# กรอบแนวคิดสำหรับการวัดคุณภาพของข้อมูลสำหรับข้อมูลเปิดภาครัฐของประเทศไทย

## Open Data Quality Measurement for Thailand Open Government Data

Phimphan Thipphayasaeng[1], Poonpong Boonbrahm[2] and Marut Buranarach[3]

School of Informatics, Walailak University, NakornsiThammarat, Thailand[1, 2]

National Electronics and Computer Technology Center (NECTEC), Pathumthani, Thailand[3]

E-mail: phimpham.th@gmail.com[1], poonpong@gmail.com[2], marut.bur@nectec.or.th[3]

**บทคัดย่อ**

บทความนี้นำเสนอกรอบแนวคิดสำหรับการประเมินคุณภาพข้อมูลเปิดภาครัฐของประเทศไทย (Open Government Data - Thailand) เพื่อตระหนักถึงคุณภาพของชุดข้อมูลเปิดภาครัฐของประเทศไทย (Open Government Data - Thailand) คุณลักษณะข้อมูลที่นำมาพิจารณาในการประเมิน ประกอบด้วย การตรวจสอบย้อนกลับ (Traceability) ความเป็นปัจจุบันของข้อมูล (Currentness) การหมดอายุ (Expiration) ความสมบูรณ์ครบถ้วน (Completeness) การปฏิบัติตามข้อกำหนด (Compliance) ความเข้าใจได้ (Understandability) ความถูกต้อง (Accuracy) และความสอดคล้อง (Consistency) ถูกนำมาใช้ในการวัดคุณภาพของข้อมูล การเลือกคุณลักษณะข้อมูล (data characteristic) ถูกออกแบบเพื่อการประเมินคุณภาพของข้อมูล พิจารณาตามลักษณะเฉพาะของข้อมูลแบบเปิด (open data) โดยคำนึงถึงประโยชน์ในการนำข้อมูลเหล่านั้นไปใช้งาน บทความนี้แปลงคุณลักษณะข้อมูล (characteristics data) ให้อยู่ในรูปแบบของสมการเพื่อสกัดผลที่ได้จากการประเมินตามเกณฑ์ให้อยู่ในรูปแบบของคะแนน ซึ่งจะแสดงให้เห็นถึงคุณภาพของข้อมูลเชิงตัวเลข ซึ่งวิธีการดังกล่าวช่วยให้สามารถเปรียบเทียบคุณภาพของข้อมูลแบบพลวัตร จากผลของการวิจัยพบว่า ผลของการประเมินคุณภาพมีความสอดคล้องกับการประเมินของผู้เชี่ยวชาญ โดยผลของการประเมินสามารถนำไปสู่การค้นพบข้อมูลเชิงประจักษ์ ตัวอย่างเช่น ชุดข้อมูลที่มีความเป็นปัจจุบัน มักจะสามารถตรวจสอบย้อนกลับ (Traceability) ได้อย่างสมบูรณ์ ในขณะที่ชุดข้อมูลจำนวนมากจากหลายหมวดหมู่ยังคงมีความล่าช้าในการเปิดเผยข้อมูล

*คำสำคัญ:* การประเมินคุณภาพของข้อมูล, ข้อมูลเปิดภาครัฐ, การเชื่อมโยงข้อมูลแบบเปิด, คุณลักษณะข้อมูล

**Abstract**

This paper proposes a framework to assess data quality of open government data of Thailand. To realize quality of datasets provided in open government data of Thailand, data characteristics including Traceability, Currentness, Expiration, Completeness, Compliance, Understandability, Accuracy and Consistency were exploited to measure data. With a nature of open data, the selected

characteristics are designed to assess all necessary aspects for usefulness in data utility. The characteristics come with the equations to calculate the quality into a set of scores to represent quality in numeric form. This can help to compare quality of datasets dynamically. From testing, the results of quality assessment correctly worked as intended. The results of assessment can lead to several findings such as the current datasets contained perfect traceability of creation score while there were delay in publication of datasets from all domains.

*Keywords:* data quality assessment, open government data, linked open data, data characteristic

## 1. Introduction

Nowadays, data play important role in computational and analytic process. Data are essentially the plain facts and statistics collected during the operations in every task. They can be used to measure many ranges of activities. While the data itself may not be very informative, it is the basis for all to be recorded within organizations. There is an urge for sharing in secretive data to be used for boarder view analysis. Non-personal data are now asked to be opened across several organizations for full accessibility and used as called Open data [1]. Open data are data that "can be freely used, modified, and shared by anyone for any purpose" [2]. This leads to collaboration, creativity and innovation [3].

