Current State-of-the-art of Research in Data Governance: A Quantitative Systematic Literature Review Using Text Mining and Latent Dirichlet Allocation

Tan Chun Kit

---

c.kit@graduate.utm.my

Article history

Received:
-

Received in revised form:
-

Accepted:
-

Published online:
-

*Corresponding author

Abstract

This paper analyzes literature in order to search for trends in data governance research across the globe. Search were performed in Google Scholar and 24 articles between 2015 and 2019 were selected in order to perform descriptive analysis, word cloud, and Latent Dirichlet Allocation analysis. The results suggested that some industries such as agricultural and economic are currently lack of research while some methodological-related research are saturated. Furthermore, unsupervised LDA analysis may not be an appropriate text mining technique to use in the given context due to the extreme similaritiy of wording within the topic of study.

Keywords: *Data Governance, Text mining, Latent Dirichlet Allocation, Word cloud*

1. Introduction

Data governance refers to the process of discovering, defining, applying and monitoring databases (Firican, 2018). It is the overall management of the availability, usability, integrity, and security of data that includes a body of governance defined policy and actionable plan to execute the policy (Firican, 2018). In another word, data governance is a system that manages the databases within an organization from the data collection process until the removal of data from the data warehouse. Diverse sectors have employed data governance in maintaining data quality such as banking and finance, healthcare, education, and the manufacturing industry. A different sector requires similar but customized data governance in order to fulfill their respective needs such as data security, data accuracy, data accessibility, and accountability.

1.1. The Need for Data Governance Research

The increasing data volume from more and more sources is happening recently due to technological advancement. These advancements had enables industries and sectors to collect data from any part of its business process. It caused data inconsistencies that have to be identified and resolved before able to perform informed decisions using a reliable, valid, and accurate set of data. Furthermore, organizations today strive to develop automated and self-serving data analytics and reporting tool. The development of these tools requires detail and consistent understanding of the data the organization is collecting, storing and analyzing. Comprehensive data governance allows a company to understand these related data without any different opinions or perspectives.

On the other perspective, the continuously increasing impact of regulatory requirements such as General Data Protection Regulation (GDPR) and The Health Insurance Portability and Accountability Act of 1996 (HIPAA) had driven the need of data governance within an organization in order to satisfy the demand of these outside regulatory board. A comprehensive data governance system and policy represent a transparent and auditable system to any regulatory board or the society, which further enhances company images.

According to Al-Ruithe, Benkhelifa, & Hameed, (2018), currently data governance researches are mostly focusing on non-cloud data governance despite the huge needs for cloud data governance. The future research is needed in order to justify some concerned aspects when adopting cloud computing: Security, Privacy, Availability, Performance and Data classification (Al-Ruithe et al., 2018)

However, due to the broad coverage of data governance-related topics and its varied application in different industries and settings, a researcher often finds it hard to identify a research gap. According to Al-Ruithe et al., (2018), a systematic literature review could summarize current research as well as pointing out the research gap for future researchers.

** Corresponding author c.kit@graduate.utm.my*

Following Al-Ruithe et al., (2018), who conducted a systematic literature review in data governance publications from 2005 to 2017, they identified a certain trend of data governance research by conducting a systematic qualitative literature review. This research intended to fill up the 2 years gap of literature review which start from 2017 to 2019. However, an additional 2 years are included in order to saturate the number of literature for the following research purpose.

