

CORE CATALYST CONDITIONS INFLUENCING DARK DATA EMERGENCE: A GROUNDED THEORY APPROACH

Ahmad Fuzi Md Ajis^{1*}, Faten Elina Kamaruddin¹, Fatin Nabihah Mohd Taha², Zati Atiqah Mohamad Tanuri³

¹Center of Information Science, College of Computing, Informatics & Mathematics, Universiti Teknologi MARA, UiTM Cawangan Johor Kampus Segamat, 85000 Segamat, Johor, Malaysia

³Faculty of Accountancy, Universiti Teknologi MARA, UiTM Cawangan Johor Kampus Segamat, 85000 Segamat, Johor, Malaysia

⁴Center of Information Science, College of Computing, Informatics & Mathematics, Universiti Teknologi MARA, UiTM Cawangan Kedah Kampus Merbok, 08400Merbok, Kedah, Malaysia

*Corresponding Author

Email: ahmadfuzi@uitm.edu.my

Received: 15 July 2024

Accepted: 10 September 2024

ABSTRACT

Most of the literature highlighted dark data research from the perspective of academic research institutions and large business firms, and none was found to be about Small and Medium Enterprises (SMEs). The impact of dark data is the risk that could harm the business enterprise. The dark data residing within the enterprise repository could disturb legal compliance with data management regulations and legislation. Therefore, grounded theory research was conducted to investigate the dark data phenomenon towards SMEs in Malaysia. The primary goal of this study is to discover how organizations prepare themselves for dark data caretaking. By using the Grounded Theory Methodology (GTM), 18 SMEs were selected and sampled through snowballing and theoretical sampling procedures. The study found that business owners prepare their organization to deal with dark data by enabling data caretaking which is comprised of a few core catalyst conditions consisting of leveraging technology, upskilling, security and controlling. Although the enabling caretaking effort facilitates the data caretaking activities simultaneously it influences the occurrences of data crisis.

Keywords: Dark Data, Core Catalyst, Leveraging Technology, Upskilling, Security Controlling

1.0 INTRODUCTION

Dark data has been defined from various perspectives as (Corallo et al., 2021) reviewed 22 publications comprised of academic and non-academic publications about dark data definition from various fields and perspectives. The systematic literature review conducted by them is to facilitate the establishment of dark data definitions according to the manufacturing industry. It seems that the establishment of the dark data definition falls under the property of searchability (Kambies et al., 2017), accessibility (Banafa, 2015), unknown existence (Lugmayr et al., 2017), uncategorized and ignored data (Intel, 2018) which influenced by its formats and led to unused of data (Trajanov, et al., 2018) yet being hoarded throughout time.

However, the demarcation of the dark data definition was not explored to its deepest epistemology.

While this demarcation is still in debate, yet dark data impact towards organizations has already started. Dimitrov et al. (2018) regard the impact of dark data as a risk that could harm the business enterprise. The dark data residing within the enterprise repository could disturb legal compliance with data management regulations and legislation. For example, in Malaysia, the establishment of the Personal Data Protection Act (PDPA), disallows data to be kept forever and a maximum of 7 years of duration for the data to be kept must be justified and unreasonable excuses for keeping the data would be penalized. Dimitrov et al. (2018) discussed that as dark data resides deep and rarely being used or in awareness unlike the active data currently being used, such data owners are open to cybersecurity threats and risks with the occurrences of personal data breached, and stolen data which put the business owners at a very risky position. The phenomenon of stolen data would jeopardize intellectual property rights and put the business trade secrets in a risky position and at the same time, decreasing business intelligence will be experienced whenever the unattended dark data leaked or fall into malicious actors. Due to data security breaches and stolen data, the reputation of business handlers could be tarnished, which would harm business profitability. Furthermore, if such enterprises are reluctant to invest in mining their dark data, they would miss out on opportunities for improvement, such as increased productivity, staff and consumer behaviour analysis to enhance services and profitability and avoid liabilities. At worst, a business enterprise will always be exposed to the risk of dark data if it neglects the numerous possibilities it could bring.

Akbar and Al-Mutahr & Nazeh (2018) emphasized how the dark data would impact the data quality. The presence of dark data, which is embedded within both structured and unstructured data, poses a challenge to accessing unstructured data. Unstructured data contains valuable information such as images, videos, and audio, which is not easily accessible through computer keyword processing. Unfortunately, this unstructured data contributes to the existence of dark data, which may consist of redundant, obsolete, and trivial (ROT) data (IBM, 2016). The barrier to data accessibility could compromise the accuracy of data analysis and lead to faulty data-driven decision-making.

Commvault, as an IT company, has experienced the impact of dark data and the costs associated with it (Commvault, 2014). Storing dark data internally requires a significant amount of storage capacity, resulting in higher storage management costs. The issue is exacerbated by the proliferation of file shares among employees using remote devices such as mobile phones, leading to a multiplication of dark data. Enterprises are increasingly aware of the vast amount of data they possess and seek to implement data management and analytics strategies to derive value from this data. However, the larger and more complex the data, the higher the associated costs, which could have been avoided at the early stages of business establishment. Costs further increase with the implementation of data extraction technology and the need to train staff on dark data awareness and technology utilization for managing it.

