UNIVERSITY OF KWAZULU-NATAL

The influence of big data on monitoring the factual quality of digital media in Southern Africa

Andile Dlomo

Student No: 210533333

A thesis submitted in fulfilment of the requirements for the degree of
Master of Commerce

School of Management, Information Technology, and Governance

College of Law and Management Studies

Supervisor: Dr Rosemary Quilling

Thesis

2022

Acknowledgement

I would like to express my deep gratitude to my primary supervisor, Rosemary Quilling, for her patient guidance, enthusiastic encouragement, and useful critiques of this research study. I would also like to thank Richard Gevers CEO of Open Cities Lab and the entire team who supported the study.

Finally, I wish to thank my family and friends for their support and encouragement throughout my study.

Declaration

I Andile Dlomo declare that

(i) The research reported in this dissertation/thesis, except where otherwise indicated, is my original research.

(ii) This dissertation/thesis has not been submitted for any degree or examination at any other university.

(iii) This dissertation/thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This dissertation/thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then: a) their words have been re-written, but the general information attributed to them has been referenced; b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) Where I have reproduced a publication of which I am author, co-author, or editor, I have indicated in detail which part of the publication was written by myself alone and have fully referenced such publications.

(vi) This dissertation/thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation/thesis and in the References section.

Signed: 

Abstract

This research study will explore how big data can drive innovation in response to dynamic change and aid society in establishing an advantage when fact-checking/monitoring new media and dealing with false information. The study emphasises that big data might answer questions and offer insights society never had access to before. In the current news media environment, the services that enable the sharing and production of large amounts of data are not sufficient to combat increasing fake news, ongoing public mistrust, and false, partisan media content for capital gains from gaining more influence in society. There is an urgent need for intervention, which big data innovation can provide. There are, however, some myths regarding the use of big data that need to be dispelled, such as the idea that an analysis of the data will ensure transparency and reliable content distribution from the developers of big data systems to the audience consuming the data. Innovating and obtaining an advantage from data is more complex than just collecting lots of data; a look at the impact big data will have on a society is vital in leveraging big data. The study explores this notion by looking at the Digital Data Genesis Capability Model. The model guides the structure and how the case study will be conducted in the media fact-checking sector. The development of the big data initiative is built on fundamental expertise. According to the findings, highly skilled employees with knowledge of both proprietary and open-source tools are essential in the development of big data systems. Furthermore, there is a high level of compatibility with the existing web environment standard and the tools being used when deploying a big data system in the web. As a result, development of a big data initiative by a technology focused organisation is only limited by their ability to implement an effective big data workflow. However, this requires detailed planning, cloud computing for hardware; software; outsourced third party services; the work on data structure built in-house; and the use of docker containers that enable mobility in the development process and the adoption of new technology when implementing the searching and querying of large datasets and streams. There was a deviation from the existing model noted. The context of the study exposed that it is possible to implement big data initiatives among more than one company as a partnership, if the companies share some business traits or the same philosophy: thus, changing the dynamic of routines and responsibility in the existing landscape.

Table of Contents

Acknowledgement i
Declaration ii
Abstract iii
List of Tables viii
Definition of Terms ix
Chapter 1: Introduction 1
1.1 Background and Context 1
1.2 Problem Statement 3
1.3 Research Questions 4
1.4 Research Objectives 5
1.5 Rationale 6
1.6 Overview of Methodology 6
1.7 Limitations 6
1.8 Summary of Chapters 7
Chapter 2: Literature Review 9
2.1 Introduction 9
2.2 The Increasing Problem of Misinformation 9
2.3 Social Media and the Spread of Misinformation 11
2.4 Combating Fake News 12
2.5 Big data 13
2.6 Big Data Analytics (BDA) Capability Impact on Formal Media 14
2.7 Big Data Analytics (BDA) Capability Technical Aspects 15
2.8 Big Data Analytics Management 17
2.8.1 Big Data Quality 19
2.8.2 Big Data Analytic Visualisation 19
2.8.3 Big Data Privacy 20
2.8.4 Information Security 21
2.9 System Performance: The outcome achieved by the big data system 22
2.10 The Academic Research into Big Data 23
2.11 Conceptual Framework 24
2.11.1 The Process of a Digital Data Genesis Generation Capability 25
2.11.2 The Digital Knowledge Innovation Hierarchy 26
2.12 Conclusion 27
Chapter 3: Research Methodology 29
3.1 Introduction 29
3.2 Research Philosophy 29

3.3 Research Design 29
3.4 Research Methodology 30
3.5 Sampling method 32
3.5.1 Research Site 33
3.5.2 Selection of the Project: Dexter 33
3.5.3 The Dexter platform 33
3.5.4 Ethical Clearance 34
3.6 Data Collection 34
3.7 Data Capturing and Editing 37
3.8 Data Analysis 37
3.9 Data Quality Control 38
3.9.1 Credibility 38
3.9.2 Transferability 39
3.9.3 Confirmability 39
3.9.4 Construct Validity 39
3.10 Limitations of the Study 39
3.11 Conclusion 40
Chapter 4: Qualitative Data Analysis 41
4.1 Introduction 41
4.2 Case Study 41
4.2.1 Participant Responsibilities 42
4.3 Stage 1: Digital Data Genesis Generation Capability 43
4.3.1 Choosing Information Technology 49
4.3.2 Integration of Information Technology 50
4.3.3 Management of Big Data 51
4.3.4 Reconfigurability of the Platform 53
4.4 Stage 2: The Output of a Digital Data Genesis Capability 54
4.4.1 Accessibility 56
4.4.2 Accuracy 58
4.4.3 Completeness 59
4.4.4 Currency 60
4.4.5 Credibility 62
4.4.6 Trustworthiness 62
4.5 Stage 3: Strategic Decision-making 62
4.6 Outcome and Firm Advantage: Factual quality of digital media 65
4.7 Conclusion 66

Chapter 5: Analysis and Synthesis 67
5.1 Introduction 67
5.2 The Dexter Project's Big Data Capability (back-end development) 67
5.2.1 Choosing IT 68
5.2.2 Integration 69
5.2.3 Management of Big Data 70
5.2.4 Reconfigurability of the Dexter Platform 72
5.3 The Link between the Big Data System and the Strategy used to Monitor the Data 72
5.4 The Big Data System Implementation Strategy applied by the Technology Support (Dexter) 75
5.5 Determine the Dexter Project's Big Data Capability Influence on the Performance of Fact-checking/Monitoring the Factual Quality of Digital Media. 77
5.6 Revision of the Prescott Model Based on the Dexter Case Study 78
5.7 Conclusion 79
Chapter 6: Conclusion 80
6.1 Introduction 80
6.2 Concluding Remarks per Research Question 80
6.3 Limitations 84
6.4 Future Study 84
6.5 Significance of the study 85
6.6 Conclusion 85
References 87
Appendix A: Ethical Clearance Approval 95
Appendix B: Interview Schedule for Open Cities Lab team 96
Appendix C: Clarification Interview Schedule 99
Appendix D: Alignment Matrix 101

List of Figures

FIGURE 1: DIGITAL DATA GENESIS MODEL 25
FIGURE 2: CASE STUDY PROCESS 31
FIGURE 3: DEXTER DATA FLOW 44
FIGURE 4: DDGC DEXTER CLOUD-BASED PLATFORM. 47
FIGURE 5: DDGC STAGE 1 - DEVELOPING THE CAPABILITY. 48
FIGURE 6: DDGC STAGE 2 - THE OUTPUT PROVIDED. 55
FIGURE 7: DEXTER REPORT SCREEN 61
FIGURE 8:DDGC STAGE 3 - DECISION-MAKING AND THE OUTCOME. 63
FIGURE 9: DDGC STAGE 1 DEXTER 68
FIGURE 10: DDGC STAGE 2 DEXTER 73
FIGURE 11: DDGC STAGE 3 DEXTER. 78

List of Tables

TABLE 1: OPEN CITIES LAB. 30
TABLE 2: SUMMARY OF PARTICIPANTS. 42
TABLE 3: BACK-END TECHNOLOGY USED. 45
TABLE 4: FRONT-END TECHNOLOGY USED. 55

Definition of Terms

Cloud computing - Cloud computing is on-demand access to computing resources—applications, servers (physical and virtual), data storage, development tools, networking capabilities, and more—hosted at a remote data centre controlled by a cloud services provider and accessible via the internet (or CSP) (Vennam, 2020).

Dexter Project - The Dexter project aims to use modern computational paradigms to monitor and analyse digital media in Southern Africa (Lab, 2021).

Digital Data Genesis Generation Capability – This is built from the organisational process of learning, integrating and co-ordinating. It takes into account the path-dependent development of IT within the organisation so that boundary conditions are understood when creating the Digital Knowledge Strategy (Prescott, 2016).

Digital Knowledge Strategy – This concerns the alignment of the data, information and knowledge that will be needed to support the business strategy (Prescott, 2016).

Fake news is described as content that is distributed with the intent to deceive people, regardless of the motivation (Nakov & Da San Martino, 2020).

Information capability is "the ability to provide data and information to users with the appropriate levels of accuracy, timeliness, reliability, security, and confidentiality; provide universal connectivity and access with adequate reach and range; and tailor the infrastructure to emerging business needs and directions" (Jaca et al., 2016, pp. 69-70).

IT capability - This is the firm's capability to detect the IT that is needed to meet business needs; to launch and integrate IT; to improve business processes cost-effectively; and to provide long-term maintenance and support for IT-based systems (Karimi et al., 2007).

Chapter 1: Introduction

1.1 Background and Context

A recent statistic shows that 4.66 billion people, as of October 2020, are active internet users. This is 59% of the global population; and of that total, 92.6% (4,32 billion) are smartphone internet users, making smartphones the most important channel for internet access worldwide (Johnson, 2021). Furthermore, according to the statistics the most-used applications by mobile users worldwide are WhatsApp, YouTube, Facebook, and Twitter. Thus, it is important to ask questions about the information management and sharing that takes place in this large- audience that varies in terms of demographics, culture, and economic living standard. Ever since 2001, there has been a substantial increase in active fact-checking initiatives all over the world (about 349 as of today), as published by Duke Reporter Lab (Stencel & Luther, 2021).

In a recent article looking at the global Covid-19 disease, the high levels of information-sharing lead to information overload in societies who consumed the content presented by different media organizations. Different media organizations were saying different things at the same time with some posing contradictory arguments challenging the legitimacy of the content presented on a daily basis (Menczer & Hills, 2020). This led to cognitive bias where the audiences consuming the different arguments became convinced that only their point of view was correct also referred to as “our in-group”. The level of trustworthiness of media organizations dropped, since no single media organization could completely dismiss challenges related to content legitimacy (Menczer & Hills, 2020). As an example: These cognitive vulnerabilities and the actions of bots and algorithms exacerbated these vulnerabilities and resulted in people believing Covid-19 was a hoax. Covid-19 presented entirely different challenges for fact-checking organizations since little was known about the disease and that which was known was being challenged.

Consequently, this posed a problem to governments around the world who depend on the media as an instrument that allows citizens to engage with politicians for service delivery, holding them accountable and responsible for their decision-making on policies and laws that govern them. This raises the question: If another pandemic were to happen, would media organizations overcome the inadequacy displayed during COVID-19, to respond and provide valuable, true and fair representation of media content to the world. The emergence of big data processing and analytics technologies allow for new innovations and calls for society to ask questions of its implication in the media space (Mayer-Schönberger & Cukier, 2013). As

industries recognize the value of digital data, we begin to understand society and social interaction in ways we could never have imagined before (GetSmarter, 2018). A definition by early scholars Ward and Barker (2013) is as follows:

> "Big data is a term describing the storage and analysis of large, and/or complex, data sets using a series of techniques including, but not limited to, NoSQL, MapReduce, and machine learning." (Ward & Barker, 2013, p. 2)

In early studies on factors that contribute to innovation in a market, Barney (1991) found that to gain a competitive business advantage within a market, capital resources, human resources and organisational resources must be made known in all key areas of revenue creation. Since then, rapid advances in all these resource areas have improved the possibility of taking innovative action. This research study will explore how big data can drive innovation in response to dynamic change and aid society in establishing an advantage when fact-checking/monitoring new media and dealing with false information. With advances in bandwidth speed and IT tools and software technologies in the consumer market, the type of data being uploaded has increased substantially. The data, collectively known as digital data, consists of images; text; GPS locations; RFID; meta-data; event logs; product view sites; social media sites; and so on (Lyko et al., 2016). Laney (2001) refers to big data as a significant amount of data that is continually being gathered, stored, and managed, as well as the changing IT technologies that enable this and the analytic tools used to comprehend the data.

Furthermore, in some instances, the data arrives in a variety of formats that do not integrate appropriately, or not at all, into most-used relational databases. The high streaming (real-time) of data has received a lot of attention in the consumer sector, more so than in any other sector. Hence, different IT architectures may be required (Prescott, 2016). Therefore, the size of the data sets, the heterogeneity of the data, and the speed of accumulation of the data make the analysis more complex. It is the joining together of the data from a variety of sources that gives big data analytic users a better, and more complete, understanding of their different audiences. Moreover, with the continued increase in the establishment of active fact-checking initiatives, an investigation should also be conducted in a developing country to determine the context of fact-checking and prevention in a developing economy and democracy. Prescott (2016) studied the use of big data analytics, looking at a single case study in the media and advertising industry, exploring a big data platform by A.C. Nielson on consumer data on television viewing and purchasing behaviour. This research project will expand on the Prescott study by looking at big data platforms in the monitoring and fact-checking sectors in Africa that are not limited by profit targets as the return of investment for the business, because the target organisation is a non-profit organisation (NPO). Since the platform adopted for the case

study is used to inform the public of the factual quality of digital media, it will allow for an in-depth study of the big data phenomenon.

Prescott (2016) emphasises that big data might answer questions and offer insights that media fact-checking institutions never had access to before. However, that alone cannot constitute a competitive advantage. For a firm to have a competitive advantage, it is important to increase its understanding of its big data environment. It needs to resolve the potential challenges that arise between the practical aspects of data collection and the development, and setting, of strategies for big data use (Prescott, 2016). In most cases, if both ideas are not understood it will lead to data overload. Innovating and gaining advantages from data is more complex than just collecting lots of data; a look at the impact big data will have on the audience is vital in leveraging big data (Prescott, 2014). As of January 2021, Twitter introduced 'Birdwatch', a new feature in Twitter to tackle misinformation on the platform by embracing the 'wisdom of crowds'. The Birdwatch feature is based on communities and aims to identify false tweets circulating on Twitter (Pröllochs, 2021). On Birdwatch, Twitter users can pinpoint tweets they feel are misleading and give feedback and context to the tweet. This also includes a rating mechanism on the feedback and notes users can add. Prescott emphasises that the findings and the model developed from the study (the Digital Data Genesis Model) should be used as a basis for research in other sectors to get a better understanding of "Digital Data Genesis Capability functions as a capability/dynamic capability, and its role in business strategy and competitive advantage" (Prescott, 2016, p. 109).

1.2 Problem Statement

In the current news media environment, the services that enable the sharing and production of large amounts of data are not sufficient to combat increasing fake news, ongoing public mistrust, and false, partisan media content for capital gains from gaining more influence in society (Ünal & Çiçeklioğlu, 2019). There is an urgent need for intervention, which big data innovation can provide. There are, however, some myths regarding the use of big data that need to be dispelled, such as the idea that an analysis of the data will ensure transparency and reliable content distribution from the developers of big data systems to the audience consuming the data (Prescott, 2016). Hundreds of websites have appeared on the internet that appear credible at face value, but are fake (Silverman, 2016).

African developing countries have attempted to combat misinformation by establishing fact-checking journalism projects. Africa Check is the name of one such project and is supported by the international news agency "Agence France-Presse" (AFP) and the Journalism Department of the University of the

Witwatersrand in Johannesburg, South Africa (Cunliffe-Jones, 2020). Africa Check runs in four countries – South Africa, Nigeria, Kenya and Senegal. The project has trained 2500 newsroom journalists in fact-checking. The project showcases a holistic approach to fact-checking, by focusing on both identifying misinformation and working with journalists, media agencies, and a media platform that aims to combat the spread of misinformation. The project initiative proves that substantial development in fact-checking has emerged in the sub-Saharan Africa region; however, not much is published about the other regions of Africa.

In South Africa, a non-profit organization developed a big data project called Dexter in 2014 with the aim of using modern computational paradigms to monitor and analyse digital media in Southern Africa (Lab, 2021). As industries recognize the value of digital data, we begin to understand society and social interaction in ways we could never have imagined before (GetSmarter, 2018). According to Kar and Dwivedi (2020), supporting studies on big data generally focus on defining the observed phenomena, but lack solid theoretical contributions. Prescott (2016) proposes a model called the Digital Data Genesis Capability to investigate the role of big data analytics in assisting firms in maximizing the benefits of big data for producing innovative, flexible, and responsive solutions. The study findings demonstrate that the capability obtained from the analysis of the big data also acts as a dynamic capability; thus helping to counteract the possible limiting effect of economic indicators, and positively contributing to the overall outcome. Prescott (2016) study model has yet to be fully tested but offers a theoretical contribution to the phenomenon of working with big data for the benefit of people and society. This study will explore the model further, looking at a media monitoring platform (Dexter) in the media industry in the Southern African region.

1.3 Research Questions

The study will explore the following question:

How does big data analytics capability influence the monitoring of the factual quality of digital media in Southern Africa?

To study the whole process, from the point of collecting the data to achieving the desired outcome, requires numerous stages, according to the digital data genesis model. The first stage looks to investigate the initial architecture decisions that form a foundation for the big data platform, including issues around data warehousing; data security; the levels of human intervention; user modelling; algorithms; machine learning; and so forth. The second stage focuses on the quality dimensions that define the output produced and the process and procedures that make up the big data platform to enable the user to analyse the data. Finally,

the process ends with an investigation of the performance of the platform in stage three, i.e., to what extent has it achieved the required outcome? This process can be deconstructed into a series of steps that will be traced during the investigation. The nature of the digital data genesis generation capability will be considered; followed by the digital knowledge strategy employed by Open Cities Lab and their client media company. The output produced by the platform that is designed to drive the strategy for decision-making enhancement is followed by the monitoring of the performance of the project, based on the outcome of the decisions taken. This gives media governing agencies a service advantage in Southern Africa in ensuring equal, ethical, and fair media transparency.

Based on this theoretical framing, the research question has been broken down into three sub-questions as listed below:

1. How does the Dexter project's big data capability influence the big data strategy used to safeguard the factual quality of digital media?
2. How does the Dexter project produce the output used by strategy makers to safeguard the factual quality of digital media?
3. What is the assessed performance of the Dexter project, in terms of safeguarding the factual quality of digital media?

1.4 Research Objectives

The following research objectives aim to address the above research questions. The objectives are aligned with the research sub-questions, as follows:

- ➢ Determine the Dexter project's big data capability.
- ➢ Determine the Dexter project's big data strategy, implemented and used as input for monitoring the factual quality of digital media.
- ➢ Determine the relationship between the Dexter project's big data capability and the digital knowledge strategy, used for monitoring the factual quality of digital media.
- ➢ Determine the performance of the Dexter project in monitoring the factual quality of digital media.
- ➢ Determine the Dexter project's big data capability influence on the performance of fact-checking/monitoring the factual quality of digital media.

1.5 Rationale

The thesis is expected to continue the work of Prescott (2016) by further exploring the Digital Knowledge Generation Capability Model he constructed for understanding big data adoption. This study applies the model in a different sector, that of digital media. This study will be like the Prescott (2016) study, in that it uses the same model as a theoretical starting point for the research. However, it will be conducted in South Africa and in a different sector and takes a more detailed look into the model stages and each component aspect. This includes the specific technological and theoretical choices at each step of the design, development, and implementation.

The sector selected for the study is faced with challenges of misinformation and fake news; and no proven solution has been discovered to combat these problems. Since big data (Torabi Asr & Taboada, 2019) has been proposed as a possible solution to solving the spread of misinformation in the existing literature, it is important to investigate big data in digital news media checking, and provides a good opportunity to learn about both cases.

1.6 Overview of Methodology

The research design chosen is a case study approach, since the study meets the criteria for an exploratory case study: The sample size is small. It is focused on a study of one platform and decisions taken in the development and use of the platform. This study used qualitative research as a research strategy and interviews as the research instrument. The interview questions were developed using the theoretical framework the study is using – the Digital Data Genesis Capability Model. Data production was achieved by using in-depth interviews with five participants. The text data was analysed by using thematic analysis. The big data platform is called Dexter project that aims to use modern computational paradigms to monitor and analyse digital media in Southern Africa (Lab, 2021).

1.7 Limitations

The study looks to explore big data in a media monitoring sector that came to the fore in the early 2000s and has been growing substantially since then. The study will investigate a platform used to help combat the issue of misinformation, and the measures implemented to curb the digital epidemic caused by advances on the internet, digital media and social media around the world. The scale of the project is big from an IS discipline perspective, as it aims to consider the full processes involved in developing and assessing the

platform outcomes. Therefore, the researcher will rely on a theoretical framework used by the previous researcher. However, the theoretical framework is new and may require empirical proof of other variables that may need to be considered for inclusion. Even so, it does capture the essence of what is being investigated.

It is important to note that accessing data of this kind was difficult, since big data platforms, when implemented correctly, provide a competitive, or business, advantage to the institution using the platform. Some businesses approached early in the research process felt that sharing any information would compromise the company's financial standing and chose to opt out of contributing to the academic study. The choice of the research site was thus largely based on their willingness to participate; and it is an organisation with a small staff complement. This means the number of people available to be interviewed, and the participants available to cross-check information provided, was also limited. Also, important to mention that the company that employs the big data technology, referred to as in the study "Partner A," is not included in the research study. However, it will include perspectives and insights from the NPO technological support, developers, and data scientists who constructed and maintain the system on Partner A's behalf. The development team at Open Cities Lab also assists Partner A with strategy development when determining which projects to use the Dexter platform for and how to best leverage the system to achieve exceptional results from start to finish.

1.8 Summary of Chapters

Chapter 1: Introduces the study, discussing the research problem and background to the phenomenon the study is investigating. It outlines the research question and sub-questions. Furthermore, it defines the objectives of the study, introduces the research methodology used, and discusses the limitations of the case study.

Chapter 2: The literature review is conducted by looking at studies on big data analytics around the topic of fact-checking and media, as well as aspects involving the working and implementation of big data and the different components that constitute a big data system. After that, a discussion of the theoretical framework used to conduct the study and how it links to big data analytics follows.

Chapter 3: The research methodology is discussed in detail in this section. The design of the case study and the reason why it was implemented using a qualitative approach are addressed. The research site is showcased, and details of the sample size and research instrument, linked to the theoretical framework, are

considered. The data collection process followed by the researcher is discussed, looking at tools and data control quality procedures used to handle data and analyse the data. Finally, the ethical considerations and limitations of the study are outlined.

