Development of Data Quality Framework for Linked Data Readiness Assessment of Thailand Open Government Data

Phimphan Thipphayasaeng[1],
Marut Buranarach[2],
and Poonpong Boonbrahm[3]
[1, 3]School of Informatics,
Walailak University, Thailand
[2]National Electronics and Computer Technology Center, Thailand
[1]phimphan.thi@gmail.com
[2]marut.bur@nectec.or.th
[3]poonpong@gmail.com

***Abstract* - This research proposes an assessment model for measuring semantic and linkable potential of open data to support an improvement towards Linked open data readiness. The model is designed to use content statistics of data for calculating score to represent difference of data potential for linking in range of** 0 **to** 1. **There are two main aspects for assessing as semantic type degree and linkability. The semantic based aspect focuses on quality of data involving in semantic meaning and frequency of existing semantic concepts. The linkability is to find a possibility to create links within datasets. Datasets in this work are the data provided in open government data of Thailand for preparation towards improving to linked open data readiness. From experiments, the results showed that the model can discriminatively generate a score to identify aspect-based quality as intended. The correlation results of both proposed aspects signified that scores of semantic type degree were positively correlated to score from linkability.**

***Keywords* - Linked Open Data, Open Government Data, Linked Data Readiness Assessment**

I. INTRODUCTION

Nowadays, data play important role in computational and analytic process. Essentially, data are the plain facts and statistics collected during the operations in every task. They can be used in many ranges of activities such as analysis and prediction. With the help from internet, many data are provided in a web. The most preferred data are those structured data ready for computational processing by machines. World Wide Web Consortium (W3C) hence coined the term about linked open data (LOD) [1] and their standard to support a development towards semantic web.

LOD is a combination of two concepts as linked data (LD) and open data (OD). Thus, LOD is defined as interlinked data released under an open license [2]. Recently, OD projects [3-5] have been established in several countries including Thailand [6] and become successful in sharing data. However, very few LD projects have been set up since LD is more complicate in development from semantic understanding and linking for more advance querying. One of the common methods to develop LOD is to enhance OD with linking relation. Nevertheless, improving OD to LOD involves several factors such as understanding of data meaning, acceptable quality of data schema, and possible linked content of data.

In this research, we aim to develop assessment model to measure a potential to become LOD from OD. Characteristics from OD essential towards LOD readiness development are observed and used in a calculation to generate a score representing likeliness to become LOD. The tested datasets in this work are datasets provided in open government data of Thailand for analyzing and preparation for further upgrading to LOD.

II. LITERATURE REVIEWS

A. Open Data and Linked Data

The term “Open data” is defined as "data that can be freely used, re-used and redistributed by anyone anytime, anywhere" [7]. Open data is to publish of non-personal, non-confidential data in reusable electronic formats and under an open license [8]. Usages of open data are such as to find implicit relation among them and to be a source of new knowledge.

Open Government Data (OGD) is an extension of open data focusing on data within government sectors. The specific definition is given as "data and information produced or commissioned by government or government controlled entities” [9]. The goal of open government data is not only to make government information public, but also to make it as useful as possible as for reuse of government data by the private sector [10-11]. Moreover, OGD allows citizens to monitor data streams and thus improves the collaborative [12], releasing social and commercial value [13] and transparency of government [14].

Many countries including Thailand have adopted the concept of OGD. Thailand Open Government Data (TOGD) project has been established and received excellent cooperation from many government sectors to provide their data [14] at https://data.go.th/ (government open data of Thailand website). TOGD is ranked in terms of amount and standardized data as the 51st place from all 94 countries in the 2016 Global Open Data Index website [4]. The data has been accumulated and gone up in number.

Lately, the concept of OGD has been suggested to include the concept of Linked data to become linked open data (LOD) for more usefulness. Hence, a star level (0-5 star) [1] as shown in Fig. 1 was suggested to classify a dataset towards usefulness of the open data. The 5 star which is the best is defined to the linked data (LD). Thus, it becomes a challenge to OGD for developing towards Linked open government data (LOGD).

