JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

RESEARCH PAPER
J Inf Sci Theory Pract 11(3): 43-57, 2023

eISSN : 2287-4577 pISSN : 2287-9099

https://doi.org/10.1633/JISTaP.2023.11.3.4

Knowledge Graph of Administrative Codes in Korea: The Case for Improving Data Quality and Interlinking of Public Data

Haklae Kim*
Department of Library and Information Science, College of Social Sciences, Chung-Ang University, Seoul, Korea
E-mail: haklaekim@cau.ac.kr

ABSTRACT

Government codes are created and utilized to streamline and standardize government administrative procedures. They are generally employed in government information systems. Because they are included in open datasets of public data, users must be able to understand them. However, information that can be used to comprehend administrative code is lost during the process of releasing data in the government system, making it difficult for data consumers to grasp the code and limiting the connection or convergence of different datasets that use the same code.This study proposes a way to employ the administrative code produced by the Korean government as a standard in a public data environment on a regular basis. Because consumers of public data are barred from accessing government systems, a means of universal access to administrative code is required. An ontology model is used to represent the administrative code's data structure and meaning, and the full administrative code is built as a knowledge graph. The knowledge graph thus created is used to assess the accuracy and connection of administrative codes in public data. The method proposed in this study has the potential to increase the quality of coded information in public data as well as data connectivity.

Keywords: administrative code, public data, knowledge graph, data quality

Received: March 5, 2023 **Revised:** March 24, 2023
Accepted: April 9, 2023 **Published:** September 30, 2023

***Corresponding Author:** Haklae Kim
https://orcid.org/0000-0002-2616-421X
E-mail: haklaekim@cau.ac.kr

1. INTRODUCTION

Public sector data are closely related to the daily lives of citizens in terms of transportation, energy, administration, weather services, and so on. Governments and enterprises worldwide are actively working to promote the use of public data in all industrial fields (Jetzek et al., 2019; Wang & Shepherd, 2020). According to the European Commission, the expected total economic value of public sector information will be approximately 194 billion euros by 2030 (Barbero et al., 2018). By adopting national artificial intelligence strategies, most countries emphasize the disclosure of large-scale public sector datasets (HolonIQ, 2020). The Korean government is actively promoting the adoption of public data releases, and provides diverse legal and institutional support for their industrial use. Korea has positioned itself as a leading country in the international evaluation of public data (World Wide Web Foundation, 2017). According to 'Government at a Glance,' published by the Organisation for Economic Co-operation and Development (OECD, 2019), Korea ranks first in the open-useful-reusable government data index and fourth in the World Wide Web Foundation's Open Data Barometer, tied with France behind Canada, the United Kingdom, and Australia (World Wide Web Foundation, 2017).

However, the quality of public data remains inadequate for various reasons (Kim, 2019). In particular, public data are often incomplete or lack significant information (World Wide Web Foundation, 2017). Platforms for providing and sharing open data are expanding, and research on data quality is ongoing (Nogueras-Iso et al., 2021). To facilitate metadata-level interoperability and apply metrics to analyze metadata quality, a number of platforms implement standards such as data catalog vocabulary (DCAT) (Maali et al., 2014) and DCAT application profile for data portals in Europe (Kirstein et al., 2019). The quality of the values contained in datasets, on the other hand, remains a challenge. For example, the Korean government has established administrative codes for the common use of information by administrative agencies. A number of codes are established for simplifying government affairs. These codes are managed by government agencies supported by legal guidelines and information systems. The codes are systematically managed by government agencies, which may use existing codes for administration purposes but may also enact new codes if necessary. Information systems operated by individual agencies can use any and all codes registered with the code management system. However, as public data is disclosed in the government information system, the gap between government and public data is widening. When data from a government information system is provided as public data, much contextual information tends to be excluded or lost. A dataset containing only code values creates a situation where users have difficulty understanding the meaning of the code; as a result, the dataset is difficult to link and combine with other datasets. Moreover, there is no way for the user to modify or enhance the value of public data to the government system.

Since public data including code values must ensure accuracy, information used internally by government and public data must be linked to each other so that they can be used as up-to-date information. To solve the current limitations associated with government codes, this study proceeds as follows: (1) a knowledge base of administrative codes for interlinking other datasets is created; (2) missing or improper values by the knowledge graph are revised and interlinked among the collected datasets. The knowledge graph can represent administrative codes in a machine-readable format and thus provide semantic relationships across public data. An ontology model is proposed to describe a set of administrative codes. Using this model, the administrative codes aggregated from the government code system are transformed into referenceable knowledge as a graph structure. The administrative codes are provided in Excel format without any relevant contextual information, making it difficult to understand and use the code in an open environment. All 314 established administrative codes were collected from the government code management system and transformed into a graph structure using our proposed knowledge model. The knowledge graph is then applied to five datasets for quality evaluation of public data; this process diagnoses the use status of administrative codes and evaluates the quality of each dataset. In particular, two metrics (completeness and accuracy) are used to determine the quality of columns matched to the administrative codes (Song & Kim, 2022; Vetrò et al., 2016).

The structure of this paper is as follows: the second section reviews previous studies. The third section describes administrative codes in Korea and methods for representing a knowledge graph. The fourth section describes the case of applying the constructed knowledge graph to public data from a quality perspective. The fifth section discusses several issues for utilizing the proposed approach. Finally, the concluding section summarizes the study and proposes future research directions.

 https://doi.org/10.1633/JISTaP.2023.11.3.4

2. RELATED WORK

A number of studies on open government data (OGD) have been conducted with respect to various subjects (Ubaldii, 2013), including strategies (Wang & Lo, 2016; Yang et al., 2015), economic values (Zeleti et al., 2016), data quality (Vetrò et al., 2016), and technical implementations (Janssen et al., 2012). However, the efficient utilization of government data presents several particular issues. Crusoe and Melin (2018) conducted a systematic literature review to identify barriers to OGD. Most of these are technical, organizational, or legal (Crusoe & Melin, 2018). In particular, technical barriers are broad in scope, ranging from publishing the data to enabling consumers to use them. Assessment and quality control of published data are critical factors in the utilization of public data. Kubler et al. (2018) proposed an open data portal quality framework that aims to evaluate open data portals across 43 different countries. Vetrò et al. (2016) proposed a theoretical measurement framework to assess dataset quality. The approach proposed here focuses on quantitatively assessing the quality of government datasets rather than evaluating open data portals.

