Dark Data Management as frontier of Information Governance

Ahmad Fuzi Md Ajis
Faculty of Information Management
Universiti Teknologi MARA
Cawangan Johor Kampus Segamat,
Malaysia
ahmadfuzi@uitm.edu.my

Siti Hajar Baharin
Faculty of Information Management
Universiti Teknologi MARA
Cawangan Johor Kampus Segamat,
Malaysia
hajar903@uitm.edu.my

***Abstract* — Information Governance (IG) decrease cost of operation, creates opportunity of new ventures, prepare liturgical evidence for litigation while committing to the information lifecycle process by abiding records management using proper and comprehensive information policy. In the meantime, IG enabling access and discovery using Information Technology (IT) infrastructure that restrict information based on privileges and perimeters with the embedment of information architecture while sustaining better business operation. Dark data is information, collected as a function of an organization's normal operations but rarely or never analyzed or used to make intelligent business decisions. By IG means, dark data would dramatically increase revenue, growth and efficiency of a business and institution compared to their competitor who don't take serious action on dark data. As a result, a Dark Data Management Model proposed comprises of 1 main core governance element, 4 main processes but defined by few activities. The model was proposed as generic approach on managing dark data since the process should be applied continuously not only to new receiving information but also stored data. Dark data mining process was not included in the proposed model as the model focus on the preventive measure which can be used to current data which falls under the category of dark data based on the value assessment of each organization.**

Keywords—Big Data; Information Governance; Dark Data Management

I. INTRODUCTION

In the business world, organizations are gathering and storing diversities of information and data from numerous sources. Since, we're in the age of Big Data, each organization faced many challenges in handling and ensuring that all the data and information being utilized. Most of them being gathered by the organization is used to improve the whole organization process including product information, marketing strategy, development of the organization, organization assets, etc.

As data and information become crucial in any businesses, the governance of the data and information should become the prioritized activities which clearly defines how information could be governed to ensure the optimization of information in business performance. Some scholars relate information governance as part of the corporate governance reflects that any information handlings or involvement in any business decision making is compulsory and unavoidable.

The decision-making process currently become more critical and meticulous since the phenomenon of the Big Data which emerge with the promise of providing valuable information which will be very beneficial in the future. Hence, enormous increment in the creation of the big data exceeds the capability of certain business to critically use information without a good practice of information governance. Abundance of information created every transactional day, collected and expanding while the mechanism of controlling it weaken.

Information governance (IG) is not a new jargon to businesses who imply information could bring better performance. IG decrease cost of operation, creates opportunity of new ventures, prepare liturgical evidence for litigation while committing to the information lifecycle process by abiding records management using proper and comprehensive information policy. In the meantime, IG enabling access and discovery using IT infrastructure that restrict information based on privileges and perimeters with the embedment of information architecture while sustaining better business operation.

There are three governance approaches as a way of governing. The hierarchical approach can be considered as the 'classical approach'. It has been around for many years and is traditionally the basic for a way to govern an organization or a state. The hierarchical approach is based on steering and control [12]. The key element of steering is 'direction', and control is the way to give the 'insurance' that rules are followed. In modern times this kind of governance is moving from command to regulation, from procuring to enabling and from benevolence to activation, but the basis consists of a centrally directed approach, including a structuration framework, using structure and agent concepts. It is a topic for discussion whether such a framework is always effective for all forms of information governance, since exchanging information, formally and informally, is not necessarily restricted to organizational borders.

The governance models of Kooiman offer a wide range of possibilities on the governance of information, each of them with its own capabilities and flaws. Optimization of the information value is related to the three (groups of) actors involved. These actors do not necessarily belong to the same organization. For example, creator and receiver may belong to different organizations and the 'governing' actor maybe part of either one of them. In this case the

'optimal' governance approach might be a co-governance form, such as a network governance approach. Other examples and approaches may apply as well. It leads to new research questions on what approach may contribute to a successful form of information exchange, for example by giving enough space to innovate.

On the other hand, IG may have different interpretation based on geographic area, as found by Soma, Termeer and Opdam (2016), in Europe IG refers to be relevant to structural changes of public administration, while worldwide, market related transformations are more relevant. IG is also established with essential elements whereby increased human interconnectedness through e- governance, self-organization, private governing and empowerment must play vital role to ensure IG successful [17].