This research also intended to test the viability of employing quantitative text mining (Latent Dirichlet Allocation technique) in identifying data governance current research trends. Latent Dirichlet Allocation technique (LDA) is a text processing techniques used to identify topic cluster in a given content (Moro, Cortez, & Rita, 2015), Moro et al., (2015) conducted a systematic literature review in banking sector using LDA technique and encounter a cluster of meaningful topic of research in banking and finance sector which suggested the uses of LDA could potentially beneficial in other sectors such as data governance sector. and hence, leads to the following research objective:

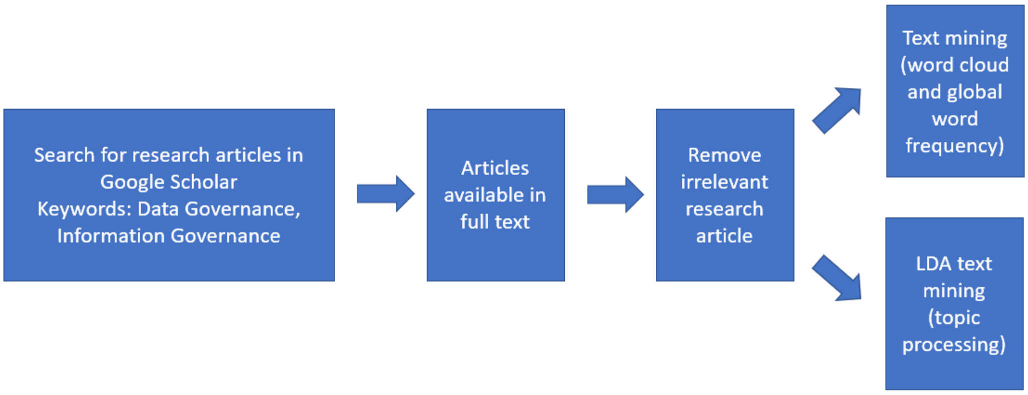
a. To conduct a quantitative systematic literature review on data governance-related research from 2015 to 2019.

b. To identify trends of data governance research from 2015 to 2019 using text mining techniques.

2. Methodology

This research used a systematic literature review (SLR) as described by Kitchenham & Charters, (2007) which is to identify, evaluate and interpret all available research relevant to a particular research topic. The purpose is to summarize the existing information and possibly pointing out the gap of research for future contributors.

This research attempts to cover data governance-related research from 2015 to 2019 by searching the article using Google Scholar only. The inclusion criteria and exclusion criteria are stated in detail and performed properly before going into the detail literature analysis in order to prevent misleading and irrelevant articles.



Search for research articles in Google Scholar Keywords: Data Governance, Information Governance
Articles available in full text
Remove irrelevant research article
Text mining (word cloud and global word frequency)
LDA text mining (topic processing)

Figure 1: Systematic literature review process flow ends with text mining

Figure 1 represents the process flow of article searching, selection, and filtering irrelevant input. The keyword used for searching is “Data Governance” and “Information Governance”. The articles with these 2 words in either title or abstract were included. Later, the articles without full-text access are removed. Lastly, irrelevant and duplicate entries were removed before putting it into text mining and review.

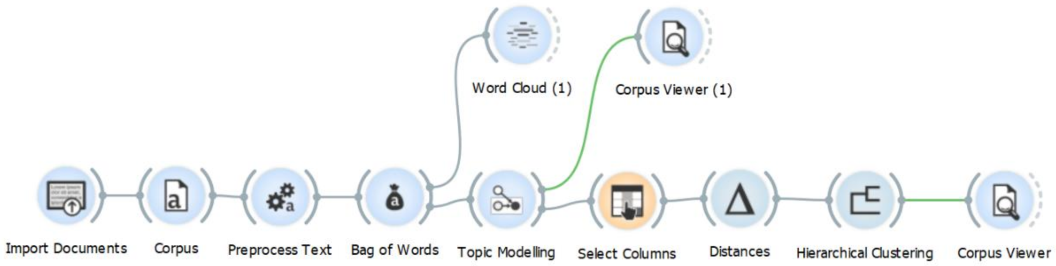


Figure 2: Orange text mining tool process flow

Figure 2 shows the process flow of text mining using the Orange text mining tool. Some words such as numerical value, “cid” “also” “Year” are removed prior to the text mining. For topic modeling, LDA technique was selected and hierarchical clustering was used in order to identify the relationship of each article to the selected topic.