Various studies have considered linked data for semantic linking of public data. Linked data interlinks among various data sources and is dependent on the quality of individual datasets. Shadbolt et al. (2012) stated that the massive influx of heterogeneous data without semantics or structure has become a problem affecting OGD. Furthermore, researchers have claimed that combining OGD with linked data technologies can leverage the scope and richness of government data because additional resources become interlinked with appropriate contexts (Shadbolt et al., 2012). Researchers are actively investigating the application of linked data technology at the national level in Brazil (Breitman et al., 2012), Singapore (Raamkumar et al., 2015), and Serbia (Janev et al., 2018). Multiple studies have been conducted in specific domains, such as statistical data (Han & Lahiri, 2019; Höffner & Lehmann, 2014), legal data (Mockus & Palmirani, 2017), and data quality and assessment (Ibáñez et al., 2019). Matsuda et al. (2018) established a unified structure for publishing and using statistical data with standard vocabularies such as resource description framework (RDF) and SPARQL (SPARQL Protocol and RDF Query Language) for the Japanese statistical center. Several researchers have presented specific ontology models to represent OGD in a semantic model (Ferneda et al., 2016; Jiang et al., 2019; Petrušić et al., 2016; Zeleti et al., 2016). For instance, Lamharhar et al. (2015) introduced a knowledge-based technique for the automatic processing of heterogeneous public administrative bodies for e-government domains. Daraio et al. (2016) proposed ontology-based data management for a comprehensive level of interoperability among different open datasets, including a data management strategy, high-quality semantic annotation, and ontology mapping.

Governments worldwide and international organizations provide various types of data and classifications as linked data, such as the Global Standards One web vocabulary (Harrison et al., 2014) and legal entity identifiers (Trypuz et al., 2016). In particular, the linked data registry developed by the Commonwealth Scientific and Industrial Research Organization provides vocabulary, ontologies, and reference resources authorized or adopted by it.[1] The Australian government provides the Classification of the Functions of Government (COFOG), defined by the United Nations as linked data based on the simple knowledge organization system (SKOS) model.[2] Moreover, the Australian government architecture framework assists in the delivery of more consistent and cohesive services to citizens and supports the cost-effective delivery of ICT (Information Communications Technology) services by the government. This framework contains a set of reference models for collaboration within and across agencies. In particular, the data reference model aims to deliver consistent data context as the basis for data governance, which can be achieved using an ontology-based approach (Australian Government Information Management Office, 2011). The Italian National Institute of Statistics and the Agenzia perl'Italia Digitale provide an official classification system in a linked data format by linking COFOG, Public Record Office Victoria (Lebo et al., 2012), and the eXtended knowledge organization system (Lodi et al., 2014). From the perspective of public data, these efforts can be considered to increase the connectivity and accuracy of data by consistently referencing codes, classification systems, and vocabularies used by governments. Kim (2018) proposed a knowledge model to represent the legal definition of administrative districts and their interrelationships in Korea and demonstrated the interlinkage of various elements such as addresses, postal codes, hospitals, and schools in an administrative-district knowledge graph.

Because there is insufficient quantitative research on

[1] http://registry.it.csiro.au
[2] https://github.com/CSIRO-enviro-informatics/cofog-a-vocab

the use of administrative codes for public data, this paper proposes a method for incorporating an administrative code into a knowledge graph to improve data quality. The proposed knowledge graph can represent administrative codes in a machine-readable format; the dataset generating the knowledge graph establishes a semantic relationship across public data.

3. A KNOWLEDGE GRAPH OF ADMINISTRATIVE CODES

3.1. Overview of Administrative Codes in Korea

An administrative code refers to a code system that can be classified in the administrative work of each level of government agency so that they can be handled easily according to a certain code (Ministry of the Interior and Safety [MOIS], 2017). In Korea, the administrative standard code is a set of administrative codes established and published according to the prescribed procedure by standardizing the administrative code required for the administrative work of each level of agency. Note that an administrative code is designed with the goal of classifying and simplifying public agency administrative tasks. The administrative standard code in Korea is a particular type of administrative code that establishes a set of codes necessary for standardizing each agency's administrative work. The term 'standard' refers to code that is the product of government legislation, and is semantically different from general standards in the ICT field. Therefore, this study refers to administrative code rather than administrative standard code.

The Electronic Government Act (MOIS, 2017) provides the legal basis of the administrative code for operating and managing administrative code sets. The purpose of this act is to implement e-government effectively by prescribing basic principles, procedures, and methods for the electronic processing of administrative affairs. According to Article 50, the Korean MOIS may establish a set of administrative codes as needed for sharing administrative information across government agencies. In addition, it may publish them in the official gazette (Article 59, Paragraph 1 of the enforcement decree of the Act) and the head of the administrative agency should comply with the codes established in accordance with Article 59, Paragraph 3. The administrative code covers central administrative agencies (including agencies reporting directly to the president and prime minister), agencies affiliated with central administrative agencies, and local government agencies. The Korean government recommends the use of the administrative code in general administrative affairs via specific guidelines: (1) all agencies should use up-to-date administrative code from the administrative code management system,[3] (2) a new informatization project should use the administrative code; if the existing codes cannot be applied, the institution must consult with MOIS in advance to derive application plans, and (3) if the administrative code has not yet been applied in the existing information system, the organization should discuss applying the code and improving interoperability between systems with MOIS.

Eleven types of administrative code were enacted in 1990 to promote e-government. These have been expanded to 314 types through several additional enactments and supplements. All currently established administrative codes are managed through the administrative code management system. The administrative code is classified into 26 subjects as administrative tasks. Note that the administrative codes represent individual subjects as strings, but there is no consistent method for identifying individual subjects. To identify them, all subjects and codes were defined by an 'S' and 'AC' (Administrative Code) and an arbitrarily assigned number. By subject (Fig. 1), S15 (personal administration) and S21 (local tax payment) were the most numerous with 29 codes (9.24%) each, whereas S2 (construction), S20 (local taxation), and S4 (transportation and logistics) had 25 (7.96%), 24 (7.64%), and 19 (6.05%) administrative codes, respectively.

Each code has a 1:1 match between code value and its meaning. The code value is detailed according to a specific subject as a simple number or blended with other characters, whereas its meaning is a name expressed in natural language. By default, each code consists of a value, a code name, and a comment as a column. The number of columns ranged from a minimum of three (97% of codes) to a maximum of 25. Note that eight administrative codes have more than three columns; as shown in Fig. 2, AC56 (national license), AC17 (construction classification), AC243 (functional category), AC192 (postal code), AC247 (job position), AC61 (international currency), and AC211 (type of disability) have more than four columns, while AC67 (organization), AC83 (road name), AC246 (job classification), AC164 (raw food material), AC298 (university major), AC162 (qualification test), AC79 (standard item), AC163 (food safety classification item), and AC242 (job

[3] http://code.go.kr

 https://doi.org/10.1633/JISTaP.2023.11.3.4

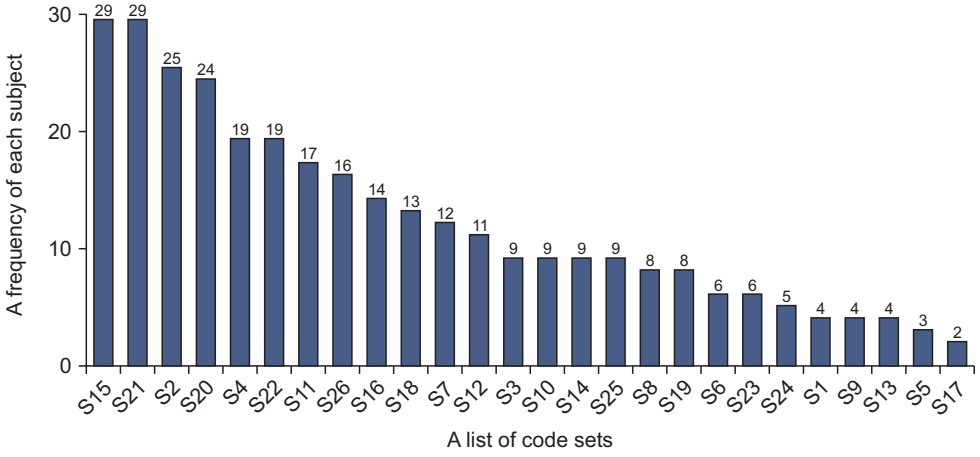

Fig. 1. Statuses of data by subject.