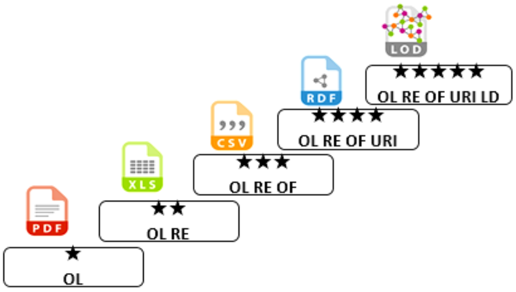

Fig. 1 A Star Level Classification and Its Meaning of TOGD

To adopt a concept of LOGD, a method of exposing and connecting data from different sources is needed to create semantic web [15]. It requires a use of URIs to semantically connect data within datasets with uniquely identified entities. The core concepts of linking datasets are to realize a semantic meaning of data and to create a connection of related or same data. Thus, it is essential that the data must be understandable and standardized. Besides, the content of the data should be common so they can be linked.

However, the current star level of datasets provided in TOGD is in range of 1-3 star level. This indicates that none of the datasets is ready for linking towards LOGD. Therefore, it is important to assess implicit characteristics of TOGD and realize its potential for LOGD development. The assessment should provide insight details for preparation towards LOGD including quality of data for linking and amount of datasets available to be linked.

B. Data Quality Assessment Model

Research on data quality assessment has been started in the area of information systems in the early 90's, and it has been extended to a different point of views. This section will briefly discuss several contexts where the quality of data is systematically assessed into tangible score. The existing researches on data quality assessment are summarized in Table I.

TABLE I
THE EXISTING RESEARCHES ON DATA QUALITY ASSESSMENT AND METHODS USED

Developer	Type of data	Assessment method	Main Aspects
Wang and Strong, 1996 [16]	Data consumers	Two-stage survey for generate list of potential data quality attribute and collect the importance of quality attribute, An exploratory factor analysis of the ratings the dimension.	Intrinsic DQ, Contextual DQ, Representational DQ and Accessibility DQ
Wand and Wang, 1996 [17]	Information system development	The analysis of the representation mapping, A comprehensive literature review.	Internal View (design, operation) and External View (use, value)
Alexander and Tate, 1999 [18]	World wide web	The guideline or the checklist method	Content quality to digital environment
Vetro et. al., 2016 [19]	Open data	Equation for scoring	Content quality and details of data distribution

From Table I, the existing assessment models were designed to measure data in terms of quality for usage. However, none of the works aims towards a development of LOD or LOGD. Among the works, Vetro's model was especially designed for open data and can be used to rationally assess quality of open data. Hence, we decided to adapt the idea of assessment from Vetro's open data quality assessment method to measure aspects related to LOGD.

III. RESEARCH METHODOLOGY

This section describes the proposed assessment model for ODG towards the concept of LODG. There are two main aspects for detection including semantic type degree and linkability. The semantic type degree focuses on quality of data schema and semantic of data. The linkability is to assess a dataset for its potential to be linked to other dataset. In linkability, we also present a type of dataset linking for classifying linking behavior among datasets.