Chapter 4: The case study is presented for data analysis in Chapter Four. In this chapter, insight gathered through the interview process is outlined and presented, looking at the big data system in relation to the Digital Data Genesis Model and the academic literature on big data development.

Chapter 5: A discussion on the case study findings is presented. This discussion is structured to address each of the objectives as outlined in Chapter One. The discussion entails looking at the different variables presented by the theoretical framework in stages 1, 2, 3, and the knowledge that exists on them with regards to big data analytics and media fact-checking. The chapter concludes by considering the over-arching question, and answers obtained, by using Prescott's model as a theoretical framing. The aspects successfully covered by the existing model are discussed. In addition, those insights gained, and modifications required, that were not highlighted by the model are presented and suggested as a modification to the existing model.

Chapter 6: In Chapter Six, the research question and sub-questions are revisited, with concluding statements highlighting the outcomes of the research study. Finally, the researcher outlines the limitations and prospects for further study.

Chapter 2: Literature Review

2.1 Introduction

The literature review was conducted through a critical review of recent studies and online research libraries. The material was gathered from online research libraries of a number databases from the following sites: https://scholar.google.com/; UKZN https://library.ukzn.ac.za/; and digital library databases: Springer; Science Direct; IEEE Xplore; and ACM library. The increased connectivity in the world, and news media technology, has led to a rise in misinformation in both developed and developing countries. The first part of the chapter explores the literature on this, with reference to big data and media fact-checking. Literature on open-source architecture for big data analytics is reviewed, as well as changes caused by big data analytics in the media world. In conclusion, the conceptual framework used in the study to unpack big data analytics is discussed.

2.2 The Increasing Problem of Misinformation

Misinformation is false information that is spread, regardless of whether there is intent to mislead (Torabi Asr & Taboada, 2019, pp. 3-4). The case study focuses on exploring the media fact-checking space by looking at the factual quality of digital media when big data analytics is applied. The study looks at how the large amount of data generated by the media outlets can be reviewed, for example, to control the integrity of the data being published. Understanding 'misinformation' is thus an integral part of this study. Common approaches for dealing with misinformation vary and depend on the level of technology adoption in the society, starting with the most basic approach: educating the public, analysing and curtailing the spread, manual checking, and lastly automatic checking (Torabi Asr & Taboada, 2019, pp. 3-4). The widespread dissemination of misinformation has a serious implication for society at large.

A study using data gathered from Hurricane Harvey (King & Wang, 2021) - a category 4 hurricane that hit Texas and Louisiana in August 2017, inflicting catastrophic floods and killing over 100 people - examines the factors that impact the virality of authentic news and misinformation on Twitter. By using predictive analytics, the study was able to offer findings on various factors that impact the virality of news: namely, that novelty of information; negative-toned news; and news with limited characters like twitter tweets spread more on social media, which affects the diffusion of information on the Twitter network during a crisis event. Through a combination of text mining, machine learning and econometrics models, the researchers were able to discover hidden patterns from the identified factors that drive the virality of real

news and misinformation on social media. For instance, authentic news that is uplifting may hold some novelty and be spread more amid shock occurrences, when there is already an overabundance of negative news reports (King & Wang, 2021). This agrees with Kar and Dwivedi (2020) findings, which suggest a need to build big data-driven theory research. According to Kar and Dwivedi (2020), studies on big data generally focus on defining the observed phenomena and lack solid theoretical contributions. A study by Prescott (2016) offers a model called the Digital Data Genesis Generation Capability Model that explores the role played by big data analytics in helping companies track how to optimise the advantage of big data for the innovative, agile, and responsive development of a data-driven product or service. Prescott's model has yet to be fully tested across all sectors but offers a theoretical contribution to the phenomenon of working holistically with data, people, and society.

In an African context, research of three data and fact-checking organizations in Sub-Saharan Africa was undertaken in South Africa. The study was based on a qualitative survey, looking at 14 respondents of three data and checking non-profit organisations – Code of Africa, Open Up and Africa Check (Cheruiyot & Ferrer-Conill, 2018). The findings explored the operation dynamic, and challenges experienced by data journalists when working in the sub-Saharan African region, from the context of the three data and fact-checking organisations. Furthermore, to keep journalists accountable the fact-checking non-profit organisations also operate as media development agencies at the periphery of journalism. This development was based on the dissatisfaction with the current local journalist's practices. A barrier identified in the media sector was the reluctance to adopt the new methodology of data-driven fact-checking practices such as data liberation, data analysis and storytelling techniques. The study demonstrates that data and fact-checking non-profit organisations function inside the framework of journalistic discourse, but that they also extend it to serve their own organisational objectives and reintroduce it by providing journalists with training (Cheruiyot & Ferrer-Conill, 2018).

In an article by Borel (2017) with the heading "fact-checking won't save us from fake news", the author illustrates the dynamics of fake news. Borel (2017) goes on to say that the struggle in the media industry is associated with the struggle for power. It is possible to gain power by using misinformation and disinformation campaigns to control audiences' behaviour. The author presents a comprehensive perspective of how media influence has evolved through time in Europe and the United States, as well as how the internet has transformed the media sector. Fact-checking is said to be a key skill for journalism and a service that is instrumental in providing information to the public. However, there are several reasons why it may not be the solution to prevent fake news. These reasons include: Firstly, the media agencies have lost the trust of readers and readers thus fall for click-bait; Secondly, the technology to build

transparent algorithms and trust-building in the media space are still needed; and finally the readers' role as consumers of news media have not developed healthy news practices; for example, "thinking before clicking" (Borel, 2017).

A recent study by Saling et al. (2021) investigated the attitudes and behaviour of a sample of subscribers to a newsletter called CoronaCheck, a publication aimed at exposing fake news about COVID-19. The study used a priori power analysis conducted using G* Power 3.1. The online survey included 1576 Australian participants. The goal of the newsletter was to push back against the plethora of misinformation occasioned by the COVID-19 pandemic. They found that, on average, 24 % were open to sharing false information and 31% had shared fake news that was later discovered to be misinformation. Their responses, when asked why they (the 31%) had shared the misinformation included: "37% said it seemed interesting; 38.3% shared to get a second opinion; and 12.4% shared for entertainment value" (Saling et al., 2021, p. 10). The circulation of disinformation i.e., information that is purposefully misleading or biased, jeopardises efforts to manage COVID-19 by lowering compliance with preventative measures, such as immunisation.

They found that while subscribing to a fact-checking newsletter like CoronaCheck may raise awareness of the issue of information accuracy, this is insufficient to prevent people from spreading potentially false information. The findings suggests that a more specific intervention is required, to prevent misinformation sharing (Saling et al., 2021).

2.3 Social Media and the Spread of Misinformation

The widespread dissemination of false information through social media is a potential threat to democracy and broader society (Allcott et al., 2019). People are increasingly turning to social media for partisan media (political media). This growing problem was investigated by Anspach and Carlson (2020) in a recent study conducted in America, looking at the impact of social commentary on media credibility when posting partisan commentary with opposing reports of political polls. An experimental survey was conducted online using a reliable and unbiased outlet (Yahoo! News) and Facebook. The conditions included the full article on Yahoo, article preview, conservative commentary, and liberal commentary on Facebook – one biased and another unbiased. The findings showed that individuals were more likely to recognise information shared on social commentary than the news article previews, despite that information being biased. The study also found that trust in mass media is at an all-time low, which may be the reason for the proliferation of fake news sites (Anspach & Carlson, 2020). More revealing was that news audiences generally distrust

social media-shared commentary news; but trust those who share news on social media less than the journalists who produce the news.

A study by Ünal and Çiçeklioğlu (2019), conducted in Turkey, explored the issue of fake news and misinformation that is gradually growing in the media industry and social media space. The study looked at a verification platform used to combat false news and misinformation. The study showed that fake news has a harmful effect on political news media and the issue of high 'echo chambers' and 'filter bubbles', which refer to people's tendency to follow news that fits the ideological views and beliefs they share, regardless of its content being true or false. And lastly, it was found that social media content which was examined and found to be false generated more audience interaction and attention than valid, accurate news. In fact the knowledge that the news came from valid sources could be seen to deter people from viewing the social media content (Ünal & Çiçeklioğlu, 2019).

2.4 Combating Fake News

A study (Vargo et al., 2018) conducted in the USA on all types of media between 2014 to 2016 drew a timeline series, modelling from big data analysis about the agenda-setting power fake news possesses in mid-line mainstream media. The analysis was generated from the Global Database of Events, Language, and Tone (GDELT), a project with the largest existing open-access database on human society, which conducts computer-assisted content analysis. According to the findings, there exists an unstable, intertwined relationship between partisan media and fake news, and those fake news websites are said to be stabilising, while gaining more freedom in addressing key agendas. In the case of American media in 2016, fake news was able to set the agenda on international relations (Vargo et al., 2018).

Another study (Zellers et al., 2019) seeks to understand and respond to neural fake news (fake news generated by artificial intelligence) by applying computer security. They used threat modelling to counter fake news before it scales, therefore detecting threats posed by adversarial attacks that are seeking to spread misinformation. A deep generative model called Grover was introduced to analyse the context of text for real news feed gathered from Crawl to serve as an adversarial defence against neural fake news (Zellers et al., 2019). The best models for generating neural disinformation are also the best models at detecting it. The findings revealed that the threats are real and dangerous, and, with the right skilled personnel and budget, any anti-misinformation newsgroup could produce more advanced generators. It is important that this knowledge be shared with the public, instead of building generators to be kept private, for the effective detection of neural fake news (Zellers et al., 2019).

Fake news has disrupted how society ensures good standards for media coverage, challenging existing measures of quality checking of news authenticity. A study by Torabi Asr and Taboada (2019) looks at this problem from the perspective of natural language processing, with the aim of creating a system to automatically detect fake news. Their findings were revealed through trying to create a robust fake news classifier from two data sets, Buzzfeed and Snopes. Although many fake news publishers exist, verifying whether the news story constitutes misinformation is hard to prove. In conclusion, Torabi Asr and Taboada (2019) emphasised that dealing with misinformation through big data solutions would allow for a study approach to fake news, articles and datasets, and not only the spread of misinformation, since modern text classification methods are not good enough to slow down ever-growing misinformation. Therefore, it is important to understand what big data is and the capability that can be derived when looking at media.

2.5 Big data

Torabi Asr and Taboada (2019) proposed looking to big data to solve the problem of the diffusion of misinformation. As a result, we must comprehend the concept of 'big data' if we want to study how it can be used and how the platforms managing the data need to be constructed and managed. A study by Gandomi and Haider (2015) looked at how global executives define big data. The findings showed a contradictory understanding of big data, where some definitions emphasised what it is, and others tried to answer what it does. An earlier study by Laney (2001) notes that big data can be defined by looking at the dimensions of volume, variety, and velocity, and this has emerged as a common definition of big data; whereas academic studies have felt that more dimensions should be added, such as veracity, variability and value. An article by De Mauro et al. (2016, p. 7) looks to redefine the definition of big data by saying that:

> *"Big data is the information asset characterised by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value."*

Since big data came to the fore, there has been a demand for leveraging big data for business growth (Gandomi & Haider, 2015). Sun et al. (2015) stated that big data is a set of data taken from heterogeneous and autonomous resources. Factors such as the size of data require processes that go beyond those conventionally used to identify, capture, manage, analyse and store data. Sun et al. (2015) further stated that most scholars emphasise the use of the three V's, namely volume, velocity and variety, to perform big data analytics. Volume refers to the size of the dataset, due to the differences between variables, the number of variables, and the observations for each variable. Velocity refers to the speed at which these data sets are collected, updated, and analysed. This also includes the rate at which the values become obsolete (George

et al., 2016). Furthermore, this includes the 'newness' of data collected by decision-makers and the capacity to analyse different data streams when it comes to improving the flexibility of decision-making and enabling real-time actions (Boyd & Crawford, 2012; White, 2011). Variety refers to the wide range of organised and unstructured data sources available, including text; audio; photos; video; networks; and graphics (Constantiou & Kallinikos, 2015; George et al., 2016). "While there are no universal benchmarks for defining the volume, velocity, and variety of big data, the defining limits are contingent upon size, sector, and location of the firm, and are subject to changing limits over time" (Gandomi & Haider, 2015, p. 139).

2.6 Big Data Analytics (BDA) Capability Impact on Formal Media

Akter et al. (2019) stated that big data analytics involves the firm's ability to provide insight by using data management strategies, talent, and infrastructure to transform the firm into a competitive force in the market. Other definitions of big data analytics focus on the ability of processes that are needed to put into place and leverage big data, and emphasise investment in the necessary resources and skills needed to align with the business strategy (Xu & Kim, 2014).

According to Gupta and George (2016) the organisation's ability and capability to identify, manage, integrate and deploy big data-based resources is known as big data analytics capability. Björkman and Franco (2017) investigated the increase in the use of big data analytics and how it affects decision-making, by looking at three newspaper companies in Sweden. The newspaper industry has changed due to several factors, including increased competition for advertising revenue. This has caused a decline in revenue, the collapse of old business models, and the loss of traditional news jobs. The study findings revealed that a huge shift in management occurs when you implement big data analytics: i.e., a shift in roles and routine and a shift in power. In addition, a high level of transparency and accuracy is required, which is less a technical issue, and more managerial. The three newspaper companies gained increased knowledge about their internal operations and external relations due to 'big data analytics', even though differences existed between them due to the strategy, size, and resources of the organisations (Björkman & Franco, 2017).

A recent study conducted in the United States describes the role of data journalists and illustrates how and why they use big data in their stories (Clark & Rodríguez, 2021). The qualitative study reveals the importance of data visualisation tools in allowing the reader to engage with the story and create their analysis based on the data journalist's work. Since the newspaper print, medium is migrating to digital, the ability to leverage the power of big data and provide in-depth analysis and visualisation may be a catalyst for a struggling industry, according to the research. Moreover, this is in line with Björkman and Franco

(2017) who speak of the shift in the role of journalist to data journalists who need skills that include creating graphics, finding and gathering data, being visually coherent, and performing statistics. These skills are necessary to help with the writing and reporting of news stories that are able to provide insights based on relevant data.

2.7 Big Data Analytics (BDA) Capability Technical Aspects

Gökalp et al. (2019) provide a systematic review of available open-source big data technologies. The conclusions, reached after a review of 241 open-source tools in Apache and GitHub, were that an abundance of tools exists to help in building the big data architecture of a business. However, tools have both strengths and shortcomings, and they therefore emphasised that businesses should try to build their own big data architecture and exploit the accessible open-source tools, instead of using commercially imposed tools (Gökalp et al., 2019). The main reason was that choosing IT and linking to existing experts could prove difficult. One could find a situation where someone who has the expertise and deep knowledge in a specific domain may not know how to utilise these tools. However, a study by Stančin and Jović (2019) looked at the comparison and description of characteristics of Python libraries in the context of data mining and big data analysis. Their findings identified six groups of libraries: Python core; data preparation; data visualisation; machine learning; deep learning; and big data. The findings suggest the use of Pandas library for data pre-processing and exploitation; for data visualisation, the three libraries Plotly, seaborn and Matplotlib; and for machine learning, Scikit-learn is endorsed. For deep learning a comparison of PyTorch, Keras and TensorFlow was given, but the decision would rely on whether a quick prototype or customisation is required. All are, however, good to use (Stančin & Jović, 2019, pp. 978-980). Hadoop streaming and PySpark are the best libraries to use when dealing with big data.

A risk exists in terms of scalability when depending on open-source solutions and in-house infrastructure. A study by McElhiney (2018) examined the scenario of scaling a website platform to manage access traffic at peak times, while saving resources using cloud computing services. The study investigated the programming development, implementation, and testing of a web service architecture to make it scalable with Amazon web services (AWS), while using the big data configuration with a shared Aurora database. The project did scale up to the size of Facebook and Twitter in the simulation, proving that the proposed model can work. This highlights the importance of big data architecture that includes cloud-based solutions in building a successful big data platform that scales as users join the platform and system requirements scale up.

Data has been stored for much of recorded history. The problem of archiving and sharing large datasets raises major issues of complexity and scalability of data. Warren and Marz (2015) book places an emphasis on the innovation and advances in scalable data systems of the past decades to overcome these issues. According to Warren and Marz (2015), properties to strive for in big data systems are as much about complexity as they are about scalability. Therefore, they further suggest the importance of a big data system being human-fault tolerant, with low latency reads and updates. A recent book by Kumar (2021) explores the most duplicated and distributed framework for processing massive data sets in batches and streams across clusters, the Apache Hadoop software library. Apache Hadoop is fault tolerant, scalable, and easy to use. The benefits of using Apache Hadoop, according to Kumar (2021, pp. 42-45), are that "It divides files into small parts and distributes them into multiple parallel processing nodes to accelerate the processing time. Hadoop leverages clusters of machines to provide ample storage and processing power at a price that business can afford."

In an earlier survey, Landset et al. (2015) discussed the Hadoop ecosystem and several tools that form part of Hadoop in an attempt to gain insight into how machine learning links into an analytical environment. At the time, the Hadoop ecosystem demonstrated the most advanced innovation when dealing with big data.

The general structure of the ecosystem can be described in terms of three layers: storage, processing, and management. The bulk of the discussion in the study is on the processing layer, which is where the actual analysis is conducted. Two processing models are identified: batch and streaming. In addition, several considerations for the evaluation of tools are outlined: latency of the project; throughput (speed); fault tolerance; usability; resource expense; and scalability. A comparison of machine-learning framework tools (Mahout, MLlib, H20, SAMOA) used a processing model and evaluation (Landset et al., 2015, p. 28). The comparison and evaluation highlighted that advantages and drawbacks exist in each tool; and when evaluating the tools, research is still needed. Even though research in machine learning for big data has focused on processing paradigms and algorithm implementation and optimisation, a lot more is expected in the future.

The finding in another study is that big data follows a value chain (Hu et al., 2014). The study conducted a literature survey and system tutorial of big data platforms. In doing so, it suggested four phases that include data generation; data acquisition; data storage; and data analytics. Data generation is the first phase of big data, and it entails generating large data from sources such as trading data, mobile data, user behaviour, sensing data, internet data, and other sources that are typically overlooked. Such information is strongly linked to people's daily lives and may include user behaviour. Individually, the data appears to be worthless;

but, by utilizing such gathered big data, relevant information such as user behaviours and interests can be identified and gathered.

In the early days, technology sources of information were entirely text-based and internet webpages. Over time this grew to include large-scale scientific experiments, e-commerce data sources, and so forth. The second phase, data acquisition, refers to gathering information and is decomposed into data collection, data transmission, and data pre-processing (Lyko et al., 2016). Technologies that corresponded with this phase in the earlier days ranged from crawlers, data integration and logfiles; to sensors, optic interconnect and RFID as the timeline progressed (Hu et al., 2014, p. 657). The third phase, data storage, is concerned with continuous storage and the management of large-scale datasets. This technology has evolved, moving away from NoSQL to the more advanced MapReduce; MongoDB; Bigtable; and so on (Hu et al., 2014, p. 657). Lastly, data analysis seeks to use analytical methods or tools to inspect, transform and model data to extract value. Early analysis included data mining, web mining and statistical analysis; and later, mobile community detection and social network analytics have been used.

The tutorial study suggested that powerful, built architecture with key technologies and different tools in the different phases in the big data value chain would allow for quality big data analytics for business to make sense of their data collected from various sources (Hu et al., 2014). According to a recent study by Dessalk et al. (2020) , this combination of procedures is known as ‘big data workflow’. The study presents the design and execution of a large data workflow approach based on software container technologies. The big data workflow approach seeks to address the issues of data process scalability; resource provisioning; scheduling; orchestration; and data management. According to Dessalk et al. (2020), the big data workflow strategy allows for the construction of big data workflows at a higher degree of abstraction, allowing for the separation of design and run-time features, while ensuring scalable workflow execution. The scalable workflow execution involves parallel data processing that compels workflow fragments to debug independently on diverse computing resources. As a result, the level of human competence required is extremely advanced in this regard, necessitating a look at the big data management components.

2.8 Big Data Analytics Management

In an early study by Tambe (2014), it was stated that the shortage of skilled personnel is not the only obstacle to the success of a big data initiative, but that changes to existing data assets, management practices, and data governance may also be needed. The research framework selected for this study explored what Tambe (2014) had suggested, i.e. that big data is more than just using data science techniques,

algorithms, and good practice to establish patterns in the data. According to the study (Prescott, 2016), managerial participation is essential for an organisation to gain value from big data. He further emphasises the important role high-quality data collected from the source plays: when all aspects are aligned with the business and knowledge strategy it will add to the transformation of the business advantage.

A data-driven culture refers to the extent to which people in an organisation make decisions based on information derived from data analysis (McAfee et al., 2012). A recent study suggests that the analytic culture is another crucial element in gaining a competitive advantage when using big data analytics (Hallikainen et al., 2020). Using a multi-industry dataset derived from 417 business-to-business firms, the impact of big data analytics on customer relationship performance and sales growth was explored in the study. According to the findings, firms with a strong analytic culture have a greater association between customer big data analytics and customer relationship performance than companies with a weak analytic culture. This is in line with the idea that big data analytic initiatives need to develop a culture of experimentation across the organisation (Alexander, 2019; Hallikainen et al., 2020).

Another study that conducted a survey of 267 service analytic professionals from the United States of America and France demonstrated the relationship between human capability and organization growth in the big data analytic environment (Akter et al., 2019). The survey revealed that the success of service analytic capability is mainly because of strong talent capability, which links to overall capability, so both information capability and technology capability are used to produce superior firm performance. According to Côrte-Real et al. (2020), strategic management perspectives, such as strategic planning and operations models, can be used to explore the value added by big data analytics and the Internet of Things. The capabilities offered by the Internet of Things and big data can be used to create and add value to a different level of decision-making. Furthermore, the study findings indicate that applying big data principles to creating a competitive advantage to post-performance, and establishing a good quality of data, can unlock the potential offered by the Fourth Industrial Revolution.

A book by Alexander (2019) investigated procedures and routines that an organisation can use for big data projects and the understanding of data routine sensitivity. This includes the re-configuring of analytic capabilities and the extent of heterogeneity of these routines to evaluate the variance within and between projects and their effect on project success, as well as the extent to which data routines are shared and integrated across other big data projects. The findings suggest that a correlation exists in the routines and big data initiative and that it is an iterative process, where the knowledge is created, shared, refined, and reconfigured based on continuous learning. Another recent study (Surbakti et al., 2020) identified four

related factors that can be used to effectively manage and deploy big data. These factors include competition, economy, internal company processes and the state of technology. As stated, there is a lot of emphasis on management taking the lead. However, it would be premature to assume that this is the deciding factor. In managing big data there are several key issues that must be considered, such as the quality of the data; how it can be visualised; privacy concerns of using big data; and the security threat that comes with high volumes of big data.