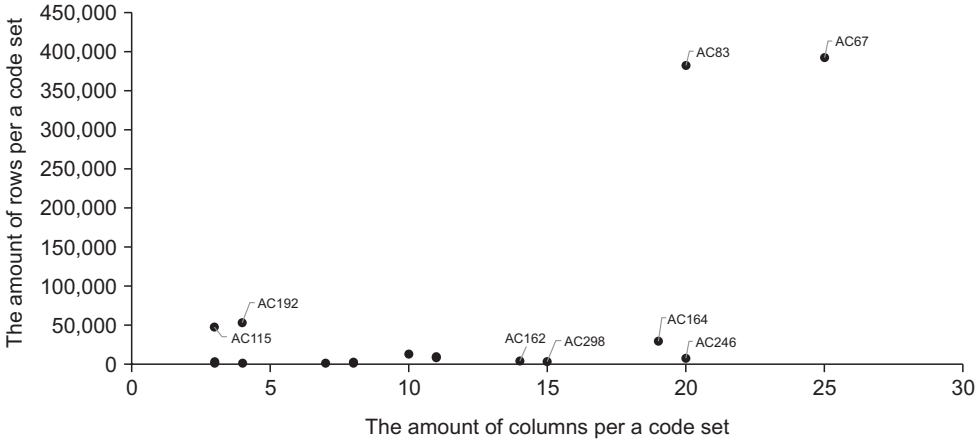

Fig. 2. Data status by the number of rows and columns. AC67 (391,377), AC83 (381,415), AC164 (28,337), AC246 (6,690, job classification), AC162 (6,690, qualification test), and AC298 (2,592, university major) have a relatively large number of codes and values.

class) have more than 10 columns. The number of rows in the datasets varied from a minimum of two to a maximum of 391,377. The number of rows is 10 or less for 43% and between 11 and 100 for 41%; approximately six (2% of code types) contain 10,000 or more codes. In particular, AC67 and AC83 have the largest number of columns and rows (25×391,377 and 20×381,415, respectively) because they contain detailed specifications of items in addition to the code values, as shown in Fig. 2. Although AC192 (postal code) and AC115 (administrative district) contain four and three columns respectively, they contain 52,284 and 46,215 codes, respectively. These datasets define all postal codes and administrative districts in the Republic of Korea. These codes are commonly used in government and national information systems.

3.2. Knowledge Model

Some existing vocabularies are repurposed to represent the administrative codes at a semantic level, namely schema.org (Guha et al., 2015), RDF (Brickley et al., 2004), and SKOS (Miles & Brickley, 2005). For example, a set of descriptive information is represented by RDF, while the relationship between concepts is described using SKOS. New terms are defined only if they do not correspond to the existing vocabulary. In schema.org, the CategoryCodeSet class describes a set of category code values. CategoryCode expresses a short textual code that uniquely identifies the value using the codeValue property. The two classes can be linked with the hasCategoryCode property, which aims to represent a code contained in a code set. To represent a set of administrative codes, the AdministrativeCodeSet and AdministrativeCode classes are defined as subclasses of the CategoryCodeSet and CategoryCode, respectively. These classes inherit the relationships defined in schema.org. Thus, the hasCategoryCode property of schema.org is reused to connect to both AdministrativeCodeSet and AdministrativeCode. In addition, each code set has one subject and one category; the Subject and Category are subclasses of skos:Concept. The category and subject properties are linked to the corresponding code set.

As shown in Fig. 3, several classes are defined to represent the characteristics of administrative codes. Each code set is maintained by at least one government body. The organization property describes an agency that creates or operates a specific code set. The expected value of this property is a type of Organization within schema.org. A management of organizational codes is classified into four types: general, competent, application, and consignment. The general organization manages overall tasks such as the designation of the competent agency (department) of the administrative code, management of enactments/revisions, and notifications. Competent organizations manage the establishment and revision of sets of administrative codes. Application organizations are responsible for applying administrative codes to an internal information system. Thus, these types are defined as types of Organization classes. Table 1 lists four types. The administrative codes are classified into four management types according to the characteristics of the competent organization. The management type is defined by the ManagementType class. The AdministrativeCodeSet class can be linked to a specific management type with the managementType property.

The URI (Universal Resource Identifier) pattern is designed to be consistent and predictable. The experiences and recommendations of linked data communities, including data publishers and consumers, are used to develop common patterns and practices (Juty et al., 2020). A vocabulary URI includes ontology, vocabulary, concept schemes, and code sets. The pattern, combining a domain address and a reference item, is an important resource. The reference item contains the reference data. In this