3. Results

Table 1: Article count throughout the process of filtering (left to right)

Article Search	Article Available in full text	Remove duplicate and Irrelevant research article	Finalized article
32	28	24	24

Table 1 shown the article count from the beginning of Google Scholar search through the end of the finalized articles that were used in this research. Initially, 32 related articles found using the keyword "Data Governance" and "Information Governance". Later, some articles without full-text availability were removed follows by the removal of irrelevant, duplicate and non-English written research. The finalized article list consists of 24 articles from 2015 to 2019.

3.1. Descriptive Analysis

Table 2 shows the list of articles included in this study. It is important to notice that no research was done within the Southeast-Asia region.

Table 2: List of articles used in this paper

Citation	Title	Year of Publication	Location	Industry
(Thompson, Ravindran, & Nicosia, 2015)	Government data does not mean data governance: Lessons learned from a public sector application audit	2015	Australia	Public
(Carratero, Freitas, Cruz-Correia, & Piattini, 2016)	A case study on assessing the organizational maturity of data	2016	Spain	Healthcare
(Dai et al., 2016)	Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking	2016	USA	Private
(Vassilakopoulou, Skorve, & Aanestad, 2016)	A COMMONS PERSPECTIVE ON GENETIC DATA GOVERNANCE: THE CASE OF BRCA DATA	2016	Norway	Healthcare
(Murtagh, Turner, Minion, Fay, & Burton, 2016)	International Data Sharing in Practice: New Technologies Meet Old Governance	2016	UK	Public
(Olaitan, Herselman, & Wayi, 2016)	TAXONOMY OF LITERATURE TO JUSTIFY DATA GOVERNANCE AS A	2016	South Africa	Public and Private

	PREREQUISITE FOR INFORMATION GOVERNANCE			
(Rasouli, Eshuis, Trienekens, & Grefen, 2016)	Information Governance Requirements for Architectural Solutions Supporting Dynamic Business Networking	2016	Netherland	Private
(Soma, Termeer, & Opdam, 2016)	Informational governance – A systematic literature review of governance for sustainability in the Information Age	2016	Netherlands	(Methodological)
(Taylor, Richter, Jameson, & Perez de Pulgar, 2016)	Customers, users or citizens? Inclusion, spatial data and governance in the smart city	2016	Netherland	(Methodological)
(Mason, 2017)	Big Data Governance: Solidarity and the Patient Voice	2016	UK	Healthcare
(Alreemy, Chang, Walters, & Wills, 2016)	Critical success factors (CSFs) for information technology governance(ITG)	2016	Saudi Arabia	(Methodological)
(Kuerbis & Mueller, 2017)	Internet routing registries, data governance, and security	2017	USA	Institutional economics
(Brown & Toze, 2017)	Information governance in digitized public administration	2017	Canada	Public
(Winter & Davidson, 2017)	Investigating Values in Personal Health Data Governance Models	2017	USA	Healthcare
(Alhassan, Sammon, & Daly, 2018)	Data governance activities: a comparison between scientific and practice-oriented literature	2017	Ireland	(Methodological)
(Nielsen, Nielsen, & Olivia, 2017)	A Comprehensive Review of Data Governance Literature	2017	Denmark	(Methodological)
(Stahl, Rainey, Harris, & Fothergill, 2018)	The role of ethics in data governance of large neuro-ICT projects	2017	UK	Healthcare

