Linked Method of Open Government Data by Datasets Oriented

Yun-Young Hwang[1], Jin-Hee Yuk[2], Sumi Shin[*3]

[1] Senior Researcher, Dept. of Intelligent Data Research, Korea Institute of Science and Technology Information, 41 Centum-dong-ro, Haeundae-gu, Busan, 48059, Republic of Korea
[2] Principal Researcher, Dept. of Intelligent Data Research, Korea Institute of Science and Technology Information, 41 Centum-dong-ro, Haeundae-gu, Busan, 48059, Republic of Korea
[*3] Principal Researcher, Dept. of Intelligent Data Research, Korea Institute of Science and Technology Information, 41 Centum-dong-ro, Haeundae-gu, Busan, 48059, Republic of Korea
yyhwang@kisti.re.kr[1], jhyuk@kisti.re.kr[2], sumi@kisti.re.kr[*3]
Corresponding author[*]: mobile Phone: +82-010-6269-8984

Article History:Received:11 november 2020; Accepted: 27 December 2020; Published online: 05 April 2021

Abstract: In order to make public data more useful, it is necessary to provide relevant data sets that meet the needs of users. We introduce the method of linkage between datasets. We provide a method for deriving linkages between fields of structured datasets provided by public data portals. We defined a dataset and connectivity between datasets. The connectivity between them is based on the metadata of the dataset and the linkage between the actual data field names and values. We constructed the standard field names. Based on this standard, we established the relationship between the datasets. This paper covers 31,692 structured datasets (as of May 31, 2020) among the public data portal datasets. We extracted 1,185,846 field names from over 30,000 datasets. We extracted 1,185,846 field names from over 30,000 datasets. As a result of analyzing the field names, the field names related to spatial information were the most common at 35%. This paper verified the method of deriving the relation between data sets, focusing on the field names classified as spatial information. For this reason, we have defined spatial standard field names. To derive similar field names, we extracted related field names into spaces such as locations, coordinates, addresses, and zip codes used in public datasets. The standard field name of spatial information was designed and derived 43% cooperation rate of 31,692 datasets. In the future, we plan to apply similar field names additionally to improve the data set cooperation rate of the spatial information standard.

Keywords: Open Government Data, Datasets, Linked Datasets, Public Data, Datasets Connection, Structured Datasets

1. Introduction

Korea was number one after receiving the public data openness index of 0.93 points from the OECD in 2019. The OECD public data index consists of three areas: data availability, data access, and government support for data utilization, while South Korea remains at the top of both areas[1]. The types of public datasets opened by public data portals are file datasets, Open API datasets, and standard datasets. The public data portal [2] provides a total of 40,305 public data sets (as of May 31, 2020). There are 34,464 file datasets, 5,669 Open API datasets, and 120 standard datasets. However, depending on the purpose of production, public datasets are produced by different production institutions and have various production and distribution forms and methods. As a result, individual researchers and institutions need a lot of time and effort to collect the required datasets and sort out the necessary datasets from the collected datasets.

The datasets registered in the public data portal have fifteen extensions types such as csv, docx, hwp, jpg and json. The datasets that can be converted into databases correspond to csv, docx, json, shp, xls, xlsx, and xml. The structured datasets that can be made into databases are 31,692 out of the datasets of public data portals, which is 78.63% of the total datasets.

The Table 1 shows the number of public datasets registered by category. It can be confirmed that the total number of public datasets registered in the table below is the same as the number of datasets classified by category. That is, a single public dataset may be classified into only one category, and it is difficult to judge the connection of public datasets through the application of multiple classification systems.

