JOURNAL OF APPLIED INTELLIGENT SYSTEMS & INFORMATION SCIENCES
VOL. 2, ISSUE 1, PP. 49-59, JUNE 2021.
Available at: www.journal.research.fanap.com
DOI: https://doi.org/10.22034/jaisis.2021.299011.1032

THE ROLE OF DATA GOVERNANCE IN IMPROVING URBAN MANAGEMENT: CASE STUDY OF TEHRAN MUNICIPALITY

Sarah Bourbour Hosseinbeigi [1] , Shirin Hamed Ahangari [2, *]

[1] *Tehran Statistics and Urban Observation Center, Tehran, Iran*
[2] *ICT Organization, Tehran Municipality, Tehran, Iran*

ABSTRACT

Today's cities face various challenges, including economic growth, environmental sustainability, social resilience, air pollution, and traffic. Given these challenges, we need information for analysis and forecasting to manage crises in specific situations. Many cities are investing in information technology research and developing policies to improve citizens' quality of life. Given the ICT trend for sustainable smart cities, the future planning process of the city is critical. Therefore, timely and reliable primary input is critical for decision-making analysis in real-time and reliable statistics, which means increasing the quality and quantity of data in all aspects. In this paper, the experiences of Tehran urban statistics and observatory center in providing open data is presented with regard to the implementation of anonymization processes using data governance measures. We describe the anonymization of some real data from municipality, which has been done by PDPC and present a checklist for classifying all information. Finally, the design of metadata storage is clarified. Essentially, one of the successful experiences is the use of researchers' analysis by providing no sensitive data as open data. This experience is based on studying global examples with the framework of DMBOK and required privacy and confidentiality of information. We used data governance in the organization to emphasize on all aspects of data management, including data production and sharing, data quality assurance, data integration, and related topics.

KEYWORDS: Data Governance, open data, data anonymity, transparency.

1. INTRODUCTION

Demand for high-quality statistical analysis continues to rise. The primary input for analysis and decision-making is real-time and reliable data which means the need for increasing the quality and quantity of data in all aspects. One of the main challenges in this field, which is also a manifestation of civil rights, is protecting citizens' privacy and observance of security considerations in the production, sharing, and publication of data. Data governance guides all data management activities with the purpose of ensuring proper data management based on policies and procedures (Thompson et al., 2015).

Data classification is the process of organizing data into appropriate categories for the use and protection of information (Henderson & Earley, 2017). The data confidentiality classification consists of various steps, such as creating a metadata repository, checking the level of data quality, and creating data identifiers needed to make decisions based on the current situation. Organizations that collect, process and disseminate personal and confidential data are required to develop and implement policies that are in line with the Personal Data

* *Corresponding Author, Email: s.hcom9@gmail..com*

RECEIVED: 10 MAY, 2021; **ACCEPTED**: 14 JUNE, 2021; **PUBLISHED ONLINE**: 22 JUNE, 2021

Protection Act (PDPA) (Chik, 2013). Designing and Implementing a Data Protection Management Program (DPMP) helps organizations build a strong data protection infrastructure that includes managing policies and processes for personal and confidential data as well as roles and the responsibilities of individuals in the organization in relation to the protection of personal data.

To reduce the risk of disseminating sensitive information, anonymization techniques i.e., converting sensitive data to anonymous data can be effective. Typically, the process of data anonymization would be "irreversible" and the recipient of the anonymized dataset would not be able to recreate the original data (El Emam & Arbuckle, 2013).

The level of confidentiality and security rules and standards are mentioned in the metadata of each data. Metadata is generated and stored in a process. In order for metadata to be of good quality, it must be managed and related processes such as the data modeling process and software production must be properly implemented. In addition, metadata standards must be considered and observed. The metadata integration process collects it from different parts of the organization. To configure and update the metadata repository, standard interfaces must be defined between the various sections to transfer information to the metadata repository.

