International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 02 | Feb 2022 www.irjet.net p-ISSN: 2395-0072

Public Data Analysis and Utilization

Duckki Lee

Assistant Professor, Department of Smart Software, Yonam Institute of Technology, Jinju, South Korea

---

Abstract - *The Fourth Industrial Revolution heralds the advent of the data era. Companies that retain and use enormous amounts of data are at the forefront of market innovation, and artificial intelligence and robotics, which are fast-growing in all aspects of the nation and society, are also data-driven. Following this trend, advanced nations, particularly the United States, realize the critical role of data in determining future competitiveness and are revitalizing the data sector and making public data more accessible. This paper analyses domestic and international trends in public data and discusses the openness and use of domestic public data. Following that, while developing commercial services that make use of public data, the issues and concerns associated with public data are highlighted.*

Key Words: Public Data, Open Government Data, Public Data Analysis, Public Data Utilization, Commercial Services using Public Data

1.INTRODUCTION

The Fourth Industrial Revolution heralds the advent of the data era. The age of oil and coal ushered in the first industrial revolution, which was followed by the age of electricity and communication, and then the age of information technology, which ushers in the age of data. The 4th Industrial Revolution is driven by intelligent information technology, which is at the heart of data as a fundamental component for intelligence, automation, and autonomy.

The era we are entering is one of a data economy [1], in which data is a critical resource in addition to land, labor, and capital, and a data society [2, 3], in which all aspects of everyday life are data-driven. Companies that retain and use enormous amounts of data are at the forefront of market innovation, and artificial intelligence and robotics, which are fast-growing in all aspects of the nation and society, are also data-driven. Following this trend, advanced Western nations, notably the US, acknowledge the critical role of data in determining future competitiveness and are actively pursuing data hegemony via strategies and increased investment in the data industry [4-7]. Additionally, these nations are striving not only to enhance economic value through the opening of public data, but also to create social value through the active use of data to address pressing issues facing the country and society, such as transportation, the environment, health, hygiene, disaster preparedness, and safety. In keeping with this trend, the Korean government likewise pursues a state-of-the-art intelligent government for the twenty-first century and promotes the data economy. Additionally, to encourage the realization of social values through the use of data, legislation, system development, and portal site establishment are being vigorously supported. The accessibility of public data is required to expand people's access to it and to promote value creation via data use in the data economy and social activities. To ensure the success of the public data opening policy, the openness of public data is insufficient; instead, a data ecosystem in which it can be disseminated, exploited, and circulated must be established [8]. South Korea has been operating a public data portal[24] since 2013 to accomplish this. By making data uploaded by all central governments and local governments freely available to all customers, including corporations and ordinary residents, the public data portal plays a critical role in the distribution and usage of data. Public data portals have been shown to have substantial outcomes in evaluations. Public data use services continue to proliferate, and in the OECD's OUR Data Index, a national assessment of public data openness, Korea topped the list three consecutive years in 2015, 2017, and 2019[9, 22]. However, there is a variety of criticism on public data portals. The quantity of data is vast, yet the essential data cannot be found [10], the format of the data is very inconvenient to use, or the data is extremely difficult to integrate owing to the different input or file formats between the data [11].

This paper analyses domestic and international trends in public data and discusses the openness and use of domestic public data. Following that, while developing commercial services that make use of public data, the issues and concerns associated with public data are highlighted.

2. Trends in Domestic and International Public Data

The term "public data" refers to information and data generated or authorized by the government. It is data that is freely supplied, reused, and distributed to anyone, and that users can use to produce their creations [12]. The World Wide Web Foundation defines public data as data that is publicly accessible online, reusable, and machine-readable, and that enables huge volumes of data to be downloaded and used as a single dataset for free [13]. According to Article 2 of Korea's Public Data Act[14], public data refers to data or information processed optically or electronically by public institutions for the

objectives specified by applicable laws and regulations. Public data is the data generated and maintained by governments and public institutions to achieve public objectives, such as conducting business and providing public services and is a critical resource with enormous potential value. Public data encompasses all information on all residents, including their resident registration, income, property, medical treatment, tax payment, and real estate, as well as information about the weather, transportation, logistics, energy, and water and sewage systems that affect their everyday life.