(Wolfert, Bogaardt, Ge, Soma, & Verdouw, 2016)	Guidelines for governance of data sharing in agri-food networks	2017	New Zealand	Agriculture
(Turel, Liu, & Bart, 2017)	Board-Level Information Technology Governance Effects on Organizational Performance: The Roles of Strategic Alignment and Authoritarian Governance Style	2017	USA	Private
(Were & Moturi, 2017)	Toward a data governance model for the Kenya health professional regulatory authorities	2017	East Africa	Healthcare
(Krimpmann & Stühmeier, 2019)	Big Data Governance	2017	Switzerland	(Methodological)
(Zhang, Gao, Yang, & Song, 2017)	Research on Big Data Governance Based on Actor-Network Theory and Petri Nets	2017	China	(Methodological)
(Alhassan, 2018)	Critical Success Factors for Data Governance: A Theory Building Approach	2019	Saudi Arabia	(Methodological)
(Fothergill, Knight, Stahl, & Ulnicane, 2019)	Responsible Data Governance of Neuroscience Big Data	2019	UK	Healthcare

Table 3 illustrate the frequency of article publication by year. It can be seen that 2017 was the peak of data governance-related paper publication year while 2018 receive no paper regarding the topic. While in 2019 (up to date: 27/4/2019), 2 research is published.

Table 3: Article frequency by year

Year	2015	2016	2017	2018	2019
Article Count	1	10	11	0	2

Table 4 shown the industry involved in the selected data governance papers. It is found that most papers focused on methodological discussion and development such as literature review and modeling. Besides that, healthcare settings received the most concerns among the academicians while the agricultural and economic sectors received little input.

Table 4: Article frequency by industry/sector

Sector	Methodological	Public	Private	Healthcare	Agriculture	Economics
Article Count	8	4	4	7	1	1

3.1 Text Mining

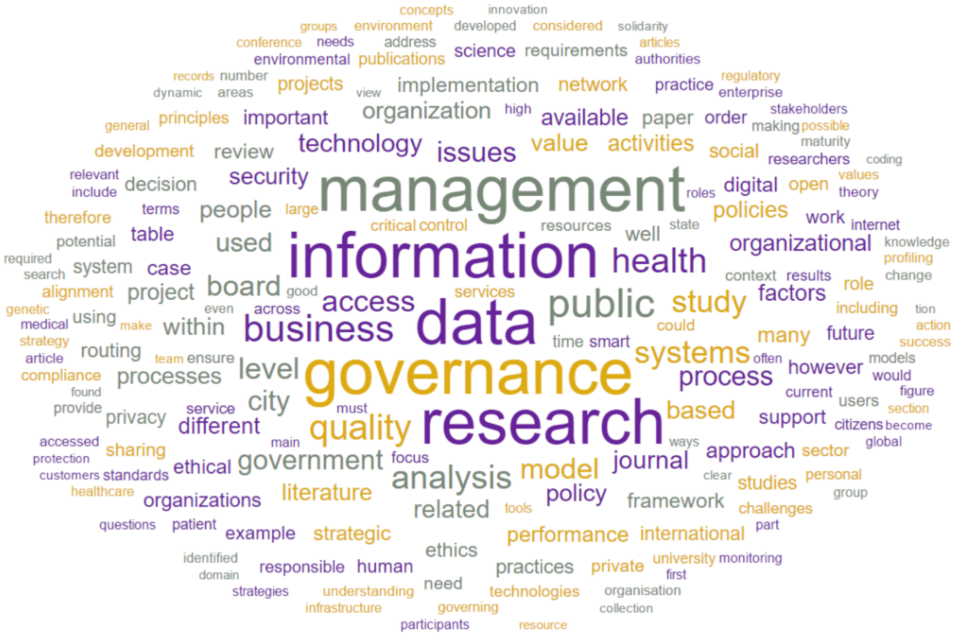


Figure 3: Word Cloud

Figure 3 shown the word cloud generated from the Orange tool. It can be seen that the most frequently used words are data governance and information governance as expected. Hence in order to make a meaningful judgment, words without pointing out a specific theme were removed for the following process.

After the manual removal of meaningless words, table 5 shown the top 10-word count among all articles. It can be seen that both public (related word: public, government) and private (business) sector receive similar concerns while most academician focuses on data governance research on healthcare setting. Methodologically, modeling, data governance policy and data security seem to be the main concerns among these researches.

