An Evaluation Framework for Linked Open Statistical Data in Government

Ricardo Matheus(✉) and Marijn Janssen(✉)

Faculty of Technology, Policy and Management,
Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands
{r.matheus,m.f.w.h.a.janssen}@tudelft.nl

Abstract. Demographic, economic, social and other datasets are often used in policy-making processes. These types of statistical data are opened more and more by governments, which enables the use of these datasets by the public. However, statistical data needs often to combine different datasets. Data cubes can be used to combine datasets and are a multi-dimensional array of values typically used to describe time series of geographical areas. While Linked Open Statistical Data (LOSD) cube software is still in an initial stage of maturity, there is a need for evaluation the software platforms used to process this open data. Yet there is a lack of evaluation methods. The objective of this ongoing research paper is to identify functional requirements for open data cubes infrastructures. Eight main processes are identified and a list of 23 functional requirements are used to evaluate the OpenCube platform. The evaluation results of a LOSD platform show that many functions are not automated and need to be manually executed. We recommend the further integration of the building blocks in the platform to reduce the barriers for the use of datasets by the public.

Keywords: Linked open statistical data · LOSD · Open data · Big data · Open government · Data cube · Evaluation · Parameters · Requirements · Agile development

1 Introduction

A large number of datasets, such as demography, economic indexes, or public policies results, are statistical types of data [1]. Often these data need to be combined to create value. Data cubes are useful for combining data [2]. Data cubes are the array of 2 or more datasets based on the Structured Query Language (SQL) join functionality [3]. Data cubes enable data analysis of for example time-series to detect trends, abnormalities, unusual patterns or can be used to compare geographic regions with each other. The authors of [4] show that data cubes can be used to aggregate unemployment and election datasets to explore the relationship between them.

Organising and reusing datasets is often found to be hard due to challenges like access to data [5], manipulation of data [6], accuracy of data [7], and a long list of other data quality issues [8–12]. Linking those datasets using the Linked Open Statistical Data (LOSD) approach enables the creation of data cubes. Statistical datasets have their peculiarities and due this reason, the W3C adopted the Resource Description Framework

M. Janssen et al. (Eds.): EGOV 2017, LNCS 10428, pp. 255–263, 2017.
DOI: 10.1007/978-3-319-64677-0_21

Data Cube (QB) vocabulary to standardise the modelling of cubes as RDF graphs [13]. While statistical data cubes platforms are still on an initial mature stage [6], there is a need to evaluate OpenCube platforms. Yet no models exist to evaluate open cubes platforms.

2 Research Approach

The objective of this project paper is to develop an evaluation framework to evaluate an open data cubes platform (ODCP). Eight main processes are identified and a list of 23 requirements are derived which can be used to evaluate OpenCube platforms and applications. Using the evaluation model six cases were evaluated. The first three cases were developed by students at Delft University of Technology (https://goo.gl/y5HgJq), whereas the other three cases have been developed within the OpenGovIntelligence project (www.opengovintelligence.eu).

Table 1. Open statistical data cube parameters, requirements and questions

Parameters	#	Requirements	Questions
Functionality	1	Functional Completeness	The set of functions on the platform covers all the needs of pilots and users to perform their specific tasks?
	2	Functional appropriateness	The set of functions on the platform covers all the needs since the beginning to the end to accomplish theirs initial objectives?
Performance	3	Resource utilization	The ODCP is able to deal with amounts (quantity) and types of resources (data)?
	4	Capacity	There is a known limit capacity of any dimension (storage, processing, etc.) that platform will face during any pilot phase?
Compatibility	5	Coexistence	The ODCP is able to perform required functions efficiently while shares a common environment with other products?
	6	Interoperability	The platform is able to create an interoperable environment for exchange and use of information?
Usability	7	Learnability	The ODCP has appropriate documentation for beginners use?
	8	Operability	The ODCP has attributes that makes easy to operate and control?
	9	User error protection	The ODCP protects users to make errors? What are the functions or attributes that helps users and/or avoid errors?
	10	Accessibility	The ODCP is prepared for the widest range of characteristic and capabilities of users?
Reliability	11	Maturity	The ODCP has reliability under normal operation.
	12	Availability	The ODCP will be available to all users at same time without losing any other requirement performance?
	13	Fault Tolerance	The ODCP has fault tolerance?
	14	Recoverability	The ODCP has data recovery function?
Security	15	Confidentiality	The ODCP has any confidentiality issues concernment?
	16	Integrity	The ODCP has any function that prevents unauthorized access, modification of system and data?
Maintainability	17	Modularity	The ODCP has the modular characteristic?
	18	Reusability	The ODCP has reusable characteristic?
	19	Analysability	The ODCP has documentation for failures and errors?
	20	Modifiability	The ODCP has characteristics of improvements without degrading existing efficient and effective characteristics?
Portability	21	Adaptability	The platform is prepared to have a adaptable language or building blocks?
	22	Installability	The ODCP can be easy installed and uninstalled?
	23	Replaceability	The ODCP has flexibility on changing to other software parts?