This article investigates the experiences of Tehran urban statistics and observatory center, a subset of the Tehran Municipality ICT Organization, regarding the implementation of anonymization processes with data governance according to some global examples. In order to increase transparency and easier access to data, using some anonymity techniques taken from PDPC Singapore, some main and sensitive urban data was published publicly in the open data site of Tehran Municipality while maintaining confidentiality. An open data website is a website with public access to which data is provided to users for free (Zuiderwijk & Janssen, 2014). To design this website, some global examples such as global databases, open data of New York and London have been studied, which will be explained later.

Identifying sensitive data requires classification, which is one of the main solutions for organizing and managing classified data, designing a metadata repository. Data Management Body of Knowledge (DAMA-DMBOK Guide) (Cupoli & Earley, 2014). using the most optimal and efficient methods, has been used as a reference for metadata repository implementation and data classification in this study.

To identify personal and sensitive data by determining the degree of importance of the data, using the checklist designed in the data governance document of Tehran Municipality, the data is scored and the level of risk is determined based on two criteria: 1) probability of occurrence and severity of consequences; 2) level of confidentiality. Each criterion has 4 levels: red, orange, yellow and green (open data).

2. DATA GOVERNANCE

Data governance is the process of determining decision-making rights and the accountability framework for creating, storing, evaluating, applying, and archiving data (Abraham & Schneider, 2019). With the increasing development of information technology, the idea of a smart city has been proposed and has become one of the strategic goals of Tehran Municipality. A lot of information in the smart city is collected and analyzed from various sources. This information, which is often collected through sensors or IoT technology, covers all urban areas such as transportation, municipal services, etc. Many organizations today need to move to data- driven functionality. Nowadays many organizations need to move forward to data governance as the purpose of data governance is to make decisions, and choosing strategies based on data (Dal Maso, 2019).

This change of culture can be done by communicating an upstream and binding document for all organizational units. One of the most authoritative global documents in this field is the DMBOK knowledge Management document (DeStefano, 2016). In this document, DAMA-DMBOK framework model, is introduced with the name of the Dama wheel as can be seen in Fig. 1. It is important to note that data governance is at the center of this wheel, and data management activities maintain consistency, balance, and coordination between other sectors. In the following sections of the article, we will describe DMBOK and how we implemented it in the Tehran urban statistics and observatory center.

Several processes of governance are performed together in big data analysis, each seeking to respond to a specific request as shown in Fig 2.

Fig. 1. DAMA-DMBOK Data Management Framework (Dach, 2016)