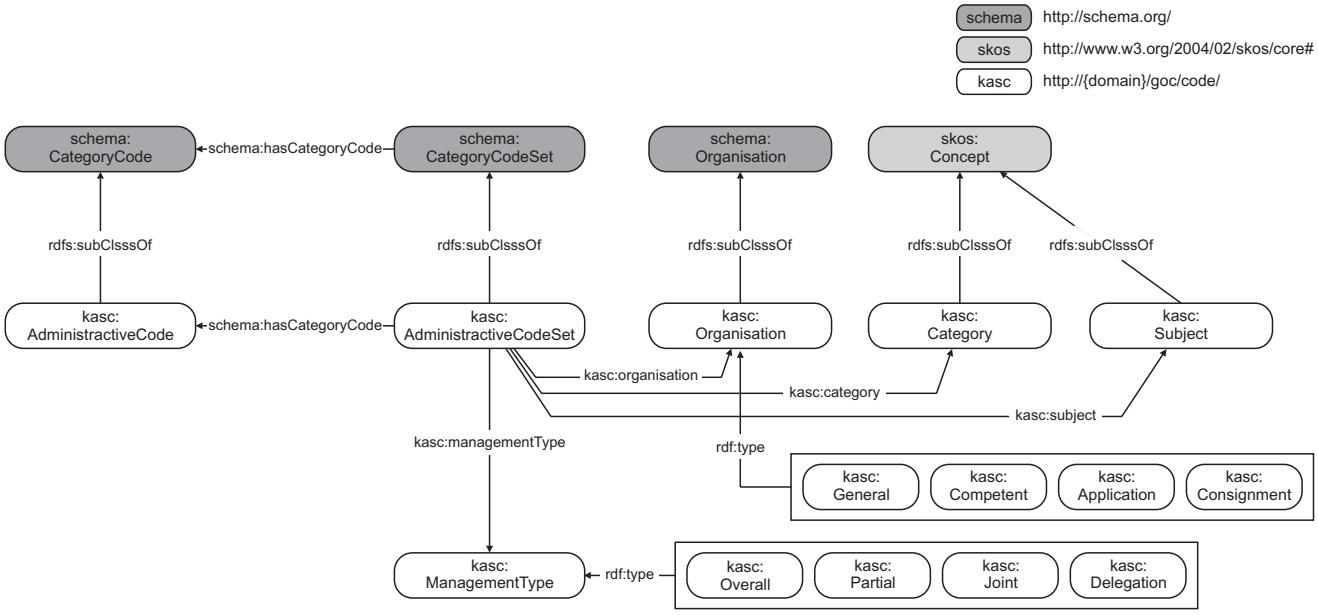

Fig. 3. Abstract model of administrative code representation. The prefix of the namespace is 'kasc' (Korean Administrative Standard Code).

Table 1. Management types and classes in the knowledge model

Management type	Class name	Description
Overall	kasc:Overall	The code is managed by one competent organization.
Partial	kasc:Partial	If the competent organizations of each part differ, the competent organization (department) for each part may be determined.
Joint	kasc:Joint	If there are multiple agencies in charge of one code, these can be designated, and the change can be managed through a mutual agreement.
Delegation	kasc:Delegation	When it is difficult for the competent organization to directly manage the change management of the lowest code value, the change management function can be delegated to the subordinate organization.

 https://doi.org/10.1633/JISTaP.2023.11.3.4

case, a reference item is a government agency, and the reference data are the administrative codes provided by the agency. This relationship is represented by the URI pattern 'gov' and 'code,' listed in a specific order. Furthermore, the vocabulary and instances of the subject are defined by classifying them into 'def' and 'id,' respectively. Therefore, the vocabulary URI is provided at http://{domain}/gov/code/def. Note that the individual ontology vocabulary is expressed in CamelCase notation such as 'http://{domain}/gov/code/def/AdministrativeCodeSet' and 'http://{domain}/gov/code/def/AdministrativeCode.' The URI for representing the dataset is defined as 'http://{domain}/gov/code/id.' However, because administrative code values have a relatively simple composition (e.g., '01,' '02,' and '03'), it is difficult to ensure the uniqueness of the code URI using only the code value. Therefore, a code identifier is defined by combining code set numbers and code values. For example, instances of the AdministrativeCodeSet and AdministrativeCode classes are defined by combining the individual code value (AC) of the administrative code with the serial number assigned to the code data. The AC creates combinations of administrative code data and code values (e.g. 'http://{domain}/gov/code/id/AdministrativeCodeSet/AC/{code set number}).

All code datasets, containing 1,011,438 individual codes, were downloaded individually as Excel files. The knowledge graph conversion was done using Python and RDFLib,[4] and the linking of the government code contained in the open data was done using OpenRefine's reconciliation service. These codes are transformed into graph-structured data by adopting the proposed knowledge model and the URI model. The administrative code as constructed by the knowledge graph contains approximately 1,011,812 entities and 20,481,472 statements. Table 2 summarizes the entities and statements for each subject in the knowledge graph.

4. EVALUATIONS

4.1. Data Collection

A small-scale evaluation was conducted to verify the use of knowledge graphs for public data. On the public

Table 2. Statistics of the knowledge graph by administrative code

Category	Type	Description	Entities	Statements
Metadata	Organizations	Agencies that manage the administrative codes	34	248
	Subject	Topics that are classified by government tasks	26	208
Administrative code	Code set	Total number of the administrative codes	314	23,496
	Code	Total number of individual codes	1,011,438	20,457,520
Total			1,011,812	20,481,472

Table 3. List of public datasets from the Korean open data portal

Dataset	Description	ID	Columns	Rows	Views	Downloads	File size
D_1	Administrative organizations in Korea	15061082	8	13,460	360	233	2 MB
D_2	Public health and medical institution	15004305	15	227	3,380	874	51 KB
D_3	Medication prescriptions	15007117	15	36,380,226	4,326	16,709	3.27 GB
D_4	Current status of the beauty industry in Gyeonggi Province	15038408	37	1,200	266	162	266 KB
D_5	National parking lots	15012896	33	14,609	-	-	4.1 MB

The ID column is an identifier assigned by the public data portal, and the ID for accessing the actual data is 'https://www.data.go.kr/data' + 'ID' + '/fileData.do.' However, D_5 has a different URL (https://www.data.go.kr/data/15012896/standard.do) owing to the type of standard data.

[4]https://rdflib.readthedocs.io/en/stable

data portal,[5] a search for the keyword 'public' was conducted and the top five datasets were selected from the results. As shown in Table 3, Dataset D_1 contains types, addresses, and phone numbers of administrative agencies such as post offices, public health centers, and administrative division offices. Dataset D_2 contains information about medical institutions, including the type, number of beds, and addresses. Dataset D_3 is composed of details of individually prescribed medicines (e.g., date of commencement of treatment, appropriate dosage, daily dose, and total dose) on one million patients per year. Dataset D_4 provides the names of businesses in the city of Icheon in Gyeonggi Province such as hair, makeup, and nail art studios, as well as licensing dates and business status. Dataset D_5 provides information on the addresses, locations, types, operating times, and fees for parking lots nationwide. D_3 is a very large dataset (approximately 3.27 GB) with the maximum number of rows, whereas D_4 is relatively small because its data are limited to a specific region. The 'Views' and 'Downloads' columns represent the numbers of users who previewed or downloaded the file contents via the portal. The portal does not indicate the number of views and downloads of D_5. D_3 has high numbers of downloads, but the other datasets have low numbers compared to the number of views. This suggests that users discover and review these datasets but generally do not download them for actual use.

Individual datasets use administrative codes ranging from 11% to 75% of the total. In D_1, for example, six out of eight columns use administrative codes, while in D_5, only 4 columns out of 36 are used. Table 4 summarizes the column to which the administrative code was applied, the matching administrative code, and the results of data quality improvement in the dataset. Matching with administrative codes proceeds in two steps. First, the column value of the dataset determines selection of the category type and comparison with the administrative code. Since the Korean column names are used as different character strings, whether or not they are semantically equivalent to the administrative code names can be determined. In fact, although the column names of the datasets are different, the same administration code is applied. For example, 'Organization type' in D_1 and 'name of management institution' in D_5 can apply the same administrative code. There are 108 columns in the collected dataset and 22 (20%) of them have administrative codes. Ten administrative codes were used in the collected datasets by applying semantic matching: AC41 (public health industry), AC67 (public organization), AC115 (administration division), AC117 (healthcare institution), AC136 (sex), AC166 (type of food hygiene business), AC192 (postal code), AC227 (type of parking lot), AC241 (job position), and AC314 (closed status of business). Ordinarily, one code value or name was applied to each dataset. Most matched columns tend to use code names rather than code values. However, C20 and C21 of D5 (provider of name and code, respectively) are the only exceptions in which both values are used together. Note that the percentage of matches across all columns in the dataset was approximately 20%, whereas 36,409,722 rows had administrative code values.