US

The United States established data.gov[15], a unified public data opening portal, in 2009, as part of an open government initiative that incorporated the principle of opening government-held public data. As of February 2022, data.gov is exposing data from 48 state governments, 48 cities, and 152 government-affiliated institutions, starting with the release of 76 data sets held by 11 government agencies[15]. Additionally, until recently, active data disclosure was ongoing, with 342,000 data sets in a variety of domains, including health, labor, education, transportation, and crime. The data is primarily available in the form of raw data sets, geodata sets, interactive data, source code, APIs and programs, and applications, and is updated monthly. Various data formats, such as xls, cvs, and txt forms, are provided to allow for the reuse of information, and RDF-type conversion is straightforward. In addition, various mashups may be reclassified, adjusted, and integrated with other datasets within the system, making the system a powerful tool.

The basis of data.gov is the 'Open Government Platform' (OGPL). In 2013, Data.gov 2.0 was introduced as an open-source data platform called CKAN[16]. As a consequence, data catalogue services have been established to connect open government data sites throughout the United States with those of other nations, states, and cities. The public data portal (data.gov) focuses on the integration and management of metadata through the utilization of CKAN. Additionally, public data is visualized using maps, and DATA USA[17] is available separately for comparing and analyzing major US cities.

Fig -1: US DATA.GOV and DATAUSA

UK

The United Kingdom has opened 51,957 datasets across 14 sectors, including economy, education, environment, defense, health, and transportation, starting with data.gov.uk[18], a public data portal established with the participation of Tim Berners-Lee, the inventor of the web and linked data. Since 2010, the UK has actively promoted an open data policy and operated an open data portal, data.gov.uk, which enables search and access to practically all public sector data. The UK government's open data policy intends to make information more accessible to the public and advance the public interest through the use of open data for policy research[19].

The British government established the Open Data Institute (ODI)[20] in November 2012 as a non-profit organization dedicated to using public data to identify new enterprises and startups. The British government is accomplishing this by providing support and developing talent for venture firms that are attempting to build new businesses via the development of technology and services connected to public data. Approximately 30 startups have been formed as a result of ODI's startup support programs. Opencorporates[21], a market leader, is the owner of the world's biggest corporate disclosure database, which has 190 million corporate records. The firm has received positive feedback from civic organizations and investors seeking corporate monitoring through the provision of data through a search portal.

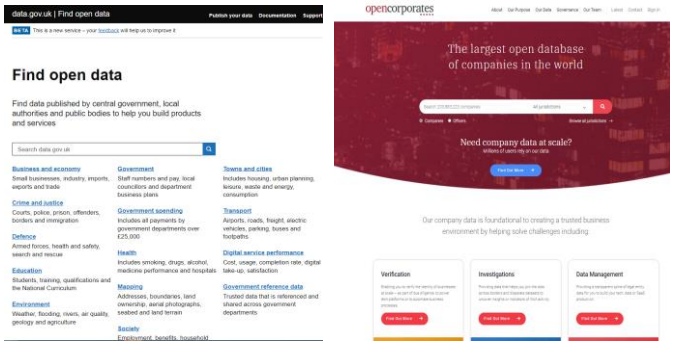

Fig -2: UK DATA.GOV.UK and opencoporates

South Korea

With the enactment of public data laws in 2013[14], South Korea has been actively promoting an open data policy. To promote the openness and use of public data in South Korea, data.go.kr[24] was founded and is being managed as a public data portal that delivers integrated information. The public data portal makes public data available in a variety of formats, including file data, open APIs, and visualization, to enable anyone to simply and comfortably utilize it, and to enable anybody to quickly and correctly find desired public data through an easy and convenient search. A data industry revival strategy headed

by data-related governmental agencies has been established and is being promoted in South Korea. As a consequence, South Korea topped the OUR Data Index for three consecutive years in 2015, 2017 and 2019, an OECD-mandated assessment of the amount of openness of public data [9, 22]. Additionally, Korea was ranked 17th in 2014, 8th in 2015, and 5th in 2016 in the WWW Foundation's Open Data Barometer (ODB)[13].

Currently, 68,865 public data sets are accessible through public data portals. These data include 51,110 file-based data, 9,020 open APIs, and 8,735 standard data; the amount of data for each field is shown in Table-1, and the national data map illustrating the proportion of data for each field is presented in Figure-1.