2.8.1 Big Data Quality

An earlier study by Janssen et al. (2017) looked at the quality of decision making from big data analytics for firms. The findings suggest that a firm's success can be attributed to the following factors: the characteristics and quality of big data sources; the quality of the big data analytic process; the big data analytic capacity and capabilities of those who gather and process big data; and the availability of a big data infrastructure are all factors that should be taken into account. They suggest that almost all the factors are still underdeveloped and lack major substance to be fully understood as key determinants of quality decisions for big data analytics. For an in-depth look at data quality dynamics, a diagram was constructed by Batini et al. (2015, p. 20), illustrating the relationship between data quality and a number of research co-ordinates pertinent to big data: data types; data sources and application areas; maps; semi-structured texts; linked open data; sensor and sensor networks; and official statistics are among the topics covered. The findings revealed that the following dimension clusters: accuracy; completeness; redundancy; readability; accessibility; consistency; and trust have evolved and changed how we look at data types, sources, and domains. Even though the study by Batini et al. (2015) was conducted years earlier, the researchers were able to identify the extent of the known research aspects that were important to big data and still prove to be an important area today. A systematic literature review by Ramasamy and Chowdhury (2020) suggested that meeting data quality dimension standards (accuracy; completeness; redundancy; readability; accessibility; consistency; trust) can prove critical for data quality assessment. It is imperative to define and analyse new dimensions connected to big data. However, the study further adds that the findings only relate to big data in a textual data format, and more research is still needed since big data occurs in several forms and with various characteristics.

2.8.2 Big Data Analytic Visualisation

Big data systems transform large and complex data streams of information into an understandable and effortless form. This form is achieved through data visualisation, which is defined as follows:

> "The presentation of data in a pictorial or graphical format, and a data visualisation tool is the software that generates this presentation. Data visualisation provides users with intuitive means to interactively explore and analyse data, enabling them to effectively identify interesting patterns, infer correlations and casualties, and supports sense-making activities." (Bikakis, 2018, p. 1).

Data visualisation is thus an integral part of data analysis. However, Mani and Fei (2017) add that effective data visualisation for big data analytics is complex compared to standard data analytics. The results of the experiment showed that, by delivering duplicate data to the visualisation system in response to the original request, it is feasible to reduce delays for future requests and improve user experience dramatically. The main goal is for the visualisation system to contain the least amount of data possible, while yet ensuring that the visualisation system can handle the majority of future requests from the user without the need for extra processing by the data processing system (Mani & Fei, 2017). The experimental study used R language for both data processing and visualisation, although other languages can be used in place of R. Since the finding was experiment-based they could not conclusively evaluate the effectiveness of data visualisation of big data analytics.

This was supported by Golfarelli and Rizzi (2020). They evaluated SkyViz as a visualisation approach based on a case study selected from a pilot application of a model-driven architecture. The approach allowed for automating the translation of user objectives declared for visualising the result of big data analytics into a set of most appropriate, practical visualisation. The finding suggests that additional features and user support are still required to have an automated big data system, mainly due to low levels of user expertise and scalability ineffectiveness. This is in line with Wahyuningsih (2020) discussion on the problems, challenges and potential of visualization of big data. The findings suggest that it is difficult to visualise large volumes of big data from existing visualisation tools with low flexibility, scalability and response time. Interactive visualisation is key and should be created by a good visualisation tool.

2.8.3 Big Data Privacy

The increased interest in big data producing profit and competitive advantage in the private and the public sector has raised privacy and moral concerns. As a result, in America communities like the National Science Foundation have established a Council for Big Data, Ethics, and Society. This includes a group of 20 scholars from diverse areas of social, natural, and computational science (Zook et al., 2017). The Council formulated the ten best ethical practices to encourage scientific and engineering researchers to address human privacy; data protection; human recognition; the possibility of harm resulting from big data analysis

initially designed for good purposes; and the lack of standards for sharing sensitive data. The fears the community stress are imperative in ensuring responsible big data research.

In Europe and America, for example, there are provisions in the European General Data Protection Regulation (GDPR) and the Californian Consumer Privacy Act (CCPA) that allow consumers to be 'forgotten', or have their consented data removed at their request (Suleh-Yusuf, 2020). In African countries, however, this is not the case. As a result, findings suggest that the privacy of personal information has been neglected in Africa in the most part. Currently, guidelines have only been put in place by governments, which do not limit large companies who deal with data mining and big data harvesting. In establishing projects and even partnership governments across the region, they escape the purview of national legislation (Suleh-Yusuf, 2020).

An earlier study by Shozi and Mtsweni (2017), conducted in South Africa, exposes the continued threat caused by social media to the privacy of people's data as Africa develops as a continent. According to Shozi and Mtsweni (2017, p. 6), "Less than half of Africa's 54 countries have data protection or privacy laws that have been passed, or bills that have yet to be passed in their respective legislatures." In the case of South Africa, during the data collection phase of this study, the POPI Act had not been enacted; but it has subsequently become law as of 30 June 2021 (Michalsons, 2020). According to the POPI Act, state information needs to be stored in South Africa. Shozi and Mtsweni (2017) believe that a more detailed investigation into the effect of social media sites on privacy laws is still needed. They stress the lack of research around data privacy and the engagement initiative in society to empower citizens to become more aware of individual privacy rights and what it means to have data ownership.

2.8.4 Information Security

Fadler and Legner (2020) explored the importance of data ownership in the big data space. An investigation was conducted by comparing three case studies of large companies. Their findings suggest that understanding data ownership means exploring data warehouses and data marts through three data ownership enterprise types: the data owner, the data platform owner, and the data product owner. The results are in line with other studies in that data ownership is a key concept in clarifying the rights and responsibilities of big data analytic (Alexander & Lyytinen, 2017). Clark and Rodríguez (2021) discuss an Act that gives the right to the United States media and press to battle secrecy through legislation called the Federal Freedom of Information Act (FOIA). The state also provides Freedom of Information rules that govern who can ask for information and who can obtain it.

2.9 System Performance: The outcome achieved by the big data system

The final aspect of big data analytics management to be considered is the system performance, or the outcome achieved by the big data analytics implementation. The reported outcomes of these implementations include both successes, or added value experienced, as well as discussing those challenges which have been experienced. Some of the benefits experienced include the ability to profile users and target advertising as well as to personalize customer experience and hence increase brand loyalty. A case study showed how big data analytics can create value by looking at a collection of big data from mobile user profiles, access behaviours, and mobility patterns from a proposed framework that supports both offline and online advertising operations (Deng et al., 2015). The organisation's team involved with sending out advertisements used the data to handle several use cases connected with addressing relevant advertisements to the customers for promotional awareness campaigns. Another study conducted an interview with six participants, looking at the strategic priority of consumer brand engagement in major retail brands in the South African retail sector (Mutendadzmera, 2015). Evidence suggests that, through big data analytics, customers' shopping and buying habits had allowed for a more personalised experience, leading to increased customer loyalty.

Further advantages obtained include the ability to use big data insight to improve functions of the business and achieve profitable outcomes and competitive business offerings. Sivarajah et al. (2020) highlighted how business-to-business enterprises can use big data and social media analytics within a participatory online environment to achieve profitable outcomes and remain sustainable through strategic operations and marketing-related business activities. This qualitative study reveals how two important functions of the business – marketing and operations – can provide the business with valuable insight into its customer base on an emotional level, using big data and social media analytic. This opens the opportunity for service proposition and business image mitigation (Sivarajah et al., 2020). Additionally, a survey of CEOs and senior managers revealed that the results are in line with prior studies (Akter et al., 2016; Wamba et al., 2017), in saying big data analytics contributes to a positive relationship between customer relationship performance and sales growth, and that firm size has no discouraging effect. Therefore, it is fair to conclude that, in the past decade, big data has been proven to positively affect business performance.

It is also important to point out big data analytics benefits vary depending on the managers approaches given the respective size of the organisation and how business uses social media channels. Dong and Yang (2020) tested the market performance of big data analytics in a recent study using a large-scale survey dataset of 18 816 small micro-medium enterprises (SMEs) and large firms in Italy. The study differs from previous studies on performance by adding that "we explain that the super-additive value arising from the

synergies of the complementary use of social media channels and big data analytics is the value creation mechanism of social media analytics" (Dong & Yang, 2020, p. 6). The study found that to obtain a competitive advantage in the marketplace, managers must proactively integrate the complementary use of social media channels and big data analytics in their businesses, which is more critical for SMEs than for large corporations (Dong & Yang, 2020).

However, even though there is value to be gained from investing in big data analytics, other recent studies suggest that this value comes with challenges. The challenges arise from the existing big data characteristics and how we look at and define big data. According to Ghasemaghaei and Calic (2020) big data is not always better data. The study addresses the significant gap in the literature concerning the influence of the main characteristics of big data (variety, velocity, and volume) on innovation performance, and their impact on firm performance. The results suggested moving away from a holistic approach of conceptually and operationally differentiating the main characteristics of big data (variety, velocity, and volume) and argued for an approach that looks at the sum of their parts (Ghasemaghaei & Calic, 2020). For instance, the study revealed that, while data variety and velocity positively enhance innovation, efficacy and efficiency, data volume has no significant impact. Furthermore, Ghasemaghaei and Calic (2020) note that data velocity plays an important role in improving firm innovation by ensuring the prompt integration of different types of data.

2.10 The Academic Research into Big Data

The academic community has established a discipline aimed at addressing challenges that are being faced, and are going to be faced, in the big data era, called data science (Song & Zhu, 2016). The data science industry landscape is confusing for employers, academic and training institution, and existing and aspiring data science professionals (Fayyad & Hamutcu, 2020). To address this the academic community has established a discipline, called data science, aimed at addressing challenges that are being faced, and are going to be faced, in the big data era (Song & Zhu, 2016). The biggest challenge has been working out how a data scientist can learn to use all the different knowledge domains in the field: CDO disciplines; the data analytic life cycle; big data technologies; model-building techniques; and so on. The findings suggest that designing programs or creating curricula is challenging since there is not yet a consensus on what data scientists are and what abilities and knowledge they must possess (Fayyad & Hamutcu, 2020, 2021; Irizarry, 2020). This implies a vulnerability could exist in terms of business continuity: For example, if a company using big data technology has a job vacancy for a data scientist, there is a moderate to high probability of not finding a data scientist that meets the technological experience needed to fill that role. The current lack

of recognised professional standards around ethics, data science roles and associated skills and knowledge increases the range of challenges that can be faced in this field and provides a large scope of aspects which require further research (Fayyad & Hamutcu, 2020). An example of one area of study is provided as an illustration below.

One of the biggest challenges in this field is dealing with adversarial attacks i.e. attacks that are designed to impact input in a specific way that results in the wrong result being provided by the data science model (Song & Zhu, 2016). Dering and Tucker (2017) proposed using a generative adversarial network (GAN) – which efficiently generates human-readable visual image objects – in big data pipelines during training to improve the veracity of data. The evaluation of a generative adversarial network required that 42 classes be trained, and the result was gathered from 32 classes. The word "generative" in generative adversarial network describes a class of statistical models that contrasts with discriminative models. Generative models are simple models which can be sampled and used to generate new data instances and, discriminative models discriminate between different kind of data instances. For example, a generative model could generate new photos of trees that look like real trees, while a discriminative model could tell a Pine Tree from a Mango Tree. The findings reveal that there are many types of generative models, suited for many purposes, and that GAN can be applied to guard against high-quality computer-generated images that make the proposed solution a valuable option for the big data system to humans.

2.11 Conceptual Framework

The Digital Genesis Data Model was used in conducting this case study (Prescott, 2016). The model is based on a resource-based view and a dynamic capability view of the process. The resource-based view is defined by Barney (1991) as resources that are valuable, rare and hard to copy perfectly, which are easily exploitable by the company, in the context of them utilising the resources to achieve the company's goals. The dynamic capability concerns the company's routines: what it can achieve in response to a change in the business environment. Also, it is designed to be very responsive to changes in the operational environment. This is achievable by using existing capabilities and routines to re-establish new, revised capabilities (Braganza et al., 2017).

Capabilities and procedures for acquisitions or mergers; research and development; business process re-engineering; quality control; and technology transfer are examples of dynamic capabilities (Teece, 2007). Three components form the foundation of dynamic capabilities (Teece, 2007, pp. 1326-1340): first of all, the analytical systems within the organization – in this case, big data tools and applications, and individual

capabilities to learn and to sense, filter, shape, and calibrate opportunities; second, business structures, procedures, designs, and incentives for taking advantage of opportunities; and third, to achieve a product advantage, certain tangible and intangible assets must be continually aligned and realigned. Dynamic capabilities are thus made up of a combination of systems, structures, resources, and organisational and individual capabilities, both tangible and intangible (Teece et al., 1997). However, understanding how the organisation becomes agile and responsive requires an understanding of how a company moves from large amounts of data through a series of steps to achieve its goal.

2.11.1 The Process of a Digital Data Genesis Generation Capability

A digital data genesis generation capability (DDGC), shown in Figure 1 (Prescott, 2016, p. 98), consists of the alignment of the data, information and knowledge that will be generated to support the business plan/strategy of the organisation (Bharadwaj et al., 2013). This includes an understanding of what data will be captured and from what sources; how the information will be made available for use and in what format; and how that knowledge will be accessed and used to create, modify, or adapt current human resource competencies and organisational capabilities. An organisation must understand what types of data need to be generated, how that data will be used, and what technology will be needed to collect that data. Is the data batch or real-time? How does this impact the current IT architecture of the organisation? How much data will be generated, and how will it be cleaned, stored and accessed? What algorithms will be needed to prepare the data for use? How will the data be made available for use? What type of reports, dashboards, or other data visualisations will be needed for the most effective use of the data?

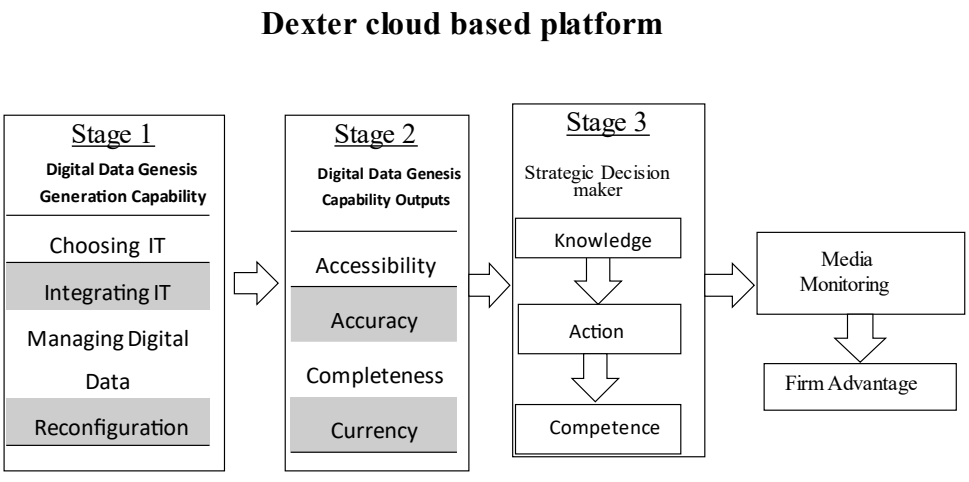

Figure 1: Digital Data Genesis Model.

Choosing information technology in big data, according to a study by Hu et al. (2014), evolves as time passes, as does its technology compatibility in the different big data value chain phases. Integration of IT from data pre-processing involves joining data from different locations and assists users with a collective view of the data. Data integration is a well-established field with two well-known approaches: the data warehouse method (ETL – extraction, transformation and loading) and the data federation method (Hu et al., 2014). With regards to managing digital data: "A big data management framework means the organisation of the information according to the principles and practices that would yield high schema flexibility, scalability and processing of the huge volumes of data, but for which traditional RDBMSs are not well suited and become impractical" (Mehmood et al., 2017, p. 6).

Nelson et al. (2005) suggest that, for the data to useful, it must be accurate, complete, and current. Accuracy implies that the data must be correct, reliable, and trustworthy. Accessibility refers to the convenience of using and locating the data. Completeness implies that the user perceives the data as up-to-date and applicable in its current context. A digital data genesis capacity entails more than just using IT to extract information from large amounts of data. It's a dynamic capability that can be used to reorganise itself. Decision-makers can leverage a digital knowledge strategy from digital data genesis capability to respond to the dynamic environment.

2.11.2 The Digital Knowledge Innovation Hierarchy

A series of procedures are used to describe how an organisation translates data into information, information into knowledge, and knowledge into employee intelligence (Sain & Wilde, 2014). As intelligence is to the individual, a capability is to the organisation; therefore, it implies a certain level of know-how. Thus, a capability consists of competencies and sub-routines, and an ongoing learning curve. They are path-dependent, and if a capability fits the requirements described previously in the resource-based framework definition, it can become a strong advantage. It is worth noting that dynamic capabilities by themselves do not provide a firm with a long-term competitive advantage unless they possess the characteristics outlined in the resource-based framework, namely, that they are valuable, rare, difficult to duplicate perfectly, and easily exploitable by the company. The output of a dynamic capability is the modification in the organisational capability (routines) and/or complementary resources (Easterby-Smith & Prieto, 2008). The following stages serve the purpose of demonstrating an understanding of the organisation's current use of data, information, and knowledge, which can allow for greater success when implementing the digital knowledge strategy (Prescott, 2016).

- Stage 1: Data is generated and stored in data repositories where it becomes information. However, this information is not shared throughout the organisation. The data is generated with no end-use in mind.
- Stage 2: Information becomes knowledge. The information is accessed for research and problem-solving purposes. This information can impact the way an employee performs their job or solves a problem. But this knowledge, unless codified, is unlikely to be shared throughout the organisation.
- Stage 3: The knowledge is codified and integrated into business processes and is disseminated throughout the organisation.
- Stage 4: With use, the routines and subroutines that make up the business processes create competency in human resources, as well as in organisational capabilities.
- Stage 5: Knowledge is used to create core capabilities that can give an organisation a competitive advantage.
- Stage 6: The organisation can innovate and respond to turbulence in the environment to sustain a competitive advantage (Teece, 2007).

As can be seen from the above hierarchy, organisational capabilities are created through the organisational process of learning, integrating, and making those processes available for use throughout the organisation. For an organisation to use data, information and knowledge to implement its business plan/strategy, there must be a culture of openness concerning the sharing and exchange of information in the organisation (Dosi et al., 2000; McAfee et al., 2012). Therefore, the ability of an organisation to implement or make use of its digital knowledge strategy is dependent upon where it falls on the digital knowledge innovation hierarchy.

2.12 Conclusion

Big data analytics has received significant research attention with regards to its implementation. At this point, most big data initiatives are profit-driven and focus on allowing an investor a competitive advantage that can be derived from talent capability, information capability, and learned routine capability, that was not possible before. Furthermore, a noticeable research trend has emerged in the past decade. After developing an understanding of what big data means, the focus has shifted to the potential it presents in relation to digital media, in terms of data mining and data analytics. Big data has grown immensely compared to its earlier days. Unfortunately, as the literature stresses, more effective approaches are urgently needed to address increasing concerns over the threat it poses in countries with low privacy infrastructure and, with the uncontrollable fake news attacks on media, it is important that governments and civic society find big data solutions to fact-checking of media to maintain healthy democracies that represent all the

citizens they govern. The next chapter provides a clear description of the research design and procedures of the study by looking at the research methodology used; the data collection; the analysis method; the research setting; sampling; and so on.

Chapter 3: Research Methodology

3.1 Introduction

This chapter will discuss the qualitative research design, and the use of a case study research methodology that allowed the researcher to gather and generate data to investigate the research question. The data to be produced will be discussed in terms of the sample instrument, the research site and the population studied. Furthermore, it will explore the preparation and design phase and the collection process of the study; and The tools and technology used in data capturing and analysis.

3.2 Research Philosophy

The study's underlying philosophical theory, interpretivism, emphasizes the notion that reality is a mental construction that can only be understood subjectively (Kroeze, 2012). Interpretivism has similarities with postmodernism in terms of its fundamental assumptions about knowledge, reality, truth, cognition, methodology, and rigor. Interpretivism is, thus, a typical postmodern epistemology in the context of Information Systems research (Kroeze, 2012). Global conceptions of who we are and what we know are altering because of this broad and profound cultural revolution (Kroeze, 2012).

3.3 Research Design

The research design used is a qualitative case study, since the study is exploratory and tries to illuminate why certain decisions were taken, how they were implemented, and finally the results (Yin, 2009, p. 17). A qualitative approach allows for in-depth interviews to be conducted on the subject being studied, thus allowing a more detailed exploration of people's ideas around the use of big data in the organisation. An exploratory research approach was used to conduct the study. The approach was chosen primarily to study the research phenomenon investigated in a prior study by Prescott's (2016). In doing so, employing the same model as a theoretical foundation. However, it will be undertaken in South Africa, in a new industry, and it will examine each component and step of the model in considerable detail. This includes the precise theoretical and technological decisions made at each stage of the design, development, and implementation.

The phenomenon is the practical implementation of a big data system which can operate as an integrated environment to meet the needs for which it is developed. The model proposes theoretical aspects which need to be accommodated for- but understanding the practical implications of these – and how they may

present themselves when required to meet different needs in different sectors – is not well known. This kind of nuanced information can only be obtained through interviews.

This approach was used to gain an understanding of opinions, motivations, and reasons behind the problem (Sekaran & Bougie, 2019). Qualitative data collecting methods capture information in the form of words generated through interviews; open-ended questions on questionnaires; focus groups; observations; and material gathered from various sources on the internet or in the library. This method of research allows researchers to obtain primary data from a variety of sources (Sekaran & Bougie, 2019).

3.4 Research Methodology

A case study is defined as:

> "An empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between the phenomenon and context are not evident" (Yin, 2009, p. 19).

The research was conducted at Open Cities Lab, located at 153A Helen Joseph Road, Glenwood, Durban, KwaZulu-Natal, in South Africa. The project-under-study contact persons from Open Cities Lab, were as follows:

Table 1: Open Cities Lab.

Name	Job Title	Company
Mr R Gevers	CEO	Open Cities Lab
Mr M Adendorff	Head Data Scientist	Open Cities Lab

A case study is most appropriate to use when the main purpose of the study is to answer 'how' and 'why' questions and when it is impossible to influence the responses of those involved in the study; also, when

the contextual conditions are as important as the phenomena being studied, and boundaries are not clear between the phenomenon and context (Hafiz, 2008).