4.2. Data Quality

Individual data sets can use codes or code names as column values. The values in the combined data set, on the other hand, may include a string whose code value is unknown, or a portion of the extracted value rather than the entire code. For example, Seoul Metropolitan City's administrative code is '1100000000'; however, depending on the institution, only two ('11') or five ('11000') digits may be indicated in front. These values reduce the data's accuracy. Two metrics (C_{mc} and A_{mc}) were used to assess the quality of each dataset: the completeness index (C_{mc}) refers to the proportion of rows in a specific column that are non-empty and have meaningful values compatible with the domain of the column, while the accuracy index (A_{mc}) indicates the percentage of cells for a specific column that have values consistent with the administrative code. The evaluation was conducted in two steps: (1) Determine whether the column value of the collected dataset corresponds to the administrative code value; and (2) re-evaluate the match by correcting some of the column values that did not correspond in step 1. Using the results of the secondary evaluation to ascertain the accuracy, all matched values were converted to the URI of the knowledge graph.

The completeness scores C_{mc} are 0.99 and 1 in the first and second evaluations, respectively, as shown in Table 4. That is, most column values do not have null values or empty strings. Column C16 of D_4 has 181 blanks, however, with a completeness score of 0.85. It includes a zip code, and the correct value can be added in the second evaluation through the address column. By contrast, the average accuracy score A_{mc} for the first evaluation was 0.65. This score was relatively low compared with the complete-

[5] http://data.go.kr

 https://doi.org/10.1633/JISTaP.2023.11.3.4

Table 4. Summary of data column match quality evaluation (the column name is the English translation of the Korean name in the original data)

Datasets	Matched columns (matched/total)	Matched ratio (%)	Columns	Name	Matched code set	1st matched result		2nd matched result	
						C_{mc}	A_{mc}	C_{mc}	A_{mc}
D_1	6/8	75	C1	Organization type	AC241	1	0.25	-	0.25
			C2	Organization sub-type	AC67	1	0	-	0.31
			C3	Representative organization name	AC67	1	1	-	-
			C4	Full organization names	AC67	1	0.64	-	0.87
			C5	Subordinate organization	AC67	1	0.68	-	0.92
			C6	New postal code	AC192	1	1	-	-
D_2	5/15	33	C7	Medical institute name	AC67	1	0.24	-	0.72
			C8	Name of related administrative agencies	AC67	1	0.77	-	0.99
			C9	Related public organizations	AC67	1	0.13	-	0.77
			C10	Types	AC117	1	0.95	-	0.95
			C11	Zip code	AC192	1	0.84	-	1
D_3	2/14	14	C12	Gender code	AC136	1	1	-	-
			C13	City and province codes	AC115	1	1	-	-
D_4	4/38	13	C14	City and county name	AC115	1	1	-	-
			C15	Name of business status	AC314	1	0.5	-	1
			C16	Road-name postal code	AC192	0.85	0.85	1	1
			C17	Information of business category name	AC166	1	0	-	0.7
			C18	Information of hygiene business name	AC41	1	0	-	0.93
D_5	4/36	11	C19	Types of parking lots	AC227	1	1	-	-
			C20	Name of management institution	AC67	1	0.05	-	0.97
			C21	Provider name	AC67	1	1	-	-
			C22	Provider code	AC67	1	0.68	-	1
Average scores						0.99	0.62	1	0.88

For clarity, column names are represented by the combination of the string 'C' and the index number.

ness score. In other words, the column value to which the administrative code is applied does not exactly match the administrative code. Accuracy is verified by applying fingerprints provided by OpenRefine's Key Collision methods to verify code values (Carlson & Seely, 2017). The Key Collision approaches are predicated on the concept of constructing an alternate representation of a value ("key") that contains just the most valuable information. The following inaccuracies can be found in columns with low accuracy:

- Type I - Unmatched categories (C1, C2, C10, C15, C17, C18): C1 of D1, C10 of D_2, and C17 of D_4 contain organizational job classifications, hospital types, and business categories, respectively. C1 contains novel composite information derived from merging

categories defined in AC241 with categories found solely in the dataset. Some categories are matched, but the AC241 categories are connected with the '_' symbol, or else new categories that are not mapped at all are generated. The administrative codes for C10 and C17 are AC117 and AC166, respectively; however, they also include categories that do not exist in the administrative codes. Meanwhile, C17 does not use the administrative code (AC166: beauty industry), despite the fact that this code exists. Instead, there is a new value in this column.

- Type II - Use of incomplete codes (C4, C5, C20, C22): AC67 includes comprehensive information about administrative agencies. Organization names, for instance, are given by splitting them into columns labelled 'full organization name' and 'organization name,' respectively. The column 'full organization name' shows all of the organization's related relationships, and higher- and lower-level links are separated by spaces. For example, because the 'National Archives of Korea' are part of the Ministry of Public Administration and Security, the full organization name is written as 'The Ministry of Public Administration and Security National Archives,' while the column 'organization name' only includes the organization name without the higher affiliated institutions. If the affiliation relationship is complex, only a portion of the values are taken and used; hence institutional names in the collected data set are inconsistently indicated. C22 comprises codes that are not in the category, and a portion of the code value is modified and used instead of the entire code value.
- Type III - Composition of complicated information (C7): In D_2, C7 is a combination of administrative districts and medical institution names. As a result, the matching rate for the value of AC67 was low (0.24) in the initial evaluation.
- Type IV - Inclusion of special characters (C2, C8, C9, C18): Some values of C2 and C18 are used by linking them with the '_' symbol. C8 and C9 have detailed parenthetical values of 52 and 192, respectively. C18 employs AC41 with the '_' symbol as well as four extended code names (82 rows) that do not exist in AC41.
- Type V - Blank values and error values (C11, C16): C11 has 35 postal codes expressed as prior address system values, while C16 has 181 values marked with blanks.