Table -1: The number of Datasets by Topics

Topics	Data Sets	Topics	Data Sets
Education	3925	Health Care	3719
Land	4270	Disaster Recovery	3512
Administration	9090	Transportation	5647
Finance	5027	Weather	5096
Industry	6473	Technology	2066
Social Services	4132	Agriculture	4201
Food	1997	Unification	964
Culture	8333	Law	413

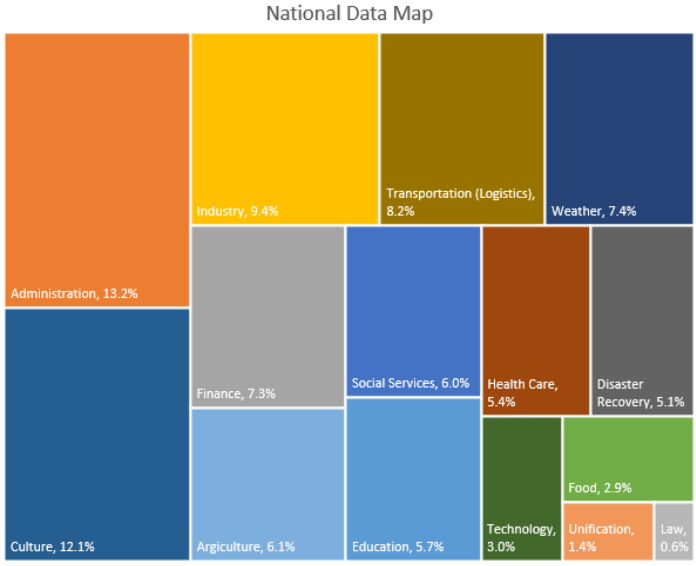

Fig -3: Korea National Data Map

3. Public Data Provision and Utilization Analysis in South Korea

In South Korea, the overall number of open data cases climbed by 12.8 times, from 5,272 in 2013 to 24,588 in 2017, 28,400 in 2018, 33,600 in 2019, 55,139 in 2020, and 67,441 in 2021.

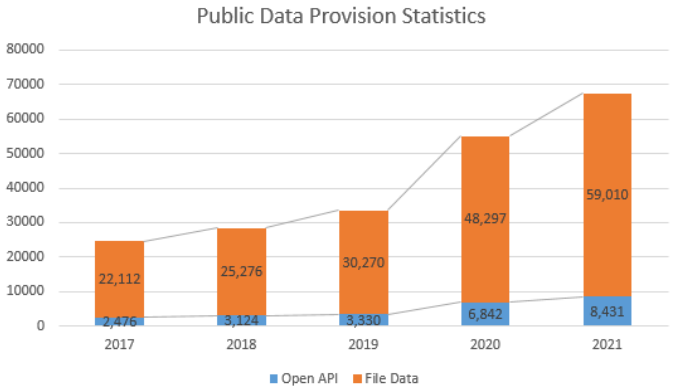

Chart -1: Public Data Provision Statistics

The number of public data openings is fast expanding as a result of the Korean government's efforts to open public data, and the number of public data uses is also exponentially increasing as more businesses utilize data. The number of public data users climbed from 13,923 in 2013 to 3,871,984 in 2017, 7,549,179 in 2018, 13,141,413 in 2019, 20,848,555 in 2020, and 33,340,436 in 2021. This exponential growth represents a 2,394-fold increase in comparison to the initial figure.

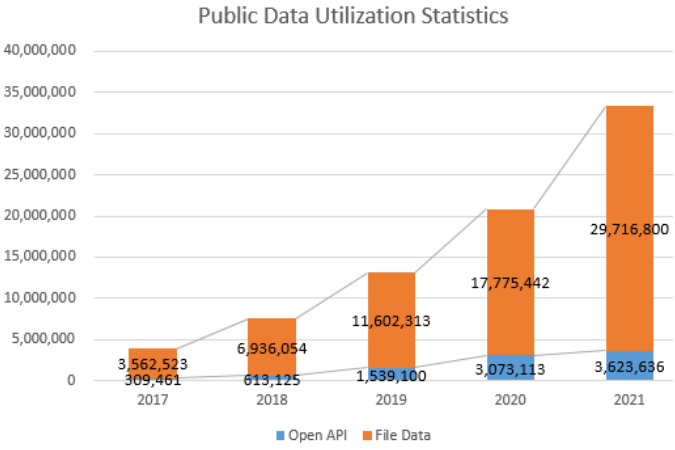

Chart -2: Public Data Utilization Statistics

Despite these efforts and accomplishments, the public data portal has come under criticism for a variety of reasons. The quantity of data is vast, yet the essential data cannot be found [10], the format of the data is very inconvenient to use, or the data is extremely difficult to integrate owing to the different input or file formats between the data [11].