The most wanted open data are government data. The open of government data brings several important benefits. The first one is transparency [1, 4]. In a democratic society, citizens have a right to know what their government is doing. They should be able freely to access government data and information and to share that information with other citizens. Another benefit is participatory governance [5]. By opening up data, citizens are enabled to be much more directly informed and involved in decision-making.

In Thailand, government organizations and agencies have joined in a participation of Open Government Data of Thailand (TOGD) project in 2013 [6]. They have provided their own data freely to access and use. More than a thousand of datasets have been published, which is operated by the electronic government agency (EGA). Similar to open government data initiatives of other countries, problems in the quality of the data have been reported. In 2010, Allison [7] published found problems of open government data. There are issues of accuracy, aggregation and precision. In 2016, Usamah [8] reported a problem of government open data

that are in several formats and are missing description or semantics. In TOGD, these reported problems have also been found and led to be less quality data.

To detect problems in data, a standard of data quality is established as preferred quality to be used for assessment criteria. In this work, we apply existing data quality assessment models with additional criteria to assess data provided in TGOD. We expect to find frequently occurred problems in low quality data. The result will assist data owners to specifically fix errors within the data and use as a guideline for creating new datasets.

The rest of this paper is structured as follows. Section II gives a summary of work related to this paper. Section III explains a model to assess data quality of TOGD in details. Section IV provides assessment results and a discussion of results. Section V gives a conclusion of the paper.

## 2. Related Work

Research on data quality has been started in the area of information systems in the early 90's, and it has been extended in different point of views. This section will briefly discuss about several contexts where quality of data is considered important to open government data quality.

In 1985, Ballou and Pazer's study [9] focuses primarily on intrinsic dimensions that can be measured objectively. They designed four dimensions that frequently appear in information quality studies: accuracy, consistency, completeness, and timeliness. In 1996, Wand and Wang [10] defined quality dimension for analyzing data quality. Their work takes an ontological approach and formally defines four quality dimensions: correctness, unambiguous, completeness, and meaningfulness. In 2002, Klein [11] developed a theoretical model of information quality problems on the World Wide Web. The model includes the factors of accuracy, completeness, relevance, timeliness and amount of data. In 2004, Scannapieco et al. [12] proposed DaQuinCIS, which is specifically designed for the cooperative Information System. In the paper, a model for data and quality data exported by cooperating organizations is given. The model mainly involves in detecting quality using correctness, completeness and consistency of data. In 2007, Kaiser et al. [13] proposed to include extra factors in quality dimension of data such as normalization and interpretability. These new factors are used with common factors such as correctness and timeliness in data quality assessment.

Similar to afore-mentioned publications, ISO/IEC 25012 [14] was developed with other factors for quality standard of data design. It proposes a data quality model using fifteen characteristics from two points of view: inherent and system dependent. Inherent data quality refers to the degree to which quality characteristics of data have the intrinsic potential to satisfy stated and implied needs when data is used under specified conditions. The characteristics in this set are specifically: Accuracy, Completeness, Consistency, Credibility and Currentness. System dependent data quality refers to the degree to which data quality is attained and preserved within a computer system when data is used under specified conditions. The characteristics in this set are: Availability, Portability and Recoverability. The common characteristics for overall are: Accessibility, Compliance, Confidentiality, Efficiency, Precision, Traceability and Understandability.

In 2016, Vetrò et al. [4] defined evaluation framework to assess the quality of open government data from a developers' perspective. Their work focuses on intrinsic data quality comprising seven different dimensions of data quality such as Accuracy, Currentness, Traceability and Expiration. This work clearly declared the focused data levels as either content level or dataset level. The content level is for data within a content cell while the dataset level refers to overall quality of dataset.

The aforementioned models compose of many characteristics of data for quality measurement as given in Table 1. The common characteristics include accuracy, completeness, currentness and accuracy. In addition to Vetrò et al. [4] data level, we found that a schema level is yet to be considered apart from content level. Since our work focuses on Open Government Data, we realize that the schema of the data is one of the important keys to link data together for data integration. Hence, quality of a schema level, such as content header, should also be assessed separately in data quality assessment.