The application of information governance found to be more enticing when technological device become more advanced and the habit of sharing information and knowledge become the lifestyle of end users. However, this habit creates major impact to the expansion of Big Data. Finding an information from the enormous amount of information stored in the gigantic digital storage, it's like finding a needle in a haystack if improper IG is in place. Implementing IG is not simply relying on the process of providing access and discovery to the information but having the whole elements and components of IG that creates the whole business operation meaningful in achieving its objectives and goal. Along with that, all the valuable information, organization almost certainly also storing an increasing amount of data that has no real tactical value at all. This unmanaged information was called as "Dark Data."

II. Dark Data Management

Dark data is information, collected as a function of an organization's normal operations but rarely or never analyzed or used to make intelligent business decisions. Most of it gets buried within a vast and unorganized collection of other data assets. Dark data also can be referred as "data exhaust," because most of the information consider as overlooked information, even though that data have valuable input to the organization. And the portions that aren't of value can be a significant drain on resources, including wasted digital storage space [15].

Information that is not managed properly can also expose an enterprise to considerable vulnerabilities. Some of the impacts of dark data are distributed and duplicated content widespread without explicit oversights and weakens security. It's too risky as the Hackers have more potential entry points and the data leaked, lost, stolen, or breached dark data can result in damaged reputations as well as loss of competitive strength towards the organization. In addition, the sheer volume of dark data impacts the costs for searching and producing appropriate information and imposes a wasted storage cost in operating budgets [15].

The organizations are advisable to overcome and manipulate their dark data into valuable to growth and in directly improve the organization weaknesses. Unfortunately, some data were misinterpreted or excluded from the silos of important and valuable information. For example, an information called as 'event knowledge' which occurred during the discussion on occasion or a decision that has been made through conferences which not recorded is another dark data which ignored and missed. Moreover, failed experiments log that being ignored because of its results can be a valuable dark data as the findings of the experiments journey reveal surprising truth.

IG QUADRANT	DESCRIPTION	RESEACHER	RESPONSIBILITY
Increased Human Interconnectedness	e-governance interconnect public services, citizens and market incentives	(Andreopoulou, 2009; Naik, 2011)	Governing Actor (M.N. Kooper et al, 2011)
Self-Organization	self-organization interconnect locals with societies at large	(Delaune, 2012)	
Private governing	private governing interconnect market and public-sector initiatives	(Gopalakrishnan et al., 2010)	
Empowerment	empowerments interconnect the weak with the stronger societal actors	(Banas, 2010).	

Fig. 2. IG awareness from local definition.

Meanwhile, non-evidence-based decision that always been made by the administration of an organization which in the end would jeopardize the whole organization operation because of ignoring the policies, circulars which referred as dark data in this case.

Those unused data can be altered and adjust when it has been handling by someone who are expertise and well trained. Here, circle of information governance is needed. Governance is generally interpreted as a hierarchical framework for guidelines, policies, responsibilities, and procedures to ensure a certain level of control within an organization. But the definition of information governance does not necessarily restrict its use to one specific framework. Information governance may vary from a set of policies, a way of working, or the creation of a space within a predefined settlement (such as an online community), or it may as well apply to a framework of strict (accounting) rules within a country (such as privacy regulations).

By IG means, dark data would dramatically increase revenue, growth and efficiency of a business and institution compared to their competitor who don't take serious action on dark data. Increased in competitive advantages due to better business intelligence data resulted from dark data mining create more opportunity for an institution operation while at the same time prepared them with greater legal compliance with minimized risk by exercising retention and disposal plan.

There were some suggestions on how to handle the dark data including encrypting, analyzing the data regularly and take actions such as classifying legacy data. Elimination of the dark data hiding in the digital cracks and corners of

organization is an important milestone when establishing an information governance strategy that can withstand inspection by auditors and compliance officers. But it is not the only step. This commitment to ensure that information governance strategy successful is not an easy task but it is achievable by having strong foundation of technology, consistent policies and processes, and employee awareness.