In the literature there is no overview of functions needed by data cubes. Nevertheless ISO/IEC 25010:2010, the standard for Systems and Software Quality Requirements and Evaluation [14] can be of help, as these present a structured list of requirements. This list of requirements will be used for evaluating statistical cubes platforms. Further, based on the description of ISO 25010:2010, we created questions to evaluate each of the requirements, as presented in Table 1.

The questionnaire was used to evaluate 6 case studies in which open data cubes were designed using the OpenGovIntelligence platform. The survey was conducted on a qualitative way to identify if the platform could be used to design statistical data cubes. The answers allowed us to evaluate the data cubes by looking at which requirements were fulfilled by the open data cube platform. Also this allowed us to identify the main issues that open statistical data cubes designers face during the design and implementation of open data cues. The requirements covered were used as an indication for the maturity of development.

3 Background

Statistical data is often organised in a multidimensional manner where a measured fact is described based on a number of dimensions. As an example, Olympics statistics can bring three different dimensions: countries (USA, GB, China), medal (gold, silver, Bronze) and year (2004, 2008, 2012) and summarised on the Fig. 1 [15]. In the example, each of the cells contains a measure referring to Olympian statistical data, but together, they form a data cube.

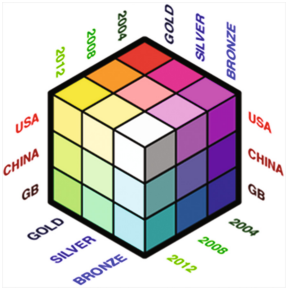

Fig. 1. Olympics Medals distributed by countries within the years.

The functionality we derived is created by adapting the Linked Open Statistical Data Cubes (LOSDC) cycle consisting of eight steps [1] modified by [16]. The steps are divided into (1) Data Cubes Creation and (2) Data cubes Analysis processes. Figure 2 shows the main steps which are described hereafter. Also the typical software tools used for supporting each step are presented.

A-**Data Cubes Creation Processes**

Step 1-Discover and Pre-process Raw Data

This first step is aimed at handling and preparing the file formats to be ready for the next steps. As an example XLS (spreadsheets file format), Comma-Separated Values (CSV) and JavaScript Object Notation (JSON) is used as an input. One of the most used tool for this step is the OpenRefine (http://openrefine.org/). This steps is needed for increasing the capacity and resilience for managing, updating and extending data because they are on an greater interoperable format (CSV, JSON) than XLS as an example.

Step 2-Define Structure and Create Cubes

The objective of the second step is to define the structure of the data cube using the Resource Description Framework (RDF) data cube vocabulary. For this own code lists or standard taxonomies created by external, supranational or international organisations like the W3C data cubes (https://www.w3.org/TR/vocab-data-cube/) can be used [13]. After this, the data in RDF format is validated. The tool used for this step is Cube Builder (https://github.com/OpenGovIntelligence/data-cube-builder) and Grafter (http://grafter.org/). This step is necessary for enabling ontology and concept scheme management.

Step 3-Annotate Cubes

The third step creates metadata about the datasets. Metadata explains the meaning of the datasets. Metadata enabled data provenance, understanding data production processes and cube structures. In this way data can be reused by others and the effort and cost for publishers to integrate with other data sources are reduces. Annotation can based on standard thesaurus of statistical concepts, validate the metadata and can