Table 5: Top 10-word count after manual removal of snot important words.

Word Count	Word
465	public
430	business
389	quality
387	health
318	access
304	model
284	government
233	policy
223	security
217	policies

The LDA analysis was set to limit the number of topics to only 4 topics and the result is tabulated as in Table 6. Unfortunately, according to the cluster, no specific and meaningful theme can be identified.

Table 6: LDA analysis

#	Topics
1	data, research, governance, information, public, health, access, patient, network, solidarity
2	solidarity, city, smart, space, woods, spatial, activism, genetics, buyx, prainsack
3	style, authoritarian, boards, bart, styles, jewer, mckay, parenting, advising, gains
4	data, governance, management, board, information, model, level, performance, strategic, business

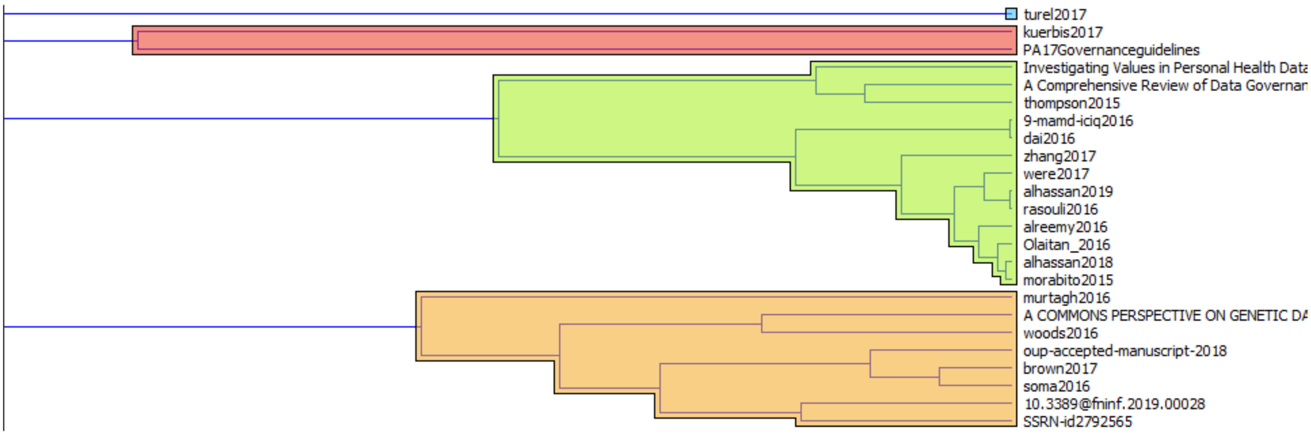


Figure 5: Hierarchical clustering

In order to identify potential themes behind the topics, hierarchical clustering was used and set to visualize 4 clusters to identify the 4 clusters under each topic. Then, each paper belongs to the cluster was studied and analyze properly. However, no significant theme was found within each theme as well as the points that separate each of the themes.

4. Discussions

The descriptive results show that research in agriculture and economics received very minimal interest in research while some sectors such as retail, manufacturing, and education receive no research in data governance. Congruence to Al-Ruithe et al., (2018), this paper suggested in recent years, little or no research was done on data governance on cloud and big data.

The LDA analysis employed in this research was not able to classify and extract meaningful themes as suggested by (Al-Ruithe et al., 2018). The issue behind may be due to Al-Ruithe et al., (2018) employed a supervised text mining by matching the word using a set of an inter-defined dictionary while this research employed unsupervised text mining. The model may be overfed by meaningless words while failing to identify meaningful patterns and words. Furthermore, the similarity between the wording and vocabulary within data governance-related research could further impact the text mining method negatively. Hence, the result suggested unsupervised text mining methods may not be an appropriate method in performing topic-specific text mining (such as data governance-related topics conducted in this paper).

