# CONSISTENCY OF OPEN DATA AS PREREQUISITE FOR USABILITY IN AGRICULTURE*

**V. Vostrovsky, J. Tyrychtr**

*Czech University of Life Sciences Prague, Faculty of Economics and Management, Department of Information Engineering, Prague, Czech Republic*

Benefits of open data are diverse and range from improved efficiency of public administrations, economic growth in the private sector. Agriculture is also an inseparable part of the private sector. These data can stimulate economic growth. Economy can benefit from an easier access to information, content and knowledge in turn contributing to the development of innovative services and creation of new business models (European Data Portal, 2016). The prerequisite of this stimulation is consistency of open data. Both of these features have to be a part of these data quality. However, this aspect of open data quality has not yet been satisfactorily resolved in the framework of international standardization of quality (Systems and Software Quality Requirements and Evaluation (SQuaRE)). The issue of possible evaluation of open data consistency in agriculture is discussed. Results suggest that open data consistency may be achieved by consistent application of the technique of data normalization of relevant data sets. Consistent application of data normalization technique of open data sets can reduce the risk of inconsistency of the open data. That is the only way to guarantee that the open data will be the benefit to private sector.

data consistency, data quality model, open data, international standardization of quality, SQuaRE

DE GRUYTER OPEN

doi: 10.2478/sab-2018-0040
Received for publication on September 27, 2017
Accepted for publication on March 18, 2018

## INTRODUCTION

Open data can stimulate economic growth (Taggart, Peltola, 2010). The economy based on open data is called open data economy (Tinholt, 2013). These data have the capability to increase economic benefit through both individuals' and companies' use of the information (Open Data Institute, 2016). Open data can also increase economic benefit through jobs creation (Tinholt, 2013). The economic benefits of open data revolve around revenue growth, cost savings and improved efficiency, and employment generation while developing skills (Table 1) (Luna-Reyes et al., 2014).

Open datra can stimulate economic growth (Huijboom, Van den Broek, 2011). The prerequisite of this stimulation must be their consistency (Kucera, Chlapek, 2014). This feature must be a part of open data quality (Zins, 2007). However, this aspect of open data quality has not yet been satisfactorily resolved in the framework of international standardization of quality (Systems and Software Quality Requirements and Evaluation (SQuaRE)).

* Supported by the Internal Grant Agency of the Faculty of Economics and Management, Czech University of Life Sciences in Prague (IGA), Project No. 20161008.

Table 1. Economic benefits of open data to private sector (source: T i n h o l t , 2013)

| | Drive revenue through multiple areas | Cut costs and drive efficiency | Generate employment and develop future-proof skills |
|---|---|---|---|
| Benefit to private sector | drive new business opportunities | reduced cost by not having to invest in corvension of raw government data<br>better decision making based on accurate information | gain skilled workforce |

**Data quality model**

The quality of open data is crucial to their usability in the government as well as the private sector (Z u i d e r w i j k , J a n s s e n , 2014). The starting point for defining the quality of open data and their consistency can be a general model of classic data quality. This data quality model is defined in ISO 25000 as the degree to which the characteristics of data meet stated and implied needs when used under specified conditions (W a g n e r , 2013). The data quality model represents a defined set of characteristics, which provides a framework for specifying data quality requirements and evaluating data quality (I S O / I E C F D I S 2 5 0 1 0 :2011).

Fig. 1 shows the interaction between the general data quality model and the system models.

This interaction is similar in the case of the open data quality model and the system models in agriculture.