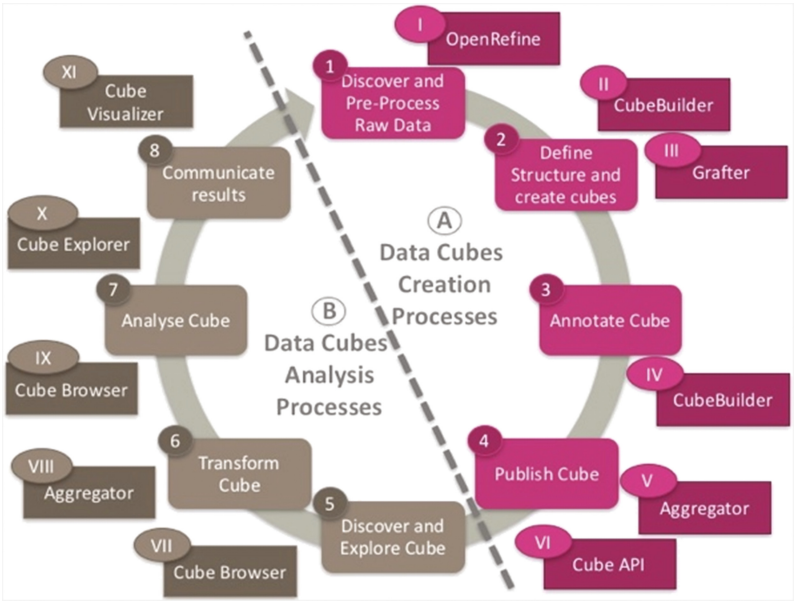

Fig. 2. Data cubes steps.

include the creation of links with compatible (external and internal) data cubes. As an example, the W3C also created the Vocabulary of Interlinked Datasets (VoID), aiming to be the connection between publishers and users of RDF datasets [17]. On the practice, OntoGov (Ontology-Enabled Electronic Government service configuration) defined a vocabulary with well-defined term that enabled automated discovery, composition, negotiation and reconfiguration of services between departments and governments [18]. The latter facilitates the analyses and even automatic combining with other datasets.

Step 4-Publish Cube
The fourth step finishes the Data Cubes Creation Process by publishing data cubes in data catalogues. This step also can use a Linked Data API (Application Programming Interface) or a SPARQL endpoint, the query language of RDFs. For this step, example of tool is the Cube API (https://github.com/OpenGovIntelligence/json-qb-api-implementation) or the aggregator (http://opencube-toolkit.eu/opencube-aggregator/).

B-**Data Cube Analysis Processes**
Step 5-Discover and Explore Cube
Based on the metadata, analysts can start to discover the cubes browsing the datasets and pivot them. This step enables the expansion of cubes, what means combining other data resources. Standardised semantic annotation helps users to find data of interest faster and easier.

Step 6-Transform Cube
The sixth step expands cubes and also allow analysts to create slices or dices, using pre-compute summarisations and other statistical functionalities. This can also help users to understand the content and structure of datasets faster and easier. The tool used on this step is the aggregator.

Step 7-Analyse Cube
This step enables statistical analysis on the cubes created using comprehensive Online Analytical Processing (OLAP) operations. The tools Cube Browser (https://github.com/OpenGovIntelligence/qb-olap-browser) and Cube Explorer (https://github.com/OpenGovIntelligence/data-cube-explorer) allow analysts to create and evaluate learning and predictive models or estimate dependencies between measures. Further, it is possible to publish the descriptions of resulting models into the Web of Linked Data. This enables the connection of data cubes with each other.

Step 8-Communicate results
This final step concludes the data cubes analysis processes and the cycle can start over again. The main objective of this step is to create visualisations and reports which can be used in policy-making efforts. As an example, analysts can create charts (bar chart, pie chart, sorted pie chart, area chart) and maps (heat maps) based on the LOSD and data cubes. The tool used for this step is the Cube Visualizer (https://github.com/OpenGovIntelligence/CubeVisualizer). The Cube visualizer is a web application that creates and presents to the user graphical representations of an RDF data cube's one-dimensional slices. It also enables non-technical users to re-use data more efficiently, in new and innovative ways without high level of technical skills.