1.1 Problem Statement and Major Purpose

The phenomenon of dark data is also faced by the government. Veritas (2017) found that the main causes of dark data growth are due to a lack of strategic addressing of data issues, users treating storage systems as data "dumping grounds," and most IT strategies and budgets being meant for data storage and processing, not its value. On the other hand, Gimple and Alter (2021) and Heidorn et al. (2018) emphasized the disconnection of data creation and usage, where both should be aligned with each other instead of going sideways. This unaligned relationship contributed to the unawareness of dark data contribution resulting from

unused data. Moreover, data become completely dark when generated data is not captured (Gimple & Alter, 2021). Generated data may come with unknown potential value, but unrecorded data deletes the opportunity of leveraging the data. Furthermore, dark data accumulates due to forsaken accuracy and traceability, where data is left unclassified. The absence of data architecture leaves the dark data unattended, and the lack of awareness of data users about dark data's existence and its impact (Angelo et al., 2021). Therefore, the study aims to discover how organizations prepare themselves to manage dark data.

2.0 LITERATURE REVIEW

Research on dark data began with the establishment of Databerg by Veritas in 2015. Databerg represents the concept of big data within an organization, with data being categorized into three layers: business-critical data, ROT data, and dark data. The study surveyed 1,475 respondents from 14 countries across Europe, the Middle East, and Africa (EMEA) to understand how they manage data. It was discovered that 54% of the data in EMEA organizations was classified as dark data, 32% as redundant, obsolete, or trivial (ROT) data, and only 14% as critical business data. The accumulation of dark data was found to be influenced by three myths of data hoarding: the belief that more data means more value, that storage is free, and that all data is equal. The study recommends that Databerg can be managed by protecting critical business data, eliminating ROT data, and shedding light on dark data. The unveiling of the Databerg by Veritas in 2015 has prompted businesses to reconsider how they handle dark data and find ways to address and benefit from it. The findings revealed by Veritas have shed light on the issue of dark data and how organizations can better manage their data and its storage.

Similarly, Veritas reports discovered the dark data characteristics of dark data as what this research is partly focussing. Bhatia and Alojail (2022) stated that dark data is the accumulation of structured, semi-structured, and unstructured data from multiple sources and in different sizes. Therefore, business leaders were mentioned to be involved in the study which is similarly involved in this research. Unfortunately, the report did not completely explain the types of data being studied, the responsibility of the respondents upon dark data, and the approaches taken to collect those data either qualitatively or quantitatively. These are some of the issues which were not discussed in the report and would influence the findings of the report findings. Differently, the nature of the data collection environment is not clearly stated as the researcher assumed that respondents of the study might come from an unfamiliar nature of the business operation and influenced the findings of the data. Moreover, the data collection populations only involve 14 developed countries in the EMEA regions and the treatments and culture of data organizations might differ in different continents and regions due to economic conditions, regional culture or even social culture. These hollow characteristics of research approaches motivate the researcher to discover dark data phenomena using more specific approaches.

Dark data management (DDM) is lacking in many areas of business due to less research being done. The emergence of social media networks and the mobility of ICT devices has created the era of dark data. There were some suggestions on how to handle the dark data including encrypting, analyzing the data regularly and taking actions such as classifying legacy data. Elimination of the dark data hiding in the digital cracks and corners of the organization is an important milestone when establishing an information governance strategy that can withstand inspection by auditors and compliance officers. Even though there are lack of research that clearly defines Dark Data Management as a whole. However, Commvault (2014) and Kevin et al. (2016) mentioned Dark Data Management is a process involving monitoring storage growth by ensuring only business-valued data being stored, holistic data capture, self-

service access delivery due to a variety of data requirements, automation of Data Lifecycle, and ensuring compliance and discovery. Significantly, dark data management could reduce 33% of redundant, obsolete and trivial data or ROT data (Martin, 2016). Although publications on dark data management are scarce, Ryan (2014) proposed a way of lighting up the darkness of the data as an attempt to manage the dark data. Four stages of dealing with the dark data were proposed comprised of identification, classification of data, controls and continuous monitoring.

While dark data is part of the types of data that exist within an organization, presumably data lifecycle management is a wide subject area which can be viewed as part of the solution to the dark data phenomenon. Committee on Earth Observation Satellites (2012) has compiled 55 data lifecycle concepts and models which comprise the framework and model of data lifecycle management from the perspective of research point of view, digital data curation, and project data management. These 55 data lifecycles were analyzed by the researcher and each model possesses a concept which closely to be identified using 6 lifecycle stages of what has been listed by IBM which comprises of (1) data capture or creation; (2) data maintenance/store; (3) data usage; (4) data share; (5) data archival and, (6) data retention or deletion (Kumar & Banyal, 2020). These 55 data lifecycle principles were analyzed by the researcher and compared with IBM's six lifecycle stages (IBM, 2013), 7 lifecycle phases of Texas A&M Transportation Institute (Miller, K., Miller, M., Moran, M. & Dai B. 2018) and 6 stages of Big Data DLM (Kumar & Banyal, 2020).