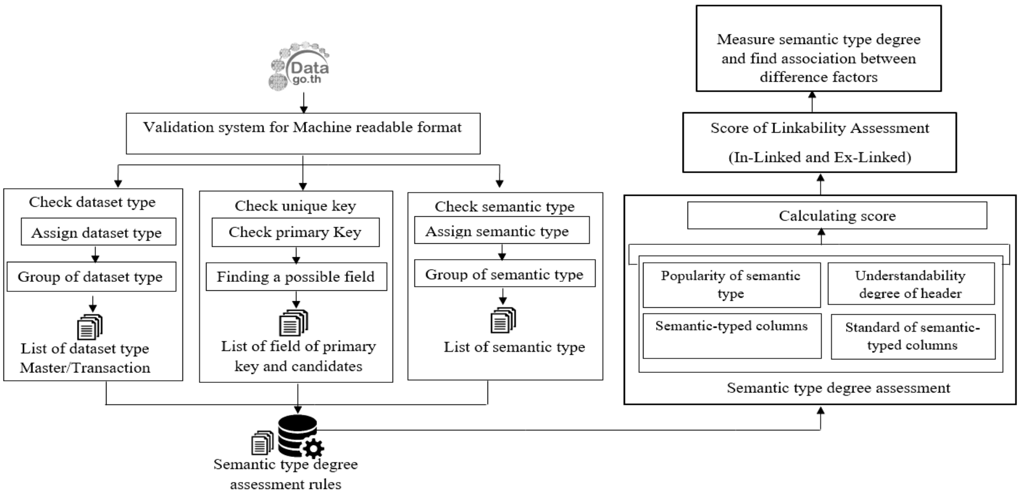
Data go.th
Validation system for Machine readable format
Check dataset type
Assign dataset type
Group of dataset type
List of dataset type Master/Transaction
Check unique key
Check primary Key
Finding a possible field
List of field of primary key and candidates
Check semantic type
Assign semantic type
Group of semantic type
List of semantic type
Semantic type degree assessment rules
Measure semantic type degree and find association between difference factors
Score of Linkability Assessment (In-Linked and Ex-Linked)
Calculating score
Popularity of semantic type
Understandability degree of header
Semantic-typed columns
Standard of semantic-typed columns
Semantic type degree assessment

Fig. 2 An Overview of a Framework to Assess Linked Data Readiness Assessment of Thailand Open Government Data

A. Semantic Type Degree

There are several features to be considered separately to consider a semantic type degree within a dataset. Detection of these features including dataset type, semantic type, and key data is necessary since they will be used to determine semantic type degree in different aspects. In this work; the term "Semantic type degree" involves a quality of data header and frequency of semantic type from data header. The data header normally can define a meaning scope of content in a column. Thus, three characteristics related to data header are designed as follows.

- *Semantic-Typed Columns (psc):* This characteristic is to find a percentage of columns that have latent semantic meaning from all columns in a dataset. The latent semantic meaning is detectable from data headers that contain conceptual meaning such as province name, project name, person name, etc. but not the columns with data index or amount-based data.

- *Standard of Semantic-Typed Columns (ssc):* This is a sequent characteristic from the first one. It is to find a percentage of the data in the semantic-typed column that are in a standard based on existing guideline or specification.

- *Understandability Degree of Header (phu):* This characteristic is to check that a column header of those containing semantic types is understandable or not. The understandability is language independent and is defined as using proper textual naming, using abbreviation in well-known or commonly accepted manner, and using readable and sound words.

Beside from quality of header and its content, we also design characteristics of semantic type usage in datasets. By tagging semantic types to data column, the columns with same semantic meaning are classified together regardless of language or used terms. In regards to semantic types, a list of semantic concepts representing meaning of data is required. The frequency of each semantic type within a dataset and as a whole is counted to represent popularity of each semantic type. The dataset that contains several popular semantic types is more likely to be linkable since it shares more common data with another. In this work, two scores are calculated regarding semantic type as follows.

- *Proportion of Semantic Type to all Datasets (Dataset-Based View) (pstd):* It is to find the frequency of existing semantic types in all datasets. The same semantic type in a dataset is possible for bilingual presentation (Thai for one column and English for another). The count for frequency is counted based on unique semantic within dataset regardless of existing amounts.

- *Proportion of Semantic Type to all Types (Semantic-Type-Based View) (psts):* Similarly to previous one, it is to count the frequency of semantic types in overall.

In total, there are 5 characteristics to be assessed for semantic type degree. Each has its own calculation for scoring in a normalized scale in range of 0 to 1 while 0 is the lowest score and 1 is the highest score. Details of calculations for semantic type degree including variable meaning and equation are given in Table II.

B. Linkability Degree

This part explains the potential to link of a dataset. The aim is to find a number of possible linked data within entire given datasets. The count then is calculated into a score representing linking potential. Thus, it is essential to classify linking types since each linking type has different points in consideration.

During a study on linkability degree, many scenarios in linking datasets were observed. Upon extensive reviewing on linked open data as a reference, it was surprising that none was mentioning of linking types for creating LOD. While attempting in linking datasets, specific patterns of linking were found and categorized as follows.