According to Yin (2009), a case study process includes six interdependent stages, as presented below:

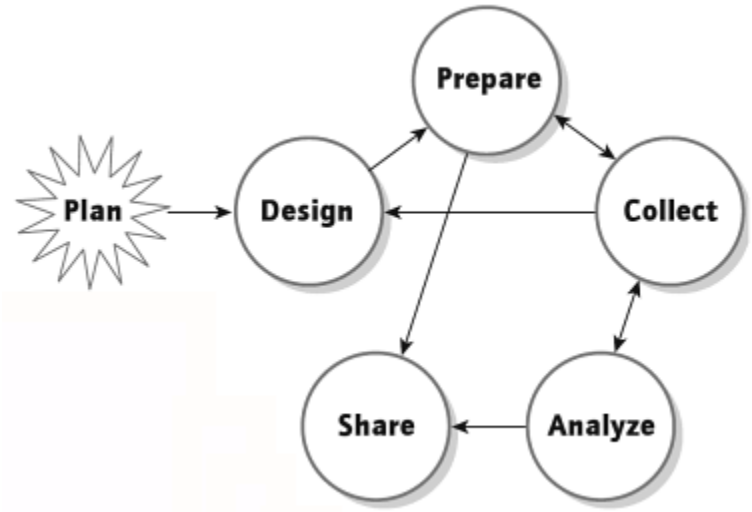

Figure 2: Case Study Process.

A case study is a complex and difficult process that first requires planning, which entails focusing and identifying the research questions and other rationale for doing a case study (Baskarada, 2014; Yin, 2009). The use of the Digital Genesis Data Model was instrumental in the construction of the research questions. Furthermore, documentation provided about the big data system and the different media, in the form of old video conference presentations about the system and general information support from the staff at Open Cities Lab, were also instrumental in providing a clear picture of what the big data system does during the planning stage. Secondly, the research design stage focused on the research interview questions. The use of an Appendix D: Alignment Matrix ensured relevant interview questions and the appropriate identification of the unit of analysis. With supervisor guidance, open-ended interview questions were constructed that would allow the participants to give in-depth insight into the subject of big data in the context of the media big data platform (Dexter) (Baskarada, 2014). The next stage of preparations included conducting a pilot study a month before the interviews. The pilot study helped give context to the dynamic nature of people working and developing big data systems, and how to fine-tune the research instrument to allow for a clear conversation by using common terminology the participants are used to.

In a study by Starman (2013), four advantages of using a qualitative case study are discussed to support and foster the view of qualitative case studies, as compared to quantitative methods specifically: conceptual

validity; the development of novel hypotheses; the investigation of causative mechanisms; and the modelling and assessment of complicated causal relationships. Conceptual validity consists of measurements and identifier indicators that best showcase the theoretical concepts that the research study is investigating. The advantage is that, in a case study, it allows for conceptual modifications with a higher validity level over fewer cases (Starman, 2013, p. 36). Case studies do not require many instances or variables. They can analyse qualitatively complex events and consider a wide range of variables. Case studies examine a significant number of influencing factors inside a single instance and assist in determining what conditions are present in a case that activate a hypothesis about the cause(s) that typically bring it about (Starman, 2013). Case studies can also be used to handle complex cause-and-effect relationships between variables, such as, examining the property of having the same effect or result from different events and processes where past events or decisions constrain later events or decisions. Moreover, a case study can hone in on real-life scenarios and put theories to the test in connection with phenomena as they occur in the actual world (Starman, 2013). The advantages of a case study research design make it suitable to explore the research question. Since the study requires a detailed gathering of information and thick analysis of this rich data it would be difficult to answer the research question using a quantitative research design. A qualitative, context-specific design, such as a case study, is necessary because big data technology is still a new market and there are a small number of active companies, NPO and government departments which will allow access to this level of detailed institutional information.

A potential disadvantage is that a case study, as a qualitative research method allows a researcher some freedom in specifying the parameters of the study. This means that a researcher's own subjective feelings may influence the case study, introducing a potential bias and resulting in a lack of scientific rigour and little grounds for generalizing results to the larger population. It thus falls to the researcher to find a well-developed theoretical framework to guide the case study. This helps ensure consistency, rigour and trust in a case study (Hyett et al., 2014). In this case study, a well-developed theoretical framework was selected to guide the study in exploring big data concepts, processes, and procedures.

3.5 Sampling method

Since the company is small, the study has five participants who took part in the development of the big data system, although this number can change, depending on the time of year or the current events in the country. However, this study will focus on the five available people who are permanently employed and work on the Dexter platform.

3.5.1 Research Site

The study was conducted at the Open Cities Lab (https://opencitieslab.org/). The study was looking at the Dexter project maintained by Open Cities Lab that seeks to leverage modern computational approaches to monitor and analyse Southern Africa's digital media to enable media governing agencies to drive accountability and equal representation (Lab, 2021).

The target population was located at Open Cities Lab, a non-profit organisation that has developed powerful machine learning and analytical methodologies within a platform called the 'Dexter Project'. The target population was the staff members who were involved in developing, implementing, and using the big data for the chosen service cycle.

3.5.2 Selection of the Project: Dexter

The research site was selected based on Open Cities Lab's specialisation in dealing with big data systems. After having had an in-person discussion with the CEO of the Open Cities Lab, the researcher met with the Open Cities Lab management team and conducted an introductory presentation, outlining the intended focus of the study. The Open Cities Lab team provided a media-based big data system for the study. The team contact information for those involved with the Dexter big data system was provided, as well as formal and informal documentation on the Dexter platform.

It is important to highlight that it was difficult finding a company willing to share and provide access to data of this nature. The companies who use big data systems in business stated that the use of their big data systems would damage the company's competitive advantage developed with their big data. Some also added that the big data platform adds close to 2 billion rand in revenue profit for them and it would not fare well for the company if information of any kind were to be used in a study. Thus, they declined the opportunity to contribute to the study. Luckily, after some time searching, an opportunity finally presented itself in the form of a non-profit organization (NPO) that specialises in big data technology and is not constrained by financial concerns. This allowed the study to continue. However, the NPO was small, with few staff members, compared to a large company. This meant the study had a small sample size.

3.5.3 The Dexter platform

Dexter is a cloud-based tool created by Open Cities Lab. It is made with Flask and a full Python stack to automatically accept digital media from a variety of nations (particularly South Africa and other African

countries) (Adendorff, 2017). After that, the system uses natural language processing (NLP) to highlight important sources, topics, and entities, allowing advanced analytics to be done on the generated data.

The big data system, ‘Dexter’ keeps a record of what is being spoken about in the media, by whom, and in what context. The system itself ingests digital media sources and extracts the utterances and names of key entities and people, allowing for a growing record of sources and their statements to be collected for analysis. The Dexter platform receives meta-data information on identified sources through a web interface, capturing demographic information, such as gender and race, as well as affiliation data (Adendorff, 2017).

Team members at Open Cities Lab identify potential problems or issues with the article or the topic being written about and make any corrections to the media articles sourced. With this curated dataset, the Dexter platform allows for descriptive analytics and analytical function to better understand which sources, affiliations, races, genders, keywords, topics, etc., are generated from the big data to display on the web content management system user interface (Adendorff, 2017). Based on this, the Dexter platform will allow third-party entities to provide feedback to the government, the media, and citizens; as well as providing policy suggestions. Since the company is small, the study interviewed all five participants who took part in the development of the big data system.

3.5.4 Ethical Clearance

The researcher applied for, and received, an ethical clearance letter from the University of KwaZulu-Natal to conduct the study at the Open Cities Lab. The application entailed a rigorous process that involved getting gatekeeper's letter from the Open Cities Lab and submitting the research proposal for evaluation. The ethical clearance is no. HSSREC/00001979/2020.

3.6 Data Collection

Following the case study methodology, numerous sources of information were employed to construct a case study database, and to maintain a chain of evidence in the data collection phase (Starman, 2013). The information about the study was sent to participants via email. Participants received an introduction, and a description of the study and the interview process. A letter of consent was attached to the initial email, to be signed before the interviews began. It is important to note that data collecting does not involve Partner A, the big data system's user. The data collection focused on technical and support components of the theoretical framework rather than an external user's day-to-day experience with the big data system.

However, perspectives and insights from technical support, developers, and data scientists which do speak to the outputs of the system are included.

Semi-structured in-depth interviews were thought to be appropriate for this study since they allowed the researcher to access the participants' in-depth impressions and ideas, as well as obtaining insight into the research problem from their perspectives (Starman, 2013). The interview schedule was developed based on all three stages of the theoretical model and was divided in two parts: the first part consisted of a set of in-depth structured questions that address the first stage. The first stage of the model looks at the following components: Choosing IT, integrating IT, and managing digital data and reconfiguration. The first part involved the staff members who had created and use the big data Dexter project. (IT staff and data scientists) The second part of the interview schedule included the second and third stage of the Digital Genesis Data Model. The second and third stage look at the content management system and the performance report of the Dexter project, i.e., the extent to which the output met the requirements of the project.

Pilot testing was conducted to gather feedback to improve the standard of the research instrument (Baskarada, 2014). The pilot study entailed evaluating aspects of the interviews and identifying misleading questions. The participant used in the pilot study was from a different industry to the one being studied. The requirements for participating in the pilot study were that the participant's company used big data on a regular basis, and that the participant must be involved in the day-to-day processes and maintenance of those big data systems. People who met these requirements were not easy to find. Fortunately, one participant who works in the banking sector was found to conduct the pilot study. The results from the pilot study exposed how interview question wording needed modification to align with terminology and wording that is used by people in the big data space. The pilot test allowed for an in-depth breakdown of questions. e.g., the question "How did you and your team go about installing and integrating the big data software necessary for enabling Dexter platform development? "Into four more detailed sub-questions, namely:

4.1 Was the data captured elsewhere and then streamed?
4.2 Is the data stored locally?
4.3 How was the big data handling (data checking) managed?
4.4 From data on the Dexter platform a monitor curator has been mentioned. What role does a monitor curator have in the big data handling?

Furthermore, gaining access to, and producing, the necessary data for the study occurred in four stages. A request was sent to the corporate affairs personnel for detailed product information, regarding the nature of

the Dexter project. Thereafter, the two Open Cities Lab team members involved in choosing the IT; integrating the IT; managing the digital data and re-configuration; and providing the data and reporting facilities to be used in product decision-making were interviewed in both parts one and two of the interview schedule. After that, the three members involved with the front-end and performance monitoring were interviewed, to answer part two of the interview schedule.

An interview schedule was developed for the first and second round of interviews. The interviewer used open-ended questions, developed to allow the participants to expand on their answers. This allowed the researcher an opportunity to ask more questions if the answers were not detailed enough initially (Sekaran & Bougie, 2019). The first interview schedule was produced using an alignment matrix (see Appendix D: Alignment Matrix) that included the theoretical framework and the research sub-questions that led to the creation of the research questions (see below). The second round of interviews was based on the analysis of the first-round findings and the aspects that needed clarification. The second-round interview schedule also served as an extension of the main interview questions to gather more in-depth information of the workings and implementation of big data from the three participants. This allowed for a chain of evidence and a link to the research sub-questions.

The data production administration entailed scheduling Zoom audio calls using email and the Calendly app to allow a participant to select a day and time most convenient to be interviewed. Then, a consent letter was sent a day before the interview was scheduled online. Each interview would take about 45-55 minutes long to complete. Thereafter, a saved recording was automatically sent after each session ended. The interview instrument was personally administered. Following the completion of all the interviews, a program called Otter (https://otter.ai/) was used to assist with the translation of the data into MSWord format, which was then placed in the Google cloud for safekeeping before being analysed on NVivo.

A second round of interviews occurred two months after the initial interviews to allow for the first round of analysis to be completed. This second interview allowed the researcher to clarify the initial analysis with the three participants who dealt with these aspects of the system and to get an in-depth analysis of the situation: The interviewer looked at the answers from the interviewees that did not provide a clear picture and, on that basis, rephrased questions to get a clearer picture of the participant's reality. In addition, the researcher kept a research/interview journal on the interaction with the five participants to record their responses to the interview questions. Any specific impressions were noted to serve as additional data.

3.7 Data Capturing and Editing

The interview data was recorded, and a transcript was created of the interviews – i.e., a typed script of everything said during the interviews, by both interviewer and participant (Creswell & Creswell, 2017). Thereafter, an online converting transcription tool, Otter.ai, was used in parallel with listening and transcribing by the researcher. The dialogue scripts were then entered into a document for NVivo 12 for processing into themes for analysis.

3.8 Data Analysis

The process of reviewing, processing, and refining data using analytical and logical reasoning is known as data analysis (Hair et al., 2010, p. 230). Analysing data to uncover links, patterns and trends is what data analysis is all about. The data is analysed to find facts that may be related to the research phenomena, as well as to demonstrate the relationship between the independent and dependent variables (Hair et al., 2010).

Thematic analysis was chosen for the examination of the data acquired from the interviews because the study followed a qualitative approach. The analysis uses the digital data genesis framework variables as a basis for thematic analysis. This allows the data to be grouped into themes that were crucial to comprehending the topics under discussion (Bhattacherjee, 2012). This was achieved by using a NVivo 12, MSWord version.

The data was ready for analysis once it had been processed, cleaned, and validated. Using the web application Otter's listening and pause feature, the transcript was double checked, and any inconsistencies were removed. Once the transcript editing was complete the transcript was downloaded into a MSWord document and the layout was changed into a dialogue script before importing it into NVivo 12. The use of NVivo codes allowed the researcher to form patterns, which could thereafter be grouped into logical categories which were given names. Since the study followed a specific model, themes and comparisons were generated from working closely with the model.

This procedure allowed the researcher to consider and re-read the data, allowing the researcher to become engaged with, and thoroughly aware of, the information. This phase allowed the researcher to create short codes that identified key aspects of the data that were crucial to addressing the study’s research questions. The entire dataset had to be coded in this manner. The researcher then gathered all the codes, as well as all the essential data extracts, for the later stages of the study. The next step was to gather information about

each candidate theme. This allowed the researcher to work with the data and consider each proposed theme's viability. In this phase, the researcher compared the suggested themes to the dataset. This was done to find potential themes that told a compelling story and answered the study's research questions. Themes were refined, split, combined, or thrown away, during this stage. The researcher then conducted a thorough examination of each topic, meticulously determining the scope and concentration of each subject, as well as the 'story' of each theme. During this stage, the researcher chose an informative name for each theme. The researcher linked together the analytic narrative and data extracts and placed the study in the context of the current literature in the last phase, which involved writing up the data analysis process (Starman, 2013).

The thick description (data) was used to write up the analysis. The role of the researcher is to both accurately describe and interpret social action (or behaviour) within a certain context, and this is referred to as thick description (Ponterotto, 2006). During the collection of data, a detail level of questioning was applied. Therefore, the thick description presents what collectively was said and understood during Nvivo analysis of each theme. The direct quotes presented in the analysis serve as representation of a collective quote instead of individual quote, even though quotes are labelled to an individual participant. For example, if Participant 4 said Dexter *"not easy. It requires training*..." and participant 1 and participant 2 also expressed the same conclusion, from that point the best individual quote is used. This is in line with what Ponterotto (2006, p. 543) said: "Thick description of social actions promotes thick interpretation of these actions, which lead to thick meaning of the findings that resonate with readers."

3.9 Data Quality Control

Data quality control is described as the set of criteria used to determine whether a study is reliable. The set of criteria used to verify whether a study is reliable are as follows: credibility, transferability, confirmability and construct validity (Anney, 2014).

3.9.1 Credibility

From the standpoint of the research subject, the credibility criterion is used to determine if the qualitative research findings are trustworthy or plausible (Anney, 2014). The real perceptions of the participants were obtained during the data collection process, which is a legitimate key to assess the credibility of the results. To establish credibility, the researcher had full access to the participants beforehand, and was able to communicate with them by email and mobile phone throughout the study (Anney, 2014).

3.9.2 Transferability

This generally pertains to the individual generalising responsibilities (Creswell & Creswell, 2017). As a result, it was vital for the researcher describe the study's research environment and key assumptions in order to show how the findings might (or might not) apply to other scenarios (Creswell & Creswell, 2017). The researcher, in this study, has thoroughly described the case context; clarified the concepts used from the theoretical model; and in the analysis chapters provides evidence to support any thematic claims made. This will make it easier for another researcher to judge the degree to which findings can be transferred to another setting.

3.9.3 Confirmability

The extent to which results can be validated by other researchers or readers is referred to as confirmability (Anney, 2014). To guarantee that the results of this study were accurate, a list of instruments and activities that needed to be performed was created, and this list was referred to on a regular basis to confirm that they had been done. A supervisor was assigned to this study to guide, oversee, and provide peer review of the researcher's work. This enabled the researcher to evaluate data collection and data analysis techniques, as well as make judgments regarding the likelihood of bias or distortion. The researcher followed these protocols when conducting this study.

3.9.4 Construct Validity

The degree to which the study investigates what it promises to investigate is known as construct validity. This was accomplished in this case via a clear chain of evidence that allows the reader to follow the investigator's progress from research topic to conclusion. The reliability was maintained by ensuring transparency through careful documentation and continual references to the qualitative case study research database through illustrative quotes (Farquhar, 2012).

3.10 Limitations of the Study

The research study focusses on the fact-checking industry which is in its early years. Therefore, not many empirical studies exist about automatic fact-checking using big data systems. This limitation will prevent the study from providing a comparison within the field of digital data. However, given the rapid growth the industry is showing, the existing gap will decrease rapidly in a reasonably short time, with more countries starting to see the negative effects of fake news and misinformation.

3.11 Conclusion

The case study used a qualitative approach that endorses interviews, observations, and document analysis as research instrument. The study used the NVivo 12 tool to interpret the qualitative data. The participants signed informed consent and the protection of anonymity was assured. Ethical clearance was received for the study. The next chapter will discuss the analysis of the data collected from this process and will provide a clear picture of the current situation when implementing a big data system.

Chapter 4: Qualitative Data Analysis

4.1 Introduction

The data collection and data preparation were handled in accordance with the thematic categories and research questions, which were aligned (Appendix D: Alignment Matrix) with the conceptual framework as new questions were added to establish more in-depth understanding of the phenomenon being discussed.

This study explored the question of how big data analytics capability influences the monitoring of the factual quality of digital media in southern Africa, using a cloud-based platform called Dexter (introduced on page 25, Figure 1: Digital Data Genesis Model.). The results obtained from the interviews are explained in more detail in the following sections. Sections 4.3 demonstrates the findings of stage 1 of the process, the development of the big data capability. Section 4.4 and Sub-Sections 4.4.1 to 4.4.5 will address the second stage of the process, i.e. the development of the digital knowledge. Finally, Sections 4.5 to 4.6 will present Stage 3: the involvement of the strategy maker and their perspective of the process and the factual quality of digital media. The case study will demonstrate the different technologies in the different stages outlined by the model that are involved in building the cloud-based platform, Dexter. However, before looking into the analysis, it is important to understand the background and history of the Dexter platform, and the different participants in the study.

4.2 Case Study

The case study is based on a big data initiative called the Dexter platform, built by Open Cities Lab for one of their partners, which will be referred to as 'Partner A' in the study. In an age when fake news is a growing concern, the usage of digital media necessitates data collection methods that exhibit exceptional honesty, responsibility, equal representation, and a lack of personal motivation. Because such measurements are inherently complicated, they should be derived, calculated, and implemented with caution. The ability to capture, process, analyse, and derive insights from digital material is at the core of constructing such accountability measures. To achieve this, Open Cities Lab developed for their Partner A, a cloud-based platform called Dexter, built on Flask, with a full Python stack, to automatically upload digital media data from several countries. After that, the system uses natural language processing to highlight important sources, topics, and entities, allowing advanced analytics to be done on the resultant data. This research study will explain the platform's philosophy, data flow, and cloud deployment, as well as how it may be

used to create accountability and provide insights into South African media. The model used will further provide a detailed roadmap for constructing a Python-based big data system, and the analyses that can be performed with it.

Open Cities Lab is a non-profit civic technology lab that implements and advocates for open data, open government, and civic technology through projects, events, workshops, and data quests. The company is the first city-focused open data/civic tech lab in Africa. The company works with government, citizens, civil society and the media, to democratise knowledge and enable informed decision-making and evidence-based planning in all sectors of society (Lab, 2021).

4.2.1 Participant Responsibilities

Open Cities Lab works in parallel with other businesses and NGOs on a sub-contracted basis on some major projects. Below is a detailed summary explaining the history of the different participants from the early development of the Dexter platform, how their roles changed during the project time span, how eventually Blackbox was sub-contracted, and what was expected as a requirement for participating in the Dexter project. The table, below, outlines the participants' company position and role(s) in the Dexter project:

Table 2: Summary of Participants.

Participant	**Position**	**Role(s)**	**Participant Summary**
P1	Data scientist	Data science lead	I started a start-up with Richard, the founder of [Open Cities Lab]. And then when I came back to South Africa, we continued work on our start-up, but then started focusing on the non-profit, Open Cities Lab, which used to be Open Data Durban. Originally, my role was lead technologist and developer. But over time, I've progressed now to become the data science lead.
P2	Senior lead developer	server-side to front-end development, including database architecture, information architecture	Currently manages the pipeline of the majority of our systems at Open Cities Lab, and this is everything from server-side to front-end development, including database architecture, information architecture flow across the system. Current tech stack is React.js with Python Flask on the backend."

P3	Full-stack developer	front-end and back-end on the desktop project	Developer at Open Cities Lab.
P4	Creator director /Designer	user interface designer	Job role included user interface and user experience designer at Open Cities lab. Year 2018 sub-contracted the designer role to a software company BlackBox.
P5	User experience engineer	Front end stack senior Developer, CEO	CEO of Black Box, who have been contracted to Open Cities Lab since the beginning of 2018 to build their front-end stack/user interface.

4.3 Stage 1: Digital Data Genesis Generation Capability

This stage deals with the back-end component of the system, including how it is managed. In establishing the big data system, the data scientist (P1) and lead developer (P2) played a key role in building algorithms, data architecture, information architecture and the type of big data analytic to be implemented, as well as third-party tooling integration to ensure the successful implementation of a big data project. This section will explain the results obtained regarding the conceptual framework 'big data capability', which carefully examines the foundation of the big data platform Dexter. The following Figure 3 illustrates the fundamental architecture data flow of the Dexter platform.