Even though the researcher has analyzed a total of 68 data lifecycle management models, and compressed those into 4 generic DLM processes, some elements seem to be missing in the process of dealing with the dark data issues, data that is considered unfindable, un-interoperable, inaccurate and un-reusable (Schembera, 2019). According to the paradigm in section 2.4, approaches or methods to deal with data that is unstructured, lies deep within the repository, with unknown existence or values were found to be the great challenge to be solved by the generic DLM process. This area was also not covered by previous studies and such knowledge is required to face the existence of dark data phenomenon. The lack of acknowledgement of the management of dark data in the literature demonstrates a significant gap in the area being studied. Scarce publications on a theory that could explain the dark data phenomenon and its management urged the researcher to conduct grounded theory research about the dark data phenomenon.

3.0 METHODOLOGY

The study employed Qualitative Research to investigate the major purpose of the research. Specifically, Grounded Theory Methodology (GTM) was applied according to Corbin and Strauss's (1990) systematic grounded theory procedures to analyze the data being collected. In a grounded theory study, research is not initiated based on a theory or framework to be proved yet it studies the area and allows relevant theory to emerge (Strauss & Corbin, 1990). Figure 1 displays the GTM procedure of the study.

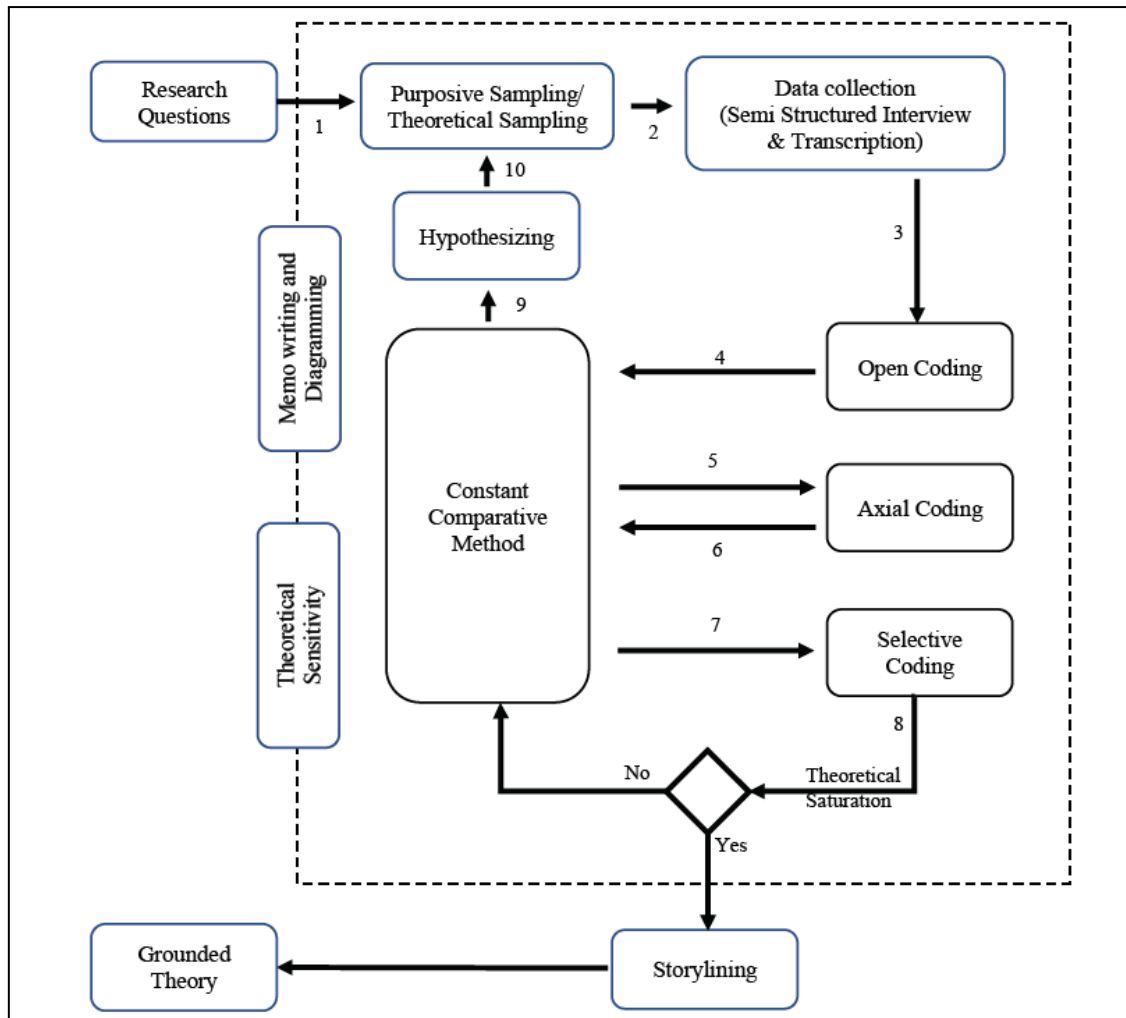


Fig. 1: Grounded Theory Methodology (Ajis, et al. 2022)