Data cleaning can be approached in a variety of ways depending on the type of inaccuracy. Quality improvement is difficult for Type I due to codes that do not exist in the administrative code, so that data refinement does not significantly increase accuracy. For example, the accuracy of C1 and C2 is relatively poor, at 0.25 and 0.31 respectively. In the event of such an issue, it is preferable to register the newly produced or extended code in the administrative code management system and update the open data. Types II and III were improved by separating data values into minimum semantic units and implementing reconciliation services with administrative code provided by knowledge graphs. For example, all C4 values are separated by spaces, and values that match the organization name represented by the knowledge graph are searched. The knowledge graph depicts the hierarchical structures and relationships between organizations in AC67. Through the reconciliation, the separated value finds a matching value and is linked to the knowledge graph's entities. By deleting embedded special characters and whitespaces, Type IV errors can be significantly reduced. Finally, for Type V, the value is either blank or contains the preceding address system's value associated to the address. First, blanked fields in C11 received the address value by searching for and entering the hospital name in the address API.[6] C16 addresses include a postal code, a road name, and a lot number. Cells with no postal code were added by utilizing the address API to look for the value of the record's street name or lot number address. The second matching result refers to the refined data. The completeness score increased from 0.99 to 1, while the accuracy score increased from 0.62 to 0.88. However, because administrative codes and other codes and categories exist in separate datasets, perfect accuracy cannot be guaranteed.

4.3. Data Interlinking

Public data can be combined or integrated with other publicly available data. However, various datasets necessitate the use of common values or standards for integration. The administrative code knowledge graph is a graph structure that gives the structure and value of administrative codes as a standard vocabulary, and it may be used to connect data based on administrative codes included in

[6]https://business.juso.go.kr/addrlink/openApi/apiExprn.do?cPath=99JA

 https://doi.org/10.1633/JISTaP.2023.11.3.4

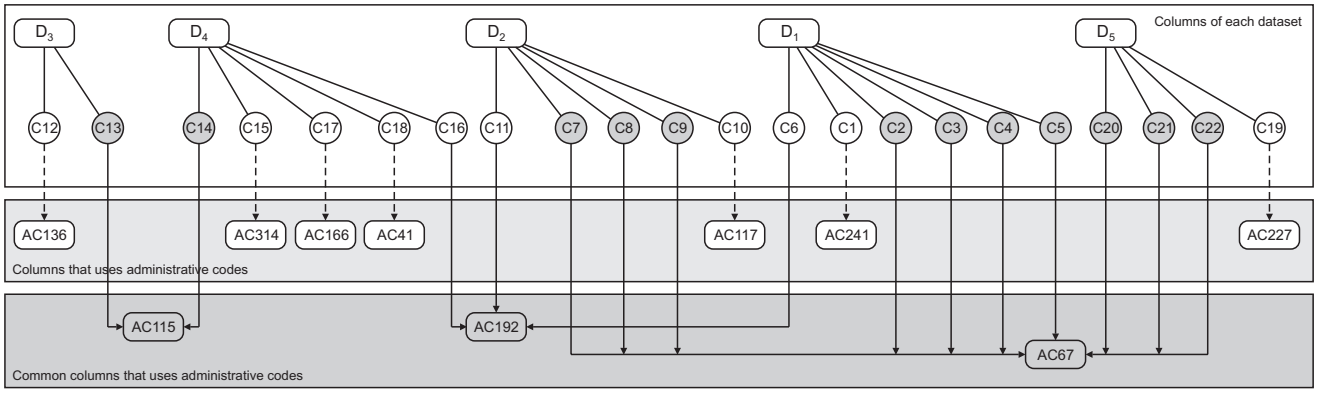

Fig. 4. Interlinking between datasets with the administrative codes.

public data.

In the five datasets studied here, 22 columns were mapped to the administrative codes. Seven codes were used only in individual datasets, and three codes were commonly used across datasets. Fig. 4 illustrates the relationships among the five datasets based on administrative codes. The established knowledge graph provides a consistent identifier of individual codes, allowing each code to be used to interlink datasets. Although the column names (in Korean) defined in datasets are different, the columns that use the same administrative code are semantically interlinked by the knowledge graph. Codes commonly used between different datasets can be used as reference points for linking them. In particular, AC67 (organization) is commonly applied to 11 columns. This code can be widely used for public data. The administrative division code for C13 of D_3 and C14 of D_4 are the same. Additionally, D_1, D_2, and D_4 commonly use AC192. For example, if relevant datasets exist with the administrative codes, use cases would be extensible.

There is a limit to generalization of the proposed method based on only five datasets. In the aforementioned example, not all datasets can be linked by an administrative code. However, independently used codes can be used to link different datasets. By constructing the administrative code as a knowledge graph, the task of individually verifying and evaluating the values of a dataset can be significantly reduced, and this can be a starting point for discovering opportunities for linking datasets. In other words, the knowledge graph provides a common criterion for linking various datasets and can present relationships between datasets in a way that computational algorithms can understand.

5. DISCUSSION

By converting administrative codes used in government administrative affairs into knowledge graphs, this research presents a strategy for improving the quality of public data and linking different data with administrative codes. Because information on administrative codes contained in public data may be lost throughout the process of disclosure to the private sector, the method suggested in this study can be used to improve the quality of public data. However, because public data encompasses information from multiple domains and public sectors, it must be augmented.

- Expansion of codified data: The administrative codes are governed by law (MOIS, 2017). Classification codes that are not registered as administrative codes, on the other hand, are commonly utilized. Searching the public data site for 'classification code' yields around 4,609 datasets, but not all of them contain administrative codes. The Korean Statistical Classification is a code system that systematically categorizes economic activities performed mostly by businesses based on their commonalities. Other than for statistical purposes, this code is frequently used in general administrative and industrial policy domains, and it is also included in public data: The Korean Standard Industrial Classification ensures the accuracy and comparability of industry-related data, while the Korean Standard Classification of Occupations is used for classifying and aggregating occupational information obtained through statistical censuses and surveys, and the Korean Standard Classification of Diseases and Causes of Death is used for classifying diseases and other health prob-

lems recorded on many types of health and vital records. Coded data identifies datasets and simplifies data linking by using common coding. As a result, it is vital to investigate linking active codes in the public and private sectors.

- Consistent updating of disclosed codes: Because administrative code was originally developed for government operations, it is difficult to examine code modifications and new information from the commercial sector in real time. Although administrative codes included in public data can be universally accessed via knowledge graphs, it is difficult to provide exact values for public data containing codes that do not fit the administrative code registration system. For instance, new code registration is restricted to public institutions, making it impossible for public data users to submit comments. If it is legally difficult to directly edit or register the code in the private sector, the provider of public data can consider modifying the code data and re-opening it by gathering opinions. On the other hand, it is necessary to check the association with a categorization system that is not registered as an administrative code throughout this process.
- Administrative code system: Administrative codes are opened in the registration system as Excel files. While individual codes include codes and code values, the majority of individual codes are represented by numbers such as 1, 2, and 3. As a result, code values cannot be utilized to identify specific codes. Because the role of administrative codes is growing more significant and its study is developing, an adequate identification system must be established. On the other hand, the meaning of the code value offers only a plain string, which is insufficient to interpret the meaning. As a result, the code value must be added and described in full. Because there is no relationship between individual codes, the method must be reviewed for thorough expression of the relationship between organizations, subjects, and codes. The SKOS-based data model suggested in this paper can serve as a starting point for expressing the link between administrative codes.