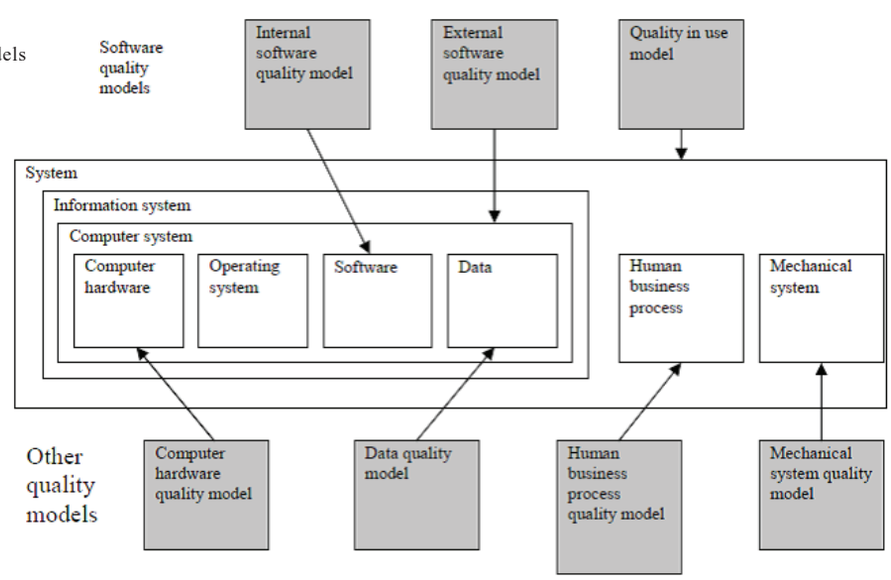
**Quality of open data**

The ISO quality model categorizes the quality into characteristics, then further subcategories into subcharacteristics and quality attributes (Fig. 2).

The above-mentioned structure of a software quality model can be applied with some variation in the issues of open data quality. Quality of a data product may be understood as the degree to which the data meet the requirements defined by the product-owner organization. Specifically, such requirements are those reflected in the data quality model through its characteristics (Accuracy, Completeness, Consistency, Credibility, Currentness, Accessibility, etc.). A data quality model is generally defined in I S O / I E C 2 5 0 1 2 :2008 Software Engineering – Software Product Quality Requirements and Evaluation (SQuaRE). This model can be used to establish data quality requirements, define data quality measures as well as to plan and perform data quality evaluations.

The international standard focuses on the quality of data as retained in a structured format within a computer system and defines quality characteristics for target data (Fig. 3, source: I S O / I E C 9 1 2 6 - 1 :2001). Data not supposed to be considered are non-target. The general data quality model defined in the standard I S O / I E C 2 5 0 1 2 :2008 includes 15 characteristics. Fig. 4 shows the quality of data product with data quality characteristics classified into main categories (source: I S O / I E C F D I S 2 5 0 1 0 :2011):

Fig. 1. System model and quality models (source: V a n i c e k , 2007



**Inherent data quality.**

It refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy the stated and implied needs when data are used under specified conditions. From the inherent point of view, data quality refers to the data itself, in particular to: data domain values and possible restrictions, relationships of data values (e.g. consistency), metadata.

**System-dependent data quality.**

This refers to the degree to which data quality is reached and preserved within a computer system when the data are used under specified conditions (ISO/IEC FDIS 25010:2011).

The aim of the present article is to define the possibility to evaluate the consistency of open data as a necessary prerequisite for their use in the private sector and hence in agriculture. There is a possible approach to addressing this issue in analogy with the existing definition of the data quality model described in the SQuaRE. The consistency of open data is an inalienable part of their quality. Currently there is an urgent need to address these aspects of quality.

## MATERIAL AND METHODS

Methods and procedures based on international standardization of software and data quality are the basis for searching for data consistency assessments. We use the measurements of quality according to the SQuaRE project. A very important characteristic of open data quality is their consistency (Janssen et al., 2012).

**Consistency**

Consistency as defined in ISO/IEC 25012:2008 is the degree to which data have attributes that are free from contradiction and are coherent with other data in a specific context of use. It can be either or both among data regarding one entity and across similar data for comparable entities (ISO/IEC FDIS 25010:2011).

The basis for minimizing the risk of inconsistency in open data are the methods and techniques of relational database technology, especially the data normalization technique.

**Inconsistency risk**

Inconsistency risk is considered proportional to the number of duplicates because update shall be performed to all occurrences of the same value in order to avoid inconsistencies. Duplications can be found for each attribute in column *i* of table *j*. Duplication score may also be calculated grouping by *k* attributes and finding duplicates over the rows. With this calculation, duplication occurs when, for the set of *k* attributes selected, two or more rows are found equal. With $\binom{n}{k}$ sets of *k* attributes for a table with *n* attributes ($k = 1, \ldots, n$), the expressions in ISO/IEC 25024:2015 becomes:

$$TDS = \sum k \sum j \sum i \, D_{ijk}$$

where:
*TDS* = total duplication score (total number of duplicates
$D_{ijk}$ = number of duplications found in set *i* of *k* attributes of table *j*

$$RI = [(TDS)/(NR * NC)]/NT$$

where:
*RI* = risk of inconsistency
*NR* = number of rows
*NC* = number of columns
*NT* = number of tables

In case of multiple tables, also the structure of tables impacts the inconsistency risk, e.g., a normalized database leads to better duplication score than a non-normalized one containing the same data. On the other hand, normalization may decrease time efficiency.