The purposive sampling method was conducted in phases whereby it involved the selection of expert sampling, snowballing sampling and inductive grounded & emergent theory sampling (Patton, 2014). Sampling was initiated with purposive sampling of a business owner of Small and Medium Enterprises in Malaysia who was chosen and regarded as the expert sample based on their experience and practices of Data Management; dominant in the Data Management procedures of the company; and involved in analyzing the company's performance. Further, the sampling uses snowballing sampling, and inductive grounded & emergent theory to collect the data. Snowballing samples were obtained from the first sample only to facilitate the inductive grounded & emergent theory sampling. Inductive grounded & emergent theory sampling also known as theoretical sampling is a procedure of sampling intended to maximize the opportunity to discover variation in concepts by densifying the emergent categories by collecting data to analyze events, not the individuals or respondents themselves which usually generalize by volumes of respondents in quantitative research. It does not predetermine who to be included in the data collection but allows the data to provide direction of whom to be sampled next, what events or phenomena to be investigated and where it took place. For this research, the theoretical sampling was sequenced in a structured table called the Theoretical Sampling Guide Matrix which was modified from Lehmann (2011). During this research, all three sizes of business operations including Microenterprises, Small Enterprises and Medium Enterprises were involved as the variation in the demographic

description supported the densification of emerging categories through range and variations in properties and dimensions.

Data collections were done through a series of semi-structured open-ended interviews towards respondents and recorded using video and audio to initiate the research and identify emerging concepts and categories as the entries towards the field of dark data. Recorded data were transcribed, analysed and coded to extract the central concept from data collected after every interview whereby constant comparative analysis was also utilized to establish the hierarchical structure of emerging categories and to find saturation of the emergent categories, before engaging with the proceeding interview. The analysis part involved three stages of coding including open coding, axial coding, and selective coding. The coding procedure was started using open coding by slicing up the data into pieces to open up the data and employing constant comparative analysis to group concepts with shared properties and identify emergent categories. Then, relationships of emerging categories were established during the axial coding phase using the coding paradigm (Strauss & Corbin, 1998) and finally, the central category of dark data management was identified in the selective coding phase.

To ensure the correct coding procedure, the researcher's theoretical sensitivity is required to be in place. Theoretical sensitivity is a pre-requisite substance that could be obtained during the pre-coding or mid-coding process to facilitate the coding process. It is also used to prevent preconception ideas from being forced onto the data. Theoretical Sensitivity obtained by the researcher through literature, professional and personal experience, and knowledge gained during the analytical process. Memo writing, diagramming and field notes were utilized in the analysis and regarded as essential substances that support the elicitation of theory in the grounded theory method in this research. The Atlas. Ti application was used to facilitate the analysis process in the management of concepts and categories along with their relationships.

In the end, storylining was used to integrate all categories and concepts by describing the phenomenon of dark data and dark data management among Malaysian SMEs to answer the research questions.

Reliability and validity of the study are also the major concerns whereby the validity of the study was established based on credibility, transferability, dependability, and confirmability (Lincoln & Guba, 1984). The credibility of the study was determined by prolonged engagement with the respondents with limited observation through virtual meetings. Other than that, triangulation and peer debriefing are also implemented to protect the credibility of the data collected and findings produced. Triangulation was executed based on data triangulation only due to the limitation of the research whereby analysed data were triangulated among different data providers which comprised of Micro, Small and Medium Enterprises. Negative case analysis was also selected and analysed to provide variation in the theory development and display the norm of research findings. In this case, few respondents represent the negative cases. Multiple mediums of data recording were used to safeguard referential adequacy for the analysis. Moreover, a few respondents were also contacted again after the analysis was completed for member checking to identify their reactions to the results of the analysis. Other than that, the transferability of the research could not be used to generalize the whole area of dark data phenomenon and its management however, the findings of the research could be tested for further research. Furthermore, the dependability of the research was also examined by observing the overlap demonstration of former research participants and whether their

experience and practices were equivalent to the latter participants. Finally, the confirmability of the research was executed to assess the accuracy of research by requiring that facts be presented in their natural state, unaltered by the inquirer's value judgments. It was accomplished by the completion of a confirmability audit, which included an audit trail and audit methodology.

4.0 FINDINGS

The data collected for the research was derived from the business owners of 18 Malaysian SMEs which comprises two major business sectors including Manufacturing and Services & Others. These business owners were considered expert samples who owned three layers of business sizes including Micro SMEs, Small SMEs, and Medium SMEs which influenced their approach to dealing with the dark data phenomenon.