TABLE II
CACULATION OF SCORE BASED ON ASPECT
FOR ASSENSING SEMANTIC TYPE DEGREE

Aspect	Metric	Variables	Equation	Scale	Reference
Semantic-typed columns	Proportion of semantic-typed columns	*ncs*: Number of columns with semantic type *nc* : Number of columns	$psc = \frac{ncs}{nc}$	[0,1]	[20]
Popularity of semantic type	proportion of Semantic type to all datasets (dataset-based view)	*ndst*: Number datasets that a semantic type appeared *nd*: Number of all dataset files	$pstd = \frac{ndst}{nd}$	[0,1]	[20]
	proportion of Semantic type to all types (semantic-type-based view)	*ncst*: Number of columns that a semantic type appeared *ncast*: Number of columns that all semantic types appeared	$psts = \frac{ncst}{ncast}$	[0,1]	[20]
Standard of semantic-typed columns	Proportion of semantic-typed columns conforming to standard	*nssc*: number of standard of semantic-typed columns *ncsc*: number of columns that is semantic-typed columns	$ssc = \frac{nssc}{ncsc}$	[0,1]	[20]
Understandability degree of header	Proportion of header understandability degree	*nuc*: Number of non-understandable column headers *nc*: Number of columns	$phu = 1 - \frac{nuc}{nc}$	[0,1]	[20]

TABLE III
CALCULATION OF SCORE BASED ON ASPECT
FOR ASSENSING LINKABILITY

Aspect	Metric	Variables	Equation	Scale
In-Linking	In-Linking type	*pi*: number of linked of pattern that same primary key *d*: number of datasets in a domain	$Ina = \frac{\sum_{i=1}^{n} Pi}{d}$	[0, ∞]
	Pattern frequency (In-Linking)	*pfi*: Pattern frequency of data categoty *d*: number of dataset in category	$PF = \frac{\sum_{i=1}^{n} pf_i}{d}$	[0, ∞]
	Unique Pattern (In-Linking)	*upfi*: Unique Pattern frequency of data category; 1 if exists; 0 otherwise *d*: number of dataset in category	$UP = \frac{\sum_{i=1}^{n} upf_i}{d}$	[0, ∞]
	Potential to in-Linking of patterns		$PInP = \frac{\sum_{i=1}^{n} pf_i}{\sum_{i=1}^{n} upf_i}$	[0, ∞]
Ex-Linking:	Ex-Linking:	*li* : number of linkable of FK to PK *d*: number of dataset in domain	$Exa = \frac{\sum_{i=1}^{n} li}{d}$	[0, ∞]

- *Merging:* Datasets contain the same schema but some or all different instances.

- *Merging (added):* Datasets in matrix format (two-dimension table) contain all same columns and rows.

- *In-Linking:* Datasets contain the primary key of the same semantic type.

- *Ex-Linking:* Datasets contain column of the same semantic type. This type is similar to In-Linking, but the difference is that the common column is primary key to foreign key or foreign key to foreign key.

For merging and merging-added, the count relies on the pattern of data columns within datasets. If the pattern is exactly the same, those datasets are able to be merged and the amount of uniquely same pattern can be counted for total frequency. However, In-

Linking is to find a common primary key (PK) of datasets while Ex-Linking is to detect the use of a PK to a foreign key (FK) of another dataset. Thus, their calculations are more complicate and involve several features in their scoring such as an amount shared pattern of header pattern and frequency of the focused semantic types. The calculation of the scores is given in Table III.

IV. EXPERIMENTS AND RESULT

A. Data Collection from TOGD

In this section, a test on applying the proposed assessment model for open data quality is described. The tested data were collected from data.go.th. (Accessed Date: 14 June 2016). The received data comprised of 1,366 datasets as each dataset per a file. The files can be categorized with their data domain from the source into 17 categories. A preprocess of data validation was performed to remove non-machine-readable data such as image and PDF format files. Moreover, datasets, which contains metadata, multiple-table or additional note among cells, were also discarded. After preprocessing, there were 161 datasets. The data were separated by its category.