In terms of the research gap of data governance-related topic, retail, manufacturing and education industry seems to be potential unexplored territory while the need of cloud data governance-related research are still in need due to the lack of research as suggested in this paper and Al-Ruithe et al., (2018). There is no methodological related gap as plenty of research has been done on both modeling, policy-making as well as data security.

The research that focuses on the methodological aspect is rather saturated based on the review, 33% of the reviewed literature investigated the methodological aspect of data governance, leaving some areas such as agriculture and economic only received 1 (4%) research respectively.

5. Conclusion

This paper attempted to perform a quantitative systematic literature review as suggested by the previous literature review. However, descriptive analysis is the only analysis that yields meaningful reviews in this study by pointing out industry and year that received a high frequency of research. The uses of unsupervised LDA text mining seem to not able to extract meaningful information due to overfed by irrelevant information. Lastly, the potential research gaps in data governance-related topics are pointed out for future researchers.

Acknowledgments

I would like to express my gratitude towards Universiti Teknologi Malaysia (UTM) for providing guidance and resources in supporting my research. I have acquired so much knowledge and understanding in this field which I believe will definitely benefit and navigate my future work. Lastly, I would like to thank my subject lecturer, Dr. Noor Hafizah for encouraging me to start writing my paper in the data governance field.

6. References

[1] Al-Ruithe, M., Benkhelifa, E., & Hameed, K. (2018). A systematic literature review of data governance and cloud data governance. Personal and Ubiquitous Computing, c, 1–21. https://doi.org/10.1007/s00779-017-1104-3

[2] Alhassan, I. (2018). Critical Success Factors for Data Governance : A Theory Building Approach. Information Systems Management, 36(May), 98–110. https://doi.org/10.1080/10580530.2019.1589670

[3] Alhassan, I., Sammon, D., & Daly, M. (2018). Data governance activities: a comparison between scientific and practice-oriented literature. Journal of Enterprise Information Management, 31(2), 300–316. https://doi.org/10.1108/JEIM-01-2017-0007

[4] Alreemy, Z., Chang, V., Walters, R., & Wills, G. (2016). Critical success factors (CSFs) for information technology governance (ITG). International Journal of Information Management, 36(6), 907–916. https://doi.org/10.1016/j.ijinfomgt.2016.05.017

[5] Carratero, A. G., Freitas, A., Cruz-Correia, R. J., & Piattini, M. (2016). A case study on assessing the organizational maturity of data management, data quality management and data governance by means of MAMD. Iciq 2018, 9.

[6] Brown, D. C., & Toze, S. (2017). Information governance in digitized public administration. Canadian Public Administration, 60(4), 581-604.

[7] Dai, W., Wardlaw, I., Cui, Y., Mehdi, K., Li, Y., & Long, J. (2016). Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking. 448, 439–450. https://doi.org/10.1007/978-3-319-32467-8

[8] Firican, G. (2018). What is Data Governance? Retrieved April 30, 2019, from Data Governance website: http://www.lightsondata.com/what-is-data-governance/

[9] Fothergill, B. T., Knight, W., Stahl, B. C., & Ulnicane, I. (2019). Responsible Data Governance of Neuroscience Big Data. Frontiers in Neuroinformatics, 13. https://doi.org/10.3389/fninf.2019.00028

[10] Kitchenham, B., & Charters, S. (2007). Source: &quot; Guidelines for performing Systematic Literature Reviews in SE &quot; , Kitchenham et al Guidelines for performing Systematic Literature Reviews in Software Engineering Source: &quot; Guidelines for performing Systematic Literature Reviews i. 1–44. https://doi.org/10.1145/1134285.1134500

[11] Krimpmann, D., & Stühmeier, A. (2019). Big Data Governance. International Journal of Service Science, Management, Engineering, and Technology, 8(3), 79–92. https://doi.org/10.4018/ijssmet.2017070105