Table 1 Responded Malaysian SMEs

SME	ID	Size	business activities	n	%
Manufacturing	M-Manufacturer IV5	Medium	Printing & Publishing	1	5.6
	S-Manufacturer IV8	Small	Food & Beverages	1	5.6
	Mi-Manufacturer IV14	Micro	Food & Beverages	1	5.6
Services	M-Services IV2; IV3; IV12	Medium	Wholesaling & retail trades, and communication service	3	16.7
	S-Services IV1; IV4; IV7; IV15	Small	Wholesaling & retail trades, health services, & Food & Beverages	4	22.0
	Mi-Services IV6; IV9; IV10; IV11; IV13;	Micro	Food & Beverages, Fitness Services	5	27.8
Others	S-Others IV17; IV18	Small	Agriculture	2	11.1
	Mi-Others IV16	Micro	Agriculture	1	5.6

N=18

4.1 Enabling Data Caretaking

Enabling data caretaking in Figure 4.6 is an effort to enable the strategies or interaction conditions which are the 6 stages of caretaking, experienced by business owners. The enabling mechanism acts not only as the lubricant towards the engine of the caretaking process but also realizing them to be properly executed. Consequently, enabling data caretaking practices allows business owners and data caretakers to facilitate the consequences of the data caretaking strategies. Enabling data caretaking is comprised of a few paradigms which are leveraging technology, upskilling, security and controlling.

4.1.1 Leveraging Technology

Business owners leverage technology as a way of taking advantage of technology superiority to support business activities whereby it does not replace their business operation but encourages simplification of data handling procedures. It involves simplifying the tedious manual operation process to be simplified with the assistance of technology and enhancing data discovery and retrieval by engaging in systematic data handling and also storing the data. The entire business owners were found to be involved with the use of technology in the business in the aspect of data caretaking although some were using technology extensively

in data handling while few relied on technology with limited data management function. The study found that the business owners were leveraging the technology by automating the data processing activity and optimizing technology features to gain benefits from the technology.

Computerizing Data Handling

Business owners were found to be computerizing data handling activity by employing technology to speed up and ease data caretaking whereby mobile applications and computer software were utilized to facilitate business activities. In dealing with the phenomenon of dark data, computerizing the data handling activity helps business owners control the increasing existence of dark data by ensuring data capture is safe and recorded with multiple backups, data analytics activities were able to be accomplished with a single button, and data usage with data storage can be monitored simultaneously which prevent the phenomenon of dark data to be happened. Computerizing data handling started with data capture, especially during business transactions. Previously data captured during business transactions became dark data due to data crises such as unrecorded transactions, and damaged and unreadable written data. The study found that the dark data were reduced with the intervention of business owners by capturing data automatically.

Later, the captured data were kept by storing them in a digital storage facility such as cloud-based storage technology. The study also found that most business owners were using multiple storage facilities to hold their data. Generally, the purpose of storing enterprise data is meant as a backup facility. Further, the data were also categorized automatically using customized computer programming to ease the process of metadata assignment and encourage data searching and retrieval.

These interventions were none other than to facilitate the process of data analysis which the analysis of the dark data will generate precious information which provides input to view the effectiveness of decision-making, performance measurement, or any operation being done during the business activities.

Computerized data handling provides opportunities the dark data can be handled properly. Even, the deletion of those unwanted data was scheduled automatically with the utilization of technology. As a result, trading transactions can be sped up and data verification can be done to reduce dark data quickly.

Digitizing searching

Computerized data handling has put business owners with the enormous size of data volume generated during business activity. The enormous size of the data influences the possibility of data discovery during data searching. As the business owners already experience the impact of dark data on data quality which comprises data accessibility, traceability, and audibility due to manual annoying data searching; the technology was used to digitize the searching whereby the business owners convert the previous process of searching which are done manually into digital searching by adopting technology platform to ensure data retrieval success.

The process started by implementing physical data conversion which mostly contributed to the dark data phenomenon to be converted into digital data. Retrospective data conversion was applied by some business owners to ensure the accessibility and retrievability of the converted dark data. Few business owners who have the technical skills in computer programming have set up their customized data filtration systems according to their business needs. Business owners' practices in digitizing searching have shone a light on dark data which previously unfound from being searched due to massive data volume and existed in a format that limits searchability and accessibility.

4.1.2 Optimizing Technology Features

A lot of software and computer applications have been developed in recent years, to accommodate the data caretaking processes. Sophisticated features of software and applications were optimized by the business owners by utilizing existing or customized technology features and exploiting them for the benefit of the business. It was also found that some business owners are involved in customizing the technology to fulfil desired objectives while others are only optimizing the use of existing technology to achieve business goals. Business owners optimizing the technology features based on the following:

Sequencing data

Business owners are using technology to do systematic handling of data where all the data will be handled according to a systematic sequence which was pre-set up in the software to reduce time consumption during business activity and increase transaction validity. As a result, dark data occurrences could be minimized as the data processes sequenced as scheduled operations. The process of sequencing data involved simplification of capturing customers' data during the ordering process. Furthermore, data sequencing enables business owners to reduce errors in data with the pre-setup automation of data compilation and analysis. A greater impact on the practices of sequencing data is the prevention of fabricated transactions whereby systematic data handling technology verifies the transaction on real real-time basis as the transaction is lively happening.

