This is the post-print version of the following article:
Kalampokis, E., Tambouris, E., Tarabanis, K. (2016) Linked Open Cube Analytics Systems: Potential and Challenges, *IEEE Intelligent Systems*, Vol. 31, No.5, pp.89-92 http://dx.doi.org/10.1109/MIS.2016.82

Linked Open Cube Analytics Systems: Potential and Challenges

Evangelos Kalampokis, Efthimios Tambouris, and Konstantinos Tarabanis, University of Macedonia, Greece

Linked Open Cube Analytics (LOCA) systems enable the performance of analytics on top of multiple open statistical data (OSD) that reside in disparate portals. We present OSD's potential and highlight problems hampering its integration and reuse. To overcome these problems, we introduce an approach for OSD integration. The proposed approach capitalizes on the data cube model and linked data technologies to enable unified access to multiple OSD published in disparate portals. Finally, we present an online analytical processing (OLAP) browser for linked data cubes as a proof of concept of LOCA systems. Throughout this article, we also outline the challenges that need to be addressed for the wide adoption of LOCA systems.

Potential and Challenges of Open Statistical Data

Many governments worldwide are opening up their data for free reuse to increase transparency and boost economic growth. As a result, there are numerous open government data portals, such as http://data.gov in the USA and http://data.gov.uk in the UK.

A considerable amount of open data is actually numeric and more specifically concerns official statistics. The importance of statistics in open government data is also recognized by the European Commission.[1]

The analysis of OSD can result in interesting and even unexplored observations about social and economic phenomena. For example, we can get interesting results regarding voting behavior by using logistic regression to analyze UK open data. Table 1 suggests that in four consecutive general elections, the probability that the Labour Party wins in a specific constituency is associated with that constituency's unemployment rate. Moreover, Figure 1 depicts that the fitted models are similar in elections in which the same party won. The Labour Party won the 2001 and 2005 elections, whereas the Conservatives won the 2010 and 2015 elections.

Table 1. *Z*-statistic and *p*-value associated with unemployment.

Election year	Unemployment	
	z value	Pr(>|z|)
2001	10.14	< 2e-16
2005	10.930	< 2e-16
2010	11.069	< 2e-16
2015	10.284	< 2e-16

This is the post-print version of the following article:
Kalampokis, E., Tambouris, E., Tarabanis, K. (2016) Linked Open Cube Analytics Systems: Potential and Challenges, *IEEE Intelligent Systems*, Vol. 31, No.5, pp.89-92 http://dx.doi.org/10.1109/MIS.2016.82

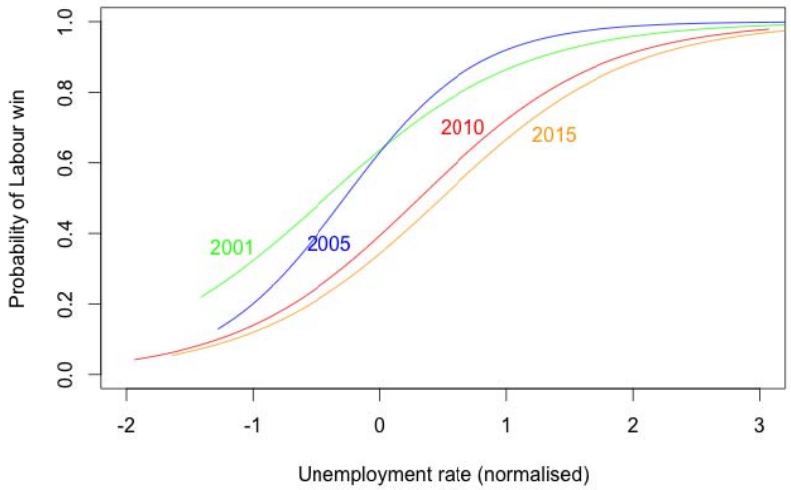

Figure 1. Labour Party wins and unemployment in the UK through open statistical data (OSD). The figure depicts four logistic regression models that predict the probability of Labour Party's win in a constituency using the unemployment rate of the constituency in four consecutive election years.

However, this type of analysis requires data that usually reside in multiple files and even in multiple portals. Following the previous example, if we search on http://data.gov.uk for datasets using the keyword "unemployment," we will come up with 122 results that provide access to 56 files and 610 links to 18 other portals (such as the Office for National Statistics and the National Archives) and by following the relevant links to more than 2,000 other files.

