), 258-268 (2012)
6. Kalampokis, E., Tambouris, E., Tarabanis, K.: Linked open government data analytics. In: Wimmer, M.A., Janssen, M., Scholl, H.J. (eds.) EGOV 2013. LNCS, vol. 8074, pp. 99-110. Springer, Heidelberg (2013)
7. European Commission: Guidelines on recommended standard licences, datasets and charging for the reuse of documents, C240/1, 24.7.2014.
8. Bhattacherjee, A.: Social Science Research: Principles, Methods, and Practices, Open Access Textbooks (2012)
9. Shields, P., Rangarajan, N.: A Playbook for Research Methods: Integrating Conceptual Frameworks and Project Management. New Forum Press Inc (2013)
10. Romero, O., Abell, A.: A survey of multidimensional modeling methodologies. International Journal of Data Warehousing and Mining 5(2), 1 (2009)
11. Tseng, F. S., Chen, C. W.: Integrating heterogeneous data warehouses using XML technologies. Journal of Information Science 31(3), 209-229 (2005)
12. Niemi, T., Hirvonen, L., and Jrvelin, K.: Multidimensional data model and query language for informetrics. Journal of the American Society for Information Science and Technology 54(10), 939-951 (2003)
13. 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)
14. Chaudhuri, S., Dayal, U.: An overview of data warehousing and OLAP technology. ACM Sigmod record 26(1), 65-74 (1997)
15. Dielman, T. E.: Pooled cross-sectional and time series data: A survey of current statistical methodology. The American Statistician 37(2), 111-122 (1983)
16. Hildreth, C.: Combining Cross Section Data and Time Series, Cowles Commission Discussion Paper, No.347, May 15, 1950
17. Shmueli, G.: To explain or to predict?. Statistical science 289-310 (2010)
18. Lenzerini, M.: Data integration: A theoretical perspective. In: Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, pp. 233-246. ACM, (2002)
19. R. Agrawal, A. Gupta, and S. Sarawagi.: Modeling multidimensional databases. In Data Engineering, 1997. In: Proceedings of the 13th International Conference on, pp. 232243. (1997)
20. J. Perez, R. Berlanga, M. Aramburu,T. Pedersen.: Integrating data warehouses with web data: A survey. Knowledge and Data. Engineering, IEEE Transactions on, 20(7), 940955, (2008)

21. T. Chiricos.: Rates of crime and unemployment: An analysis of aggregate research evidence. Social Problems 34(2), 187212 (1987)
22. Kalampokis, E., Karamanou, A., Tambouris, E., Tarabanis, K.: Towards a Vocabulary for Incorporating Predictive Models into the Linked Data Web. n: Proceedings of the 1st International Workshop on Semantic Statistics (SemStats 2013) within 12th International Semantic Web Conference (ISWC2013), vol.1549, Sydney, Australia, CEUR-WS (2013)
23. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Semantic Services, Interoperability and Web Applications: Emerging Concepts 205-227 (2009)
24. Tambouris, E., Kalampokis, E., Tarabanis, K.: Processing Linked Open Data Cubes. E. Tambouris, M. Janssen, H. J. Scholl, M. Wimmer, K. Tarabanis, M. Gasc, B. Klievink, I. Lindgren, and P. Parycek. (eds.) EGOV2015, LNCS, vol. 9248, pp.130-143. Springer (2015)
25. Cyganiak, R., Reynolds, D., and Tennison, J.: The RDF Data Cube Vocabulary. W3C Recommendation. World Wide Web Consortium (W3C), 16th Jan 2014
26. Petrou, I., Papastefanatos, G., Dalamagas, T.: Publishing census as linked open data: a case study. In: Proceedings of the 2nd International Workshop on Open Data, Ser. WOD 2013, pp. 4:14:3. ACM, New York, NY, USA (2013)
27. Meroo-Peuela, A., Ashkpour, A., Rietveld, L., Hoekstra, R., Schlobach, S.: Linked humanities data: the next frontier? In: A Case-study in Historical Census Data. Proceedings of the 2nd International Workshop on Linked Science 2012, vol. 951 (2012)
28. Kalampokis, E., Roberts, B., Karamanou, A., Tambouris, E., Tarabanis, K.: Challenges on Developing Tools for Exploiting Linked Open Data Cubes, In: Proceedings of the 3rd International Workshop on Semantic Statistics (SemStats2015) within the 14th International Semantic Web Conference (ISWC2015), vol.1551, 11-15 Oct 2015, Bethlehem, Pennsylvania, USA, CEUR-WS (2015)
29. W. Kim, J. Seo.: Classifying schematic and data heterogeneity in multidatabase systems, Computer 24(12), 1218 (1991)
30. C. Batini, M. Lenzerini, S. B. Navathe.: A comparative analysis of methodologies for database schema integration, ACM computing surveys (CSUR) 18(4), 323-364 (1986)
31. S. Berger, M. Schrefl.: FedDW global schema architect: Uml-based design tool for the integration of data mart schemas., in: I.-Y. Song, M. Golfarelli (Eds.), DOLAP, ACM, Maui, Hawaii, USA, 2012, pp. 3340.
32. Kalampokis, E., Nikolov, A., Haase, P., Cyganiak, R., Stasiewicz, A., Karamanou, A., Zotou, M., Zeginis, D., Tambouris, E., Tarabanis K.: Exploiting Linked Data Cubes with OpenCube Toolkit. In: Proceedings of the ISWC 2014 Posters and Demos Track a Track within 13th International Semantic Web Conference (ISWC2014), vol.1272, 19-23 Oct 2014, Riva del Garda, Italy, CEUR-WS (2014)
33. P. E. R. Salas, F. M. Da Mota, K. K. Breitman, M. A. Casanova, M. Martin, S. Auer.: Publishing statistical data on the web. International Journal of Semantic Computing, 6(4), 373388 (2012)
34. Kalampokis, E., Tambouris, E. and Tarabanis, K.: ICT Tools for Creating, Expanding, and Exploiting Statistical Linked Open Data, Statistical Journal of the IAOS, [in press] (2016)