6. CONCLUSION

This paper proposed a method for expressing governmental administrative codes in a machine-readable graph as well as methods for improving the quality of public data. Administrative codes are widely used in government information systems, and public data disclosed by the government contain a significant number of administrative codes. Public data contains these codes; however, most datasets remain obscure due to use of abbreviations without sufficient metadata. Data users find it difficult to identify these values within public datasets, thus reducing the quality of public data and preventing their widespread use. Furthermore, if code added or changed by the institution itself is included in public data, determining the accuracy of the code becomes difficult. If government codes had a standard representation for referencing, this would be the primary tool for interlinking among various public datasets and for improving data quality. Additionally, the effort required by users to clean up the data could be minimized.

In this study, 314 administrative codes established by the Korean government were transformed into a knowledge graph. The proposed ontology model is designed to represent recent information applied in related laws and administrative code systems. The vocabulary of the ontology model is designed to incorporate administrative code into a knowledge graph and is extensively repurposed. For example, code sets and codes are defined by extending classes from schema.org. According to the evaluation, approximately 20% of all columns used administrative codes. When it comes to the quality of public data, accuracy tends to be lower than completeness. Although the completeness index was close to 1 in the evaluation results, the accuracy was relatively low at 0.62. Two reasons for this can be identified: (1) the code value is either incomplete or expressed as a string rather than a code; (2) the agency is using the administrative code by itself, either expanding or changing it. In particular, there is a considerable amount of research on diagnosing data quality, but there is a limit to how to refine and improve the actual data. This study diagnoses data quality and improves it by applying the constructed knowledge graph. After applying the value defined in the knowledge graph, the accuracy improved to 0.88. On the other hand, the knowledge graph can be applied to semantically connect disparate datasets. In the five datasets, administrative codes such as AC136 are used only for D_3, whereas A67 was used with ten columns in three datasets. Thus, A67 can serve as a reference between different datasets, i.e., it can be interpreted that public data containing government codes share an identifier that can be linked to each other. The connections between datasets will increase as the scope of public data is expanded. The established knowledge graph structurally expresses the

 https://doi.org/10.1633/JISTaP.2023.11.3.4

administrative codes with the latest information in the administrative code system and semantically defines the relationships among them.

As public data becomes more open and various data can be linked to each other, it is critically important to provide accurate data. Various codes or classification systems established by government legislation are widely used in public data. Therefore, it is necessary to represent the data included in the public data accurately. In particular, data from the public sector are in high demand by governments, companies, and individual citizens. This means that public data can be linked to various data sources in the future. Future research must study public data to include government codes and link them to the created knowledge graph. To do this, the columns contained in the open data must be examined and matched to government codes. In contrast, in addition to the officially used government codes, there are classification systems used as standards by national organizations or private businesses; therefore, it is vital to assess the inclusion of classification systems in public data. Specifically, this task is not a one-time activity, but involves public and private cooperation, and it is desirable to discuss the vocabulary design of the knowledge graph and standards for data interlinking together.

ACKNOWLEDGEMENTS

This research was supported by the Chung-Ang University Research Grants in 2021.

CONFLICTS OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

Australian Government Information Management Office (AGIMO). (2011). *Australian government architecture reference models. Version 3.0.* AGIMO.

Barbero, M., Bartz, K., Linz, F., Mauritz, S., Wauters, P., Chrzanowski, P., Graux, H., Hillebrand, A., de Vries, M., Innesti, A., Ypma, P., Tenge, E., Jakimowicz, K., & Osimo, D. (2018). *Study to support the review of Directive 2003/98/EC on the re-use of public sector information: Final report.* European Union.

Breitman, K., Salas, P., Casanova, M. A., Saraiva, D., Gama, V., Viterbo, J., Magalhaes, R. P., Franzosi, E., & Chaves, M. (2012). Open government data in Brazil. *IEEE Intelligent Systems,* 27(3), 45-49. https://doi.org/10.1109/MIS.2012.25

Brickley, D., Guha, R. V., & McBride, B. (2004). *RDF vocabulary description language 1.0: RDF schema.* https://www.w3.org/TR/2004/REC-rdf-schema-20040210/

Carlson, S., & Seely, A. (2017). Using OpenRefine's reconciliation to validate local authority headings. *Cataloging & Classification Quarterly,* 55(1), 1-11. https://doi.org/10.1080/01639374.2016.1245693

Crusoe, J., & Melin, U. (2018, September 3-5). Investigating open government data barriers. In P. Parycek, O. Glassey, M. Janssen, H. J. Scholl, E. Tambouris, E. Kalampokis, & S. Virkar (Eds.), *Proceeding of the 17th International Federation for Information Processing Working Group 8.5 International Conference, EGOV 2018* (pp. 169-183). Springer.

Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). The advantages of an ontology-based data management approach: Openness, interoperability and data quality. *Scientometrics,* 108(1), 441-455. https://doi.org/10.1007/s11192-016-1913-6

Ferneda, E., Cruz, F. W., do Prado, H. A., da Veiga Guadagnin, R., dos Santos, L. C., dos Santos, D. L. N., & da Costa, O. L. (2016). Potential of ontology for interoperability in e-government: Discussing international initiatives and the Brazilian case. *Brazilian Journal of Information Science: Research Trends,* 10(2), 47-57. https://brapci.inf.br/index.php/res/v/14486

Guha R. V., Brickley, D., & MacBeth, S. (2015). Schema.org: Evolution of structured data on the web: Big data makes common schemas even more necessary. *Queue,* 13(9), 10-37. https://doi.org/10.1145/2857274.2857276

Han, Y., & Lahiri, P. (2019). Statistical analysis with linked data. *International Statistical Review, 87(Suppl 1),* S139-S157. https://doi.org/10.1111/insr.12295

Harrison, M., Beideman, R., Barthel, H., Gray, S., & Traub, K. (2014). *HTTP uniform resource identifiers to associate a web resource with a GS1 key and optional application identifiers.* GS1.

HolonIQ. (2020). *50 National AI strategies - The 2020 AI strategy landscape.* https://www.holoniq.com/notes/50-national-ai-strategies-the-2020-ai-strategy-landscape

Höffner, K., & Lehmann, J. (2014, September 4-5). Towards question answering on statistical linked data. In H. Sack, A. Filipowska, J. Lehmann, & S. Hellmann (Eds.), *Proceedings of the 10th International Conference on Semantic Systems (SEM '14)* (pp. 61-64). Association for Computing Machinery.

Ibáñez, L. D., Millard, I., Glaser, H., & Simperl, E. (2019, October 26-30). An assessment of adoption and quality of linked

data in European open government data. In C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, & F. Gandon (Eds.), *Proceeding of the 18th International Semantic Web Conference (ISWC 2019)* (pp. 436-453). Springer.