Table 1. Number of Datasets by Classification

Classification	File Data	Open API	Standard Data	Total
Administration	5,138	553	9	**5,700**
Science and Technology	1,435	411	1	**1,647**
Education	1,715	297	9	**2,021**
Transportation Logistics	2,636	665	30	**3,331**
Land Management	1,828	446	10	**2,284**
Concentrated Fisheries	1,770	499	4	**2,273**
Cultural Tourism	6,746	829	20	**7,595**
Law	128	11	-	**139**
Medical	1,729	262	4	**1,995**

[*]**Corresponding author:** Sumi Shin
Principal Researcher, Dept. of Intelligent Data Research, Korea Institute of Science and Technology Information, 41 Centum-dong-ro, Haeundae-gu, Busan, 48059, Republic of Korea . sumi@kisti.re.kr

Social Welfare	2,289	393	8	**2,690**
Industrial Employment	3,461	456	1	**3,918**
Food Health	860	222	1	**1,083**
Disaster Safety	1,590	170	16	**1,776**
Finance	850	186	-	**1,036**
Diplomatic Security	356	120	-	**476**
Environmental Weather	1,882	329	7	**2,218**
Total	**34,464**	**5,669**	**120**	**40,305**

In addition, public datasets are created for each institution, and it is difficult to confirm connectivity between field names or data values that have the same meaning and have different open formats. For example, in the case of public data related to "flooding", the information of the observatory is composed of fields such as coordinates, road name address, street address, and point number of the observatory. It acts as an obstacle to the data connection.

In order to solve these problems, we defined a list of data open to public data as a dataset, and based on the metadata of the dataset and the linkage between the actual data field names and values, connectivity between datasets was defined. For this purpose, the public data portal dataset is collected, and the metadata and actual data field names and values are extracted. Derive common metadata values and field names for datasets and standardize them to determine connectivity between datasets.

2. Materials and Methods

We defined the public dataset [3] for the public data portal as shown in Figure 1. The public dataset consists of metadata and raw data. The metadata describes raw data. The raw data is composed of fields and real values according to the fields. The main fields of metadata are classification, production organization, registration date, usage range, update frequency, and explanation information. Many conventional methods have proposed a method for linking datasets based on metadata, but this method has a limit in finding a meaningful link.

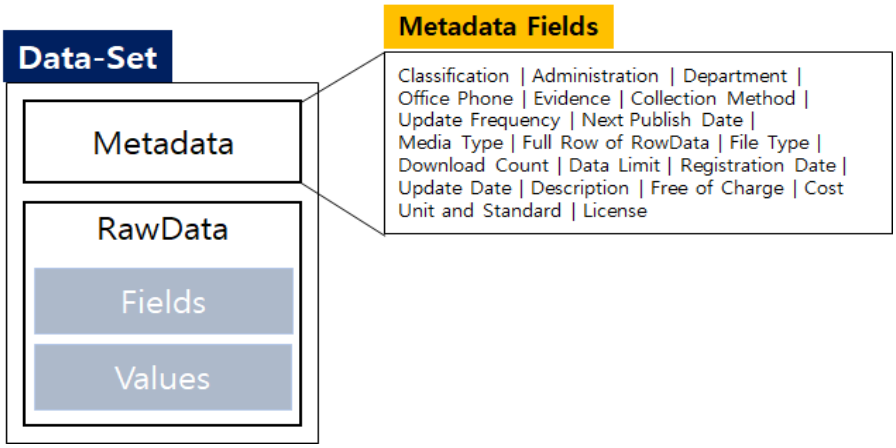

Figure 1. Definition Datasets

In order to increase the connection between datasets with such a dataset structure, our team analyzed the values of raw data fields and extracted common fields name as shown in Figure 2. This figure shows the extraction process of linked datasets. We have collected all datasets of public data portal. After then we analyzed the collected datasets and raw data of them. It is defined the standard field name by analyzing raw data field name. And then, the linkage between the datasets is defined based on the defined standard field names. Finally, we verify linkage between datasets and derivate the linkage rate.