Information governance nowadays is not just content or data management, records and retention schedules and mere storage but the scope become large. Information governance is a strategic approach to the management of all enterprise information starting from creation to final disposition and guides the organizations to control costs and eliminate the risks more effectively. Acting on the dark data gathering inside organization is an essential part of a governance program. Directly, it can be apply in the field of records management, information management, enterprise content management, privacy (data protection), freedom of information, corporate governance, information risk, information security and e-discovery [19].

III. Dark Data Management Model

A few literatures have been reviewed and compared to recognize the possible way to manage dark data either to prevent it from piling up or to avoid dark data from created.

As a result, a Dark Data Management Model proposed comprises of 1 main core governance element, 4 main processes but defined by few activities (Figure 1).

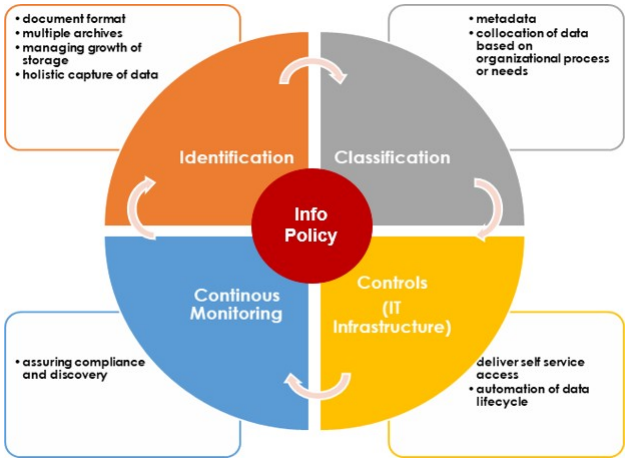

Fig. 2. Proposed model for Dark Data Management

The main core governance element of Dark Data Management is Information Policy. It is the policy that act as a rule that regulate how such information being created, used, stored, access, communicate or distributed internal or external organization [6, 16, 18].

Identification which is the process of identifying what data should be kept including the storage facility issues. It involves determining the document format should be stored, creating multiple archives based on departments exist in the organization or the values which deemed important to the operation of the institutions, managing growth of storage by only stores information that has values to the organization, and holistic capture of data which requires every data and information created must be assessed to recognize its value for business and operation to be kept [3, 5, 11].

Second process involves classification procedure where it involves assignment of metadata towards information of data stored in the institutional repository to ensure collocation of information done properly. The collocation process should be done according to the organizational process or needs to motivate discovery and familiarity with the information [8, 6, 10, 11].

The process continues with the control part whereby it concerns with the security and integrity of the information by maximizing legal compliance. Self-service access is proposed as method of utilizing any information as the IT infrastructure must be ready to record logs of the user footprint on each document. Moreover, automation of data lifecycle should be executed according to retention and disposal schedule to minimize risk of legal claim [3, 5, 12]. Finally, continuous monitoring is needed to ensure compliance and discovery of the information for the sake of the organization operation [11].

IV. Discussion & Conclusion

Information Governance is no longer an aspirational objective strategy, which it was when the term was first in vogue a few years ago. New technologies allow you to access, understand, control, and act on a wide variety of dark data to ensure successful usage of dark data, mined from the organization repository.

Equally important, avoiding dark data in organization must be strictly implemented to achieve greater benefits. With these advances, and an increasing need for such solutions of dark data, now is unmistakably a coming of age for Information Governance. However, the existence of dark data in the big data composition could cost a fortune to be handled by the organization in the long run. Therefore, the model was proposed by integrating few scholarly propositions.

The model was proposed as generic approach on managing dark data since the process should be applied continuously not only to new receiving information but also stored data. Dark data mining process was not included in the proposed model as the model focus on the preventive measure which can be used to current data which falls under the category of dark data based on the value assessment of each organization. This could be the gap of the research that can be studied in the future.

Acknowledgment

This project is financially supported by the Ministry of Higher Education Malaysia under Fundamental Research Grant Scheme (600-IRMI/FRGS 5/3 (054/2017).

REFERENCES

[1] Abualigah, L. M. Q. (2019). Feature Selection and Enhanced Krill Herd Algorithm for Text Document Clustering. Studies in Computational Intelligence.

[2] Abualigah, L. M. Q., & Hanandeh, E. S. (2015). Applying genetic algorithms to information retrieval using vector space model. International Journal of Computer Science, Engineering and Applications, 5(1), 19.