4 Open Cubes in Practice: Case Studies

This paper selected six cases to evaluate its implementation of statistical data cubes. The first three cases were developed by students at Delft University of Technology (https://goo.gl/y5HgJq). The other three cases have been developed as part of the OpenGovIntelligence project (www.opengovintelligence.eu). The six applications are:

1. The "world most suitable country to live" (http://kossa.superhost.pl/sen1611/app/);
2. The "Gender Inequality in Europe" (http://raditya.me/genderinequality/paymentgap/mapview/);

Table 2. Open statistical data cube platform benefits and challenges

Requirements	Benefits	Challenges of development
Functional Completeness	The purposed platform has tools to open and link datasets. Also has functionalities to browse, expand, analysis and visualisation	The platform is hard to use for beginners and management level
Functional appropriateness	All the steps on the Fig. 1 (data cube cycle) can be realised if tools used properly	The platform has no manual or documentation for proper use. Examples of usage could be created to encourage and inspire usage
Resource utilization	The platform is able to deal with the amount and quantity of datasets. They are yet on the scale of Megabytes (MBs)	Quantitative analysis will be conducted on the next round of evaluation to identify if Gigabytes (GBs) scale can be processed
Capacity	This limit was not yet reached. Capacity is not an issue because datasets are on the scale of MBs	No challenges of development identified on this requirement
Coexistence	The platform has no limitations of Coexistence with other products/tools functioning in the same environment	No challenges of development identified on this requirement
Interoperability	The platform was created interoperable by default. Considering data source on an interoperable format such as tabular CSV, and output data (Data Cubes format, RDF and Turtle (TTL))	JSON API could be enhanced to increase interoperability level of platform
Learnability	If proper documentation and manual be created, the usage of platform can be done easier than learning from the scratch	Proper documentation and manual should be created to increase the capacity building (skills), for beginners and management level
Operability	Data cubes were created based on easy of usage attributes	If proper documentation, manual and also examples of usage be created, the operation of platform can be easier than current status
User error protection	Currently the platform has some functions that verifies user actions and disable options that would lead the user to commit errors. As example, some users cannot insert another CSV file after first upload	Any challenges of development for user error protection was mentioned
Accessibility	Not all the accessibility attributes had been developed on the platforms	Platform still has no accessible functions that should be adjusted according to user profile. An example, Linked data conversion will be allowed for service creators not consumers
Maturity	The initial version of platform is offering demo services on a testing server. For normal	There were no maturity issues described by any student or technical expert partners

(*continued*)

Table 2. (*continued*)

Requirements	Benefits	Challenges of development
	conversion, loading and visualisation operation the % of Uptime is around 98%	
Availability	The probability of the "Linked Data" service to be available shall be *at least* 99% of the time. In essence, the system *shall* be available "24 × 7" except for scheduled downtime related to configuration or system upgrades	There were no availability issues described. Except when system was down due human error (power off server)
Fault Tolerance	The known degree to which a system, product or component operates as intended despite the presence of hardware or software faults	There were no fault tolerance issues identified by any student or technical partner
Recoverability	The known degree to which, in the event of an interruption or a failure, a product or system can recover the data directly affected and re-establish the desired state of the system	Platform has no data recoverability functionality
Confidentiality	Statistical datasets has no issues about confidentiality, security, privacy, etc.	This requirement had no issues identified
Integrity	There is a function defining the level of user and what actions they can perform on the platform. Open access to system only provided to users of the pilot data dashboards	No complains observed about this requirement
Modularity	Platform is a loosely coupled application	No complain about this requirement
Reusability	Platform components can be reconfigured to work with other data sources/applications. As an example, the R libraries can be used	No issues identified for this requirement
Analysability	A list of errors code and documentation can enhance and encourage the use by external people to OGI Project	Students complained about no error code list or documentation
Modifiability	Platform is able to keep running the current version and suffer improvements on the background, being easily replaced by new version	No issues identified on this requirement
Adaptability	Platform components were written in different languages, but all components are communicating via APIs	No complain about this requirement
Installability	Platform components are web services and can be easily installed/uninstalled and run at any environment	Users had issues while performed local installation due Java, mainly Macintosh OS X users. Suggestion to create web version based in only one server for a group of users
Replaceability	Platform components can be reconfigured to work with other data sources/applications. As an example, visualisation service can be run using any Fuseki instance containing data cubes	No issues identified on this requirement

3. The "Best places for automotive industry install your plants in Europe";
4. The "Environmental monitoring centre" of The Flemish Government (Belgium);
5. The "Irish System of Maritime tourism, search and rescue" from Galway (Ireland);
6. The "Real Estate Market Analysis Dashboard" from Estonian Ministry of Economy (Estonia).

All cases took similar approaches of development, but have different objectives and audiences. Using the 22 requirements a questionnaire was designed to evaluate the benefits and identify the challenges of the data cube. The questionnaire was filled in by 40 students and 6 technical experts of the OGI Project. The benefits and challenges of the platforms are summarized in Table 2.