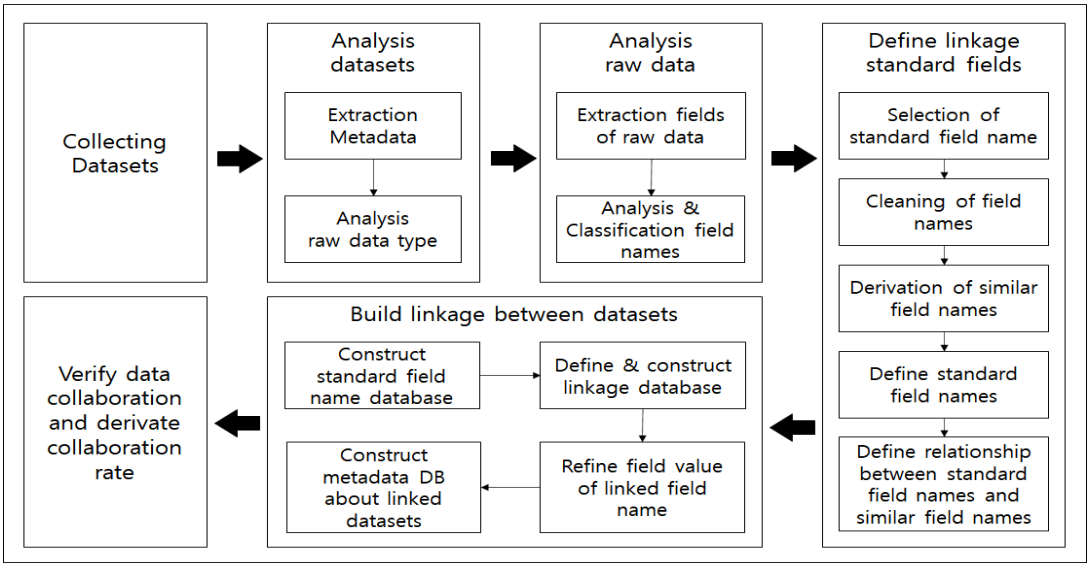

Figure 2. Extraction process of linked datasets

During the process of collecting the dataset, metadata is extracted and raw data types are analyzed. We have found that there are fifteen different types of raw data, which is very diverse. Among these types, we selected seven types of structured raw data that can be converted into database. The number is 31,692 datasets composed structured raw data. These datasets are 78.63% of the total public datasets. We build linkages between datasets, targeting only structured datasets.

In the third step, we extracted the field names in raw data and derived similar concepts of them. As a result of analyzing the field names, we found that field names are classified into space-related information, time-related information, measurement information, code information, telephone numbers, management numbers, and etc. information. The field names related to spatial information were the most common at 35%. This paper verified the method of deriving the relation between datasets, focusing on the field names classified as spatial information.

The defining linkage standard fields consists of file detailed courses such as selection of standard field name, cleaning of filed names, derivation of similar field names, define standard field names, and define relationship between standard field name and similar field names.

In the process of defining linkage standard fields, we analyzed the values of common fields, defined fields with the same meaning, and standardized the fields. The standardization of common fields is a plan to analyze and expand the field names of the public data portal based on the common terms of the Government Master Data (GMD)[2] and public datasets of the Ministry of Science, Technology and Information and Communication. Figure 3 shows an example of proceeding with the cleaning work for the field names to be standardized for the field names collected for the first time. The cleaning process is that removes double quotes, spaces, single quotes, and parentheses from the primary collected field names. In addition, all of them are converted to lowercase.