In Thailand, there was also a study on data factors and processes influencing benefit realization from IT in organizations [15]. The work aims to study factors in IT data from opinions of stakeholders in IT organizations. The results can be summarized that the most important factors are usefulness, accuracy and track of updates. Moreover, the study of data warehouse in actual use for provincial water work [16] also confirms that specific quality characteristics of data (such as accuracy, accessability and consistency) are relatively affected the outcome of result, and they are the key for successful application development.

Table 1. Characteristics of data used in existing data quality models

| Aspect | Traceability | Currentness/Timeliness | Expiration | Accuracy/Correctness | Completeness | Compliance | Understandability | Consistency | Unambiguous | Meaningfulness | Trustworthiness | Relevance | Amount of Data | Availability | Credibility | Portability | Recoverability | Accessibility | Confidentiality | Efficiency | Precision |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ballou and Paze (1985) | | ✓ | | ✓ | ✓ | | | ✓ | | | | | | | | | | | | | |
| Wand and Wang (1996) | | | | ✓ | ✓ | | | | ✓ | ✓ | | | | | | | | | | | |
| Klein (2002) | | ✓ | | ✓ | ✓ | | | | | | | ✓ | ✓ | | | | | | | | |
| Scannapieco et al. (2004) | | | | ✓ | ✓ | | | ✓ | | | ✓ | | | | | | | | | | |
| Kaiser et al. (2007) | | ✓ | | ✓ | | | | | | | | | | | | | | | | | |
| ISO/IEC 25012 (2008) | ✓ | | | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Vetrò et al. (2016) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | | | | | | | | | | | | | |

## 3. Assessment Model for Open Government of Thailand

This work aims to assess quality of data provided in Open Government Data of Thailand (TOGD) to identify how much quality of each dataset. First, we will need to define characteristics for consideration. Second, assessment calculation is defined to represent an assessment score for each characteristic.

### 3.1 TOGD Data Assessment Model

From the existing works, there are many characteristics to be considered for data quality assessment. Some of them were particularly designed for internal-used database such as confidentiality and accessibility, and they are not fit to a concept of Open Data. Hence, we have to select among those to design a list of characteristics of data for TOGD.

For TOGD, the provided data are government details about official projects, responsible objects, and declarative report. The data are already open and freely accessible; thus, availability and accessibility are pointless to include. From observation, the most problems of data are accuracy, understandability and consistency for all types of data. Furthermore, datasets providing timely report such as monthly updates of project progression should include characteristics as traceability, currentness and expiration. With the reason, we include a list of characteristics for TOGD assessment model as shown in Table 2.

Table 2. Selected characteristics for TOGD assessment model

| Characteristic | Metric | Level |
|---|---|---|
| Traceability | Track of creation | Dataset |
| | Track of updates | Dataset |
| Currentness | Percentage of current rows | Cell |
| | Delay in publication | Dataset |
| Expiration | Delay after expiration | Dataset |
| Completeness | Percentage of complete cells | Cell |
| | Percentage of complete rows | Cell |
| | Percentage of complete schema | Schema |
| Compliance | Percentage of standardized cell | Cell |
| | eGMS Compliance | Dataset |
| | Five star Open Data | Dataset |
| Understandability | Percentage of columns with metadata | Cell |
| | Percentage of schema in comprehensible format | Schema |
| | Percentage of cell in comprehensible format | Cell |
| Accuracy | Percentage of accurate cells | Cell |
| | Percentage of accurate schema | Schema |
| Consistency | Percentage of Consistency cells | Cell |
| | Percentage of Consistency schema | Schema |

In details, we explain each characteristic below.

- Accuracy represents a correctness in terms of given data and specification of a data field. The values, that do not belong to a specification, can be checked by counting improper data of the type. For example, a field is set to integer but contains a string. This also includes field restrictions such as a symbolic usage instead of proper value, language, and white spacing amount.

- Completeness: this aspect examines an amount of compete cells from all cells. The missing value in a cell or header will be counted and calculated for percentage. This aspect will be split up for content level (cell) and schema level (header).

- Compliance is for only data content. This will check data wording comparing to standard asked by standard control group to reduce synonymy issue. For example, a term for Bangkok can be deployed in data in Thai as 'กทม', 'กรุงเทพมหานคร', and 'กรุงเทพฯ', but the standard asks to keep the term as 'กรุงเทพมหานคร'; hence, those in variant can be considered as non-preferable and will decrease the standard level of the dataset.