[3] Al-Sai, Z. A., & Abualigah, L. M. (2017, May). Big data and E-government: A review. In Information Technology (ICIT), 2017 8th International Conference on (pp. 580-587). IEEE.

[4] Abualigah, L. M., & Khader, A. T. (2017). Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. The Journal of Supercomputing, 73(11), 4773-4795.

[5] Andreopoulou, Z. Adoption of information and communication technologies (ICTs) in public forest service in Greece. Environ. Prot. Ecol., 10, (2009): 1194–1204.

[6] Lugmayr, A., Stockleben, B., Scheib, C., Mailaparampil, M.A. (2017) Cognitive big data: survey and review on big data research and its implications. What is really "new" in big data*?*. *Journal of Knowledge Management, 21(1)*, pp.197-212. Retrieved from doi: https://doi.org/10.1108/JKM-07-2016-0307

[7] Banas, P. International ideal and local practice – access to environmental information and local government in Poland. Environ. Policy Gov., 20, (2010): 44–56.

[8] Commvault. 5 ways to illuminate your dark data . USA. (2014).

[9] Cook, C. N., Hockings, M., & CArter, R. Conservation in the dark? The information used to support management decisions . Front Ecol Environ , (2010): 181-186.

[10] Crawford Technologies. Shedding Light on Dark Data: Extending the Value of Document Archives. USA. (2017).

[11] Compert, C, et al. Information Governance Principles and Practices for a Big Data Landscape . US :IBM Redbooks. (2014).

[12] Delaune, G. Rapa Nui on Verge: Easter Island's struggles with integration and globalization in the Information Age. Berkeley Plan. J., 25, (2012): 126–139.

[13] ESI. Information Governance: Executive briefing book. 2017. 10 October 2017. https://www.esi.ac.ma/Dossiers/20140709040704.pdf

[14] Giordano, A.D. Performing Information Governance: A Step-by-Step Guide to Making Information Governance Work. IBM Press Pearson. (2014)

[15] Gopalakrishnan, S., Kessler, E., & Scillitoe, J. Navigating the innovation landscape: past research, present practice, and future trends. Organ. Manag. J., 7, (2010): 262–277.

[16] Kevin, N. M., Wanyaga, F. M., Kibaara, D., Dinda, W. A., & Ngatia,
J. K. Dark data: Business Analytical tools and Facilities for illuminating dark data. Scientific Research Journal , (2016): 1-10.

[17] Kooiman, J. Governing as governance. London: Sage publications. (2003).

[18] Kooiman, J. (Ed.). Fish for life, interactive governance for fisheries. Amsterdam: Amsterdam University Press. (2005).

[19] Mullon, P.A., Ngoepe, M. (2019). An integrated framework to elevate information governance to a national level in South Africa. *Records Management Journal*. Retrieved from doi: https:// doi.org/10.1108/RMJ-09-2018-0030

[20] M. R. Anderson, D. Antenucci, V. Bittorf, M. Burgess, M. J. Cafarella, A. Kumar, F. Niu, Y. Park, C. R´e, and C. Zhang. Brainwash: A data system for feature engineering. In CIDR 2013, Sixth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 6-9, 2013, Online Proceedings, 2013.

[21] D. Barbosa, H. Wang, and C. Yu. Shallow information extraction for the knowledge web. In 29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013, pages 1264–1267, 2013.

[22] F. Chen, X. Feng, C. Re, and M. Wang. Optimizing statistical information extraction programs over evolving text. In IEEE 28th International Conference on Data Engineering (ICDE 2012), Washington, DC, USA

[23] Oracle, "Oracle: Big Data for the Enterprise," Oracle, Redwood, White Paper 2013.

[24] Schniederjans, M.J., Schniederjans, D. G. & Starkey C. M., Business Analytics Principles, Concepts and Applications, 1st ed., Gill Editorial Services, Ed. New Jersey, Upper Saddle River: Pearson Education, Inc., 2014.

[25] Paul Zikopoulos, Chris Eaton, Diroos Dirk, Tom Deutsch, and George Lapis, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, 1st ed., Steve Sit, Ed. New York, United States: MCGRAW-HILL, 2012.