COLLECT_HEADER

DATASET_SEQ	META_SEQ	META_SUB_SEQ	DETAIL_URL_PK	HEADER_NAME	HEADER_NAME_SORT	HEADER_IDX	SYN_SEQ
DATASET_0000002	META_00000123	SUB_000000003	15010506	"Park Name"	Park Name	1	
DATASET_0000002	META_00000123	SUB_000000003	15010506	Location Road name Address	Location Road name Address	2	SYN_00002
DATASET_0000002	META_00000123	SUB_000000003	15010506	Location Lot Number	Location Lot Number	3	SYN_XXXX
DATASET_0000009	META_00000127	SUB_000000010	15010254	Road Address	Road Address	1	SYN_00002
DATASET_0000009	META_00000127	SUB_000000010	15010254	(Lot Number)	Lot Number	2	SYN_XXXX

Cleaning

Standardization

STD_HEADER

DATASET_SEQ	META_SEQ	META_SUB_SEQ	DETAIL_URL_PK	HEADER_NAME	Standard Name	HEADER_IDX	STD_SEQ	SYN_SEQ
DATASET_0000002	META_00000123	SUB_000000003	15010506	Park Name	Park Name	1		
DATASET_0000002	META_00000123	SUB_000000003	15010506	Location Road name Address	Road Address	2	STD_00001	SYN_00002
DATASET_0000002	META_00000123	SUB_000000003	15010506	Location Lot Number	Lot Number	3	STD_00007	SYN_XXXX
DATASET_0000009	META_00000127	SUB_000000010	15010254	Road Address	Road Address	1	STD_00001	SYN_00002
DATASET_0000009	META_00000127	SUB_000000010	15010254	Lot Number	Lot Number	2	STD_00007	SYN_XXXX

Figure 3. Example of standardization and cleaning field name

Figure 4 shows how to preserve the relationship between standard and similar field names. The standard field name table has standard field names, standard units, and classification values. The similar field name table includes a similar field name, sequence number of standard field names in standard field name table, a standard unit, and a conversion formula into standard units.

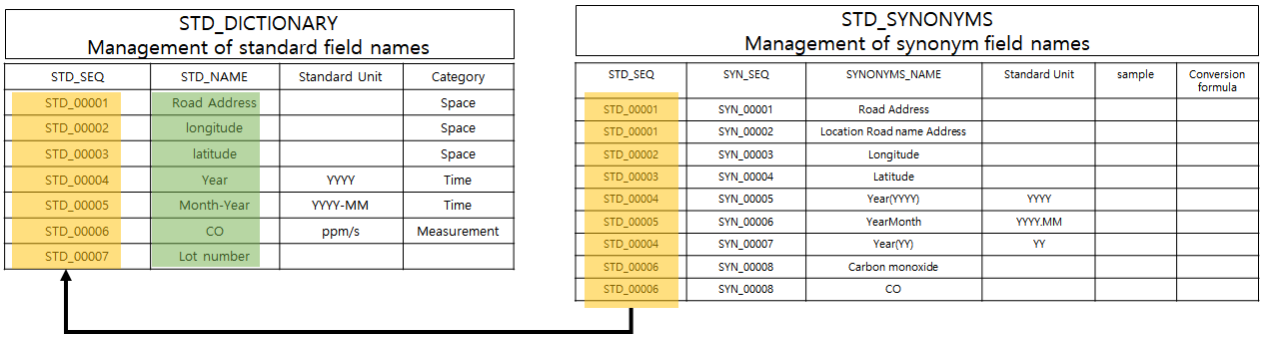

Figure 4. Relationship between standard field names and synonym field names

The next is the dataset linkage construction stage. We construct linkage metadata for each dataset, and store and manage the information of the linked fields in a database.