Systematic data management with the assistance of technology exhibits the reduction of falsified data which contributed to the phenomenon understudied. Although business owners enjoy the significant advantage of the ease of data handling processes, these advantages still contribute to the data crisis, especially the capturing of voluminous data due to the speedy data capturing.

Complementing technological difference

The study exhibits that business owners leverage the differences in technology features by complementing technology with each other. Business owners were found to be adopting multiple technology platforms due to the technological features differences since each platform features fulfilling different objectives and purposes of their business activities and complementing each other's limitations. Specific functions urged the business owners to integrate different software functions to fulfil data needs by the business activities. Certain computer applications were integrated because of their data handling function and business operation, while some were integrated for data sharing. Many business owners did not integrate the technology used for the business; however, they utilized the software by optimizing the features offered by each software to fulfil their information needs however redundant data is imminent and occurs every time a transaction happens which influences the accumulation of dark data. Complementing technological differences provides insights into how different technologies complete each other limitations. However, the disintegration of multiple data handling software would contribute to the phenomenon of dark data. Its implementation may provide advantages towards dark data management but its utilization under an uncontrolled environment would create chaotic data handling events.

Customizing data tools

Customizing data tools is a situation of business owners who developed their technology to facilitate business activity whereby specification of the technology is set up by the business owner. Innovations made by business owners were mainly done to facilitate data searching and retrieval and reduce the time consumption of manual data searching methods. Moreover, data could also be linked with the network to enable data sharing and ease the process of business monitoring.

Although customizing their data management system is considered an ideal skill for business owners to have, however, only a few business owners can customise their own data handling technology. Nonetheless, the flexibility to modify data handling tools enables data owners to develop software features to the maximum extent possible.

Minimizing physical storage

Another effort to optimise technology features is a situation of business owners keeping all data using cloud-based storage whereby it reduces the storage of physical data. However, minimal impact was found on minimizing physical storage since the majority of the subjects were still relying on physical data stored within the enterprise premises.

4.2 Upskilling

The existence of technology employment does not guarantee successful management of dark data without prior knowledge of leveraging the technological features. Upskilling refers is a situation of a business owner obtaining advanced skills which relevant in the present setting and in the future whereby the upskilling was provided by a domain expert, community of practice or acquired during business operation. In this study, the data suggested that upskilling is required to be acquired by mastering the domain knowledge.

4.2.1 Mastering Domain Knowledge

Mastering domain knowledge is a situation of business owner learning specific knowledge on specific skillsets of discipline or field whereby it is used to support the dark data management processes. Two core domain knowledge indicated by the data were learning ICT knowledge and data skill knowledge.

Learning ICT

The data suggest that business owners are compulsory to have ICT knowledge in the present era of technology as technology helps a lot in running business activities especially when it comes to data management. ICT knowledge not only was told to support business growth but also become an essential skill that must be possessed by any business owners. Some business owners were found to be learning ICT on their own while some got the knowledge by reading and taking formal classes. The ICT knowledge is believed to be the foundation of dark data management. Therefore, the data indicated that ICT knowledge is required to enable business owners and data caretakers to be able to handle dark data properly.

Acquiring Data Skills

Another domain of knowledge to be mastered as indicated by the findings is acquiring data skills. It is a situation of the owner acquires knowledge of dark data handling and caretaking whereby is sourced from formal training or gained through personal experience. The majority of the business owners were found to be learning the data skills through their own experience of fixing data handling mistakes during business activities or getting inspiration from others' successful encounters with dark data. Only few engaged in formal training and learn from coaches. Without any knowledge of data skills, the data revealed that the occurrences of dark data massively happened and attracted unnecessary risk to the business such as losing valuable data, unjustified decision making and decreasing profitability survivability.

4.3 Security

Security of the data dealing with the safety protection either preventing data from being lost, stolen or unusable. Security of the data is determined by the safeguarding effort planned by the business owners.

4.3.1 Safeguarding

Safeguarding refers to an effort to ensure the protection of data-keeping safety. The study found that the business owners were safeguarding the data by limiting access to data, gradually assessing the protection and planning for data disaster intervention. Although these practices are as important only a few business owners embarked on the safeguarding effort.

Limiting access

Limiting access is the practice of forbidding unauthorized access to enterprise data. The study suggested that limiting access can be implemented by hardening security features including encrypting data storage, limiting access to staff only, and restricting access including trade secret data. Consequently, limiting access enhances data security by preventing access by unauthorized parties and some were limiting access only to business owners.

Assessing protection

Schedule protection assessments were also found to be the routine of ensuring data safety. Major purpose of assessing protection is to prevent data from lost or inaccessible. The business owners execute the assessment activities by determining data platform healthiness by practising data risk assessment, assessing the strength of security protection by hiring hackers and preparing data recovery plans. Thus, the finding indicated that enabling data protection is not sufficient to protect the security of dark data without assessing the strength of the data protection.

Data disaster planning.

Another security feature that business owners employ is data disaster planning. Data disaster would expose data to be damaged, inaccessible, and unusable. Therefore, the data exhibit that data disaster planning could prevent data access failure and loss of data. The data disaster plans were found to help recover missing or lost data by executing automatic data backup and the use of printed data recording in the event of system failure.