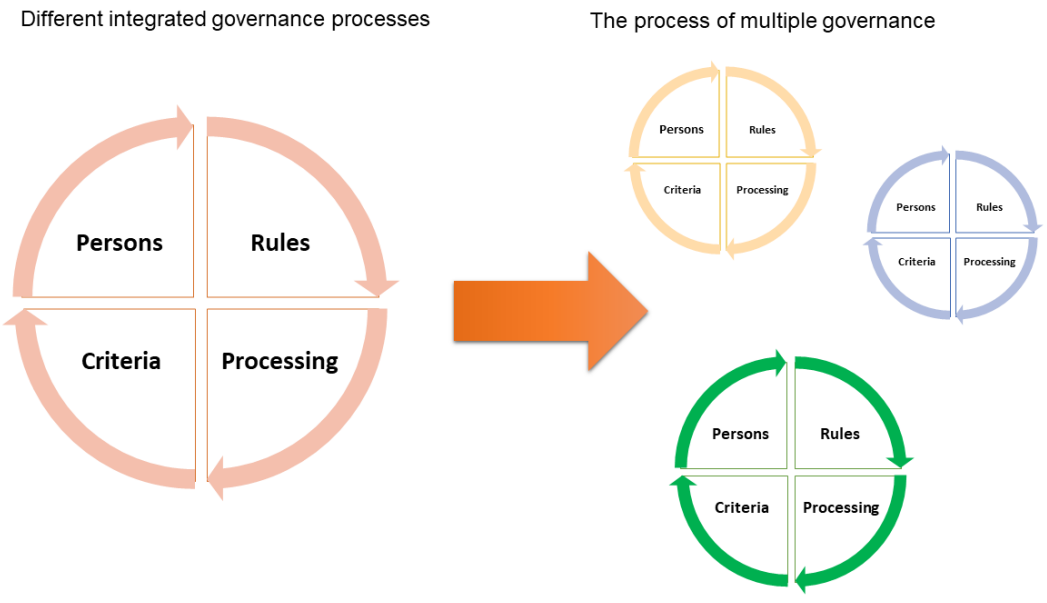

Fig. 2. Data governance change with big data analytics

2.1. Metadata management

The importance and sensitivity of data should be considered in the metadata analysis. Metadata is a description of the data consisting of the database name, data model, business processes as well as technical aspects, rules, and constraints associated with data and their physical and logical structure. With the help of metadata, the organization can understand its data, system, and workflow and use it optimally. Metadata helps to evaluate the quality of data (Kamandi., 2019). ISO / IEC 11179 standard also provides a framework for defining metadata components to facilitate data transfer. This standard is presented in 6 sections: a framework for creating and normalizing data elements, basic features, rules and guidelines for data definitions, naming and identifying data elements, identifying and registering data elements (Dandawate, 2013). A robust metadata management process has proven to be essential for successful information management.

The data characteristics which are provided by metadata are explained in Table 1.

Table 1. Some information about the organizational data which is provided by metadata

Metadata items	Explanation
Data name	Usually, for each business-level information item, several technical-level information tables are used. The data name is one of the items mentioned in this section. It is essential to the business level to use words meaningful to experts and clients in that specific field. For example, it accounts for renovation and urban development costs, air pollution indicators, density sales, etc.
Applications	What is the usage of this data?
General Organization	This information is more relevant to which department or organization.
Data owner	In this paper, we call the organization that gathers the data, data owner. However, the data owner is part of the formal organizational structure and cannot be a person.
Information producer	Which unit is responsible for creating information?
Quality level	The quality level of the data is measured based on the rules and indicators specified in the metadata. It needs to be planned so that the calculation of the quality level is done automatically.
Data Classification	According to the official confidentiality instructions of the relevant organization
Scheduled maintenance information.	How long should this information be kept? If there are special considerations regarding the deletion of information, it should be mentioned.

2.2. Data security

One of the most important considerations when publishing open information is data protection. The appearance of the big data age has provided significant opportunities for social advancement and posed many threats to information security for society making the protection of personal information privacy a great concern. Sufficient level of technology for personal information security is necessary to ensure security and privacy protection of big data. Still, we also need to strengthen citizens' privacy to implement information security and privacy. If big data is not adequately protected for the user, privacy and data security are directly threatened. Data protection can be divided into anonymous identifiers and privacy protection (Zhang, 2018). For those data related to the indicators that cannot be published publicly, we used PDPA rules to anonymize sensitive data.

For example, we anonymize some information in air pollution, metro, waste, fresh markets, and traffic cameras first and then published on data.tehran.ir website.

A flowchart of the data anonymization process can be presented as shown in Fig. 3 (Allen & Gledhill LLP., 2018).