## RESULTS

For the design of our solution, we use an example of open data on bee varroasis incidence in the Czech Republic (CR). These data resulting from the relevant laboratory analysis should be available at all district offices of the Regional Veterinary Administration in

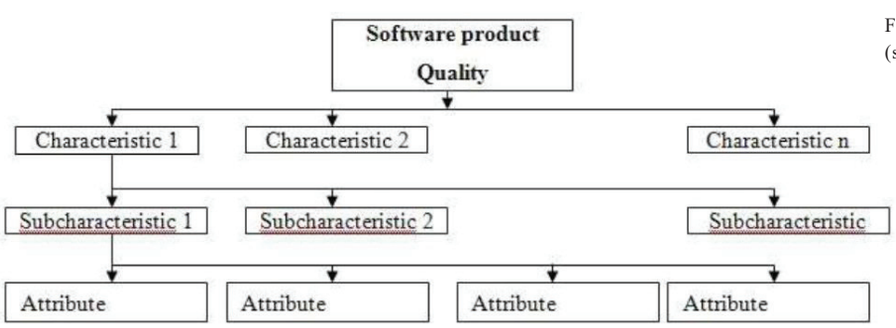



Fig. 2. Tree of quality model hierarchy (source: ISO/IEC 9126-1:2001

Table 2. Data of State Veterinary Administration CR relating to the laboratory analysis of bee samples of bees in relation to bee incidence

| ID_District | District | Honeybee_habitat | ID_beekeeper | Beekeeper_date of birth | Name_beekeeper | Number_of_ varroa |
|---|---|---|---|---|---|---|
| CZ0202 | Beroun | Mořinka | 11123 | 25/6/1959 | Kalaš Josef | 1 |
| CZ0202 | Beroun | Lážovice | 11123 | 25/6/1959 | Kalaš Josef | 0 |
| CZ0202 | Beroun | Skřipel | 11155 | 2/3/1963 | Ulm Aleš | 13 |
| CZ0202 | Beroun | Vižina | 11155 | 2/3/1963 | Ulm Aleš | 159 |
| CZ0202 | Beroun | Kotopeky | 11155 | 2/3/1963 | Ulm Aleš | 110 |
| CZ0202 | Beroun | Otmíče | 11183 | 1/10/1947 | Dvořák Jan | 0 |
| CZ0202 | Beroun | Bykoš | 11191 | 2/5/1980 | Adam Jiří | 2 |
| CZ0524 | Rychnov | Olešnice | 11205 | 10/1/1961 | Toman Karel | 187 |
| CZ0524 | Rychnov | Olešnice | 11205 | 10/1/1961 | Toman Karel | 11 |
| CZ0524 | Rychnov | Opočno | 11207 | 12/3/1967 | Ruml Vojtěch | 121 |
| CZ0524 | Rychnov | Záhoří | 11208 | 1/12/1948 | Karas Jan | 0 |
| CZ0524 | Rychnov | Osečnice | 11208 | 1/12/1948 | Karas Jan | 12 |
| CZ0524 | Rychnov | Pěčín | 11208 | 1/12/1948 | Karas Jan | 72 |
| CZ0524 | Rychnov | Podbřezí | 11212 | 7/4/1985 | Haleš Martin | 155 |

Table 3. Partitioned table with modified structure to minimise duplicity

| ID_District | District |
|---|---|
| CZ0202 | Beroun |
| CZ0524 | Rychnov |

Table 4. Partitioned table with modified structure to minimise duplicity

| ID_District | ID_beekeeper | Number_of_varroa |
|---|---|---|
| CZ0202 | 11123 | 1 |
| CZ0202 | 11123 | 0 |
| CZ0202 | 11155 | 13 |
| CZ0202 | 11155 | 159 |
| CZ0202 | 11155 | 110 |
| CZ0202 | 11183 | 0 |
| CZ0202 | 11191 | 2 |
| CZ0524 | 11205 | 187 |
| CZ0524 | 11205 | 11 |
| CZ0524 | 11207 | 121 |
| CZ0524 | 11208 | 0 |
| CZ0524 | 11208 | 12 |
| CZ0524 | 11208 | 72 |
| CZ0524 | 11212 | 155 |

the CR (see Table 2). These data can serve as a basis for a beekeeper's decision to start beekeeping in the specified area (habitat).