4. Considerations for Designing Public Data-based Commercial Services

This chapter covers the considerations that must be made when developing commercial services that make use of public data included on public data portals.

4.1. How to Access Public Data

Currently, there are two general methods for using public data.

The method using the Open API

It is a sort of service mashup that has been in the limelight since the Web 2.0 era and has the advantage of not requiring an individual or corporation to make a separate database to build a service based on the given data. Currently, a variety of public data sources are available in XML and JSON formats, and the majority of service APIs released in the last few years are supported in JSON format for increased transmission speed and processing efficiency.

The following issues arise while developing commercial services utilizing the open API.

- Even if the cost per call is free, individual approval for each API must be acquired.

- If it is necessary to freeze data at a certain moment in time, unintended new data may be utilized or old and new data may be combined if version control of the open API is not correctly implemented.

Since the service to be provided is now reliant on the service quality of the open API, it becomes more difficult to regulate service quality. In comparison to other drawbacks, one of the greatest impediments to creating commercial services is the difficulty in managing service quality owing to failure. One frequently used method for resolving this issue is to cache the open API call results made by the service in a memory database, etc. In this case, performance and stability may be assured like that of constructing its database for regularly occurring API call results.

In general, if cost reduction is not the main objective while establishing commercial services, it is preferable to construct its database based on the data rather than relying on public data APIs.

The method using file data

If the quantity of data is sufficient to distribute as a file, then each piece of data may be distributed as a file. In this case, the advantages of the API-based method are applied as disadvantages, while the shortcomings of the API-based method are implemented as advantages.

One of the main advantages is that it is possible to create a database with the same content as the data available through the open API. This enables simple management of service stability since it is not dependent on the stability of external services.

When data is distributed in the form of a file, it is often in the form of a compressed MS Office Excel or CSV file. CSV files are often utilized when dealing with data in the context of creating commercial services since the data is processed and turned into a database by machines rather than humans.

4.2. Considerations for Designing Public Data-based Commerical Services

In this section, we will discuss the factors to consider while building a service using the CSV-type file data that was selected for commercial service development owing to its simplicity of use.

The following issues are significant since many public data have trouble inputting or refining data with manpower. Because errors are more likely to be discovered during the service stage of unstructured data processing than during the processing stage, it is frequently required to consider them during the data processing stage.

Error in the CSV file itself

While the CSV file's properties make it simple to parse, errors in the information structure are common because data fields are divided by a separator such as a comma. For example:

- If the data included in a field contains a separator as content, even if it is distinguishable by a human cognitive ability, a mechanical parser will have difficulty distinguishing it, resulting in improper data processing or errors during processing.

- Because field division is entirely dependent on the existence or absence of a separator, a separator may be accidentally added or omitted during the creation or processing of a CSV file.

Data Encoding Issues

In contrast to other file formats, CSV files do not require a specific character encoding scheme. As a result, without additional information relating to the CSV file's metadata, work such as inferring the CSV file's encoding information is performed. If the encoding is not consistent or is difficult to manage in a particular development environment, pre-encoding conversion work should be undertaken.

Classification of Omissionable Information

Not all elements of public data have the same data schema. Both the MS Office Excel file and the CSV file are comprised of a sum of the data's schemas in this case. As a result, some fields are left blank for each data element, while others are filled with an empty value (i.e., represented by a continuous separator).

In the case of a NoSQL database that is well-suited for unstructured data response, there is no significant issue with storing data elements that lack certain fields. However, in the case of SQL databases, it is required to determine which fields are allowed to be null-set. Even when a NoSQL database is used, the client must be aware of the nullable field during the result parsing process.

However, because such information is excluded in the CSV file, it is unavoidable to collect data by inspecting fields that can be null-set when the data is processed.

Absence of Type Information

Each data field has a specific type, but this is not specified explicitly. Even if the field is simply named "ID" or "code," if the field is a type of serial number (e.g., ID in the form of year + month + day + high order), it is possible to keep this type during processing for future database storage and query.