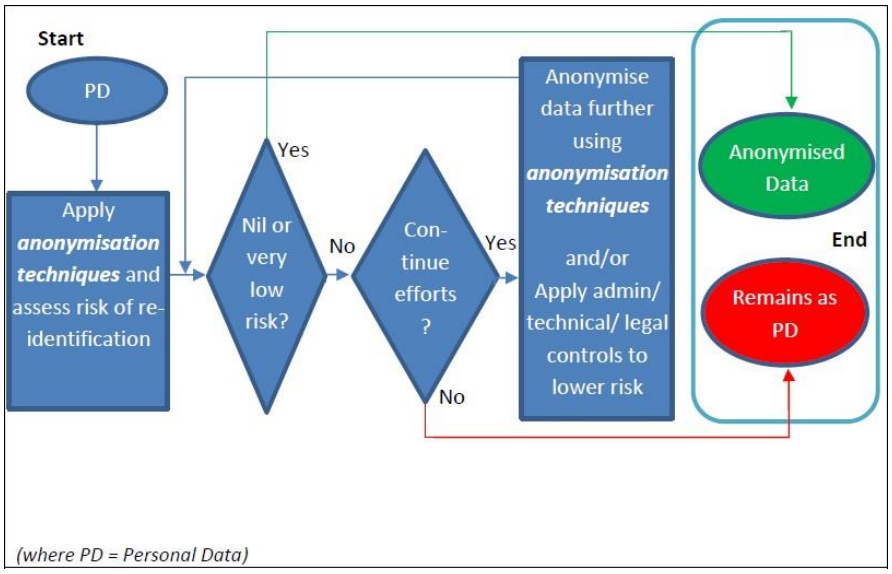

Fig. 3. The process of data anonymization

Another way to secure data is to classify it. Data classification is the process of organizing data into appropriate categories for the use and protection of information. In information security, data is labeled based on its sensitivity level, making it easier to find, track and protect information. Data classification significantly contributes to data governance, risk management, compliance, and information security. Fig. 4 shows how we classified our data in the organization for data governance initiative. Based on this classification, managing publishing, access, storage, backup, and other items related to data management is determined.

Class of data	The data class in the data governance document
Confidential data	Red grade
Sensitive personal data	Orange grade
Personal data	Yellow grade
Open data	Green grade

Fig. 4. Data Classification in data governance

2.3. Data anonymization

Data anonymization requires a good understanding of the following elements:

a. **Anonymization purpose and its advantage:** The process of anonymization, regardless of the techniques used, reduces the original information in the dataset to some extent. And hence, generally, as the size of anonymization increases, the amount of data precision and clarity is reduced. Therefore, there is a trade-off between security and privacy.
b. **Characteristics of anonymization techniques:** Each type of data needs different methods of anonymization. For instance, character masking is usually used on direct identifiers and aggregation is

used for indirect identifiers. Also, if the attribute value is continuous, data perturbation works much better. The various anonymization techniques also modify data in significantly different ways. We need to mask character; we need to replace the value across multiple records by the aggregation for some attributes. For some, we replace the entire attribute with unrelated, but consistent information like pseudonymization, and for others we remove the attribute entirely so that we call attribute suppression.

c. **Inferred information:** It may be possible for specific information to be inferred from anonymized data. For instance, masking may hide personal data, but it does not hide the length of the original data in terms of the number of characters. The inference problem is not limited to a single attribute but may also apply across attributes, even if all have had anonymization techniques. Therefore, the anonymization process must note every possibility before deciding on the basic methods and using the methodologies.

d. **Expertise with the subject matter:** Anonymization techniques reduce the "identifiability" of one or more individuals from the original dataset to a level acceptable by the organization's risk portfolio.

e. **Competency in anonymization process and techniques**: Anonymization is complex. So, we need an expert person in this field.

f. **The recipient:** The recipient should also be aware of the anonymization methods. In particular, the recipient's expected use of the anonymized data may impose limitations on the applied techniques because the utility of the data may be lost beyond acceptable limits. Extreme caution must be taken when making public releases of data and requires a much more robust form of anonymization than shared data under a contractual arrangement.

g. **Tools**: Due to the complexity and computation required, software tools can aid in executing anonymization techniques. Note that even the best tools will need adequate inputs (e.g., appropriate parameters to be used) or may contain limitations; hence human oversight and familiarity with the tools and data are still required.

2.3.1 Basic data anonymization techniques

Data anonymization refers to the method of preserving private or confidential information by deleting or encoding identifiers that link individuals to the stored data. It is done to protect the private activity of an individual or a corporation while preserving the credibility of the data collected and exchanged.