- Understandability includes a naming in header and wording used in data. Header should be checked that the column name should be a proper word as a nominal. A header with all symbolic and alphabetical rendering such as AA, B&, and 1 is not understandable for both human user and machine.

- Currentness is about how current data are. This indicates the ratio between the delay in the publication (number of days passed between the moment in which the information is available and the publication of the dataset) and

the period of time referred by the dataset (week, month, and year).

- Expiration indicates the ratio between the delay in the publication of a dataset after the expiration of its previous version and the period of time referred by the dataset (week, month, and year).
- Consistency refers to the appearance of data that may be synonymously inconsistent. The variant of data will reduce quality in overall.
- Traceability is to examine about creation and updates so they can be traceable.

### 3.2 Assessment Calculation

Since there are three data levels, we separate assessment calculation into three parts. The first part is data level. The data level refers to content given in a cell and is the most important and major part of dataset. The second part is schema level that is about a term used in a header of the data table. The header of the data table can indicate understandability and scope of data. Quality of header can relatively affect users about meaning of data. Last, dataset level represents overall quality of data. The dataset level is mostly used for time-based data to assess for currentness and expiration. Formula for each of assessment calculation is given in Table 3 for data, schema and dataset level.

Table 3. Metrics defined

| Aspect | Metrics | Variables | Formula | Scale | reference |
|---|---|---|---|---|---|
| Traceability | Track of creation | s: Source<br>dc: Date of creation | $c = 2s + dc$ | [0, 3] | Vetrò A. et al. (2016) |
| | Track of updates | lu: List of updates<br>du: Dates of updates | $tu = lu + du$ | [0, 2] | Vetrò A. et al. (2016) |
| Currentness | Delay in publication | da: Date of information availability<br>dp: Date of publication<br>sd: Start date of the period of time referred by the dataset<br>ed: End date of the period of time referred by the dataset. | $dp = 1 - (\frac{dp - da}{ed - sd})$ | (-∞, 1] | Vetrò A. et al. (2016) |
| | Percentage of current rows | ncr: Number of not current rows<br>nr: Number of rows. | $pcr = \left(1 - \frac{ncr}{nr}\right) * 100$ | [0%, 100%] | Vetrò A. et al. (2016) |

| Aspect | Metrics | Variables | Formula | Scale | reference |
|---|---|---|---|---|---|
| Expiration | Delay after expiration | ed: Expiration date<br>cd: Current date<br>sd: Start date of the period of time referred by the dataset<br>ed: End date of the period of time referred by the dataset. | $dae = 1 - (\frac{cd - ed}{ed - sd})$ | (-∞, +∞) | Vetrò A. et al. (2016) |
| Compliance | Percentage of standardized columns | ns: Number of columns with associated standards<br>nsc: Number of standardized columns | $psc = 1 - (\frac{ns}{nsc}) * 100$ | [0%, 100%] | Vetrò A. et al. (2016) |
| | eGMS compliance | s: Source<br>dc: Date of creation<br>c: Category<br>t: Title<br>d: Description (if applicable)<br>id: Identifier (if applicable)<br>pb: Publisher (if applicable)<br>cv: Coverage (recommended only)<br>l: Language (recommended only) | $egmsc = s + dc + c + t + 0.2(d + id + pb + cv + l)$ | [0–5] | Vetrò A. et al. (2016) |
| | Five star Open Data | the level of the 5 Star Open Data model | the value assigned depends on the level of the scheme in which the dataset is. | [0, 5] | Vetrò A. et al. (2016) |
| Completeness | Percentage of complete cells | nr: Number of rows<br>nc: Number of columns<br>ic: Number of incomplete cells<br>ncl: Number of cells | $ncl = nr * nc$<br>$pcc = 1 - (\frac{ic}{nd}) * 100$ | [0%, 100%] | Vetrò A. et al. (2016) |
| | Percentage of complete rows | nr: Number of rows<br>nir: Number of incomplete rows | $pcpr = 1 - (\frac{nir}{nr}) * 100$ | [0%, 100%] | Vetrò A. et al. (2016) |
| | Percentage of complete schema | nh: Number of header column<br>ich: Number of incomplete header column | $pcs = 1 - (\frac{ich}{nh}) * 100$ | [0%, 100%] | |
| Understandability | Percentage of columns with metadata | ncm: Number of column with metadata<br>nc: Number of columns | $pcm = 1 - (\frac{ncm}{nc}) * 100$ | [0, 100] | Vetrò A. et al. (2016) |
| | Percentage of cell in comprehensible | ncuf: Number of columns in understandable format<br>nc: Number of columns | $pcuf = 1 - (\frac{ncuf}{nc}) * 100$ | [0%, 100%] | Vetrò A. et al. (2016) |