In this section, we propose the procedure for calculating the inconsistency risk in different database implementations (Tables 1–4) for $k = 1$ and $k = 2$.

For simplicity we use the following data attributes: ID = Id_District, D = District, H = Honeybee_habitat, Ib = ID_beekeeper, Bd = Beekeeper_Date_of_Birth, N = Name_beekeeper, V = Number_of_varroa.

Subsequently we count the number of duplication for Table 1:

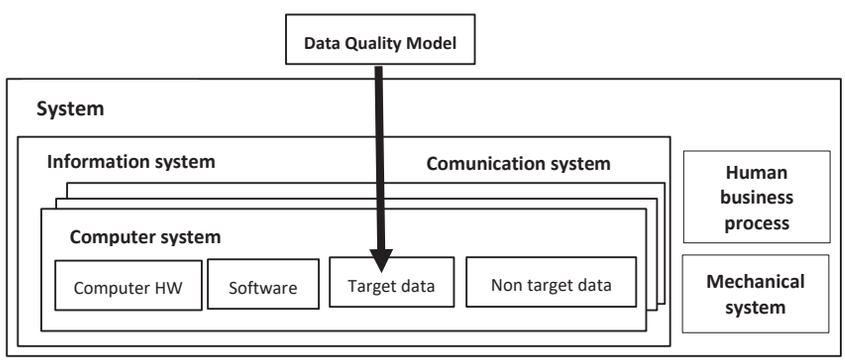



Fig. 3. Quality of data as retained in a structured format within a computer system (source: ISO/IEC 9126-1:2001

Table 5. Partitioned table with modified structure to minimise duplicity

| ID_beekeeper | Beekeeper_Dat;e of Birth | Name_beekeeper |
|---|---|---|
| 11123 | 25/6/1959 | Kalaš Josef |
| 11142 | 2/3/1963 | Ulm Aleš |
| 11183 | 1/10/1947 | Dvořák Jan |
| 11191 | 2/5/1980 | Adam Jiří |
| 11205 | 10/1/1961 | Toman Karel |
| 11207 | 12/3/1967 | Ruml Vojtěch |
| 11208 | 1/12/1948 | Karas Jan |
| 11212 | 7/4/1985 | Haleš Martin |

Number of duplications ND1 (k = 1, ID, D, H, Ib, Bd, N) = 14+14+2+10+10+10 = 60

Number of duplications ND2 (k = 2, IdD, IdH, IdIb, IdBd, IdN, DH, Dib, DBd, DN, Hib, HBd, HN, IbBd, IbN, BdN) = 14 + 2 + 10 + 10 + 10 = 46 + 2 + 10 + 10 + 10 = 78 + 0 + 0 + 0 + 10 + 10 + 10 = 108 + 10 = 118

Number of rows NR = 14, Number of columns NC = 6, Number of tables NT = 1

Risk of inconsistency RI = [(ND1 + ND2) / (NR * NC)] / NT = [(60 + 118) / (14 * 6)]/1 = 2.11

The three decomposed tables (Tables 3–5) are based on the original Table 2.

Table 3 is a partitioned table with modified structure to minimize duplicity:

Number of duplications (k = 1, Id, D) = 0 + 0 = 0

Number of duplications (k = 2, IdD) = 0

Number of rows = 2, Number of columns = 2

Table 4 is a partitioned table with modified structure to minimize duplicity:

Number of duplications (k = 1, Id, Ib) = 14 + 10 = 18

Number of duplications (k = 2, IdIb) = 10

Number of rows = 14, Number of columns = 2

Table 5 is a partitioned table with modified structure to minimize duplicity

Number of duplications (k = 1, ID, Bd, N) = 0 + 0 + 0 = 0

Number of duplications (k = 2, IdBd, IdN, BdN) = 0 + 0 + 0 = 0

Number of rows = 8, Number of columns = 3

Risk of inconsistency RI = [(0 + 0) / 4 + (18 + 10) / 28 + (0+0) / 24] / 3 = 0.33

It is evident that the variant of the three sub-tables (Tables 3–5) based on the original table (i.e. Table 2) has a better score of inconsistency risk than the implementation in Table 1.