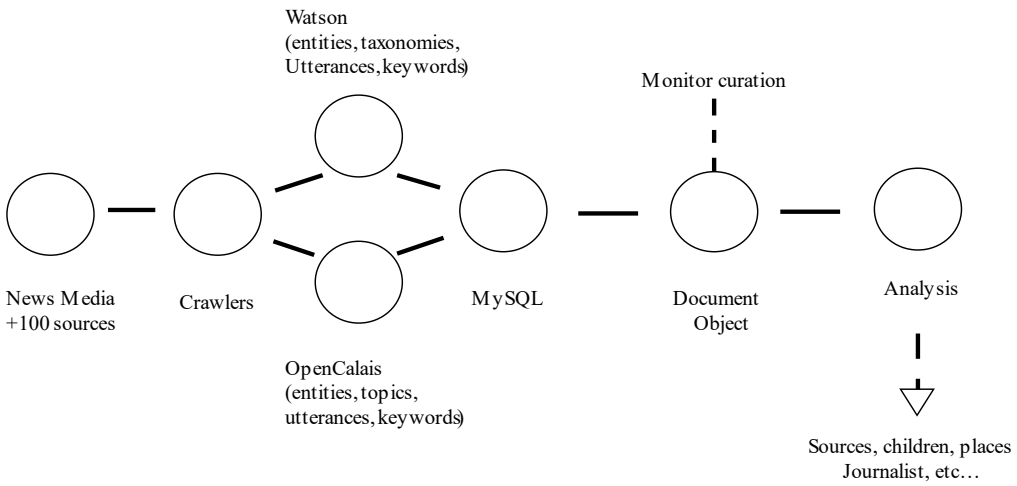

Figure 3: Dexter Data Flow.

The information architecture (data flow) of Dexter begins with a look at a document to find two primary things of interest: sources and entities. A source is somebody who has been quoted on something, either named or unnamed. An entity is something that is being spoken about, for example people, organisations, or places (Adendorff, 2017).The document is generated from more than 108 news media sources, with varying publication based on the size and reach, which are then data scraped using bots also called data crawlers that serve the purposes of retrieving information from the media sources. i.e., the type of data extracted by data crawlers (Date of publication, article text, URL, news publication, titles, authors and so on). For example, news24 a well-established media house may publish 200 articles a day and a local news media may publish four times a day, causing varying velocities in the news media being uploaded into the Dexter platform. In some cases, for instance in an election year, the Dexter platform would experience a high velocity of news being uploaded.

The digital raw text extracted by the bots (data crawlers) is uploaded into natural language processing systems, IBM Watson and Open Calais (Adendorff, 2017). After that, the next step is to send the processed data to MySQL database as a document object as the fundamental unit for analysis.

The Dexter platform is a full Python stack system, meaning that back-end to the front-end is built using Python code (Adendorff, 2017). The case study examines the Dexter system as the context for answering the research question and the table below, with Python infrastructure terminology and descriptions, is created to assist with the explanation of the system.

Table 3: Back end Technology Used.

Terminology	Description
Flask	This is a web micro-framework, built on Python, that is running on EC2.m4 Large for Dexter (Flask, 2021)
EC2.m4 Large	Amazon Elastic Compute Cloud (Amazon EC2) is a web-based service that provides safe, scalable processing capability. Its purpose is to make developers more accessible to web-scale cloud computing (Amazon, 2021).
Object-relational mapper (ORM)	This is a library of code that automates the conversion of data from relational database tables to objects that may be used in application code: e.g., SQLAlchemy and Alembic.
Tasks (Celery)	Celery is a simple, adaptable and dependable distributed system that can process large volumes of data while also providing operations with the tools they need to keep it running.
Natural language processing (NLP)	NLP is a branch of computer science – specifically, an artificial intelligence branch – concerned with computers' capacity to read text and spoken language in the same way that people can. NLP combines computational linguistics, which is rule-based human language modeling, with statistical, machine learning, and deep learning models. For natural language processing NLTK (Natural Language Tool Kit), TextBlob is used in Dexter. The NLP systems IBM AlchemyAPI (Watson) and OpenCalais are used in Dexter.
Web scraping	This is the procedure for obtaining data from the Internet, e.g. the Beautiful Soup Python package (Boppana & Sandhya, 2021).
Templating pyHAML	This is an implementation of HAML for Python. HAML (HTML Abstraction Markup Language) is a templating system that makes HTML cleaner, since a web document should not contain inline code. HAML offers HTML the freedom to contain some dynamic material. Similar to template systems like eRuby and other web languages like PHP, ASP, and JSP.
Analysis	Pandas, an open-source Python library, is used for data analysis and manipulation. It is built on top of the Python programming language. Pandas is a quick, powerful, versatile, and easy-to-use data analysis and manipulation tool. Additional analysis is performed using Metabase to provide business intelligence(BI). These are tools that allow you to pose questions about your data

	and present the responses in understandable formats, such as a bar graph or a table.
Heroku	Heroku is a cloud platform that allows businesses to create, distribute, monitor and scale apps. It is a quick method to get from a concept to a URL, skipping the associated infrastructure constraints. Heroku is a platform as a service (PaaS) provider. (Heroku, 2021).
Docker Containers	Docker Containers are an abstraction that groups code and dependencies together at the application layer. Multiple containers can run on the same machine and share the operating system kernel, each executing as a separate process in user space. Containers take up less space than virtual machines (container images are often tens of megabytes in size), can run more applications, and require fewer virtual machines and operating systems than virtual machines (Docker, 2021).
Amazon Web Service	This is the world's most comprehensive and broadly-adopted cloud platform, that offers infrastructure as a service (Amazon, 2021)
Newstools	This is a toolbox for African journalism observers. It is an open source, easy-to-use toolbox of analytical software to keep media institutions honest, increase media professionalism, and encourage great journalism (Newstools, 2021).

The table terminology relates to the back-end technology used to build the Dexter platform. The programming language Python is used, in a stacked Python approach, with supportive open-source libraries and a cloud-based architecture with Heroku that runs on Amazon Elastic Compute Cloud. This explanation of the technology used to manage the digital data, helps to illustrate the complexity of the infrastructure and processes required to establish such systems.

In Figure 4, an outline of the Dexter platform is shown using the digital data genesis model to provide an illustration of how each technology is used relevant to the big data value chain. The diagram separates the technology used to develop Dexter into three stages and each stage will be discussed in different sections.

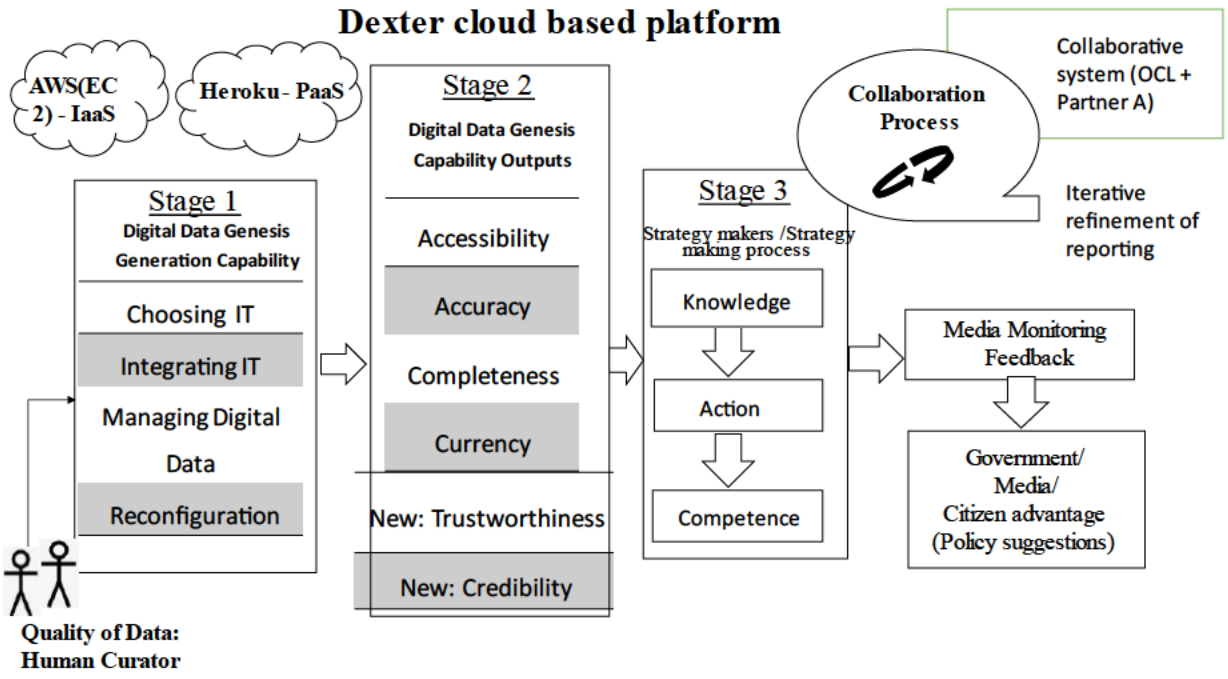

Figure 4: DDGC Dexter Cloud based platform.

Below is the first stage diagram Figure 5 outlining Stage one of the model with the associated technologies used to build the big data capability for Partner A. Stage 1 deals with understanding the backend of the Dexter platform. Backend development refers to web development operations performed on the backend of applications. Backend development, as opposed to frontend development, focuses on server-side web application logic and integration, as well as activities such as writing APIs, generating libraries, and working with system components. A study by Stančin and Jović (2019) suggests that, to be able to implement a working data flow architecture successfully, which incorporates different types of libraries in the fields of Python core, data preparation, data visualisation, machine learning, deep learning and big data, a lot of work needs to be done to ensure compatibility. In a big data backend development developers must create code that allows a database and an application to communicate. The challenges are on controllability (in a build, and when deploying) and the granularity or level of detail in the data of a new proposed system. In other words, the petabyte of data being sourced requires steps to be followed by the developer/ data scientist to ensure secure database and application communication during deployment and builds of the big data solution.

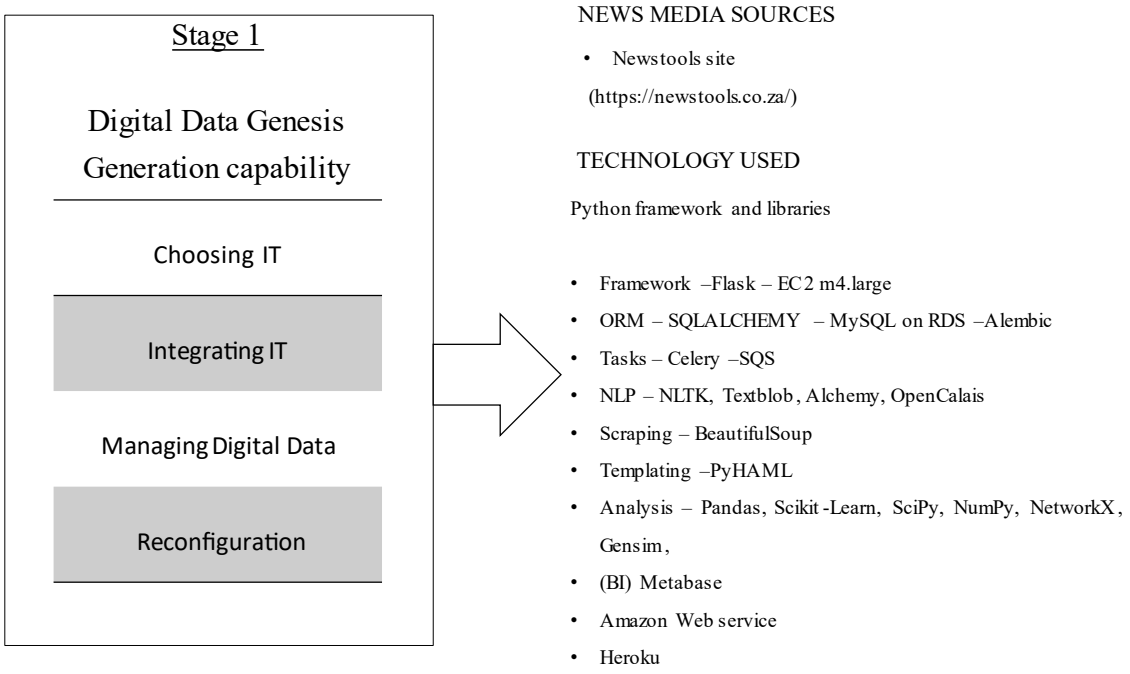

Figure 5: DDGC Stage 1 Developing the capability.

Furthermore, the challenge of defining the level of detail in a set of data must be considered e.g., the extent to which time needs to be delimited as minutes, hours, days, weeks months, and years. In a big data solution with a large volume of structured data, unstructured data, and semi-structured data this becomes a multifaceted decision. Therefore, the use of docker technology is important especially when looking at cloud-native development. Docker is a containerization platform that is open source. It allows developers to bundle programs into containers, which are standardized executable components that combine application source code with the OS libraries and dependencies needed to run the code in any environment. Containers enable deploying distributed programs easier and have become more popular as companies move to cloud-native development and hybrid multi - cloud settings (Dessalk et al., 2020). Therefore, the process of extracting data is an important part of controlling the quality of the data entering the system (Hu et al., 2014). As pointed out by P1 in his comment on how Amazon Elastic Computer Cloud (information as a service), Python stack and Heroku (platform as a service) are structured together by using docker containers: "*Dexter runs as a flask application inside a docker container. A docker container using the Heroku-style deployment methodology. Amazon web service instance from local. The system is containerised. It's run with continuous integration that will accept or reject and will test the system, deploy before system, confirms system, and will reject the push if something's wrong".* The dependence on cloud solutions eliminates several issues, including that of scaling a platform. It is also important to point out that the use of Heroku as a platform as a service is a preference choice made by the Open Cities Lab team based

on the benefits and ease of use the platform provides. More importantly, the use of docker containers dramatically improves cloud development. Docker containers enable mobility. Containers have the added benefits of running anywhere, providing the perfect environment for continuous integration and continuous deployment (CI/CD) (Docker, 2021).

A comment by P2 explained the benefits of Heroku with further integration with regards to the analytical framework. P2 said that: "…*analytical framework that we work with sits as a separate application, which runs on Heroku*". This is in line with Hu et al. (2014) findings when exploring data acquisition. Extracting data involves several practices to ensure data quality, such as using known data integration techniques like a data warehouse method which allows extraction, transformation and loading of data without compromising data integrity. Hence, the following sections will explore further the different tasks in this stage, as outlined in the diagram of Stage 1 Figure 4.

4.3.1 Choosing Information Technology

Prescott (2016, p. 97) suggested that choosing information technology refers to being able to pre-process and upload digital data at the source. Furthermore, he emphasised that a current understanding of emerging IT technologies that generate and capture data digitally is critical for IT personnel and specialists. They must also be able to apply that knowledge to business problems in order to solve them. When choosing infrastructure, evaluation occurs, as stated by P1 "*The project was taken over for migration since it meets infrastructure criteria" …* "*because we were also using flask, which is what it's built on. We continue to keep it in that format. So that's why any advancements that we've done to it, I've also maintained the use of flask, and it was deployed on Amazon's web services. And I mean, that's a common infrastructure.*"

However, when it comes to the analytical infrastructure, a different approach was implemented, according to P2 "*We've started building an analytical infrastructure. We custom designed the information architecture around working with bundles of data, so collections of data rather than just working with standard simple queries. So, we started thinking about using collections of documents as ongoing, to run analyses on*." In addition, P2 further adds the prerequisites expressed in terms and conditions for taking charge of the Dexter platform system: "*We were chosen to take the project to focus on data science. And we're unique in that space. And so, I think we just kind of relied on ourselves in that regard*." "*And we haven't outsourced any of the data structuring. So yeah, it's always been inhouse*". This is in line with Gökalp et al. (2019) findings that businesses should try to build their own big data architecture and exploit the accessible open-source tools instead of using commercially tools that cover basic scenarios and limit growth; for instance, out of

241 open-source tools on Apache that were evaluated, none were found to be the best solution for a particular component in the architecture.

4.3.2 Integration of Information Technology

In the Stage 1 of the model, integration is the next step. The IT technology that was chosen for pre-processing and capturing digital data must be integrated into key business processes. P1 and P2 both agreed that the Dexter platform provides descriptive analytics and analytical functions. P1 further described the Dexter platform as a descriptive-analytical tool, "(Dexter*) does many tasks. It runs natural language processing on around 2000 articles every night, and then acts as a content management system so that* (Partner A) *can edit the information that gets extracted. And then, finally, we do have these analytical functions that run on the stored articles that run things like natural language processing, topic detection, monitoring keywords over time, that sort* (of thing)". P1 suggested that, in the future, "*We are looking to enhance it to do predictive analytics, but it's tricky. We still need to frame that as a question to be analysed*." For instance, predictive analytics concentrate on forecasting future likelihoods and trends, as compared to descriptive analytic that aims to use, to good advantage, historical data to describe what occurred. Moreover, it is important to note that integration involves a high level of co-operation from all stakeholders. In addition, P1 added that: "*Data was captured elsewhere in the project. Dexter is in partnership with another system called Newstools, which is developed by* (Partner A), *assembled up in Johannesburg. And they do the web scraping…of the news from the various news sites and bundle that into a feed*." P1 explained why the Newstools was created in the first place in his comment reflecting on the history of the Newstools site "*It was specifically designed for Dexter. It was seen that there needed to be scrapers, at least at the time, which were not very common for South African media. And it was deemed worth it to set up a dedicated system for managing southern African and African news scraping. So, it wasn't so much chosen as it was part of the design"*. These findings suggest, after additional analysis of the Newstools site (https://newstools.co.za/), that news feed from source is collected from different news agency sites around southern Africa, then bundled together to be web scrapped by Newstools for uploading into the Dexter platform for machine learning and data manipulation, before analysis to solve a specifically focussed topic that is framed by a researcher at Partner A.

With regards to data storage, which involves securing and keeping data in a safe and protected server or cloud storage server, P2 commented that "*The data is not stored on any local server. It's stored on Amazon web service, and it's stored within a database. And for the Newstools, that data would be stored in the cloud*." P1 also confirmed that "*The whole of the system is on various AWS servers, even the Newstools*

site. Pretty much everything here is on AWS, and each part might be on separate servers, but it's all on there." The document database used is MySQL. However, as reported by P1, a shift is needed for scalability reasons. This supports McElhiney (2018) conclusion regarding using scalable tools to build a big data system. P1 stressed the importance of adopting scalable technologies: "*In time, we're going to want to move the document database away from MySQL, and onto something like a MongoDB or NoSQL with a network type database, just to handle the scaling because it's getting big now. The text has got over two million articles, which makes it quite unwieldy to use with just standard relational databases*."

4.3.3 Management of Big Data

Managing digital data involves controlling the quality and the accessibility of the data. "Information capability of a firm consists of the processes and routines necessary for receiving, storing, and disseminating the digital data" (Prescott, 2016, p. 97). P1 highlighted that the Dexter platform is based on text analytics and explained in more detail the reality of building a system for this format of data: "*Because text data is so big and develops over time, so much, it has to be run in continual, almost real-time. It's a daily process where the processes are running most of the day…and making sure that it's continually streamed through*".

Docker containers are far more lightweight than virtual machines in terms of resource requirements: They share the host kernel, network stack and file system drivers, and generally do not run complex services such as systemd or sshd. Docker containers only run the packaged application and are quicker and easier to use to set up big data environments than virtual machines. Therefore, they have a distinct advantage over big data ecosystems that use virtual machines. However, in some rare exceptions you may need an alternative solution, as suggested by P1. This could occur during periods of high database processing which then requires high memory usage, According to P1, "*we take a data dump, if we need high memory… at which point, we can then flip it into something like a Python disk, or one of the …sort of… the distributed graph, streaming memory approaches. We power up a server, a temporary server, with very high memory because to do some of the analytics, particularly the topic monitoring, …. However, the advantage of having this thing in kind-of a containerised docker type of system means that you can switch out the underlying server quite efficiently.* "

Moreover, the use of docker containers enables easy use of other technologies without compromising compatibility. P2 adds that digital data is uploaded on the Dexter platform using newstools, an easy-to-use toolbox of analytical software that performs "*pre-processing of importing data (news media sources)",* which involves the component of capturing the relevant digital data using bots/ data crawlers. Furthermore,

to ensure quality data P2 adds that "*What we receive from Newstools every day is being monitored constantly by the application. And there are live dashboards to (keep) track of when there is a dip in articles, or when we have received a lot more throughout the day. We make sure that Dexter can handle them on coming in".*

The machine learning performed by the Dexter platformed is on the historical data that is uploaded and not on the adjusted data that includes user inputs performed by the human staff hired to check and edit flagged and incorrect data through the user interface. P1 adds that "*Machine learning isn't checking what the user is doing and suggesting new alternatives. But what it does is look at the media. For example, it does machine learning to determine what are the current topics in the news, which are done algorithmically*," and further commented that, *"There's machine learning in running the analysis, because they're often machine learning-based analyses*".

Machine learning happens after the newstools data capturing phase. Third-party machine learning natural language processing (IBM Watson and Open Calais) is used to perform deep learning. These extract meaning and metadata from unstructured text data. P1 further adds that, *"The machine learning is in the natural language processing of the articles before they enter the database to extract key individuals. That is where the machine learning happens*." This is further supported by P2: "*The Dexter platform runs a lot of natural language processing to extract out the important information from within the news media articles sources*."

The next component discussed looked at the process performed to provide enhanced insights from the Dexter platform. P1 commented on the work done to incorporate an open-source business intelligence tool that allows for an easy and rich toolbox of presentation tools. The setup of Metabase, according to P1, "*is running on its independent server as part of the analysis. It gives a full business intelligence method. This is an open-source business intelligence system and we have been prototyping it for these recent elections (November 2021) to use for real-time analysis.*" (Ponterotto, 2006, p. 542) says: "A central feature to interpreting social actions entails assigning motivations and intentions for the said social actions."

The Dexter platform involves an additional human intervention architecture component that takes on a checking responsibility according to P1 of "*basically going in and checking that the initial data has been captured correctly. This isn't done in every article, because that's a bit intractable to have a human eye do it*." In some instances, P1 added, when news media is at a high in the calendar year – that usually happens during elections – Partner A would require further assistance with human auditing and checking of the data:

"…*when there are specific times of the year, that is usually around the elections. That is when* (Partner A), *whose system it is, we just are the tech support for it. They will hire a bunch of human monitors, to review articles to make sure that the analyses on the articles are properly cleaned and edited. And we have been building systems to try and identify an auto-clean, but that's ongoing work*". Additionally, P2 noted, that when discussing the human involvement aspect of the system, human capacity is involved in "*data checking*" and further commented that "*It's automated for the most part, as well* as (human intervention) *human curated as well. We always check in and make sure that everything is up to date. If anything falls through, we always backdate and post-process that data. They can now go in and tweak what applications, models, and machine learning are pulled out."* P2 further supports P1's comment on human hired staff involvement in the Dexter platform: *"…(a) simple example is they can tweak sources, they can add sources, they think that was missing from the machine learning algorithms*". This is a minimal human intervention role but is a vital aspect.

The Dexter platform's security is built on existing security solutions performed on the cloud native development used. These include the security solutions provided by Amazon web service and docker containers. This maintains the data integrity according to P2 "…*all handled by Amazon web service. We have set up multiple levels of security over and above the password authentication."* As the system grows, the more complex and procedural it becomes. Furthermore, P1 and P2 both had noted that "*The security is handled by a senior data scientist who has complete access and backed-up by senior developers at Open Cities Lab*."