| Aspect | Metrics | Variables | Formula | Scale | reference |
|---|---|---|---|---|---|
| | format | | | | |
| | Percentage of schema in understandability format | nhuf: Number of header column in understandable format<br>nh: Number of header column | $psuf = 1 - \left(\frac{nhuf}{nh}\right) * 100$ | [0%, 100%] | |
| Accuracy | Percentage of syntactically accurate cells | nce: Number of cells with errors<br>ncl: Number of cells | $pcuf = 1 - \left(\frac{nce}{ncl}\right) * 100$ | [0%, 100%] | Vetrò A. et al. (2016) |
| | Percentage of structure accurate schema | nhe: Number of header column with errors<br>nh: Number of header column | $psas = 1 - \left(\frac{nhe}{nh}\right) * 100$ | [0%, 100%] | |
| Consistency | Percentage of consistency cell format | nce: Number of cells with distinct values<br>ncl: Number of column | $pccf = \left(\frac{ncdv}{ncl}\right) * 100$ | [0%, 100%] | |
| | Percentage of consistency schema | nce: Number of header column with distinct values<br>nh: Number of header column | $pcs = \left(\frac{nhd}{nh}\right) * 100$ | [0%, 100%] | |

## 4. Results and Discussion

### 4.1 Setting

To examine a capability of the proposed method, we collect and test it with data of five domains from data.go.th. (accessed date: 14 June 2016) as follows.

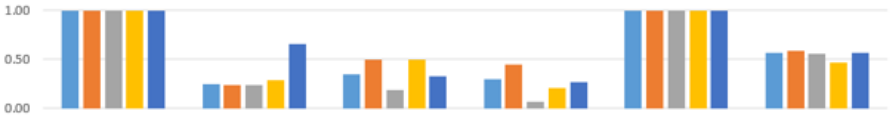
- Law, Crime and Justice
- Transportation and Logistic
- Government Budget
- Economy, Finance and Industry
- Society and Welfare

A preprocess of data validation was performed to remove non-machine-readable data such as image and PDF format files. Moreover, datasets, which contains multiple-table or additional note among cells, were also discarded since they are not pure data. After preprocessing, there are 139 datasets in total.

### 4.2 Results

The results of assessment are shown in three figure separated by level. Since the results are given in dataset level, content level and schema level with many datasets, we calculate results into several representations. For each assessment score, we calculated for average (AVG), standard deviation (SD) and Percentage (PCT). The normalized results in range of 0 (minimum) to 1 (maximum) are given in Figure 1-3.

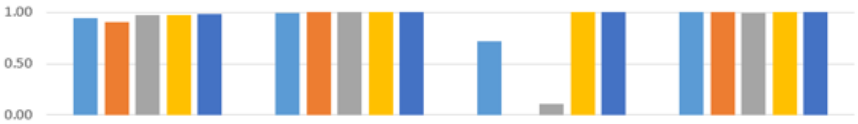
Dataset Level




| Category | Traceability | | Currentness | Expiration | Compliance | | | |
|---|---|---|---|---|---|---|---|---|
| | Track of creation | Track of updates | Delay in publication | Delay after expiration | eGMS Compliance | | Five star Open Data | |
| | PCT. | PCT. | PCT. | PCT. | AVG. | SD. | AVG. | SD. |
| Law, Crime, and Justice | 1.00 | 0.25 | 0.35 | 0.30 | 1.00 | 0.00 | 0.57 | 7.33 |
| Transport and Logistics | 1.00 | 0.24 | 0.50 | 0.45 | 1.00 | 0.00 | 0.59 | 4.36 |
| Government Budget and Spending | 1.00 | 0.24 | 0.19 | 0.07 | 1.00 | 0.00 | 0.56 | 7.25 |
| Economy, Finance, and Industry | 1.00 | 0.29 | 0.50 | 0.21 | 1.00 | 0.00 | 0.47 | 9.89 |
| Society and Welfare | 1.00 | 0.66 | 0.33 | 0.27 | 1.00 | 0.00 | 0.57 | 7.04 |