[12] Kuerbis, B., & Mueller, M. (2017). Internet routing registries, data governance, and security. Journal of Cyber Policy, 2(1), 64–81. https://doi.org/10.1080/23738871.2017.1295092

[13] Mason, P. H. (2017). Big Data Governance: Solidarity and the Patient Voice. Journal of Bioethical Inquiry, 14(4), 571–574. https://doi.org/10.1007/s11673-017-9812-y

[14] Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. Expert Systems with Applications, 42(3), 1314–1324. https://doi.org/10.1016/j.eswa.2014.09.024

[15] Murtagh, M. J., Turner, A., Minion, J. T., Fay, M., & Burton, P. R. (2016). International Data Sharing in Practice: New Technologies Meet Old Governance. Biopreservation and Biobanking, 14(3), 231–240. https://doi.org/10.1089/bio.2016.0002

[16] Nielsen, O. B., Nielsen, B., & Olivia, " A. (2017). A Comprehensive Review of Data Governance Literature. Issue Nr, 8(8), 3. Retrieved from http://aisel.aisnet.org/iris2017http://aisel.aisnet.org/iris2017/3

[17] Olaitan, O., Herselman, M., & Wayi, N. (2016). TAXONOMY OF LITERATURE TO JUSTIFY DATA GOVERNANCE AS A PREREQUISITE FOR INFORMATION GOVERNANCE. 586–605.

[18] Rasouli, M. R., Eshuis, R., Trienekens, J. J. M., & Grefen, P. W. P. J. (2016). Information Governance Requirements for Architectural Solutions Supporting Dynamic Business Networking. 9586, 184–189. https://doi.org/10.1007/978-3-662-50539-7

[19] Soma, K., Termeer, C. J. A. M., & Opdam, P. (2016). Informational governance - A systematic literature review of governance for sustainability in the Information Age. Environmental Science and Policy, 56, 89–99. https://doi.org/10.1016/j.envsci.2015.11.006

[20] Stahl, B. C., Rainey, S., Harris, E., & Fothergill, B. T. (2018). The role of ethics in data governance of large neuro-ICT projects. Journal of the American Medical Informatics Association, 25(8), 1099–1107. https://doi.org/10.1093/jamia/ocy040

[21] Taylor, L., Richter, C., Jameson, S., & Perez de Pulgar, C. (2016). Customers, Users or Citizens? Inclusion, Spatial Data and Governance in the Smart City. Ssrn, (Kip 13759). https://doi.org/10.2139/ssrn.2792565

[22] 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. https://doi.org/10.1016/j.giq.2015.05.001

[23] Turel, O., Liu, P., & Bart, C. (2017). Board-Level Information Technology Governance Effects on Organizational Performance: The Roles of Strategic Alignment and Authoritarian Governance Style. Information Systems Management, 34(2), 117–136. https://doi.org/10.1080/10580530.2017.1288523

[24] Vassilakopoulou, P., Skorve, E., & Aanestad, M. (2016). A commons perspective on genetic data governance: the case of BRCA data. Proceedings of the 24th European Conference on Information Systems, ECIS.

[25] Were, V., & Moturi, C. (2017). Toward a data governance model for the Kenya health professional regulatory authorities. TQM Journal, 29(4), 579–589. https://doi.org/10.1108/TQM-10-2016-0092

[26] Winter, J. S., & Davidson, E. (2017). Investigating Values in Personal Health Data Governance Models Twenty-third Americas Conference on Information Systems. 1–10.

[27] Wolfert, S., Bogaardt, M., Ge, L., Soma, K., & Verdouw, C. (2016). Guidelines for governance of data sharing in agri-food networks. (October), 1–11. https://doi.org/10.5281/zenodo.893700

[28] Zhang, S., Gao, H., Yang, L., & Song, J. (2017). Research on big data governance based on actor-network theory and Petri nets. Proceedings of the 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design, CSCWD 2017, 372–377. https://doi.org/10.1109/CSCWD.2017.8066723