TABLE IV
OVERALL SEMANTIC TYPE DEGREE ASSESSMENT RESULT BY DATA CATEGORY

Data Category	Semantic-typed columns		Popularity of semantic type				Understandability degree of header		Standard of semantic-typed columns	
	PSC		PSTD		PSTS		PHU		SSC	
	AVG	SD	AVG	SD	AVG	SD	AVG	SD	AVG	SD
Religion, Art, and Culture	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Transport and Logistics	0.30	0.33	0.25	0.01	0.09	0.00	0.80	0.30	1.00	0.00
Government Budget and Spending	0.18	0.11	0.92	0.27	0.34	0.13	0.98	0.06	0.98	0.14
Economy, Finance, and Industry	0.21	0.07	0.25	0.01	0.12	0.06	0.89	0.10	1.00	0.00
Healthcare	0.09	0.07	0.50	0.36	0.17	0.12	0.98	0.03	0.36	0.33
Law, Crime, and Justice	0.28	0.12	0.26	0.10	0.08	0.03	0.88	0.34	0.88	0.34
Society and Welfare	0.17	0.18	0.21	0.12	0.07	0.04	0.95	0.08	0.79	0.43
Energy, Natural Resource, and Environment	0.22	0.11	0.29	0.08	0.10	0.03	0.79	0.09	1.00	0.00
Agriculture and Irrigation	0.24	0.21	0.40	0.43	0.14	0.14	1.00	0.00	0.79	0.40
Location, Tourism, and Sport	0.49	0.16	0.41	0.22	0.14	0.07	0.94	0.13	1.00	0.00
Science, Technology, and Innovation	0.33	0.00	0.88	0.00	0.27	0.00	1.00	0.00	1.00	0.00
Information and Communication Technology	0.21	0.08	0.42	0.29	0.14	0.08	0.96	0.07	1.00	0.00
Education	0.10	0.08	0.09	0.09	0.03	0.03	0.27	0.18	1.00	0.00
Politics and Government	0.25	0.22	0.45	0.41	0.15	0.14	0.66	0.46	0.60	0.55
Map	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Climate and Disaster	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Digital Object Identifier and Standard Code	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

B. Assessment Results

In this part, data from TOGD were evaluated regarding semantic type degree to inform possibility in making linked open data. All datasets were assessed in scores following the proposed aspects. The results for semantic type degree by its data category are given in Table IV. The results of linkability from overall are given in Table V.

TABLE V
OVERALL SCORES OF FOUR LINKING TYPES FROM ALL DATASETS

Linking Types	Current Score
Merge	8 pattern (72 dataset)
Merge (added)	2 pattern (18 dataset)
Overall score of In-Linking	0.082
Overall score of Ex-Linking	6.532

From the results of semantic type degree, the calculated scores showed that datasets in different data category had different semantic degree quality. The proposed model can discriminate the high and low quality of semantic type degree with the calculated score. From observation, the datasets with low score accordingly contained issues with non-standard content and incomprehensive header while the high score ones relatively had less issued content. This indicates that the proposed model works as intended.

TABLE VI
CORRELATION OF SEMANTIC TYPE DEGREE AND LINKABILITY ASSESSMENT

Characteristics	In-Linking	Ex-Linking
Semantic-typed columns	0.42	0.50
Popularity of semantic type :in dataset-based view	0.85	0.85
Popularity of semantic type : in semantic-type-based view	0.88	0.79
Understandability degree of header:	0.66	0.58
Standard of semantic-typed columns	0.55	0.43

For linkability degree, the results showed the very few number linkable datasets. There were 8 patterns that are commonly used in which results in merge type linking while there were 2 patterns for merge-added type. The In-Linking score signifies that there are very few datasets containing common PK; thus, the score is obviously low. Last, the score of Ex-Linking is implied that this linking type was tentatively available more than an In-Linking one.

Moreover, we are interested in correlation between semantic type degree and linkability degree; hence, we calculate correlation between the two results. We expect that the score of semantic type degree should correlate to the linkability. Thus, we obtain the positive correlation between these two scores as shown in Table VI.