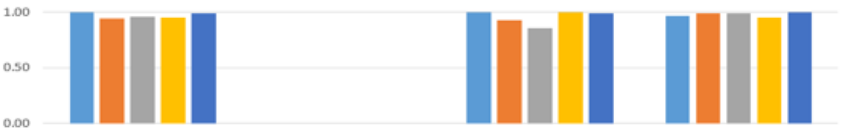
Figure 1. Show the result of calculating the data level

Content Level




| Category | Completeness | | | | Compliance | | Currentness | |
|---|---|---|---|---|---|---|---|---|
| | Percentage of complete cells | | Percentage of complete rows | | Percentage of standardized cell | | Percentage of current rows | |
| | AVG. | SD. | AVG. | SD. | AVG. | SD. | AVG. | SD. |
| Law, Crime, and Justice | 0.94 | 7.98 | 0.99 | 0.59 | 0.72 | 43.03 | 1.00 | 0.00 |
| Transport and Logistics | 0.90 | 8.17 | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| Government Budget and Spending | 0.97 | 6.29 | 1.00 | 0.00 | 0.11 | 12.60 | 0.99 | 3.25 |
| Economy, Finance, and Industry | 0.97 | 5.82 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 |
| Society and Welfare | 0.98 | 2.84 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 0.00 |

Content Level




| Category | Understandability | | | | Accurate | | Consistency | |
|---|---|---|---|---|---|---|---|---|
| | Percentage of cell in comprehensible format | | Percentage of columns with metadata | | Percentage of accurate cells | | Percentage of Consistency cells | |
| | AVG. | SD. | AVG. | SD. | AVG. | SD. | AVG. | SD. |
| Law, Crime, and Justice | 1.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.97 | 7.94 |
| Transport and Logistics | 0.94 | 6.82 | 0.00 | 0.00 | 0.93 | 7.01 | 0.99 | 1.45 |
| Government Budget and Spending | 0.96 | 12.64 | 0.00 | 0.00 | 0.86 | 17.50 | 0.99 | 1.66 |
| Economy, Finance, and Industry | 0.95 | 11.24 | 0.00 | 0.00 | 1.00 | 0.00 | 0.95 | 4.79 |
| Society and Welfare | 0.99 | 1.95 | 0.00 | 0.00 | 0.99 | 1.95 | 1.00 | 0.00 |

Figure 2. Show the result of calculating the content level


| Schema Level | Completeness | | Understandability | | Accurate | | Consistency | |
|---|---|---|---|---|---|---|---|---|
| | Percentage of complete schema | | schema in comprehensible format | | Percentage of accurate schema | | Percentage of Consistency schema | |
| Category | AVG. | SD. | AVG. | SD. | AVG. | SD. | AVG. | SD. |
| Law, Crime, and Justice | 0.98 | 4.40 | 0.90 | 24.28 | 0.95 | 11.73 | 0.97 | 5.82 |
| Transport and Logistics | 0.99 | 2.31 | 0.91 | 23.74 | 1.00 | 0.00 | 0.98 | 3.13 |
| Government Budget and Spending | 0.99 | 5.21 | 0.62 | 26.91 | 0.99 | 5.21 | 0.98 | 4.61 |
| Economy, Finance, and Industry | 0.53 | 33.75 | 0.53 | 33.75 | 0.55 | 35.90 | 0.77 | 41.02 |
| Society and Welfare | 0.98 | 4.30 | 0.95 | 9.38 | 1.00 | 0.00 | 0.96 | 6.77 |

Figure 3. Show the result of calculating the Schema level

### 4.3 Discussion

From the results, content level from all domains returned acceptable scores in average while dataset level yielded the lowest for some characteristics. The results also showed that four domains, namely Law, Transportation, Budget and Economy, suffered from average of about 27% score for Track of Update characteristic. This can be implied that dataset in these domain are rarely up-to-date. Moreover, Delay in Publication characteristic also received unsatisfied scores for all domains. This showed that most of the dataset did not get update immediately within the expected date. In overall, we found that datasets from Open Government Data of Thailand were not up-to-date, and these issues should be concerned.