The General Data Protection Regulation (GDPR) (Voigt & Von dem Bussche,2017). outlines a specific set of rules that protect user data and create transparency. While the GDPR is strict, it permits companies to collect anonymized data without consent, use it for any purpose, and store it for an indefinite time—as long as companies remove all identifiers from the data.

The following are some anonymization methods:

1) Attribute Suppression
2) Record Suppression
3) Character Masking
4) Pseudonymization
5) Generalization
6) Swapping
7) Data Perturbation
8) Synthetic Data
9) Data aggregation

In the following, we will explain how we implemented character masking and data aggregation for our data anonymization.

- *Characters Masking*

Character masking is the change of the characters of a data value, e.g., by using a constant symbol (e.g., "*" or "x"). Masking is typically partial, i.e., applied only to some characters in the attribute. When the data value is a string of characters and hiding part of it is sufficient to provide the extent of anonymity required.

Table 2-a. Customer orders data before anonymization

Postal Code	Favorite Delivery Time Slot	Average No. of Orders Per Month
100111	8 pm to 9 pm	2
200222	11 am to 12 noon	8
300333	2 pm to 3 pm	1

Table 2-b. Customer orders data after anonymization

Postal Code	Favorite Delivery Time Slot	Average No. of Orders Per Month
10xxxx	8 pm to 9 pm	2
20xxxx	11 am to 12 noon	8
30xxxx	2 pm to 3 pm	1

Depending on the nature of the attribute, we replace the appropriate characters with a chosen symbol. Depending on the attribute type, we may decide to replace a fixed number of characters (e.g., for credit card numbers) or a variable number of characters (e.g., email address). In some cases, for sensitive data, we needed to mask data so that even the length of the data could be hidden.

Example:
This example in Table 2 shows an online grocery store conducting a study of its delivery demand from customer historical data to improve operational efficiency. The company masked the last four digits of the postal codes, leaving the first two digits, which correspond to the "sector code" within Singapore.

- *Data aggregation*

Convert a dataset from a list of records to a list of summarized values is called data aggregation. When individual records are not required and aggregated data is sufficient for the purpose, we use data aggregation. A detailed discussion of statistical measures is beyond the scope of this paper; however, typical ways include using totals or averages, etc. It might also be helpful to discuss with the data recipient about the expected utility and find a suitable compromise.

In the example provided in Table 3, a charity organization has the records of the donations made and some information about the donors. The charity organization has assessed that aggregated data is sufficient for an external consultant to perform data analysis, hence performs data aggregation on the original dataset. If the data collected has a single record in each category, identifying a donor can be easy for someone with additional knowledge.

3. CREATING OPEN DATA

After classification and anonymization, it is necessary to know how to publish data publicly. To solve this challenge, Open Data Framework is useful and comes in handy to creating public data.

There are three major benefits associated with release of government data to the general public as open data. Open data is seen as key to improving the effectiveness and efficiency of government policy and services. For example, data on crops, weather and geography might be analyzed to improve current approaches to farming and industry, or data on hospital admissions might be analyzed alongside demographic and census data to improve the efficiency of health services in areas of need. The other major benefit is that open data helps to create transparency and accountability, as a greater proportion of government decisions and operations are being shared with the public (Hardy & Keiran,2017). In this section we discuss some important examples that we used for preparing our open data.

Table 3-a. Original dataset of the charity organization

Donor	Monthly Income ($)	Amount donated in 2016 ($)
Oonor A	4000	210
Donor B	4900	420
Oonor C	2200	150
DonorD	4200	110
DonorE	s500	260
DonorF	2600	40
DonorG	3300	130
DonorH	5500	210
Donorl	1600	380
Donor J	3200	80
Oonor K	2000	440
DonorL	5800	400
DonorM	4600	390
DonorN	1900	480
DonorO	1700	320
Oonor P	2400	330
OOROF	4300	390
Oonor R	2300	260
OonorS	3500	80
Oonor T	1700	290