This situation is not restricted to the keyword "unemployment" or to the UK data portal. It is the norm rather than the exception. It is therefore clear that we need a more advanced way of organizing and publishing OSD that will facilitate combining and exploiting them in analytics. Interestingly, these scattered datasets and files provide complementary views of the same phenomenon. For example, in our unemployment scenario, various portals provide unemployment data at different geospatial levels (for example, regions, constituencies, boroughs, and wards), time periods, or population groups (for example, youths, women, and so on).

Integration of Open Statistical Data

A possible way to connect these complementary views and integrate the respective data is to employ the traditional data cube model, which was initially introduced in data warehouses.

According to this, a data cube (or simply "cube") is defined as (M, D), where $M = \{m_1, m_2, \ldots, m_k\}$ is a set of k measures and $D = \{d_1, d_2, \ldots, d_n\}$ is a set of n dimensions. A dimension $di \in D$ comprises a set of objects $O_i = \{o_{i,1}, o_{i,2}, \ldots, o_{i,m}\}$, where m denotes the size of O_i. The objects of a dimension can structure hierarchical relationships among them at different levels. These levels are defined as the attributes $A_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,l}\}$ of d_i. An ordered set of attributes defines a hierarchy $h_{i,s}$ of d_i. The cube comprises all cells $c_k = (t_{c,k}, v_{c,k})$, where $t_{c,k} \in \prod_i O_i$—meaning the generalized Cartesian product of O_i for the n dimensions of the cube—and $v_{c,k}$ is a tuple of values for M.

The use of the cube model is important not only for combining data about the same phenomenon (such as unemployment) and providing the user with a unified view but also for combining data about diverse phenomena (such as unemployment and election results) and exploring the relationship between them. In the latter case, we need an integrated view of data that describe these phenomena. These data should be relevant to the problem we want to study and appropriate for a particular method of analysis.

To be able to create this integrated view, however, we need to start from the big picture of all the available data about the phenomena of interest and search for the parts that will solve a problem at hand.

This is the post-print version of the following article:
Kalampokis, E., Tambouris, E., Tarabanis, K. (2016) Linked Open Cube Analytics Systems: Potential and Challenges, *IEEE Intelligent Systems*, Vol. 31, No.5, pp.89-92 http://dx.doi.org/10.1109/MIS.2016.82

Taking into account, however, that different portals provide parts of bigger cubes, we need to rigorously define cube integration on the Web if we want to harness OSD's full potential.

Cube integration has been studied in data warehouse literature for almost two decades.[2] Traditionally, an organization identified a collection of measures that were important to its operation. These measures were organized in a data warehouse in which a logical cube was centrally constructed and the model was populated with actual data. In this environment, cube integration was not very important, and thus operations such as drill across were loosely defined.

In the emerging open data environment, we must formally define all the different ways of combining cubes in the same sense that relational algebra formally describes different ways of combining tables.[3] If X, Y, and Z are three sets of cubes, we need to formally describe the characteristics of cubes $x \in X$ and $y \in Y$, so that y is compatible to join with x, and the characteristics of the resulted cube $z \in Z$ in relation to x and y in each of the different types of join. We can do so by formally defining, respectively, binary relations that link cubes x and y, and operators that map from (x, y) to z.

Moreover, linked data technologies provide fertile ground to implement the cube integration operations and enable the performance of data analytics on top of multiple OSD. Linked data have been proposed as the most effective way for opening up data on the Web because they facilitate data linking and integration.[4] The W3C adopted the Resource Description Framework Data Cube (QB) vocabulary to standardize the modelling of cubes as RDF graphs.

At the moment, few public organizations that produce official statistics publish their data as linked data cubes, meaning linked data that are modelled as cubes using the QB vocabulary. These include the Scottish government, the Department for Communities and Local Government in the UK, the Italian National Institute of Statistics, the Irish Central Statistics Office, and the European Commission's Digital Agenda. Although these organizations employ linked data technologies and use the QB vocabulary, they create islands of linked data cubes because they often follow different practices. For example, they define the unit of a measured variable at different levels (that is, qb:DataSet, qb:MeasureProperty, and qb:Observation) and using different type of properties (for example, qb:AttributeProperty and qb:MeasureProperty).

An open issue in this area is to bridge these islands. It is essential for the community to identify and agree on a set of best practices that will ensure interoperability among linked data cubes coming from disparate sources.

Linked Open Cube Analytics

Linked data cubes have the potential to solve data interoperability and thus enable the realization of innovative data analytics scenarios. A type of system that enables users to perform analytics on top of multiple linked data cubes is important toward this direction. We call this type of system a LOCA system.