For content level, overall scores were impressive since all scores in this type of all domains received over 77% while many of the scores were with maximal score. In this level, the best domain was society domain that was given with at least 98% score in all characteristics. In the schema level, datasets of economy domain suffered with the lowest score among all tested domains. From examining through the datasets of economy domain, we found that datasets in this domain severally lacked table headers, and this caused the score to be relatively low. Furthermore, we also noticed that datasets from budget domain often used ambiguous terms and abbreviations in the headers and obtained about 60% score for understandability characteristic in schema level.

From the results, we found that the scores of all characteristics worked as intended. Datasets with low score contained issues as same as characteristic standards mentioned while the datasets with high score are more likely to follow the standards.

## 5. Conclusions

This paper presents an assessment method for open government data of Thailand. Characteristics to be assessed are Traceability, Currentness, Expiration, Completeness, Compliance, Understandability, Accuracy and Consistency. In this work, we also design to separately assess data quality to content level (content in cell), schema level (label in cell header) and dataset level (overall quality) to represent difference in attributes since the focused part in each level is different. With the separation in characteristics and levels, the calculated assessment results can be traced and clearly identify the cause of low quality for creators to fixate and be aware in future data creation. Testing results of datasets of five selected domains from open government data of Thailand showed several findings. The given scores worked as intended as to represent quality of data, and low score indicate low quality and vice versa. In the future, we plan to cover all datasets from open government data of Thailand to see overall results. Moreover, we plan to include the characteristics indicating linkability among datasets as a measurement for readiness to develop linked open data.

## Acknowledgements

We would like to thank the Electronic Government Agency (EGA) for providing datasets for analysis and testing.

## 6. References

[1] Srimuang C., Cooharojananone N., Tanlamai U. and Chandrachai A. (2017). Open government data assessment model: An indicator development in Thailand. 19th International Conference on Advanced Communication Technology (ICACT).

[2] Open Definition. “The Open Definition.” Internet: http://opendefinition.org, [Nov. 9, 2016].

[3] Hofmokl J. (2010). The Internet commons: toward an eclectic theoretical framework. International Journal of the Commons, 4 (1), 226-250.

[4] Vetrò A., Canova L., Torchiano M., Minotas C.O., Iemma R. and Morando F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. Government Information Quarterly, 33 (2): 325-337.

[5] Davies T. G., Ashraf Bawa Z. (2012). The promises and perils of Open Government Data (OGD). The Journal of Community Informatics, Vol.8 (2).

[6] Government Open Data of Thailand website, Internet: https://data.go.th/, [14 June 2016].

[7] Allison B. My data can't tell you that. In D. Lathrop, & L. Ruma (Eds.). (2010). Open government — Collaboration, transparency, and participation in practice, O'Reilly Media, Inc, 257-265.

[8] Usamah A. (2016). Outstanding Challenges in Recent Open Government Data Initiatives. International Journal of e-Education, e-Business, e-Management and e-Learning, 6, 91-102.

[9] Ballou D.P. and Pazer H.L. (1985). Modeling data and process quality in multi-input, multi-output information systems. Management Science, 31 (2), 150-162.

[10] Wand Y. and Wang R. (1996). Anchoring data quality dimensions in ontological foundations. Comm. ACM, 39.

[11] Klein B. D. (2002) When do users detect information quality problems on the World Wide Web? American Conference in Information Systems.

[12] Scannapieco M., Virgillito A., Marchetti M., Mecella M., and Baldoni R. (2004). The DaQuinCIS architecture: a platform for exchanging and improving data quality in Cooperative Information Systems. Inform. Syst, 29, 551-582.

[13] Kaiser M., Klier M., and Heinrich B. (2007). How to Measure Data Quality? - A Metric-BasedApproach . ICIS 2007 Proceedings, 108.

[14] ISO/IEC-FDIS-25012. (2008). Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model.

[15] Promthong N., Porrawatpreyakorn N., Viriyapant K., Boonyapalanant A. (2016). Factors and Process Influencing Benefit Realization from IT in Organizations. The Science and Technology RMUTT Journal, 6 (2 ), 84-101.

[16] Mahattanasin P., Chirawichitchai N. (2015). Data warehouse to decision support systems for provincial waterworks authority. The Science and Technology RMUTT Journal, 5 (2), 135-144.