Table 3-b. Anonymized dataset of the charity organization using aggregation

Monthly Income ($)	No. of Donations Received (2016)	Sum of Amount donated in 2016 ($)
1000-1999	4	1470
2000-2999	5	1220
3000-3999	3	290
4000-4999	5	1520
5000-6000	3	870
Grand Total	20	5370

3.1. Publish open data

An open data portal is a publicly accessible website on which data published for public download is placed. Free information dissemination organized efforts to facilitate open access to scientific data. The concept of open access as an organized movement dates back to the 1950s (Committee on the Scientific Achievement of Earth-Spatial Observations, National Research Council, 2008). The exploitation of "big data" databases holds significant promise for the innovation, efficiency, and growth of commercial and industrial applications such as retail, transportation, and energy. Increasing public awareness has increased the demand for transparency by governments and organizations (Caruso & Nicol et al., 2013).

The Open Government Working Group (2007) (Voigt & Bussche, 2017) suggested that if data is implemented in a general manner and accordance with the following eight principles for better performance, that government data will be considered as open data: (1) Completed data (2) Initial data (3) Timely data (4) Available data (5) Machine processable data (6) Non-discriminatory data access (7) Non-proprietary data formats and (8) Unlicensed data.

Table 4. Experiences of different countries in the field of open data

country	Country Open data Website Address
Austria	***http://data.gv.at/***
Belgium	***http://data.gov.be/***
Brazil	***http://dados.gov.br/***
Canada	***http://www.data.gc.ca/***
Denmark	***http://digitaliser.dk***
Estonia	***http://pub.stat.ee/px-web.2001/Dialog/statfile1.asp***
Finland	***http://data.suomi.fi/***
France	***http://data.gouv.fr/***
Germany	***https://www.govdata.de/***
Greece	***http://geodata.gov.gr/geodata/***
Ireland	***http://www.statcentral.ie/***
Italy	***http://www.dati.gov.it/***
Japan	***http://datameti.go.jp/data/***
Netherlands	***http://data.overheid.nl/***
Norway	***http://data.norge.no/***
Portugal	***http://www.dados.gov.pt/pt/inicio/inicio.aspx***
Slovakia	***http://data.gov.sk/***
Spain	***http://datos.gob.es/***
Sweden	***http://öppnadata.se/***
England	***http://data.gov.uk/***
United States	***http://www.data.gov/***

These principles provide the best way to implement free data so that governments become more effective, transparent, and relevant to citizens' lives (Open Working Group, 2007). Due to the lack of an evaluation method, these eight principles have been accepted as appropriate and valuable guidelines in evaluating the performance of open data implementation. We consider city-owned data to be available data if it is public in a way consistent with these principles (Alanazi & Chatfield, 2012).

To make the most amount of data available, we need to make it understandable for both human and machine usage. The European Union has set up a working group specializing in data mining and text mining. Many researchers, knowledge-based companies, research institutes, libraries, and small and medium-sized companies are involved in such projects.

As mentioned earlier, different countries have taken steps to provide free information. Table 4 shows some of these countries and their free information gateway as good examples (Caruso et al., 2013).

- *The World Bank*

The mission of the World Bank Development Data Group is to provide high-quality national and international information to customers inside and outside the group and to enhance the capacity of member countries to produce and use statistical information. As part of the International Statistical System, the data group works with other organizations on new statistical methods, data collection activities, and statistical capacity-building programs. Most of the data is obtained from the statistical systems of the member countries, and the quality of global data depends on the performance of these national systems. The World Bank is working to help developing countries improve national statistical methods' capacity, efficiency, and effectiveness. Without better and more comprehensive national data, it is impossible to formulate effective policies or monitor progress towards global goals (Malpass, 2020).

- *London Data store*

The Greater London Authority (GLA) created the London Data Store as the first step in the free dissemination of London information (London Data Store, 2021). Raw data usually do not produce results until they are presented in a meaningful way, and most people do not have the tools to turn this data into meaningful information.