4.3.4 Reconfigurability of the Platform

The benefit of building platforms for a business that are not constrained by the firm's path dependencies and asset position, is that it allows further reconfiguring of organisational capabilities to match, or create, changes in the market (Prescott, 2016). The use of docker containers provided developers the freedom to explore open-source technologies (Dessalk et al., 2020). The approached used by Open Cities lab enabled the developers to design large and sophisticated applications to be delivered quickly, and reliably. It is important to note, it also allows a company's technology stack to evolve.

This is expressed by P1 in his comment on the Dexter platform ability to be reconfigured: "*Dexter runs as a flask application inside a Docker container…* (Dexter) *runs an app with an Amazon queue, the queue can always be accessed independently. And then analytical framework that [Open Cities lab team] work with sits as a separate application, which runs on Heroku, which mean that it's handling data of article analysis is from a read only copy of the database so as not to interfere with the main database*" In support of P1's

statement, P2 explained the business cases that allow Dexter to be reconfigured: "*It's as flexible as* (Partner A) *would like it in terms of… they have allowed the third party outside access, where people are allowed to view certain parts of Dexter, allowed to do certain bits of analysis. The flexibility is based on them* (Partner A). *And the analysis. They'd like to run OCL (object constraint language*)." In other words, the Dexter platform is easy to manage and test; is loosely coupled, independently deployable and arranged according to business capabilities. This makes it flexible in terms of how it can be configured and used. These findings support McElhiney (2018) suggestion for constructing a big data architecture. As in this study, McElhiney (2018) recommended the Amazon web service as a cloud solution for implementing a big data platform that scales up and provides easy configuration capability.

4.4 Stage 2: The Output of a Digital Data Genesis Capability

The factor, 'factual quality', is the focus of this study, which is linked to the research statement: "The influence of big data on monitoring the factual quality of digital media in southern Africa". For the Dexter platform to provide 'factual quality' it must showcase a high level of high-quality data that are accessible for analysis and use in decision-making by key stakeholders. The analysis capability of the Dexter platform allows for strategic decision-makers to fact-check the quality of media. To drive media accountability, the platform must perform focused analysis on targeted areas. Examples of the focused analysis areas include children; media diversity ratings; investments; and journalist and media house reporting. These were presented by P1 in a talk, hosted by Python conference South Africa (2017), called: "Large-scale media analysis for driving accountability" (Adendorff, 2017). In the case of children, the platform would output information on questions like, "How much of the news is centred on children?"; "Are the voices of children being directly heard?"; "In what light are children being represented?"; and "In what capacity are the children being referenced?" This section will provide insight into the quality of digital data, which is determined by looking at the following quality dimensions: accessibility; accuracy; completeness; trustworthiness; and currency. In the following Figure 6 presents the overall discussion of this section as mentioned.

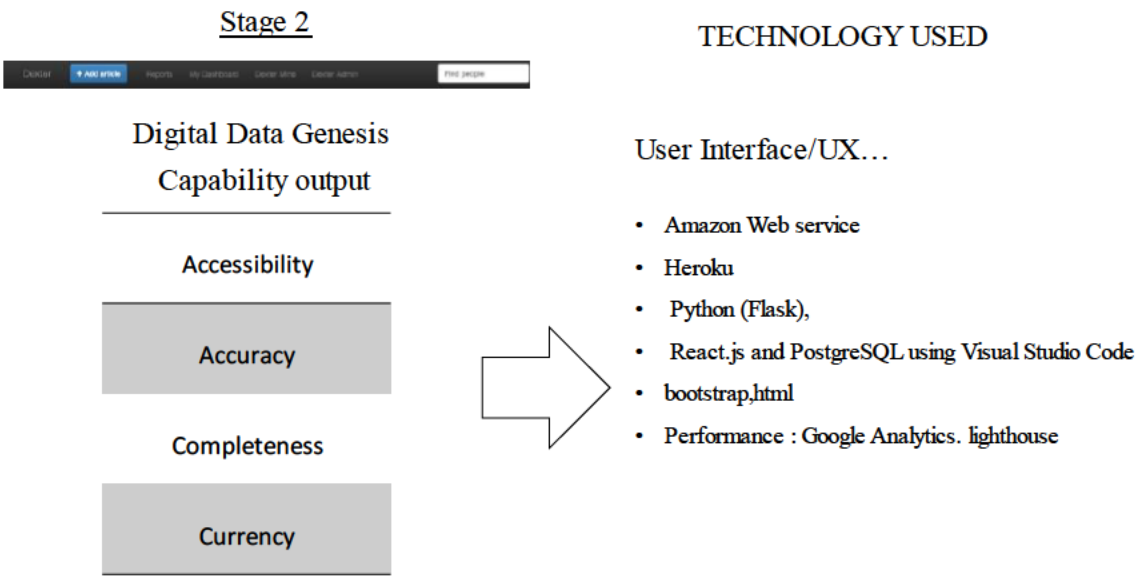

Figure 6: DDGC Stage 2 The Output provided.

Figure 6 illustrates Stage 2 of the Digital Data Genesis Model, and on the right are the technologies used on the Dexter platform, for which a descriptive outline is presented in Table 4:

Table 4: Front end Technology Used.

Name	Description
React.js	This is a free and open-source front-end JavaScript framework for designing user interfaces based on user interface components (React, 2021).
Bootstrap framework	A free and open-source cascading style sheet (CSS) framework for front-end web development that is responsive and mobile-first (Bootstrap, 2021).
Lighthouse	An open-source, automated web page quality-improvement technology. (Lighthouse, 2021).
Google Analytics	A Google web analytics application that tracks and reports website traffic. It's currently part of the Google Marketing Platform as a platform. (Google, 2021)
Rollbar	Rollbar is a bug tracking and monitoring tool for businesses of all kinds that is hosted in the cloud. JavaScript, Python, .NET, Drupal, WordPress, and Pyramid are just a few of the programming languages and frameworks that Rollbar supports.

4.4.1 Accessibility

Accessible data is easily located in the most portable way available for usage, and in a reliable condition (Prescott, 2016). Even if the data is of excellent quality, if it is not made readily accessible to people who need it, the objective of collecting data digitally will lose value and usefulness, because accessibility is a critical element in information use. According to the whole development team, the Dexter platform is only accessed in the web environment. The main reason, agreed by all, but explained most clearly by P2 from a code and development point of view, is that: "*A mobile app for Dexter is certainly not viable. On the CMS – content management system – side of things. There's just a lot of information that needs to be curated.*" Also, from a user experience, engineering viewpoint, P5 added that: "*I think there are benefits to having it as a web tool, as opposed to an app. Just going into an app there are limitations, especially if you want to eventually release it formally through a store, or you've got a website. You've got a better chance of releasing it as a progressive web app*".

According to P4, the user interface of the Dexter platform is "…*not easy. It requires training. It's quite a complicated beast, because of what the outputs push out. Someone from the public generally wouldn't know what Dexter can do."* Another revealing comment by P1 about the user interface is that "*the first version was fairly user-friendly. You do need to know what's going on. It's not immediately intuitive, but it's nice and clean and uses a simple bootstrap. It's easily responsive*." Furthermore, P5 commented on the type of people involved with using the system, "*It's not designed for the public. It's designed for people to work on, for people like journalists, and people* (who) *usually go to articles a lot. For someone who doesn't know any better now – that doesn't have a tech background or doesn't have a background when it comes to journalism – they might find it very… not easy to use*". P2 commented on the level of aptitude required to use the Dexter platform, "*Understanding the way the system has pulled out bits of data, what is the source? What is the utterance? What does each part of Dexter mean, from the post-processing? … requires a bit of training, but for the most part, I think somebody could look around and start to see the power of it straight from the CMS – content management system*."

According to the team tasked with user training, P1 and P4 stated that the system is built for seamless process flow: "*It's very simplistic when it comes to the user interface. You don't want to create any cognitive overload in simple processes, you want a user to feel completely at ease until they get to a point where they want to thank you."* Since Dexter was built for researchers and monitors based at the site company (Partner A), P2 added that: *"There does need to be a training component because … not so much necessarily from the interface, but more just that it's a complex system. It's dealing with complex topics, like article sources,*

and the understanding of what an utterance is, what primary and secondary sources (are). It's very much built for purpose. It does require training, and it's quite nuanced".

According to P5, who specialises in designing user interfaces, *"the forms always have validation. There's a lot of cross-validation in the system to check the process before it makes changes to the database. So, it is a robust user interface. You cannot make a catastrophic error …*" When adding new changes to the database the Dexter graphical user interface mention by P5 allows for editing of article to be performed by human staff hired by Partner A and even the team at Open Cities Lab support team. The same sentiment was understood from the development team who not only built the system but regularly use the system with Partner A.

In this section, the overall performance of the system was spoken about extensively by all participants. The first finding to highlight comes from P5 who commented about the nature of the raw data being used in terms of the size and space of the bytes: "*Because this is such a database-heavy and asset-low application, there are very few assets on here. It's not loading pictures; it's not loading fancy kinds of graphics and images and threads and things like that. It's loading text data.*" This suggestion that the big data analytic is text-based supports Ramasamy and Chowdhury's (2020) finding that most big data platforms use a text data format.

P2 explained the importance of using the latest libraries and development frameworks: *"We've kind of shifted to React.js as a front-end development framework and this allows us to pull in data more dynamically, rather than the usual static HTML."* He added that, since volume is a challenge in processing big data, more emphasis is directed to optimising the time an operation takes to load *"…the performance metrics and tracking. For this, we mainly look at … not so much on the user interface side, but application, processing of the heavy data. I think for the most part the user interface can handle* (it) *and React.js can very easily handle what we push into it. It's more about those bigger queries and running on the server-side of things. "*This is in part what Mani and Fei (2017) said when reflecting on how complex effective data visualisation is for big data analytics, compared to pure data analytics. Furthermore, it was noted by P5 that the use of third-party tools like Google Analytics tools also help in improving the system performance. "*We will use Google Analytics so we can understand, or we can gauge, quite a lot of a user's experience by mimicking the device that they use and testing for that device.*" P1 and P2 emphasised that, overall, "*In terms of performance metrics on the user interface, we've not done a lot of extensive analysis, like looking at click heat mapping and things like that. So, it's much more from anecdotal feedback from the users themselves*."

In terms of ensuring availability of the system, P1 suggested: "(*we) make sure that we put a lot of monitoring on the app to make sure that it doesn't go down."* This is done by a data scientist from the client company, for practical reasons, as highlighted by P5: "*There's a specific rough kind of threshold of articles that we want to be scraping every week. And if it starts falling below the average mark, it starts an alert. So we go and check which articles aren't getting scraped."* He went into detail by explaining that the scraping of articles every week is a huge part of the system. He gave an example of where some level of monitoring of the system was necessary to ensure availability: "*What sometimes happens is newspapers will be sold, and they'll change the holding company. And they'll have a slightly different URL. In this case, (we) will backdate the scrape, and from there it goes and finds those changes and it'll scrape them.*" This is in line with Hallikainen et al. (2020) suggestion of having an analytical culture in the organisation. As demonstrated in the study findings, good performance is stronger in companies with a strong analytic culture, than in companies with a weak analytic culture.

P3 also commented on loading speed and downtime of the Dexter platform by highlighting online tools that check for website performance. Lighthouse, for instance was mentioned, and he added that: "*We use an audit tool from (the) Chrome browser app*". This app runs four categories of audit: it audits the performance; your best practice; accessibility; and good performance. However, this is done from the user interface client-side. On the server-side of the system P1 commented that: "*We've got a cloud monitor. It's like a set of robots that will just monitor whether or not the system is up or not."* From an analytical framework perspective, P1 added that: *"We('re) introducing elements like Rollbar to look at ensuring continual real-time error tracking and monitoring".* The result of work done on the system was described by P4 as enhancing accessibility: *"For instance, loading times… chunking 10,000 articles and getting an analysis out of it will take a much shorter time*". Furthermore, P2 noted that: *"As far as downtime (goes), we've got alerts on Amazon web service checking the app, and Heroku as well*."

4.4.2 Accuracy

Accuracy, in the context of data, is about having correct data that shows a true reflection, and high correlation, between data that was sought to be pre-processed and captured and the data that was generated (Prescott, 2016). However, accuracy is defined as the degree to which a set of measurements is accurate in relation to its actual value (Menditto et al., 2007). The "accuracy" discussed in this section is in relation to what the big data system was initially built to achieve and how it has evolved since its first deployment. According to P4 the Dexter platform was designed "…*at inception was to allow for insights to be analysed*

from the media, and marketing". A concept mentioned a lot in the discussion regarding the extent of the system accomplishing what it set out to do initially was "*iterative processing."* As P2 remarked on the state of continuous development of the Dexter platform: *"It is a nice reporting system. So, it is close to what they wanted. And it is continuously being built and updated."* However, factual data quality is imperative to successful big data solution and in the second interview P1 added that they perform analysis on the incoming data to "*check for plagiarism or journalism.".* Additionally, human staff is hired to check and edit flagged data of articles with problems and inconsistent data to ensure accuracy in the data uploaded.

Moreover, P1 further added on the expansion of the system performing more than reporting big data information: *"Our work is particularly focused on the addition of an analytical subsystem, but we are also currently porting …, rebuilding Dexter in a second version,"* and that the *"current system fulfils its role as being a great catalogue and store of information. We are continuously adding new analytical and output functionality to support new endeavours that the client company does.*" Therefore, it was understood that a big data system is not only evolving but can also meet other business requirements that are beyond the existing system functionalities. In other words, the system had not reached it maturity level but continued growing through expansion to subsystems.

4.4.3 Completeness

From the user's perspective, completeness refers to the data including all the information required for its intended function, but specifically to how comprehensive the information is. When looking at data completeness in the Dexter platform, consider whether all the media information a user requires is available: A user may only require an article title, URLs, date of publication but article edition may be optional for a specific research area of interest. Users must believe that the data is current or up-to-date in order for them to trust it (Prescott, 2016). In the context of media, fake news does pose a problem for newsfeed sources. However according to P1: *"Dexter is both designed to monitor fake news and real news. We would then keep track of which news source is fake or which is legitimate. Most of the news sources that are generated have been selected by hand and chosen for their data sources. They are considered truth media, for whatever that means. We want to know what is being spoken about in South Africa's major media."* Furthermore, on the topic of fake news, P1 added: *"Whether that's fake news or not is something that gets unpacked. But of primary importance is to know what people are talking about. In some circumstances that might be fake news for analysis, but we would know that what we're looking for is fake. It will be selected because it is known to be fake, for example."* This supports a study conducted on USA media. It examined how fake news websites are maturing and gaining the power to set key agendas in the country. They are

intertwined with partisan media, in ways that allow fake news to set the topic and force normal partisan media to try to prove the news is fake (Vargo et al., 2018). P5 commented on the frequency of the analysis: *"The analysis is often done on a user request basis. But in terms of every night, the sources indexed see if any of them have changed over time. This ensures an up-to-date context of media, for instance… if you want to know about the president, you can just say 'the presidency' and then it is going to pull the data from today*" (P5).

4.4.4 Currency

Currency refers to how current information is. It is timely if it was acquired within the last hour - unless new information has arrived that renders previous information irrelevant (Prescott, 2016). According to P4, currency is maintained, in that, as a user" *You have a first visual on the landing page, where you're able to see if that date … (it) should be today's date. It should be like, the exact date that you're on."* This is illustrated by the red circled areas on the Dexter report screen, as shown in Figure 7 below. Furthermore, a health monitor section on the landing page is displaying information based on the algorithms performed on the news data sources being uploaded on the backend, this is presented on the screen shot below in a dotted square. "*There's the health monitor on that same landing page that tells you that it's been adjusting enough articles that everything's working*." For instance, in Figure 7 the landing page user interface shows article problem algorithms performed under the oval grey shape labelled "dashboard of article search". The article problem section includes output of articles of the following categorise on articles coming in: articles missing a topic, articles per country on that given search, articles missing an origin, and so on. Furthermore, P2 added that: "*Every night, our crawlers and our queue on the application runs checks for new imported data and processes*". This ensures the context of the newsfeed is kept current and up to date.

Figure 7 shows Dexter platform user interface landing page showing a news media insight report screen.

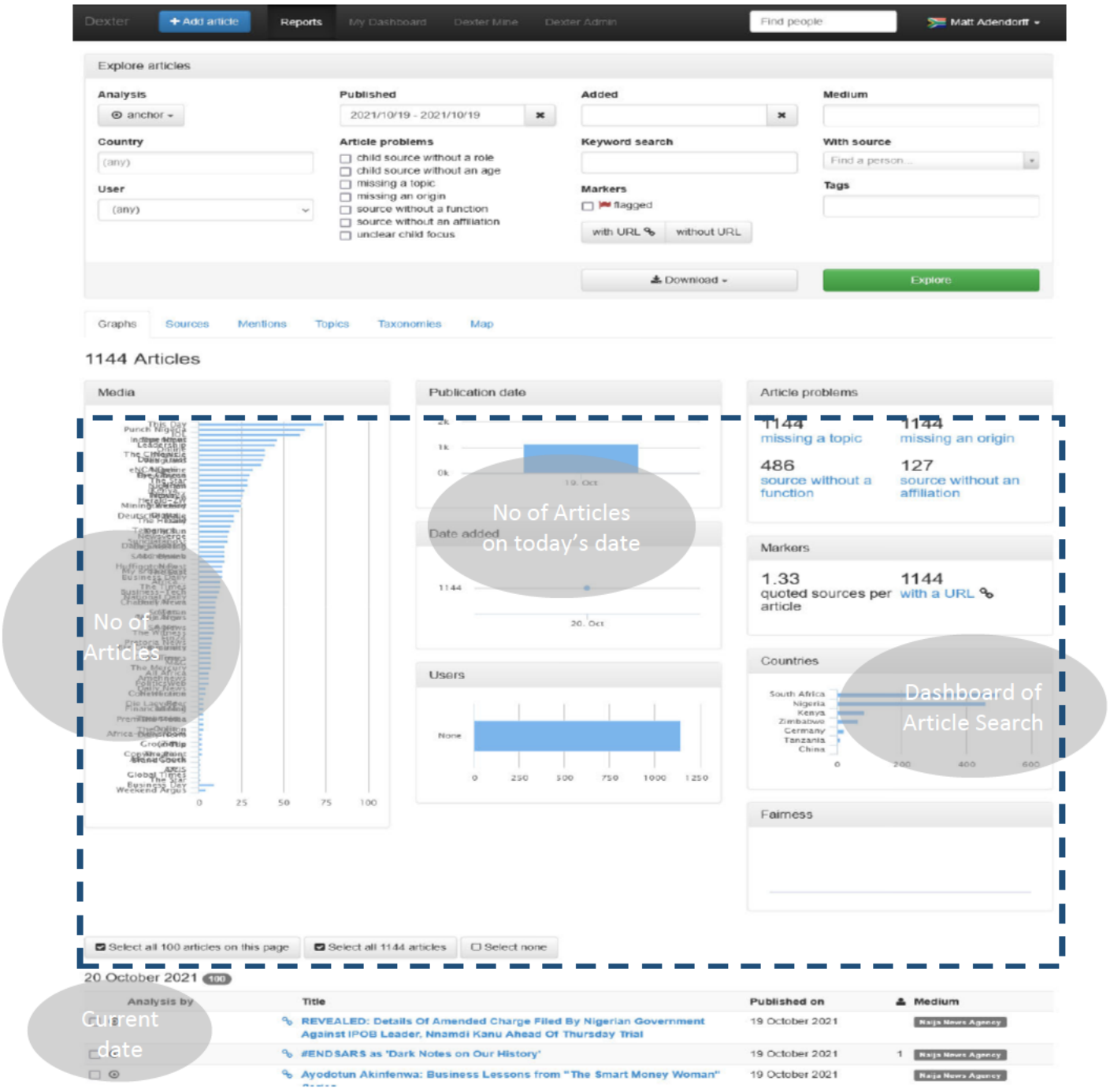

Figure 7: Dexter Report Screen

4.4.5 Credibility

Credibility is a new entry to stage 2 in the research model identified during the study. Individual traits and the perceived quality of the communicator, as well as the news item itself, are used in credibility research, which is based on many aspects such as fairness, objectivity, correctness and believability (Henke et al., 2020). In terms of credibility, P1 added that: "*We make sure that it's the right news source. And we do run analytics on the news sources that come in, to check for plagiarism or journalism.*"

4.4.6 Trustworthiness

Trustworthiness is a new entry to stage 2 in the research model identified during the study. Trust is defined as an actor's readiness to be susceptible to the acts of another actor, the trustee, based on the expectation that the latter will perform a specific activity vital to the trustor, regardless of the trustor's ability to supervise or control the trustee (Henke et al., 2020). According to P1 in order to find a solution to the challenge of trustworthiness of data source an online tool (Newstools) was created earlier as a separate part and integrated into a big data system later known as Dexter. Currently the trustworthiness of the data imported into the Dexter platform is thus ensured by the Newstools application which was specifically created to manage this requirement The Dexter platform performs a flagged operation if it suspects fake news. All the publishers' detailed information are provided to the users at Partner A so that they have full disclosure when wanting to verify trustworthiness. The insights presented to users at Partner A (researchers) allows them "… *to know what they're* (Media) *talking about. And the purpose of the text is to analyse and to start to identify pieces that are wrong and to lobby for it or to say that you are not asking the right people the right questions."*

4.5 Stage 3: Strategic Decision-making

Strategic decision-making depends on the capability of the Dexter platform to provide quality fact-checking of media. High-quality information allows for a dynamic environment that can be taken advantage of to reconfigure operational capabilities in a company. This means when a company is faced with a turbulent environment, strategic decision-makers can develop solutions that affect the business routines, learning mechanisms, co-ordination, and sensing elements of the business. This section will look at the strategic decision-maker, outcomes, and business advantage. Stage 3 is shown in Figure 8 which illustrates the interaction of the Dexter platform system, strategic decision-makers, and the outcome of the analysis.