V. CONCLUSION

This research presents an assessment model for measuring semantic and linkable potential of open data to support an improvement towards linked open data readiness. The model includes two main aspects as semantic type degree and linkability. The former focuses on quality of data involving in semantic meaning and frequency of existing semantic concepts. The latter is to find a possibility to create links within datasets. An output of assessment is a calculated score to discriminate difference of data potential in range of 0 to 1. Datasets in this work are the data provided in open government data of Thailand for preparation towards improving to linked open data readiness.

From experiments, the results showed that the model performed smoothly and were capable to discriminatively generate a score as intended. The assessment results of testing datasets indicated that the dataset required adjustment in semantic quality to improve linking potentials while shared semantic types were low in frequency. The correlation results of both proposed aspects signified that scores of semantic type degree were positively correlated to score from linkability.

For improvement, we plan to research of a method to increase linkability of the given datasets. Furthermore, a method to notify or justify inappropriate data in terms of linked open data will be invented for reducing a chance of new dataset with low quality.

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Berners-Lee, T. (2006). “Linked data – design issues”. <http://www.w3.org/DesignIssues/LinkedData.html>.

[2] Ontotext. (2017). “What are Linked Data and Linked Open Data?”. <https://ontotext.com/knowledgehub/fundamentals/linked-data-linked-open-data/>.

[3] EU open data portal. (2017). <https://data.europa.eu/euodp/en/home>.

[4] Global open data index. (2017). “Place overview”. <https://index.okfn.org/place/>.

[5] NASA Open Data Projects. (2017). <https://www.nasa.gov/nasa-open-data-projects>.

[6] Government Open Data of Thailand website. (2016). <https://data.go.th/>.

[7] Open Data Handbook. (2016). “What is Open Data?”. <http://opendatahandbook.org>.

[8] Ubaldi, B. (2013). “Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives”. OECD Work, Paper, Public Gov, 22.

[9] Open Knowledge Foundation. (2016). “Open government data”. <http://opengovernmentdata.org>.

[10] Bates, J. (2012). “This is what modern deregulation looks like: Co-optation and contestation in the shaping of the UK’s Open Government Data Initiative”. The Journal of community information, Vol. 8(2).

[11] Robinson, D., Yu, H., Zeller, W.P., and Felten, E.W. (2009). “Government data and the invisible hand”. Yale J. Law Tech., Vol. 11, pp. 160-175.

[12] Rojas, L., Bermúdez, G., and Lovelle, J. (2014). “Open data and big data: A perspective from Colombia”. In: Uden, L., Fuenzaliza Oshee, D., Ting, I.H., and Liberona, D. (eds.) Knowledge Management in Organizations, Lecture Notes in Business Information Processing, Springer International Publishing, Vol. 185, pp. 35-41.

[13] Alexopoulos, C., Zuiderwijk, A., Charalabidis, Y., Loukis, E., and Janssen, M. (2014). “Designing a second generation of open data platforms: Integrating open data and social media”. 13th International Conference on Electronic Government (EGOV), Dublin, Ireland, pp. 230-241.

[14] 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).

[15] Bizer, C., Heath, T., and Berners-Lee, T. (2009). “Linked data-the story so far”. International journal on Semantic Web and Information Systems, Vol. 5(3), pp. 1-22.

[16] Wang, R.Y. and Strong, D.M. (1996). “Beyond accuracy: what data quality means to data consumers”. Journal of Management Information Systems, Vol. 12(4), pp. 5-33.

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

[18] Alexander, J.E. and Tate, M.A. (1999). “Web wisdom: How to evaluate and create information quality on the web”. Mahwah, NJ: Erlbaum.

[19] 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, Vol. 33(2), pp. 325-337.

[20] Thipphayasaeng, P., Boonbrahm, P., and Buranarach, M. (2017). “A Framework to Assess Dataset Linking Readiness of Open Government Data”. The 2nd IEEE International Conference on Science and Technology, Rajamangala University of Technology Thanyaburi.