## DISCUSSION

The results presented in this article were based on the data normalization technique and the new series of international standards named Systems and Software Quality Requirements and Evaluation (SQuaRE) compared to other studies (Song et al., 2014; Marik, 2016; Hanna et al., 2017; Phansalkar, Dani, 2017). It is obvious that the high quality application of this technique will pose increased demands on the appropriate data quality management implemented by the suppliers of the relevant open data. Suppliers are entities that provide open data, although this activity is not necessarily their primary purpose or source of profit. Publishing these data must be a part of their wider strategy of increasing customer confidence and enhancing their integrity (Howard, 2013). This corresponding business model mainly includes companies that provide open data to help improve customer decision-making on the market. It seems logical that the open data supliers should have this knowledge and ability to consistently apply the data mormalization technique. This assumption will be crucial to exploiting the potential of open data in the private sector and thus also in agriculture.

There is a need to evaluate further issues. According to ISO/IEC 25012:2008, data quality can be measured from 'inherent' and 'system dependent'

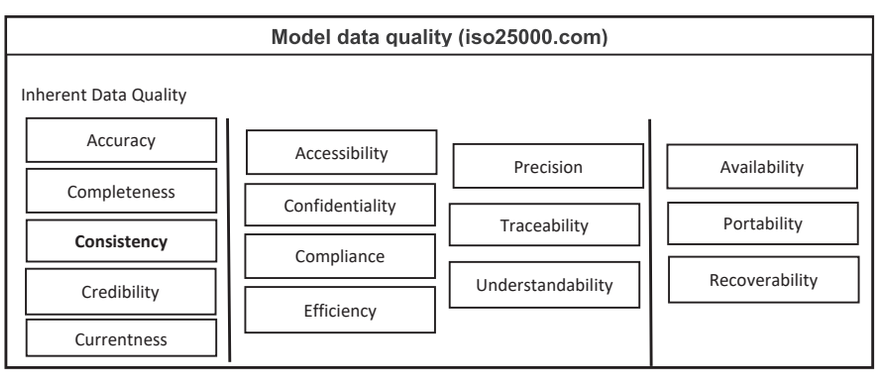



Fig. 4. Quality of data product (source: ISO/IEC FDIS 25010:2011

points of view. This issue leads to that there may be a correlate with other quality models and quality entities.

## CONCLUSION

The above method of evaluating the consistency of the tables indicates the great importance of normalizing open data datasets. It is evident that this normalization of open data must be a part of management of open data operated by their suppliers. Data quality management is defined as the business processes that ensure the integrity of an organization's data during their collection, application, aggregation, warehousing, and analysis (AHIMA, 2012). These suppliers are the subjects who supply open data, although this activity is not necessarily their primary objective or the source of their profit. Publishing the data could be a part of their broader strategy to increase the trust of their customers and to strengthen their integrity (Huijboom, Van den Broek, 2011). This business model includes primarily the companies which provide open data for customers to better decide on the market (Vanroekel, Todd, 2014). In the agricultural sector, this subject should be covered by the Ministry of Agriculture, and its primary objective should be the transparency of the whole sector (Charvat, 2014). Consistent application of the data normalization technique of open data sets can minimize the risk of inconsistency of the open data. That is the only way to guarantee that the open data will be a benefit to both private and agriculture sectors.

## REFERENCES


AHIMA (2012): *Pocket glossary of health information management and technology.* AHIMA Press, Chicago.

Charvat K, Esbri MA, Mayer W, Campos A, Palma R, Krivanek Z (2014): FOODIE – open data for agriculture. In: IST-Africa 2014 Conference Proceedings, Mauritius, 1–9.