Figure 2 presents a generic architecture of LOCA systems comprising a number of subsystems. Given a linked data cube in a local RDF store, the main role of the Compatibility Explorer subsystem is to search remote RDF stores on the Web and identify cubes that are compatible to join with the initial cube. This subsystem comprises diverse components to address all the different types of join, as they will be formally described in binary relations. The Integrator subsystem creates a new integrated cube by joining two or more compatible cubes. Thereafter, the Cube Analytics subsystem exploits the resulted cube in order to perform various types of analytics, such as OLAP exploration and statistical models creation. Following the UK elections example, a user could start with the election results data. A LOCA system can search on specific RDF stores for linked data cubes that could be analyzed together with the initial cube. These cubes could describe various phenomena, such as unemployment, poverty, and criminality. Thereafter, the system can perform statistical analysis on top of all available data and, finally, come up with the model that better explains the Labour Party's wins. The architecture finally defines the Transformer subsystem, which is responsible for computing derived values that might be important for a specific type of analysis.

This is the post-print version of the following article:
Kalampokis, E., Tambouris, E., Tarabanis, K. (2016) Linked Open Cube Analytics Systems: Potential and Challenges, *IEEE Intelligent Systems*, Vol. 31, No.5, pp.89-92 http://dx.doi.org/10.1109/MIS.2016.82

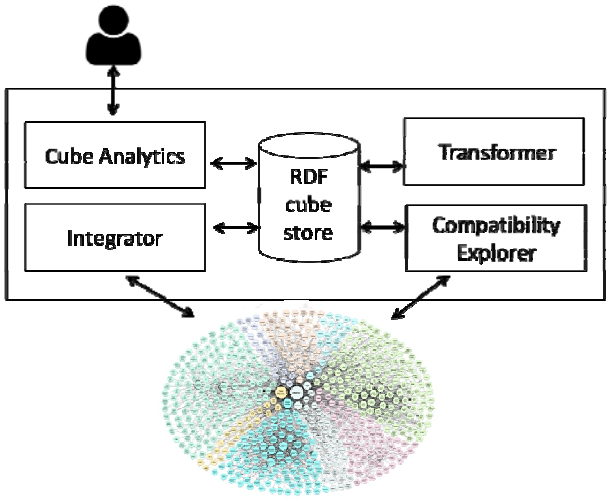

Figure 2. Generic architecture of Linked Open Cube Analytics (LOCA) systems. LOCA systems enable users to identify, merge, and perform analytics on top of compatible linked data cubes on the Web.

Different LOCA systems can be developed from the generic architecture presented in Figure 2 based on the exact needs and requirements. Thus, not all subsystems are required in all LOCA systems. For example, in an environment in which all needed data cubes are stored in the internal RDF store, the Compatibility Explorer subsystem can be restricted to searching only that store. Furthermore, the components of the Cube Analytics subsystem largely depend on the required analysis methods.

A proof of concept of a LOCA system is the OLAP browser for linked data cubes shown in Figure 3 (http://opencube-toolkit.eu/opencube-olap-browser). A typical OLAP browser lets users interactively view and analyze data from different perspectives and with multiple granularities. The presented proof of concept enables OLAP operation to be performed on top of integrated views of multiple linked data cubes coming from disparate sources.