The London Database Privacy Policy is protected by the Data Protection Act (DPA) 2018 and the European Union's General Data Protection Regulation (GDPR) of privacy and personal data processing

- *New York Open Data*

New York, Open Data (NYC Open Data) makes many public data available by various New York City organizations including businesses, education, environment, and health (NYC Open Data, 2021). NYC Open Data contains information that is collected and maintained by the City government. Both New York State and the federal government also maintain data related to New York City that may not appear on NYC Open Data. As part of the site's process to improve government access, transparency, and accountability, a catalog of access to a machine-readable data repository is provided.

3.2. Tehran open data portal

According to the Law on Dissemination and Free Access to Information of Iran, every person has the right to access general information unless prohibited by law. Public institutions are forced to make the data subject to this law available to the public in the shortest possible time and without discrimination.

In the Tehran urban statistics and observatory center, to freely disseminate information, like the big cities of the world, we have published the statistical data of Tehran in the Tehran open data portal (Tehran data portal, 2021). This portal operated under the supervision of the Information and Communication Technology Organization of Tehran Municipality and was launched in December 2016.

The main idea of this portal, like similar examples, is to create a space for quick and easy access for researchers, academics, analysts, etc., to the statistics and information of Tehran.

The information is presented in 18 categories and 60 sub-categories in charts, tables, and shapefiles. So far (June 2021), more than 1400 data sets have been uploaded, including Tehran city statistics information used as a machine reader.

4. CONCLUSION

In 2019 all data in municipality we collected by municipality ICT organization of the Tehran Municipality. In this paper some explanation about data governance framework in municipality were presented and the process for creating open data were described. The goal of data governance is to make data as a decision helper for the city managers and policy makers. First, the definition of the life cycle of data was described then the method for defining data classification was explained. The result of classification is to clarify the access level of each data. In this paper, for transparency the green level is defined for non-sensitive data that can be published publicly. For the sensitive data in municipality, PDPC Singapore rules were followed to anonymize data and change its level to the green one. Also designing the metadata for efficiently using municipality data was described. The next step for achieving data governance is to predict the future in order to prevent possible future crises and optimal urban management.

REFERENCES

Thompson, N., Ravindran, R., & Nicosia, S. (2015). Government data does not mean data governance: Lessons learned from a public sector application audit. *Government information quarterly*, *32*(3), 316-322.

Henderson, D., & Earley, S. (Eds.). (2017). *DAMA-DMBOK: data management body of knowledge*. Technics Publications.

Chik, W. B. (2013). The Singapore Personal Data Protection Act and an assessment of future trends in data privacy reform. *Computer Law & Security Review*, *29*(5), 554-575.

El Emam, K., & Arbuckle, L. (2013). *Anonymizing health data: case studies and methods to get you started*. O'Reilly Media, Inc.

Zuiderwijk, A., & Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison.

Government information quarterly, 31(1), 17-29.
Cupoli, P., Earley, S., & Henderson, D. (2014). Dama-dmbok2 framework. *DAMA International*.
Abraham, R., Schneider, J., & Vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, *49*, 424-438.
Dal Maso, A. (2019). *The evolution of Data Governance: a tool for an improved and enhanced decision-making process*. Bachelor's thesis, Università Ca'Foscari Venezia.
Kamandi, A (1398). *DMBOK data management knowledge body* (First edition). Tehran: Sharif University of Technology Publications.
DeStefano, R. J. (2016). *Improving enterprise data governance through ontology and linked data*. ETD Collection for Pace University.
Dama Dach. (2016). DAMA-DMBOK Functional Framework, Available at: <https://damadach.org/dama-dmbok-functional-framework/>
Voigt, P., & Von dem Bussche, A. (2017). *The EU general data protection regulation (GDPR), A Practical Guide.* Cham: Springer International Publishing, 10, 3152676.
NYC Open Data (2021). New York Open Data Website, available at < https://opendata.cityofnewyork.us/data/>.
London Data Store (2021). Privacy Policy, available at https://data.london.gov.uk/about/privacy-policy
Tehran open data portal, Website, available at < https://www.data.tehran.ir >