Janev, V., Mijović, V., & Vraneš, S. (2018). Using the linked data approach in European e-government systems: Example from Serbia. *International Journal on Semantic Web and Information Systems,* 14(2), 27-46. http://doi.org/10.4018/IJSWIS.2018040102

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management,* 29(4), 258-268. https://doi.org/10.1080/10580530.2012.716740

Jetzek, T., Avital, M., & Bjorn-Andersen, N. (2019). The sustainable value of open government data. *Journal of the Association for Information Systems,* 20(6), 702-734. https://doi.org/10.17705/1jais.00549

Jiang, S., Hagelien, T. F., Natvig, M., & Li, J. (2019, January 30-February 1). Ontology-based semantic search for open government data. In B. Chou, & K. Li (Eds.), *Proceeding of the 2019 IEEE 13th International Conference on Semantic Computing (ICSC 2019)* (pp. 7-15). IEEE.

Juty, N., Wimalaratne, S. M., Soiland-Reyes, S., Kunze, J., Goble, C. A., & Clark, T. (2020). Unique, persistent, resolvable: Identifiers as the foundation of FAIR. *Data Intelligence,* 2(1-2), 30-39. https://doi.org/10.1162/dint_a_00025

Kim, H. (2018). Interlinking open government data in Korea using administrative district knowledge graph. *Journal of Information Science Theory and Practice,* 6(1), 18-30. https://doi.org/10.1633/JISTaP.2018.6.1.2

Kim, H. (2019). Analysis of standard vocabulary use of the open government data: The case of the public data portal of Korea. *Quality & Quantity,* 53(3), 1611-1622. https://doi.org/10.1007/s11135-018-0829-z

Kirstein, F., Dittwald, B., Dutkowski, S., Glikman, Y., Schimmler, S., & Hauswirth, M. (2019, September 2-4). Linked data in the European data portal: A comprehensive platform for applying DCAT-AP. In I. Lindgren, M. Janssen, H. Lee, A. Polini, M. P. Rodríguez Bolívar, H. J. Scholl, & E. Tambouris (Eds.), *Proceeding of the 18th International Federation for Information Processing Working Group 8.5 International Conference, EGOV 2019* (pp. 192-204). Springer.

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of metadata quality in open data portals using the analytic hierarchy process. *Government Information Quarterly,* 35(1), 13-29. https://doi.org/10.1016/j.giq.2017.11.003

Lamharhar, H., Chiadmi, D., & Benhlima, L. (2015, December 11-13). Ontology-based knowledge representation for e-government domain. In M. Indrawan-Santiago, M. Steinbauer, I. Khalil, & G. Anderst-Kotsis (Eds.), *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services (iiWAS '15)* (pp. 1-10). Association for Computing Machinery.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. (2012). *PROV-O: The PROV ontology.* https://www.w3.org/TR/2012/CR-prov-o-20121211/

Lodi, G., Maccioni, A., Scannapieco, M., Scanu, M., & Tosco, L. (2014, October 19). Publishing official classifications in linked open data. In S. Capadisli, F. Cotton, A. Haller, A. Hamilton, M. Scannapieco, & R. Troncy (Eds.), *Proceedings of the 2nd International Workshop on Semantic Statistics co-located with 13th International Semantic Web Conference (SemStats@ISWC 2014).* IEEE.

Maali, F., Erickson, J., & Archer, P. (2014). *Data catalog vocabulary (DCAT).* https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/

Matsuda, J., Mizutani, A., Asano, Y., Yamamoto, D., Takeda, H., Ohmukai, I., Kato, F., Koide, S., Harada, H., & Nishimura, S. (2018, November 26-28). Publication of statistical linked open data in Japan. In R. Ichise, F. Lécué, T. Kawamura, D. Zhao, S. H. Muggleton, & K. Kozaki (Eds.), Proceedings of the Semantic Technology: 8th Joint International Conference, JIST 2018 (pp. 307-319). Springer.

Miles, A., & Brickley, D. (2005). *SKOS core vocabulary specification.* https://www.w3.org/TR/swbp-skos-core-spec/

Ministry of the Interior and Safety. (2017). *Electronic Government Act No. 14914.* https://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=45844&type=part&key=4

Mockus, M., & Palmirani, M. (2017, May 17-19). Legal ontology for open government data mashups. In P. Parycek, & N. Edelmann (Eds.), *Proceedings of the 7th International Conference for E-Democracy and Open Government (CeDEM)* (pp. 113-124). IEEE.

Nogueras-Iso, J., Lacasta, J., Ureña-Cámara, M. A., & Ariza-López, F. J. (2021). Quality of metadata in open data portals. *IEEE Access,* 9, 60364-60382. https://doi.org/10.1109/ACCESS.2021.3073455

Organisation for Economic Co-operation and Development (OECD). (2019). *Government at a glance 2019.* OECD.

Petrušić, D., Segedinac, M., & Konjović, Z. (2016). [Semantic modelling and ontology integration of the open government systems]. *Technical Gazette,* 23(6), 1631-1641. Croatian. https://doi.org/10.17559/TV-20150514115428

Raamkumar, A. S., Thangavelu, M. K., Kaleeswaran, S., &

Khoo, C. S. G. (2015). Designing a linked data migrational framework for singapore government datasets. *arXiv.* https://doi.org/10.48550/arXiv.1504.01987

Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., & Schraefel, M. C. (2012). Linked open government data: lessons from Data.gov.uk. *IEEE Intelligent Systems,* 27(3), 16-24. https://doi.org/10.1109/MIS.2012.23

Song, C., & Kim, H. (2022). Considerations in releasing public data: The case of local governments in Korea. *Journal of Information Science*. https://doi.org/10.1177/01655515221106636

Trypuz, R., Kuziński, D., & Sopek, M. (2016, July 6-9). General legal entity identifier ontology. In O. Kutz & S. de Cesare (Eds.), *Proceedings of the Joint Ontology Workshops 2016 Episode 2: The French Summer of Ontology co-located with the 9th International Conference on Formal Ontology in Information Systems (FOIS 2016)*. International Association for Ontology and its Applications.

Ubaldii, B. (2013). *Open government data: Towards empirical analysis of open government data initiatives.* Organisation for Economic Co-operation and Development.

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to open government data. *Government Information Quarterly,* 33(2), 325-337. https://doi.org/10.1016/j.giq.2016.02.001

Wang, H. J., & Lo, J. (2016). Adoption of open government data among government agencies. *Government Information Quarterly,* 33(1), 80-88. https://doi.org/10.1016/j.giq.2015.11.004

Wang, V., & Shepherd, D. (2020). Exploring the extent of openness of open government data – A critique of open government datasets in the UK. *Government Information Quarterly,* 37(1), 101405. https://doi.org/10.1016/j.giq.2019.101405

World Wide Web Foundation. (2017). *OpenData barometer: Global report.* (4th ed.). World Wide Web Foundation.

Yang, T. M., Lo, J., & Shiang, J. (2015). To open or not to open? Determinants of open government data. *Journal of Information Science,* 41(5), 596-612. https://doi.org/10.1177/0165551515586715

Zeleti, F. A., Ojo, A., & Curry, E. (2016). Exploring the economic value of open government data. *Government Information Quarterly,* 33(3), 535-551. https://doi.org/10.1016/j.giq.2016.01.008