4.4 Controlling

The findings indicate controlling as one of the guidance mechanisms whereby it governs the processes of data caretaking. Controlling the caretaking was found to be executed by structuring policy on dark data management and harnessing integrity among workers to develop trustworthiness among workers.

4.4.1 Structuring Policy

Structuring policy is the process of business owners constructing standard operating procedures to prevent any data threats such as human errors, fraudulent business transactions, and data loss. Structuring policies was done by business owners by updating procedures.

Updating procedures

Procedures of data handling were found to be updated regularly by business owners to prevent any careless data handling and prevent data errors from emerging. Indirectly the updated procedures not only prevent dark data phenomenon but also improve business operations to be more structured and minimise dark data derived from human error.

Therefore, the findings indicate that the business owners update their standard operating procedure (S.O.P) not only about the business operation but also dark data management. Generally, most of the business owners did not possess any data handling S.O.P. before they experienced the impact of dark data. Such impact of the dark data made them realize that data handling is crucial and needs to be properly acknowledged. Consequently, structuring

policy influences the business owners to execute the strategies accordingly which results in proper management of the dark data.

Harnessing integrity

Harnessing integrity is also one of the core controlling attributes in the enabling caretaking effort. It is the practice of being honest and demonstrating a persistent and unwavering devotion to strong moral and ethical ideals and values, regardless of the circumstances. Updating the procedure would assist the process of harnessing integrity among data caretakers by instilling an integrity culture.

Integrity Culture

Integrity culture is a practice of believing that honesty is the approach that guaranteed data handling procedure is complying by the data caretakers. Majority of the business owners practice an integrity culture to avoid undesirable trouble caused by dishonesty. Some business owners use a religious approach to maintain an integrity culture in the company which believes in fostering business sustainability, protecting data and handling integrity.

Other than the religious aspects in promoting integrity culture, putting the trust of staff by the business owners also harnesses the integrity culture among workers. As a result, data handling integrity protects business survival and influence profitability. In a nutshell, integrity culture not only protects the data handling integrity but the company as a whole. As integrity culture is embraced by business owners and data caretakers, caretaking initiatives can be successfully implemented, and dark data occurrences can be suppressed effectively.

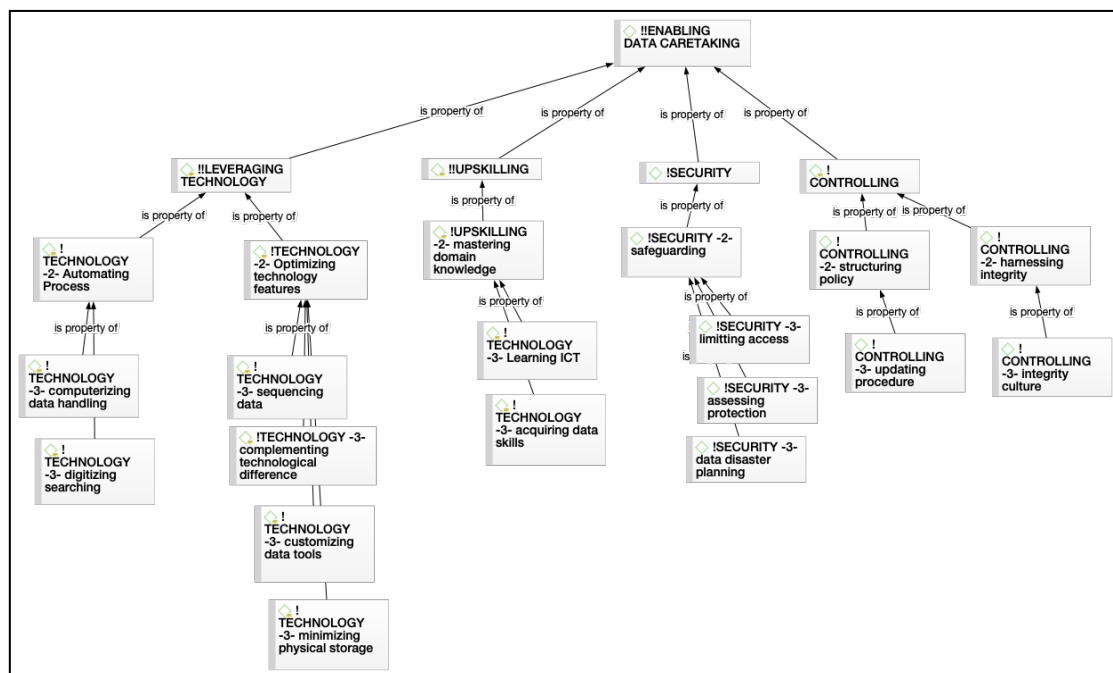


Fig. 2: Model of Dark Data Caretaking Enablers

5.0 DISCUSSION & CONCLUSION

Enabling data caretaking involved a few attributes including technology, upskilling, security and control. Technology was used to automate the data processing operation and simplify the manual process (Akbar et al., 2018). Furthermore, upskilling ICT knowledge and data skills were suggested by the data to be a prerequisite and mastered by the business owners to intervene the dark data management. While dark data possessed precious value to the enterprise, protection of the data safety was found compulsory as it would prevent unnecessary risk to the company (Commvault, 2014; Dimitrov, et al. 2018). Each of the

attributes was then monitored and administered by the controlling effort using standard operation procedures and integrity culture (Al-Sai & Abualigah, 2017). Although the enabling caretaking effort facilitates the data caretaking activities simultaneously it influences the occurrences of data crises whenever the drawbacks happen such as failing technology, lack of knowledge on required skills, breached security protection and non-compliance with the controlling mechanism.