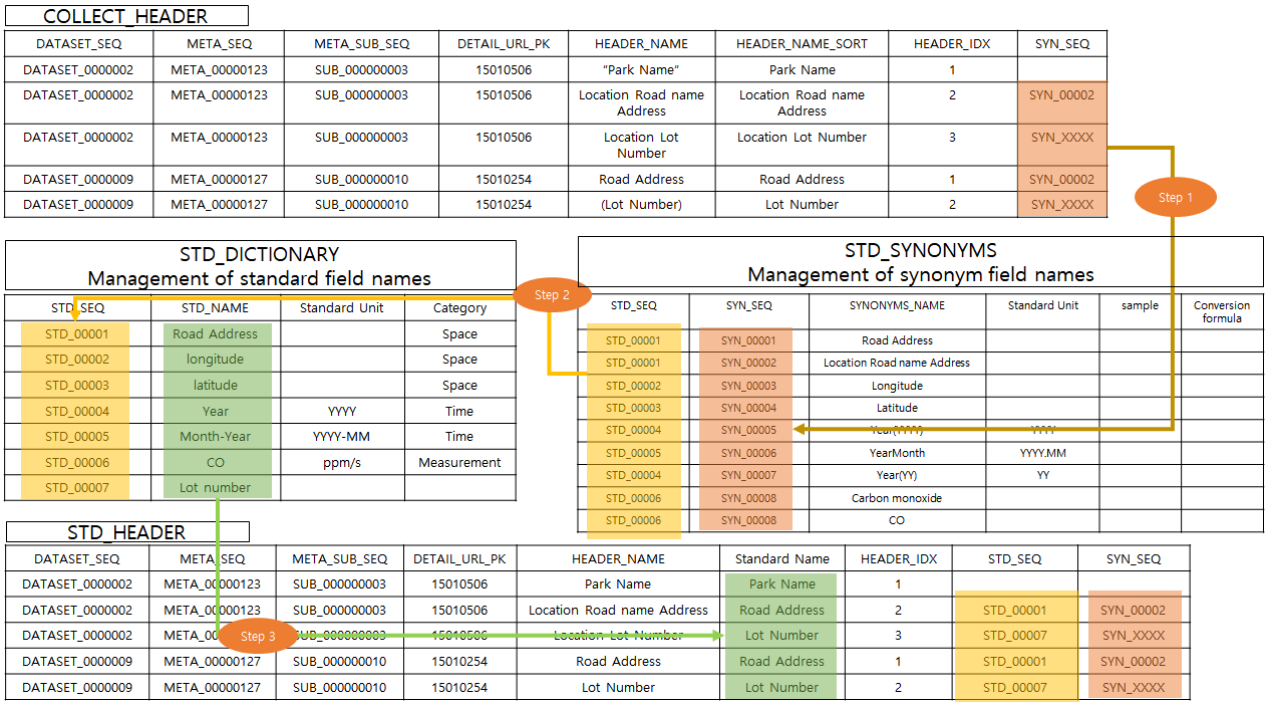

Figure 5. Linked datasets

Figure 5 shows a real-world example where the connection to the dataset has been extended through the standardization of raw data field names. The process of building a collaboration goes through three steps: step 1 is that give a similar word sequence for each extracted field names, step 2 is map a sequence of standard representative words using the given synonym sequence, and step 3 is change similar words in source data to standard representative words and save when creating standard entry table. Secondly, we are standardizing the raw data values of fields selected as common fields. Standardization of raw data values is based on units, representation formats, etc. After standardization is completed, an additional connection between datasets will be secured through a process of complementing null and error values.

3. Result and discussions

Our team constructed standard data of space for verifying the method of deriving the relationship between datasets, and constructed dataset linkage based on the process described above [5].

To construct the spatial standard data, we extracted all the field names that make up the regional pending data set, and extracted the field names that match the terms of the pre-defined spatial concept. The term of the main defined spatial concept is based on the coordinates defined in the National Spatial Information Portal [6] and the above-mentioned longitude information and the term used in the Road Name Address Open API [7] of the Ministry of Public Safety. It consisted of 38 words related to road name address, lot number address, administrative area, coordinates, upper/hardness, and mailbox address. The term of spatial concept is used for extracting field names related to the spatial concept, and is used for generating a candidate group that can establish a spatial related relation. The defined spatial concept terms can be extended based on field name similarity.

We constructed a standard data of space for verifying the relation between datasets and deriving the relation between datasets utilizing different spatial information. Spatial standard data was designed to have a mapping relationship between coordinates (UTM-K) based on road name address/hardness, mailbox address, province, lot number address.