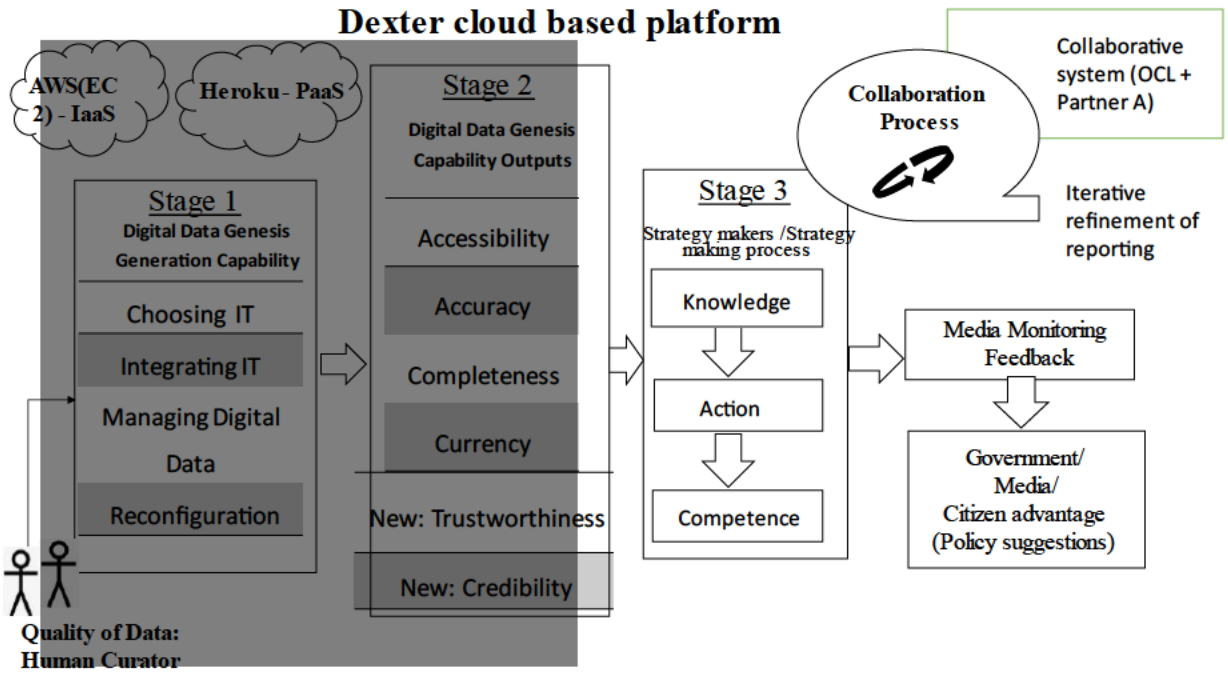

Figure 8:DDGC Stage 3 Decision making and the outcome.

Figure 8 comprises the newly added capability output stated in Sections 4.4.5 and 4.4.6 and seen in Stage 2 of the diagram as credibility and trustworthiness, which were determined in the data analysis of the research.

Alexander (2019, p. 39) discovered that strong collaboration is a key competence in developing a successful big data application. According to P4, a high level of engagement is imperative: *"We work more in developing knowledge and understanding the problem. Where the partner wants something, we validate that what they want is something that is needed, and we build that thing. "*P2 also said that: "*We provide a lot of guidance on the data side of things: On what would be useful and what can be pulled out from Dexter. So yeah, it's just a constant collaboration between us and the client company* (Partner A). *And in this case, they're not a client. They're more like a part of us."* According to P1 actions performed by them at Open Cities Labs is determined by collaborating with Partner A. The team provides technical support knowledge to assist in the determination of the action required to complete a given project or task. This directly effects competency of Partner A to complete a given project. "(*Analysis is) defined through the work of the team for Partner A. They define that through their research engagements with journalistic bodies and research institutions, as well as their continual policy work."* the competency of the users at Partner A is supplemented by the readily available technical support provided by Open Cities Lab who work collaboratively with Partner A staff to develop appropriate queries, big data visualization, project feasibility

consultation and so on. In terms of "action" the model suggests that decision-makers at Partner A can leverage a digital knowledge strategy developed from digital data genesis capability to perform actionable initiatives like media monitoring feedback on government, citizens, policy lobbies and so on.

It was noted by P1 that, from a content management system point of view, "*Trends that the Dexter team wanted … we made sure that those things, by default, would be visible online… like summary information. For example, the top of Google Trends presents themselves immediately (on the landing page) and then you can adjust the research by looking at specific keywords within the data sets."* P2 added that the "*daily dashboard*" (see Figure 7) is used as a tool to share important information about the big data analysis of the new data which has been uploaded and the new articles coming in. It is important to highlight that client staff and Open Cities Lab senior staff are granted access to the above-mentioned user interface functions because "*They are a trusted source and people who are decision-makers*." Currently, the system has limitations "*in terms of someone from the outside looking in and using Dexter and having access. That's not a thing at the moment*" (P2). However, the development team does develop "*bespoke analysis interfaces* (i.e., an analytical subsystem)". For example, this service was provided to Bloomberg and Partner A on request. P1 noted that: *"Our work is particularly focused on the addition of an analytical subsystem".*

The analytical framework of Dexter uses standard reporting tools and advance analytical open-source tools that are integrated into the big data system. The analytical framework was designed to be, in P1's words, "… *user friendly. It's got to be user-driven in terms of how it's structured. And that comes from deep engagement with the team that uses it, to unpack what makes their job better.* "This makes data explorations attractive to all kinds of users looking for insights into what the media is saying.

With regards to the Dexter platform user interface reporting capability, P1 commented that the analysis framework allows standard analyses such as chart types, report types and downloaded summary information in xml format: *"Every one of the analyses that we built into the analysis framework can be downloaded either as a table or as a set of charts. Then from the main text interface, people can download a whole subset selection of data in XML format that includes their basic information about the article and the attached information."* Furthermore, for advance presentational tooling a third-party open-source tool was integrated called Metabase, this is according to P2 *"We have expanded the aspects of the analysis. We have connected a business intelligence tool, called Metabase, to the analysis part, which allows you to create a kind of real-time dashboard building and analysis as an option, a full business intelligence capability"*. Metabase online tool provides beautiful visualizations.

In terms of the Dexter platform allowing customised presentation of analysis, *"It's as customisable as the client company would like, in terms of… They run specific monthly analysis, and we've automated a lot of that … they've run at specific times of the year."* It was added that analysis reporting is "… *customisable in the ability for the user to change the underlying data, which will then update the charts. The different parts can be downloaded as required, but it's up to the user to curate their input"* (P2). and with the added benefit of using Metabase an easy-to-use online tool to generate business intelligence that is customisable using any presentation layout. Furthermore, P4 comment that: "(Dexter analysis reporting) *been geared towards how Dexter and the client's company team work now. "Further* upgrades were projected by P1: *"We're currently upgrading the current Dexter into a new version; in particular, the interface is to be more effective. We want it to be efficient, so people shouldn't have to wait forever. But if they do, we need to indicate to them how long they need to wait. And data needs to be clear and simple" (P1).*

As maintained by the leader of the team, P1, "*The primary users of that are either the client company's main users directly, (or) us when we're doing analyses for them."* The role shift performed by Open Cities Lab of handling security functionality and analysis presentation on behalf of Partner A is in line with previous studies (Björkman & Franco, 2017; Clark & Rodríguez, 2021), who suggested that a shift in the routine and roles is imperative for a successful big data initiative. P4 agreed: "*Sometimes on some of the big projects, like elections, for instance, it needs some serious computing power (query optimization). "*Open Cities Lab would assist with programming more targeted and specialist queries to extract the necessary data for specific situations like elections when more complex calculations and interrogation of data may be required. *"And some of the findings and outputs are just too complicated…* Likewise for developing the business intelligence queries and filters to show the answers to specific queries are beyond the capabilities of the average user *"We only help them with the huge, complicated stuff.*" This collaboration ensures that competency of users at Partner A are substantively complemented by the Open Cities Lab staff assistance with creation of queries, charts, and BI analysis. The actions performed by Partner A are directly boosted by the continuous technical assistance provided by Open Cities Lab, especially in cases were Partner A issues a request for assistance. Furthermore, this ensures a high level of competency in the use of the Dexter platform because the "user" role is a dual role handled by both sets of stakeholders.

4.6 Outcome and Firm Advantage: Factual quality of digital media

A firm advantage that is derived from a big data system that does descriptive analysis of media in the southern region of Africa can reveal inefficiencies and flaws that would be impossible to capture through conventional means. However, there are performance factors mentioned by P1 and P2 that were considered

important when developing and using the Dexter platform, such as speed: *"We'll be able to monitor how many articles are coming in per day, and how many articles are coming in for each media house. And we have certain thresholds that will tell us when not enough articles have come through, or if there's been a blockage in the news feed or a break*." (P1) Another factor evaluated was efficiency: "*In terms of metrics, we're specifically looking at standardised, predictable levels, acceptable levels of incoming media article count."* (P2) Also, some level of benchmarking is said to be applied on the Dexter platform, noted by P1 as, "*In terms of the benchmarking, the speed of queries is one of the big ones for us. So, the time it takes to load subsets of articles, the time it takes to process those articles, as well as then the user interface loading … all the processing of that information…."*.

Clark and Rodríguez (2021) findings suggest that organisations are becoming more efficient, but a possible downside is a potentially rapid, dramatic change of roles. Open Cities Lab is a non-profit, non-partisan organisation that uses action research, co-design, data science, and technology, as well as civic participation, to help cities and urban places become more inclusive (Lab, 2021). They can meet partners' demand for role changes when it resonates with the vision and mission of the NPO – i.e. to contribute to building accountability and trust in civic space. P3 emphasised that: "*…regular user sessions with client/partner companies... understand their uses and their needs and how well it's been functioning,*" and that, "*Feedback from the partner is going to be quite important to tell us if we're on the right track"*.

4.7 Conclusion

The qualitative results obtained from the Open Cities Lab team involved in the development and daily analysis of the Dexter cloud platform showed that they were experts and knowledgeable in the use of big data, reflecting the amount of experience the team has in handling a project of this nature. The evidence shows how data-centric and technically competent the staff were with the concepts and techniques used in implementing big data in a project, and when working with digital data and big data platforms. The Dexter platform was inherited by the team at Open Cities Lab and required some level of flexibility and adaptation.

Big data and front-end developers commented openly with regards to the Dexter platform, given how little is known empirically about the implementation of big data, as shown in the previous literature studies. Furthermore, the big data software developers showed a thorough understanding of the topic and the platform being studied. That sentiment was also felt when interviewing the front-end developers. Even though they are not involved in the back-end code work and implementation, they show a solid understanding of the ins and outs of the system, and the architecture and third-party solutions were evident.

Chapter 5: Analysis and Synthesis

5.1 Introduction

As a result of the findings, which were presented in the previous chapter, a model diagram evaluating the Dexter platform is discussed in-depth in this chapter. This is crucial in responding to the research question; and since the research question was broken down into sub-questions and then to objectives in Chapter 1, a clear representation with that in mind is outlined in detail.

Sections 5.2 and 5.3 look at the objectives relating to the back-end development (Stage 1). Secondly, Sections 5.4 and 5.5 explore the aspects of big data presentation and use, which is associated with front-end development (Stage 2). Lastly, Sections 5.6 deal with the performance (Stage 3) objectives. Finally, a revised model is presented.

5.2 The Dexter Project's Big Data Capability (back-end development)

Big data platforms are made up of IT capability, which means that, when forming a business model, it is a requirement to align technology. Therefore, it is safe to say that it is impractical to develop any big data platform without considering the impact technology will play in constructing it. Karimi et al. (2007) define IT capability as the firm's capability to detect the IT that is needed to meet business needs; to launch and integrate IT; to improve business processes cost-effectively; and to provide long-term maintenance and support for IT-based systems.

Stage 1 of the Prescott Model, as it refers to the Dexter platform, is shown in the diagram in Figure 9. Stage 1 deals with the back-end formulation that includes four aspects: choosing IT; integration of IT; management of digital data; and reconfiguration. These are reflected in the sub-headings of Sections 5.2.1 to 5.2.4 in the discussion.

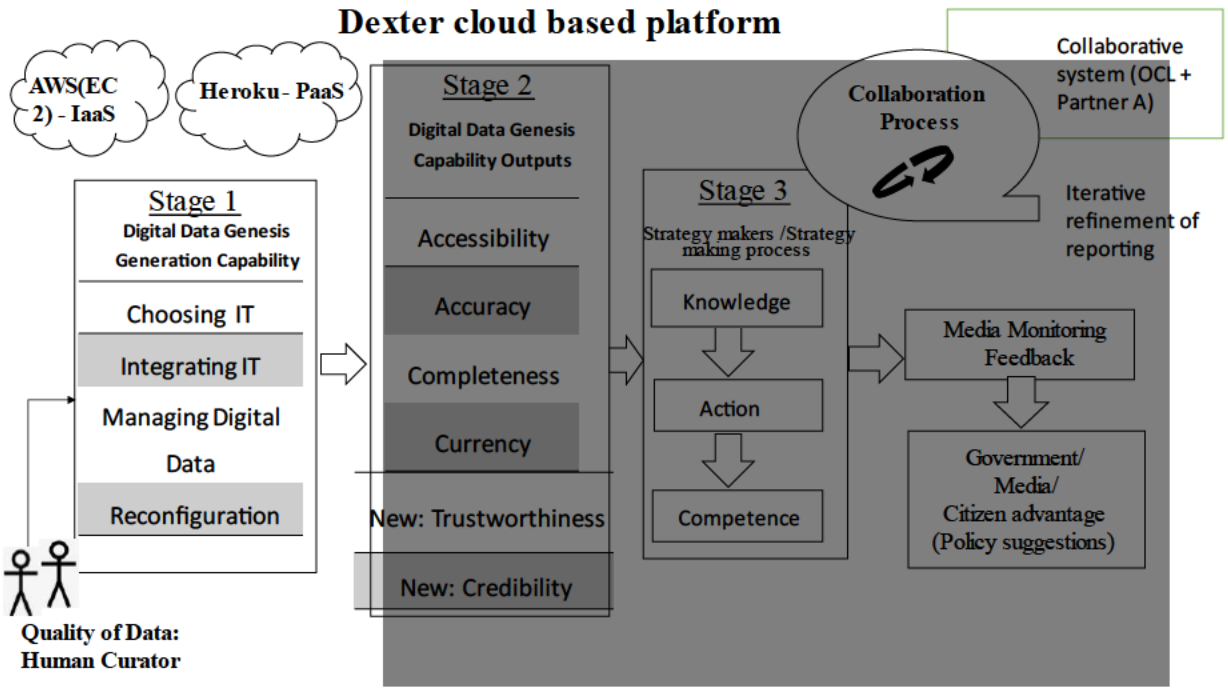

Figure 9: DDGC Stage 1 Dexter

5.2.1 Choosing IT

Since Dexter was inherited, and from the analysis, hardware and software for the Dexter platform were selected, based on existing criteria that match the technology and skills at Open Cities Lab. A process of code migration then followed. The criteria included the use of Amazon Web Service as an infrastructure, and Python Flask. During the development of the Dexter platform, a lot of new technologies were adopted: for instance, a platform as a service Heroku that sits on Amazon Web Services, which is used as infrastructure as a service. As noted by P2, this allowed for a cloud native development approach to building an analytical infrastructure. Much is written about the advances in technology in the different big data value chain phases (Gökalp et al., 2019; Hu et al., 2014; Kumar, 2021; Landset et al., 2015). The use of docker, a well-known platform for managing solutions, is based on container technologies (Dessalk et al., 2020).

It is important to point out that the use of the relational database, MySQL, poses a scaling problem, as previously mentioned (Hu et al., 2014; Strohbach et al., 2016). During the second-round interview, P1 said that in the future they would be looking to moving to a NoSQL database like MongoDB to better manage the large volume of data Dexter emits and processes for analysis purposes. Referring to the process of analysing data, P2 mentioned that instead of standard simple queries, a custom design approach is used that takes in a collection of data for analysis. According to P2, the data structure is built in-house. In some of

the literature, mention is made of encouraging organisations in the big data space to build in-house big data capability. This factors in with the reality that leading technologies in the big data space are built in house using open source solutions by the open source communities (Gökalp et al., 2019; Kumar, 2021; Landset et al., 2015); and with so many projects continually being built and maintained by different communities, the challenge of integrating these technologies, tools and data structures into one eco-system becomes a great task. A moderate risk exists in terms of scalability when depending on open-source solutions, however, it is a rare occurrence and usually happens to libraries with few developers maintaining the open-source solution. In the case of the Dexter platform all the open-source tools used are fully supported. In the following section, a discussion on integration of information technology is explored in the case of Dexter platform.

5.2.2 Integration

Big data integration depends largely on the type of analytics that will be implemented. Open Cities Lab performs descriptive analytics on the Dexter platform. Analytics begins with asking easy questions; for instance, in descriptive analytics you are more interested in what happened, then why it happened. Then more complex aspects are considered, to look at what is likely to happen (predictive analytics). Finally, the most advanced analytics focus on what should be done (prescriptive analytics) (Hu et al., 2014). Dexter works by looking at the news from various sites that is sent to an open-source tool called Newstools, which was designed as a form of data warehouse for the Dexter platform. Newstool thus integrates the data from numerous sources which is then available for use by the Dexter platform. P1 mentioned in a second session interview that Newstools was built with the purpose of serving raw data into the Dexter platform when it was initially built. Both Newstools and Dexter were developed for Partner A. The tool conducts verification and validation of the news feed. The data sources provided by Newstools get uploaded using bots also called data crawls into the Dexter platform. Natural language processing is performed on this stored data, and according to P1 the processed data is stored separately to the uploaded data on Amazon Web Service. Newstools data is also stored on the Amazon Web Service. The developed data integration strategy reduced complications and ensured a smooth flow of information between key collaborators. When it comes to data warehouses, this resulted in a more effective and accurate analysis.

Dexter uses Amazon Elastic compute cloud (EC2.m4 large) for cloud computing and runs Heroku on the web service to allow developers to work in an environment of their choice that is most comfortable for them to build the project. Heroku allows Python developers to deploy, execute, and manage their applications. This application integration is possible according to P1 with Docker containers. Containerized application

development is a process that starts with the developer. Containers and Docker are chosen by the developer because they minimize friction in deployments and IT operations, allowing everyone to be more agile, productive, and faster from start to finish. A host is where a container runs. The host could be local or distant. When an operator launches “docker run”, the container process is isolated from the host, with its own file system, networking, and isolated process tree (Dessalk et al., 2020). This allows for easy integration with the online cloud platforms; and since AWS is among the best in the world in secure and reliable networks, it is a viable option to use. This kind of infrastructure and tool integration allows for a quick-development life cycle and not much time is spent on the setup and configuration of the underlying components to the project infrastructure. The drawback, as pointed by Gökalp et al. (2019), of working with a big data system is that a lot of ‘learning by doing’ is involved. Using open-source tools and addons requires research by the engineers to be able to utilise the tools selected and ensure development will not cripple the project. Substantial learning time required to adapt to using these new tools. It is a dynamic way of working, but the use of an easy setup environment assists the engineers to be able to meet client expectations.

5.2.3 Management of Big Data

The Dexter platform is both automated and involves human intervention. The option of having human capacity perform data auditing of the news feed sources shows that machine learning algorithms from natural language processing systems (NLP) can perform the function of providing a seamless big data platform but cannot be relied on completely without the phase of human auditing and checking. Some form of data review or auditing is required because algorithmic processes are not perfect. It is also necessary to have humans who can react to flagged items to provide input to allow for corrective action if required.

From the analysis, Dexter runs descriptive analytics and analytical functions that run natural language processing, topic detection and monitoring keywords over time on the stored articles in Amazon (EC2.m4 Large). According to P1 and P2, Dexter performs machine learning that results in the analysis. The use of open-source library and tools like SQLAlchemy, Celery, Alembic, OpenCalais and IBM Watson to assist and perform natural language processing and machine learning provides a manageable structure for the Dexter platform. For instance, SQLAlchemy library turns function calls into SQL queries and converts Python classes to tables in relational databases (SQLAlchemy, 2022). Each open-source library and tool serves a different function in the current Dexter platform information architecture.

According to P1 there is a growing need to change the database structure in the future. This is due to the increasing volume of the data. The changes would require changing the relational database MySQL database structure to NoSQL database like MangoDB. In such a case, libraries such as SQLAlchemy would be replaced since the library is focused on relational databases. Additionally, a new open-source business intelligence platform called Metabase was introduced to expand on Dexter platform existing analysis capability.

Human invention is said to be minimal on the Dexter platform and happens as a form of checking and editing to ensure data integrity. Their role is to ensure the factual correctness of the news media sources. The system checks that the output is aligned 'factually' with the input sources; in other words, that the news media source data has in no way been corrupted or manipulated; but it does not determine the 'truth' of facts. This is also true when editing and making changes the system may have missed. Issues of this nature, which may constitute fake news, can only be determined by a human judgement call in the current version of Dexter. According to P1, they are expanding the Dexter platform and adding additional features based on Partner A specifications. However, at this point machine learning, natural language processing is not fully operational without human intervention. A study by De Oliveira et al. (2021) demonstrates the use of Natural Language Processing (NLP) to identify fake news. This is current research focused on the success of various ML techniques. From an operational perspective OCL would first need to consider how the current academic knowledge in this area has developed before they can determine the potential relevance of this information and how it might impact future iterations of the platform

The security of the Dexter platform is based on Amazon Web Service (AWS), a cloud-based service, and is not based on Open Cities Lab premises. AWS is among the top cloud platforms that enterprises are utilising to build their robust big data and analytics solutions around the world. With regards to security, AWS provides the best security features to safeguard against instances of hacking and sensitive data (Amazon, 2021). Moreover, the authentication is easily maintained and distributed to the authorised personnel. In the case of the Dexter platform, these are the data scientist and the senior developers. The management of a big data system is also considered when choosing IT and deciding how it will be integrated. Therefore, the four components in Stage 1 do not operate independently but are inter-related as the cloud-based infrastructure has built-in features that govern how access is granted and information security is handled, thus considering the interoperability of components.

5.2.4 Reconfigurability of the Dexter Platform

Configuration of resources is based on dealing with big data integration and interoperability challenges. Among the most important challenges are accommodation of the sheer scope of data, data inconsistency from heterogeneous sources, query optimization at each level of data integration and mapping components (existing or new schema). Additional challenges include dealing with inadequate resources for implementing data integration (financial resources, skilled professionals) and scalability of the platform. However, if the configuration of both hardware and software is performed in a way that allows flexibility in terms of configuration it then allows for reconfiguration if this should be required to improve performance or to allow for additional functionality. The purpose is achieved in terms of building a platform that is independent of its key components, by using docker containers. Software container technologies are new, and studies conducted on containerisation software suggest that proposals for, and experiments in, the use of containerisation of big data systems are promising (Corodescu et al., 2021; Dessalk et al., 2020). This case study demonstrates the advantages of using containerisation in a big data system. According to P1, Dexter runs as a flask application inside a docker container. The use of docker containers enables easy deployment to the cloud, more control, granular focus and a microservices-based method focused on efficiency (Docker, 2021). In other words, dockers are tools designed to make it simple to create, deploy and run applications by using containers. Containers permit a developer to package up an application with all the parts it needs, such as libraries and other dependencies, and ship it all out as one package.