5 Discussions and Conclusions

More and more statistical data have been disclosed by organizations, which enables people from around the world to use these data. Yet data cube platforms are not a mature technology yet. This paper purposed a model for evaluation open statistical data cubes using a list of 23 requirements derived from the ISO 25010:2010 standard for Systems and Software Quality Requirements and Evaluation. Based on this list of 23 requirements, a questionnaire was developed which was used to evaluate six cases which makes use of the same platform for processing LOSD using open data cubes. The questionnaire was filled in by 40 persons and using this benefits and challenges of using open statistical data cubes were determined. The identified benefits include ease of use, the easy creation of open cubes when available in linked data format, and the flexibility of open cube platform to integrate with other software for enable the use of functionalities provided by other software. Challenges of development identified include no single platform for covering all steps, a lack of proper documentation, no guidelines for open data cube creation (which blocks capacity building and learning skills), fragmentation of tools, need for much manual work, and, installing and running issues with software which is needed to run OpenCube. The results show that Open Cubes can be used, but that there is still a lot of manual effort necessary and a variety of tools are needed that are not build to interoperate with each other. We recommend the further integration of the building blocks in the platforms to reduce the barriers for use of LOSD by the public.

Acknowledgement. Part of this work is funded by the European Commission within the H2020 Programme in the context of the project OpenGovIntelligence (www.opengovintelligence.eu) under Grant Agreement No. 693849.

References

1. Kalampokis, E., et al.: Creating and utilizing linked open statistical data for the development of advanced analytics services. In: Second International Workshop for Semantic Statistics, SemStats2014. CEUR-WS.org. (2014)
2. Sterling, T.D., Pollack, S.V.: Introduction to Statistical Data Processing. Prentice Hall, Englewood Cliffs (1968)
3. Gray, J., et al.: Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Data Min. Knowl. Discov. **1**(1), 29–53 (1997)
4. Kalampokis, E., Tambouris, E., Tarabanis, K.: Linked open cube analytics systems: potential and challenges. IEEE Intell. Syst. **31**(5), 89–92 (2016)

5. 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)
6. Kalampokis, E., et al.: Exploiting linked data cubes with opencube toolkit. In: Proceedings of the 2014 International Conference on Posters and Demonstrations Track, vol. 1272. CEUR-WS.org (2014)
7. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manag. Inf. Syst. **12**(4), 5–33 (1996)
8. Wang, R.Y., Ziad, M., Lee, Y.W.: Data Quality, vol. 23. Springer, Cham (2006)
9. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Commun. ACM **45**(4), 211–218 (2002)
10. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data quality in context. Commun. ACM **40**(5), 103–110 (1997)
11. Matheus, R., Janssen, M.: Transparency of civil society websites: towards a model for evaluation websites transparency. In: Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance. ACM, New York (2013)
12. Matheus, R., Janssen, M.: Transparency dimensions of big and open linked data. In: Janssen, M., Mäntymäki, M., Hidders, J., Klievink, B., Lamersdorf, W., Loenen, B., Zuiderwijk, A. (eds.) I3E 2015. LNCS, vol. 9373, pp. 236–246. Springer, Cham (2015). doi:10.1007/978-3-319-25013-7_19
13. Cyganiak, R., Reynolds, D., Tennison, J.: The RDF data cube vocabulary. W3C Recommendation (January 2014) (2013)
14. ISO/IEC: ISO/IEC 20510:2010 Systems and Software Engineering-Systems and Software Product Quality Requirements and Evaluation (SQuaRE)-System and Software Quality Models. International Organization for Standardization, Geneva (2010)
15. SWIRRL: How the Olympics Explains Multidimensional Data (2016). https://medium.swirrl.com/how-the-olympics-explains-multidimensional-data-8e58b127edb2-.glchsby71
16. Matheus, R., Janssen, M., Praditya, D.: Project Deliverable: D4. 1-Pilots and Evaluation Plan-v1, p. 100 (2016)
17. Alexander, K., et al.: Describing linked datasets with the void vocabulary (2011)
18. Tambouris, E., Gorilas, S., Kavadias, G., Apostolou, D., Abecker, A., Stojanovic, L., Mentzas, G.: Ontology-enabled E-gov service configuration: an overview of the OntoGov project. In: Wimmer, M.A. (ed.) KMGov 2004. LNCS, vol. 3035, pp. 122–127. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24683-1_13