Because there is no type of information for individual fields at the moment, it is required to make assumptions during the data processing process and infer the type by examining if the assumption holds for all fields. If the type is decided in this manner, it is vital to account for the likelihood of problems in the future when data is updated and assumptions are not established.

4.3. Public Data Verification

Due to the characteristics of XML and JSON, verifying public data provided over an open API is simple. However, due to a variety of previously identified issues, data rticonstruction or batch processing is challenging.

Specifically, a malfunction is detected during the service process, rather than during the batch processing process, where a data error is found. Additionally, such an error is extremely likely to be discovered only after the user reports it. As a result, it is critical to detect errors in data before the database's input or processing process.

5. CONCLUSIONS

Data is becoming an essential and critical factor of the Fourth Industrial Revolution. Following this trend, advanced nations such as the United States and the United Kingdom recognize the value of data, develop diverse strategies to rejuvenate the data industry, and repeatedly make efforts to open public data. The paper investigated trends in domestic and overseas public data, as well as the opening and utilization of domestic public data. Additionally, problems that arise while establishing commercial services were covered, as well as points to consider, using genuine public data. The consumers of public data range from ordinary citizens to professionals, and the purposes for which they are used vary significantly, ranging from simple information retrieval to commercial service development. As a result, it is difficult to deliver data in a specific format or with specific content for a single consumer. It is required to give a variety of data in a variety of formats. Additionally, to improve the use of public data in the development of actual commercial services, it will be necessary to open public data that reflects the issues and considerations presented in this paper.

REFERENCES

[1] D. Newman, “How to Plan, Participate and Prosper in the Data Economy”, Gartner, 2011, https://www.gartner.com/en/documents/1610514/how-to-plan-participate-and-prosper-in-the-data-economy

[2] D. Reinsel, J. Gantz, J. Rydning, “Data Age 2025: The Evolution of Data to Life-Critical. Don’t Focus on Big Data; Focus on the Data That’s Big, An IDC White Paper, Apr, 2017.

[3] D. Reinsel, J. Gantz, J. Rydning, “Data Age 2025: The Digitization of the World. From Edge to Core, An IDC White Paper, Nov, 2018.

[4] Joint Ministry of Relations, “2021 National Key Data Open Plan, Public Data Strategy Committee, April 2021

[5] Joint Ministry of Relations, “2021 Public Data Provision and Use Revitalization Implementation Plan (draft), Public Data Strategy Committee, April 2021

[6] Joint Ministry of Relations, “Data Industry Revitalization Strategy: I-KOREA 4.0 Data Sector Plan, I-DATA, 4th Industrial Revolution Committee, 2018

[7] Joint Ministry of Relations, “Measures to revitalize the data platform. - From system platform to user service platform –“, 4th Industrial Revolution Committee, June 2021

[8] R. Pollock, “Building the (Open) Data Ecosystem”, Open Knowledge International Blog, 2011

[9] OECD, “OECD Open, Useful and Re-usable data(OURData) Index: 2019, Mar, 2020

[10] NIA, "Research on Legislative Improvement for Public Data-Based Industrial Ecosystem Creation", Research Report of NIA, 2017

[11] Tae-Yeop Kim, "The Current State of Public Data Opening Policies and Future Tasks", Issues and Points, No. 1455, April 2018, National Assembly Legislative Research Office

[12] OECD, "Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact", OECD, 2018

[13] World Wide Web Foundation, https://webfoundation.org/

[14] ACT ON PROMOTION OF THE PROVISION AND USE OF PUBLIC DATA, https://elaw.klri.re.kr/eng_mobile/viewer.do?hseq=47133&type=part&key=4 https://www.data.gov/

[15] https://www.data.gov

[16] https://ckan.org/

[17] https://datausa.io/

[18] https://data.gov.uk/

[19] Great Britain, DBIS(Department for Business, Innovation and Skills), "Seizing the data opportunity: A strategy for UK data capability", 2013

[20] https://theodi.org

[21] https://opencorporates.com/

[22] Open Government Data – OECD, https://www.oecd.org/gov/digital-government/open-government-data.htm

[23] M. Young, the Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[24] https://www.data.go.kr/

BIOGRAPHIES

Duckki Lee is currently an Assistant Professor in the Department of Smart Software, Yonam Institute of Technology in South Korea. His research interests include big data system, public data analysis and utilization.