In this paper, the 73 datasets provided by public data portals related to issues (fine dust, earthquake, inundation) in the Busan in South Korea, the 31 datasets provided by specialized institutions (Busan City, AirKorea[8], Meteorological Agency[9], and etc.). We measured the cooperation rate between datasets based on the spatial information.

Table 1. shows the number of items related to spatial field names for 104 types of datasets related to the Busan area. The total number of field names is 14,281, and there are 1,301 spatially related field names. The association data set candidates having items related to the spatial terms extracted from Table 1. were extracted using the called frequency tf (t, d)[10], and the association information was constructed based on the spatial standard data.

Table 2. Number of spatial term related fields names

No	Space term	Count of Fields
1	Latitude/longitude	186
2	Coordinates (UTM-K)	161
3	Address	530
4	Road name/lot number	284
5	City Name/State, City Name	119
6	Bopjeongdong	21
Total Count of Fields		1,301

There are 34 types of datasets related to pending issues in the Busan area, where association information was constructed based on spatial data, with a cooperation rate of 32.69%. This result is expected to lead to a high cooperation rate when similar terms related to the spatial term are expanded by extracting the item names that exactly match the spatial term and constructing the cooperation information. It Also, although the item name matches the space term, the actual value is a code managed by the provision destination, and related information could not be constructed. In the future, we plan to obtain the management code for the recipients and expand the standard data of the space to further build the association information.

4. Conclusions

In this paper, in order to enhance the connection with the data of the public data portal, we define the public dataset, propose the linkage between the raw data fields of the public dataset, and the linkage method with the raw data value. In addition, we proposed a method for deriving relations between datasets based on spatial data, and described the results of verification by applying it to actual datasets. Spatial data is used as an important base data for solving regional concerns, but it uses various terms and data values to help users derive associations with their datasets. It requires a lot of effort. This research team plans to expand the related information by constructing a space related item by using open API argument information. In the future, we plan to further verify the connection of public datasets based on social issues such as inundation, earthquakes, and fine dust.

Acknowledgements

This research was conducted with the support of Open Data Solutions (DDS) Convergence Research Program funded by the National Research Council of Science and Technology "Development of solutions for region issues based on public data using AI technology -Focused on the actual proof research for realizing safe and reliable society-".

References

1. Jacob, A.R.P., Cecilia, E. Barbara, U. (2020). OECD Open, Useful and Re-usable data (OURdata) Index: 2019. OECD Policy Papers on Public Governance, 1
2. Ministry of the Interior and Safety. (2020). Public Data Portals. https://www.data.go.kr/
3. Haklae, K. (2018). Metadata Analysis of Open Government Data by Formal Concept Analysis, Journal of Korea Contents Association. 18(1). 305-313. DOI : 10.5392/JKCA.2018.18.01.305
4. Administrative Safety Department. (2020). Government Master Data. Name of Site or Board. https://www.gmd.go.kr/
5. J. Kopke & J.Su (2016). Towards Ontology Guided Translation of Activity-Centric Processes to GSM, Lecture Notes in Business Information Processingbook series (LNBIP), 256. DOI : 10.1007/978-3-319-42887-1_30
6. Ministry of Land, Infrastructure and Transport. (2020). National Spatial Data Infrastructure Portal. http://www.nsdi.go.kr/
7. Ministry of the Interior and Safety. (2020). Road Name Address Portal. https://www.juso.go.kr/
8. Korea Environment Corporation. (2020). AirKorea Portal. https://www.airkorea.or.kr/
9. Meteorological Agency. (2020). Meteorological Portal. https://www.weather.go.kr/
10. Anne, D. R., Avik, S., Paul G. (2004). Frequent Term Distribution Measures for Dataset Profiling. Proceedings of the 4th International Conference on Language Resources and Evaluation.