Europen Data Portal (2016): Benefits of open data. Creating value through open data. https://www.europeandataportal.eu/en/using-data/benefits-of-open-data. Accessed 10 July, 2017.

Hanna F, Droz-Bartholet L, Lapayre JC (2017): Toward a faster fault tolerant consensus to maintain data consistency in collaborative environments. International Journal of Cooperative Information Systems, 26, article No. 1750002. doi: 10.1142/S0218843017500022.

Howard A (2013): Governments looking for economic ROI must focus on open data with business value. O'Reilly Radar. http://radar.oreilly.com/2013/02/roi-open-data-economy-value.html. Accessed 16 September, 2014.

Huijboom N, Van den Broek T (2011): Open data: An international comparison of strategies. European Journal of ePractice, 12, 4-16.

ISO/IEC 9126-1:2001. ISO/IEC 9126-1. Software engineering – Product quality – Part 1: Quality model. International Organization for Standardization, Geneva.

ISO/IEC 25012:2008. ISO/IEC 25012. Software engineering – Software Product Quality Requirements and Evaluation (SQuaRE) – Data quality model. International Organization for Standardization, Geneva.

ISO/IEC FDIS 25010:2011. ISO/IEC FDIS 25010. Systems and software engineering – Systems and Software QualityRequirements and Evaluation (SQuaRE) – System and software quality models. International Organization for Standardization, Geneva.

ISO/IEC 25024:2015. ISO/IEC 25024. Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – Measurement of data. International Organization for Standardization, Geneva.

Janssen M, Charalabidis Y, Zuiderwijk A (2012): Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29, 258–268.

Kucera J, Chlapek D (2014): Benefits and risks of open government data. Journal of Systems Integration, 5, 30–41.

Luna-Reyes LF, Bertot JC, Mellouli S (2014): Open government, open data and digital government. Government Information Quarterly, 31, 4–5.

Marik R (2016): On design of data consistency verification. In: Proc. 17th Internat. Conference on Mechatronics – Mechatronika (ME 2016), Prague, Czech Republic, 1–8.

Open Data Institute (2016): Research: The economic value of open versus paid data. https://theodi.org/article/research-the-economic-value-of-open-versus-paid-data/. Accessed 1 November, 2016.

Phansalkar SP, Dani A (2017): Selective data consistency model in No-SQL data store. In: Phansalkar SP, Dani A (eds): Privacy and security policies in big data. IGI Global, Hershey, 124–147.

Song J, Sierra SC, Rodriguez JC, Perandones JM, Jimenez GDC, Bujan JO, Garcia RM, Galdon AS (2014): Data consistency management in an open smart home management platform. In: 2014 European Modelling Symposium, Pisa, Italy, 366–371. doi: 10.1109/EMS.2014.51.

Taggart C, Peltola V (2010): OpenCorporates: Why we're crying out for this database of companies. The Guardian. http://www.theguardian.com/news/datablog/2010/dec/20/open-corporates-chris-taggart. Accessed 14 February, 2014.

Tinholt D (2013): The open data economy. Unlocking economic value by opening government and public data. Capgemini Consulting. http://www.capgemini-consulting.com/resource-file-access/resource/pdf/opendata_pov_6feb.pdf. Accessed 4 November, 2016.

Vanicek J (2007): Software quality measures validation in the Czech Republic. Agricultural Economics, 53, 94–100.

Vanroekel S, Todd P (2014): Continued progress and plans for open government data. https://www.whitehouse.gov/

blog/2014/05/09/continued-progress-and-plans-open-government-data. Accessed 9 May, 2014.

Wagner S (2013): Software product quality control. Springer, New York.

Zins C (2007): Conceptual approaches for defining data, information, and knowledge. Journal of the American Society for Information Science and Technology, 58, 479–493. doi: 10.1002/asi.20508.

Zuiderwijk A, Janssen M (2014): Open data policies, their implementation and impact: A framework for comparison. Government Information Quarterly, 31, 17–29.

*Corresponding Author:*

Ing. Jan T y r y c h t r, Ph.D., Czech University of Life Sciences Prague, Faculty of Economics and Management, Department of Information Engineering, Kamýcká 129, 165 21 Prague 6-Suchdol, Czech Republic, phone: +420 224 382 074, e-mail: tyrychtr@pef.czu.cz