ACKNOWLEDGEMENTS

This research publication was funded by the Universiti Teknologi MARA Cawangan Johor under the fund of Geran Bestari Fasa 1/2023.

REFERENCES

- Ahmed, W., & Ameen, K. (2017). Defining big data and measuring its associated trends in the field of information and library management. *Library Hi Tech News*, 9, 21-24.
- Baragan, S. P. (2020). Appraisal and retention of information in the private sector: a case study. [Doctoral Dissertation].
https://eprints.qut.edu.au/199783/1/Salvador_Barragan_Thesis.pdf
- Bhatia, S., & Alojail, M. (2022). A Novel Approach for Deciphering Big Data Value Using Dark Data. *Intelligent Automation & Soft Computing*, 33(2).
- Birks, M., & Mills, J. (2014). *Grounded theory: A practical guide* (2nd ed.). Sage.
- BNM (2013). *Circular on new definition of Small and Medium Enterprises (SMEs)*.
https://www.bnm.gov.my/documents/20124/761700/Appendix1-Circular_on_Definitions+_for_SMEs.pdf.
- Charmaz, K. (2008). *Constructing grounded theory* (2nd ed.). Sage Publications.
- Commvault (2014). *5 ways to illuminate your dark data*. Commvault Systems.
- Corallo, A., Crespino A. M., Vecchio, V. D., Lazoi, M. & Marra, M. (2021). Understanding and defining dark data for the manufacturing industry. *IEEE Transactions on Engineering Management*.
- Delve. (2010). *The essential guide to coding qualitative data*. <https://delvetool.com/guide>
- Dimitrov, W., Siarova, S., & Petkova, L. (2018). *Types of dark data and hidden cybersecurity risks*. doi:10.13140/RG.2.2.31695.43681.
- Erik J. Martin, (2016). *Dark data: Analyzing unused and ignored information*. econtentMag.com.
- Gharehchopogh, F.S., & Khalifelu, Z.A. (2011). *Analysis and evaluation of unstructured data: text mining versus natural language processing*. 5th International Conference on Application of Information and Communication Technologies (AICT).
- Gimpel, G. (2020). Dark data: The invisible resource that can drive performance now. *Journal of Business Strategy*.
- Gimpel, G. & Alter, A. (2021). Benefit from the internet of things right now by accessing dark data. *IT Professional*, 23(2), 45-49.
- Glaser, B. G. & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for Qualitative Research*. Aldine.
- Glass, R., & Callahan, S. (2014). *The big data-driven business: How to use big data to win customers, beat competitors, and boost profits*. Wiley.
- Guetat, S., & Dakhli, S. (2015). The architecture facet of information governance: The Case of urbanized information systems. *Procedia Computer Science*, 64, 1088-1098.
- Hand, D. J., (2020). *Dark data: Why what you don't know matters*. Princeton University Press.
- Highquest Solution (2016). *Dark data making your organisation data-enabled?*.
<https://doczz.net/doc/8987800/white-paper-dark-data-making-your-organisation-data>
- Hitachi (2013). *Big Data - Shining the light on enterprise dark data (EDD)*.
<https://www.hitachivantara.com/en-us/resources.html>.

- IBM (2013). *The fundamentals of data lifecycle management in the era of big data: How data lifecycle management complements a big data strategy*. <http://hosteddocs.ittoolbox.com/TheFundamentals.PDF>.
- Intel (2018). *Datumize and Intel transform dark data into operational insight for manufacturing and logistics*. <https://www.intel.sg/content/dam/www/public/us/en/documents/solution-briefs/atomize-dark-data-in-manufacturing-and-logistics-solution-brief.pdf>.
- Kantardzic, M. (2020). *Data mining: Concepts, models, methods, and algorithms*. New Jersey.
- Kumar, K & Banyal R. K. (2020). Data life cycle management in big data analytics. *Procedia Computer Science*. 173, 364–371.
- Liu, Y., et al. (2019). *A framework for image dark data assessment*. Doi:10.1007/978-3-030-26072-9_1.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: a revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
- Miller, K., Miller, M., Moran, M. & Dai B. (2018). *Data management life cycle: Final report*. Texas, Texas A&M Transportation Institute.
- Schembera, B. & Duran, J. M. (2020). Dark data as the new challenge for big data science and the introduction of the scientific data officer. *Philosophy & Technology*, 33, 93–115.
- Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.
- Urquhart, C. (2013). *Grounded theory for qualitative research: A practical guide*. Sage Publications.