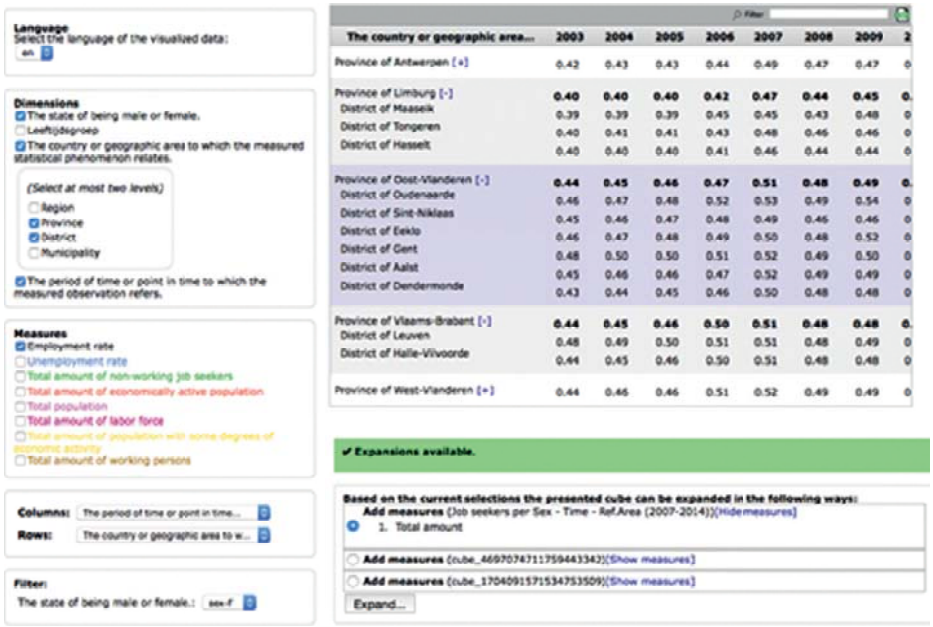

Figure 3. Online analytical processing (OLAP) browser for linked data cubes. The left area describes the structure of the cube, the right upper area presents the values of the cells, and the lower right area indicates cubes that are compatible to join with the current version of the cube as it has been specified by the selections of the user in the left area

This is the post-print version of the following article:
Kalampokis, E., Tambouris, E., Tarabanis, K. (2016) Linked Open Cube Analytics Systems: Potential and Challenges, *IEEE Intelligent Systems*, Vol. 31, No.5, pp.89-92 http://dx.doi.org/10.1109/MIS.2016.82

Figure 3 presents a screenshot of the browser. The screen is divided into three areas. The left area describes the initial cube's structure, which may contain multiple measures, dimensions, and attributes. The user can select part of these structural elements and thus explore part of the initial cube. The actual values of the cells are presented on the right side of the interface based on the user's selections. At this point, the user can perform all the typical OLAP operations (such as drill down and roll up).

Most importantly, however, the browser's Compatibility Explorer component identifies linked data cubes that are compatible to join with the current version of the cube as it has been specified by the user's selections. This is presented to the user at the lower right area, which provides information on all the available compatible cubes. The user can select any of these compatible data cubes and perform OLAP operations on top of the merged data cube that the Integrator created. In the specific instantiation of a LOCA system, we define "join of cubes" as expanding an initial cube using data from a second one. We assume that a cube can be expanded by increasing the size of one of the sets that define a cube. Therefore, cube $x \in X$ can be expanded by adding one or more elements into: the set of measures M_x, the set of objects of an attribute $a_{i,j}$ of a dimension $d_i \in D_x$, the set of attributes A_i of a dimension $d_i \in D_x$, or the set of dimensions D_x.

Finally, the browser includes the Aggregator component, which is an instantiation of the Transformer subsystem. The Aggregator computes aggregations of cells across dimensions or hierarchies. Its role is twofold. First, given an initial cube with n dimensions, the aggregator creates $2n - 1$ new cubes, taking into account all the possible combinations of the n dimensions. Second, given a cube and a hierarchy of a dimension, the aggregator computes values of cells for all attributes of the hierarchy that are at a higher level than the original cube's attribute.

The evaluation of the browser with data from the Flemish and Scottish governments revealed that we still need to overcome challenges related to performance, especially in the case of exploiting cubes from multiple RDF data stores.

We believe that the Web could be a platform for performing statistical analysis on top of combined data. We expect this article to trigger and contribute to a discussion on the development of common practices for creating linked data cubes, enabling the easy and automatic combination and exploitation of open statistical data on the Web.

Acknowledgments

Part of this work was funded by the European Commission within the H2020 Programme in the context of the project OpenGovIntelligence (http://OpenGovIntelligence.eu) under grant agreement no. 693849.

References

1. Guidelines on Recommended Standard Licenses, Datasets and Charging for the Reuse of Documents, European Commission, C240/1 2014.
2. R. Agrawal, A. Gupta, and S. Sarawagi, "Modeling Multidimensional Databases," *Proc. 13th Int'l Conf. Data Eng.*, 1997, pp. 232–243.
3. E.F. Codd, "A Relational Model of Data for Large Shared Data Banks," *Comm. ACM*, vol. 13, no. 6, 1970, pp. 377–387.
4. L. Ding, V. Peristeras, and M. Hausenblas, "Linked Open Government Data," *IEEE Intelligent Systems*, vol. 27, no. 3, 2012, pp. 11–15.

Evangelos Kalampokis *is a postdoctoral researcher at the University of Macedonia and the Centre for Research & Technology-Hellas, Information Technologies Institute. Contact him at ekal@uom.gr.*

Efthimios Tambouris *is an associate professor in the Department of Applied Informatics at the University of Macedonia. Contact him at tambouris@uom.gr.*

Konstantinos Tarabanis *is a professor in the Department of Business Administration at the University of Macedonia. Contact him at kat@uom.gr.*