The flexibility to change legacy systems and adapt to new system requirements enables the platform to act dynamically in providing valuable reporting at the right moment. However, a question of scalability in the relational database was highlighted as an area of concern in the future. Currently, the Open Cities Lab team are looking at MongoDB to possibly replace the MySQL database and document object that utilizes collection of documents. The platform is backed up in most parts and is kept on the Amazon Web Service. According to P1 and P2 the back-end development can be reconfigured in minutes if the need arises, due to the use of docker containers. Overall, the Dexter platform seems to be constructed with the latest advances in technology because of its use of docker containers, Amazon EC2, Heroku, etc. This is possible with a talented skilled staff that is open to learning new technologies.

5.3 The Link between the Big Data System and the Strategy used to Monitor the Data

The Dexter platform is a content management system (CMS) that allows strategic decision-makers access, at some level of automation, to the tasks required to effectively manage media content to find out what

people are talking about. The Dexter platform is found on the web. It is an online data tool that makes use of machine learning to digest the news daily. It reveals trends in graphical format (data visualisation) on specific topics; people; places; or any keyword that a researcher uses to extract data from the media reports.

The findings show that CMS has been managed for high performance using React.js as a front-end development framework with the latest library for building user interface components. The quality dimensions investigated, as outlined in the model diagram as Stage 2 (Figure 10), are accessibility; accuracy; trustworthiness; credibility; completeness; and currency.

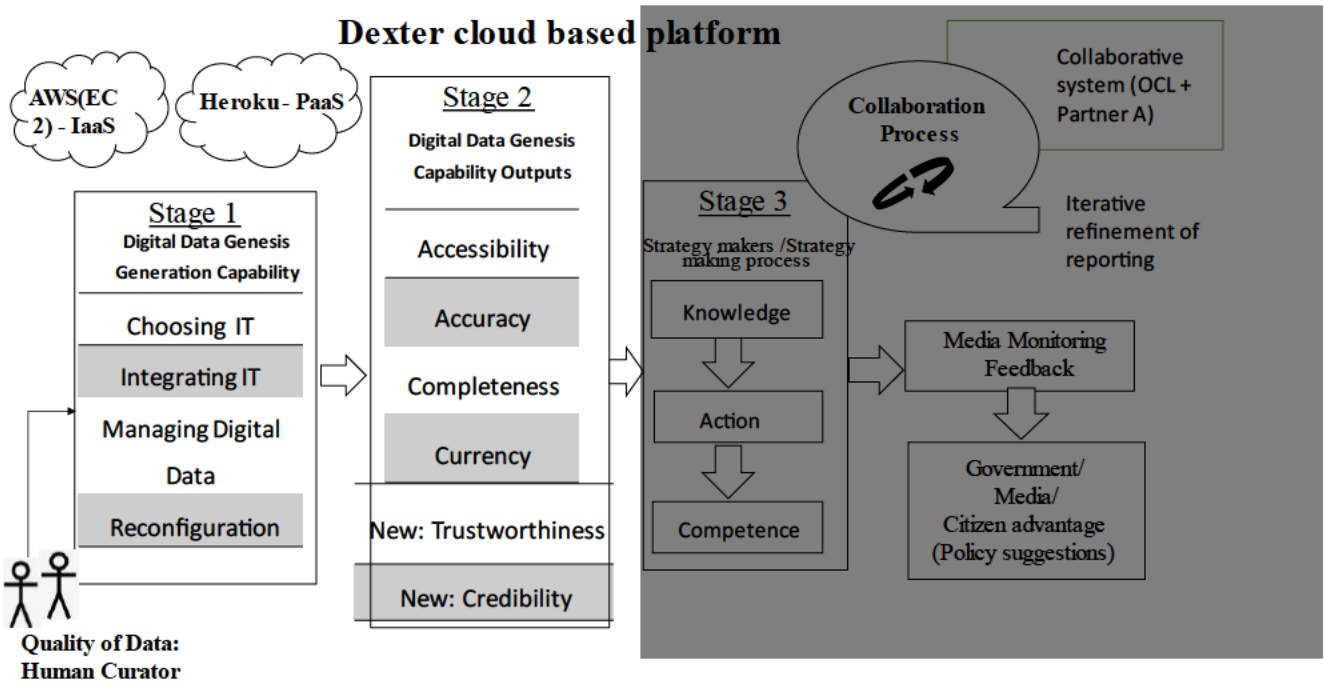

Figure 10: DDGC Stage 2 Dexter

The first four quality dimensions are concerned with output produced by the Dexter platform. The accessibility of the Dexter platform is through a website URL (https protocol) and is easily accessed on the web by authorised personnel. The users are researchers and senior managers from Partner A and Open Cities Lab. The users given access to the website undergo user interface training developed by the Open Cities Lab team involved in the development of the big data platform. Furthermore, the CMS is optimised using performance metrics and online optimisation tools like Lighthouse and Google Analytics. Also, it is worth noting that later in the continual development of the Dexter platform, a decision was made by Open Cities Lab to subcontract the front-end and interaction design development to a company specialising in design and front-end development to further assist in, not just the Dexter system development, but other

systems as well. According to P3, his company was subcontracted in 2018 for the specific front end development specialization the company used in its web development that aligned with the technology at Open Cities Lab. The NPO Open Cities Lab has grown substantially since its inception in 2014 and was expanding its own capability to handle more projects. This approach demonstrates an emphasis on building a human-centric big data system and exposes the dynamics of building a big data system that includes user expectations.

The Dexter platform's accuracy in terms of its ability to provide the information required by its users, requires that the platform needs to be continually updated in accordance with the user requirements (see Section 4.4.2). The platform uploads and processes news media data. Due to the nature of the media data, the tech support at Open Cities Lab works on expanding the analytical process into sub-analytical systems for Partner A's new endeavours.

To ensure that the completeness of the data visualisation represents what is being spoken about in southern Africa's major media. indexing is used to keep the content up to date. Cataloguing was mentioned: for instance, categories like 'fake news' and 'normal news' will be packed separately to be unpacked in the analysis. In other instances, the human staff becomes involved whenever data needs checking and editing to ensure data integrity is maintained. The currency of the Dexter platform is in 'real-time' and is accessible via the user interface and live dashboard displays.

The quality dimension proposed in the model does not consider the data coming into the system, but rather focuses on data processing and data presentation performed by the Dexter platform. Because of this, two additional quality dimensions were added to the model, namely trustworthiness and credibility, to investigate the news media used as data sources, as illustrated in Figure 10. The years of work done on Newstools was to ensure reliable and trustworthy content is uploaded to Dexter. This data has served as research evidence which has led to actions that have lobbied media houses to focus on media inconsistency and to hold the media accountable for what there are saying to the public. This historical context creates trust in the data provided by Newstools and emphasizes the importance of introducing trustworthiness into the quality dimension in the development of a big data system. So far, the system has focused on representing the data as accurately as it is presented in the referenced media sources, but has not, historically, tried to determine if it is 'true'. If news is fake, the users of the system will need to determine this. The system does not automatically check for factual correctness of content but instead provides the ability to flag issues, for example, articles missing a topic, source without an affiliation, source without a function, articles missing an origin and so on., which could be of concern; or it indicates potential abuse,

misuse, or misrepresentation of facts. When it comes to credibility, a standard check is made on news sources for plagiarism on the Newstools website. The credibility of the data in the Dexter system is based on the credibility of their source. There is a distinction made when referring to media houses i.e., a distinction of data from major news media houses as opposed to that from upcoming news outlets. In the Southern African region, a few media houses have established their credibility with their audiences. However, South Africa ranks number 1 out of 164 countries in the world as the most unequal society (Sulla et al., 2022). In addition, with digital technology still out of the reach of much of the population, even in urban areas, it is impossible to give a clear indicator of the potential of the media to drive specific agendas in the Southern African region. Overall, this discussion has lent valid insight into the importance of quality dimensions with regards to the Prescott (2016) model. However, more still need to be understood in this area.

5.4 The Big Data System Implementation Strategy applied by the Technology Support (Dexter)

The fact-checking of digital media by Dexter is done through a web content management system that allows Partner A users the ability to interact with descriptive analytic content and analytical function outputs, as based on different techniques. These include keywords, dashboard displays, etc., that look at what is currently happening in the media. P1 and P2 commented that Dexter is a dynamic content management system that requires continual updating because it relies on media news sources for its input data. The strategy makers are Open Cities Lab staff as they deal with the big data. However, they do not necessarily determine the problem to be analysed, i.e., the topic or the issue that is being considered, and for which the media data must be analysed. The topic or problem of interest is determined by Partner A. Open Cities Lab primarily serves as their technology support. For example, in a scenario where Partner A is looking to lobby or provide direction to the social issue of gender-based violence within the country, the key stakeholders that will interact with the system for a given period will include a project team from Partner A of researchers, as well as a programme manager, tech support consultant and developers at Open Cities Lab. At some point all the above-mentioned stakeholders will interact with the Dexter platform at different levels during the projects either online or through generated reports. The project issue under consideration will be defined by the team at Partner A and technology support at Open Cities Lab will advise and perform tasks aimed at providing insight and reporting on those problems using the Dexter platform on request by Partner A.

Open Cities Labs develops analytical subsystems for Partner A project teams. These subsystems present the generated analysis from natural language processing algorithms based on specific criteria. The strategy

implemented in these cases involves extension of the analytical capability to provide analytical subsystems that expand the availability of factual, quality digital media to Partner A projects. Analytical subsystems are a system that is part of a larger system; in this case, Dexter is the larger system.

The user interface and data visualisation become an important component of the Dexter platform when generating reports for the project team at Partner A involved in a project. According to P1 and P2, Dexter uses visuals that are commonly used, like tables and charts that can be easily understood and allow for a whole set of data, in XML format, to be downloaded. This is in line with Wahyuningsih (2020) findings on the need to build interactive visualisations.

The performance of the Dexter platform has a powerful, built-in analytical framework with data visualisation tools that are simple and common for interpretational analysis. Initially, the landing page automatically shows trends that Partner A project team requests as summary information, before doing any processing of a standard report.

The project team member who had undergone professional training of Dexter platform, go on to check the daily dashboard, a powerful component to get started with when engaging with the platform. Running a query in a big data system, like Dexter, will take more time than running a local database query (Mani & Fei, 2017). However, with the use of React.js and the cloud solution infrastructure (Amazon EC2 and Heroku), Dexter queries are much faster due to the use of this newer technology. P1 and P2 mentioned the monitoring of articles and an article-average target that is set every week which alerts the developers at Open Cities lab if it starts falling below this average.

The project team from Partner A can download reports in several formats and graphical diagrams. With the inclusion of Metabase, a business intelligence, open-source tool, the users can generate valuable reports in real-time. The customisation of querying is not set up by default. However, Metabase (BI) provides support for advanced SQL customisation and using Metabase (BI) requires no experience with SQL to create visualisations. The tool is free and self-hosted (Metabase, 2021). The investment in data visualisation by Open Cities Lab illustrates the importance of having effective data visualisation for a big data system, as was emphasised by Mani and Fei (2017). In conclusion, the platform is built for purpose and allows for extension and proper visualisation which project teams are common with using when interpreting the data.

5.5 Determine the Dexter Project's Big Data Capability Influence on the Performance of Fact-checking/Monitoring the Factual Quality of Digital Media.

Ensuring the correctness of the data requires a human intervention process to check and edit the data to maintain data integrity. The process improves the quality of data and thus the performance of the system. This function is made available to the users of the system to perform this process on the content management system. Use of the user interface builds trust and allows a manual, but tedious, process that is necessary for verifying and maintaining consistency in the correctness of the reporting capability. The engineers are aware of the disadvantages of using natural language processing and built the Dexter platform with that in mind.

The fact-checking of digital media using Dexter is dependent, as reported by the participants, on the following performance factors: speed; efficiency; changing evolution of job roles; benchmarking; and partnership building. Speed and efficiency work is centred on the technology used, as well as human capacity, to find the best solutions to solving big data challenges. The use of a cloud-based infrastructure approach with the latest Python, React.js, and library dependencies allows for speed and efficiency. Effectiveness and efficiency are recognised as important determinants in the deployment of a big data system (Mani & Fei, 2017). An article count of 1500 articles a day, and the fact that a dip in that count causes a red flag notification, is used to track effectiveness and the establishment of standards to allow for benchmarking article query load time and UI loading speed when producing data visualised for analysis (Wahyuningsih, 2020). Furthermore, the collaborative approach that allows those working on Dexter to also work directly with Partner A employees creates a greater synergy between the requests and the delivery of results and reports to meet those queries. The shift in roles, and adding or changing responsibility, is a determining factor in bridging the skills shortage and the overall professional role shift due to technology developments. These are new developments of the existing model; and due to the changes, a new revised model is discussed below, looking at the changes uncovered in the case study.

5.6 Revision of the Prescott Model Based on the Dexter Case Study

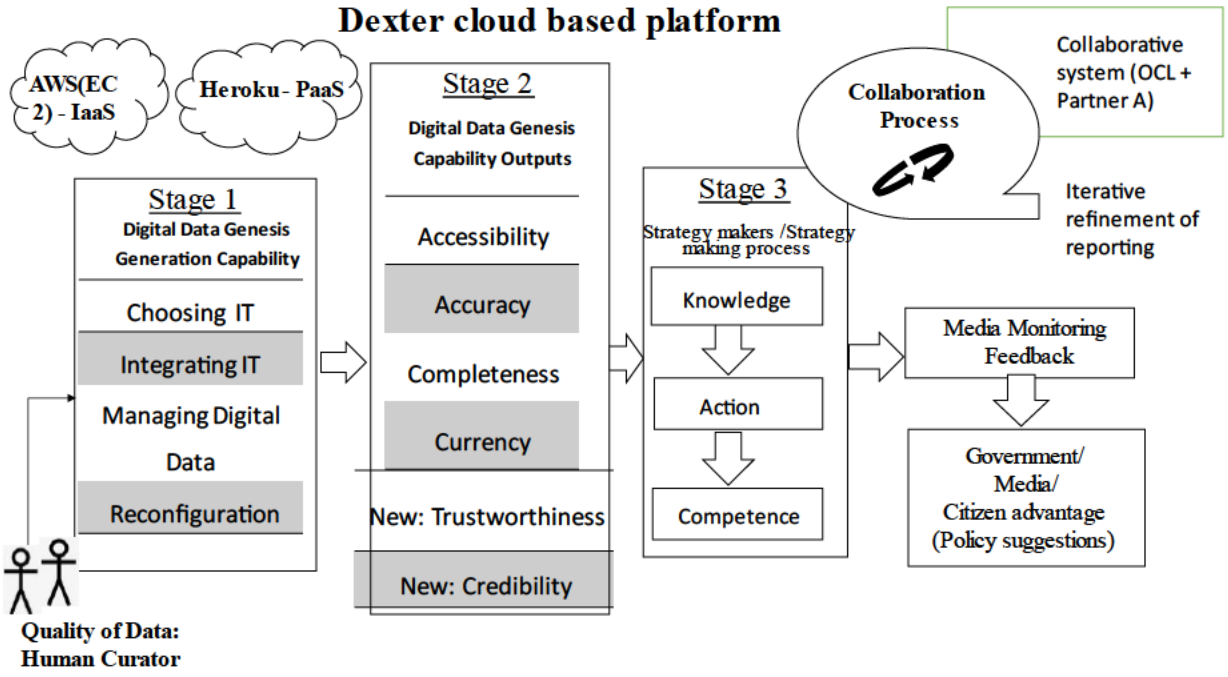

Figure 11: DDGC Stage 3 Dexter.

In the diagram of the model Figure 11 a revised digital genesis data model is presented, because of the findings of this case study. The findings address several areas in the development of a big data platform. The first issue is the information technology used in building the Dexter platform. The details of the information technology used provide a comprehensive understanding, not only of the implementation of big data information architecture, but of the diversity available in terms of cloud-based solutions, in-house structures and development, and open-source tools. This big data workflow is complex. However, with the use of docker containers, the team at Open Cities Lab was better equipped to manage everything without compromising any added development while scaling the big data system. This allows for a big data development that is not rigid and limiting.

Furthermore, human capability is complemented by web application frameworks that provide instant development. In the Dexter platform's case, the flask was implemented, thereby avoiding the necessity of writing extensive lines of code. Python flask is among the most-used Python web application frameworks, and, with pre-defined templates and an intuitive coding interface, developers can save hours of strenuous

coding effort. This, of course, serves as a starting block, and opens opportunities for developers and data scientists to expand on building a big data platform.

The technology support for Dexter takes a different approach, compared to normal tech support. It is more specialised in terms of sharing expertise and guidance. A collaborative and consultative approach is taken by Open Cities Lab. The team are specialists in the use and development of the big data system. Therefore, the model extends to include a collaboration process that entails this specialist technical support and Partner A staff. Partner A staff create the scenarios to be explored for monitoring the factual quality of digital media in southern Africa. An iterative refinement of reporting output is thus produced. The collaboration process considers the complexity and constraints of the big data system, as well as the available avenues the system can pursue in the short- and long-term, in the creation of output.

5.7 Conclusion

In conclusion, the model was discussed in depth in this chapter by looking at all the stages in relation to the findings and the literature in the field of big data. The findings also allowed for the existing model to be updated to include the new development based on the analysis, which will form the basis of the final chapter.

Chapter 6: Conclusion

6.1 Introduction

This chapter will discuss the conclusions drawn from the results of the qualitative case study, by referring to the research question and sub-questions as originally stated at the start of the study. In addition, it will discuss the limitations of this study and avenues for further research.

6.2 Concluding Remarks per Research Question

This concludes a discussion based on the research question that has been broken down into sub-questions that correspond with the theoretical framework used to answer the research question.

The development of any information system requires decisions to be made relating to hardware; software; telecommunication; databases; data warehousing; human resources; and procedures. The development of a big data information system requires a similar decision-making process, with more advanced approaches to the complexity of a big data workflow that not only works but is affordable and can be developed by an existing workforce.

The purpose of the study was to answer a question that explores how the big data analytical process, from choosing IT intentionally, generates or uploads data digitally at the source and integrates that technology into the appropriate business processes; then, managing the capability output data once uploaded and stored for the strategic decision-making team to utilise, within the context of accountability and equal representation of digital media in southern Africa. Because of the numerous challenges in data processing and the management of big data, this study intended to use a model called the Digital Genesis Data Model that tries to simplify those challenges and provides a framework to analyse the processes involved in the development, implementation, and ongoing use of a big data system (Dexter). The goal was to address the research question as phrased, but also to assess the ability of the model to provide a practical framework for analysing such systems. The main research question the case study explores is:

> How does big data analytics capability influence monitoring of the factual quality of digital media in southern Africa?

The following are the conclusions drawn for the research sub-questions in relation to the Dexter cloud-based platform case study:

1. How does the Dexter project's big data capability influence the big data strategy used to safeguard the factual quality of digital media?

The complexity of big data workflow cannot be understated, especially if a big data project is to scale in production. In the Dexter platform, a tool, called docker containers, is used to orchestrate the big data workflow. Given that Dexter uses cloud development, the use of docker containers enables the team at Open Cities Lab to switch out the underlying server quite efficiently.

The Dexter platform is built on Python code and the web application framework that Open Cities Lab uses is flask. That allows Dexter to run on Heroku, which is a platform as a service. Cloud services, in the form of Amazon Web Services (EC2.m3 large) are used to handle Dexter. Another tool was developed prior to the creation of Dexter, on which it is predicated. Newstools became a part of the design of Dexter and sits on Amazon Web Service (EC2.m3 large). Newstools' development was necessary to perform the work of scraping news media sources for input into Dexter.

An analytical framework is then implemented using natural language processing and other tools to move the processed data to MySQL database. The data structure was built in-house by the Open Cities Lab team. The use of a relational database becomes problematic, as emphasised by many researchers. This also became clear in this case study, due to the wait time experienced and the overloading of the database. The plan by the team at Open Cities Lab is to move to MongoDB in the future. The human intervention in this stage of the process will edit and check the data, to ensure data integrity, after which the analysis process can start.

The analysis process involves using several tools like Pandas and NumPy that can analyse big data. The BI tool, Metabase, was used to generate analysis on the big data for normal and advanced querying. Most of the tools used in Dexter are open source. Many advantages come with open-source tools. For example, the team has the added advantage of adjusting the code to do more advanced programming and can customise the code to meet project needs. When looking at the steps involved to orchestrate the big data workflow of Dexter, a clear picture is evident of the level of influence the big data capability has on the big data strategy used to safeguard the factual quality of digital media.

2. How does the Dexter project produce the output used by strategy makers to safeguard the factual quality of digital media?

The output produced depends on the aspects of quality that are closely associated with the overall success or failure of the Dexter platform. The quality dimensions investigated were taken from the model. Accessibility has the most impact of the four dimensions due to the advancement of web development technology in recent times. Accessibility is ensured by successful content management system's graphical user interfaces that can handle functionalities that deal with large amounts of data. Building a platform that focuses on the target audience was a key finding, as well as making sure the users undergo training in the accessibility quality dimension. The team at Open Cities Lab uses responsive web development tools to build the content management system, such as react.js, python stack, amazon web servers for cloud computing and so on. This approach dramatically improves the performance and usability of the Dexter platform. In addition, the system went through performance auditing, using Lighthouse and Google Analytic, to maintain good user interface and interaction performance. This also happens on the server side using robot monitors. Other quality dimensions include accuracy, completeness, and currency of data, all of which are important as well for a successful Dexter big data system. The quality dimensions were studied from the perspective of the engineers who designed and support the Dexter platform and who collaborate with Partner A to provide required output. Despite this, it was clear that the quality factor is important. The model was expanded to include two new quality dimensions: trustworthiness and credibility. Open Cities Lab's years of effort to ensure that solid and trustworthy data is uploaded to the Dexter platform has proved the importance of having consistent data from dependable data sources. According to the data analysis this is possible with Newstools site. However, this alone is not sufficient to ensure a successful big data system. Specialised technical support to handle the complex big data workflow that consists of various steps, with multiple technologies and many moving parts, becomes the determining factor.

A collaborative process that involves Open Cities Lab's specialised technical support and Partner A staff is implemented when performing descriptive analysis or analytical functions. The Open Cities Lab team offers guidance when Partner A launches a new endeavour, since media is dynamic and changes with the trends. This specialised team technical support oversees building analytical subsystems, the updating of Dexter and emergency support. Partner A would approve a proposal and enlist Open Cities Lab's specialised technical support to assess feasibility and deliver a realistic solution for the proposal by using the Dexter platform, which might involve query optimization, new machine learning algorithms, and other tasks. For smaller projects, the Dexter platform's graphical user interface is sufficient for Partner A employees to execute a project without the assistance of Open Cities Labs' expert technical team. According to P2 Open Cities Labs does conducts a feedback session on the usability of the Dexter platform for users at Partner A

on a project basis, especially for beginners in training. The more experience a user gets the more productive and efficient he/she is in completing a tasks or project.

3. What is the assessed performance of the Dexter project, in terms of safeguarding the factual quality of digital media?

The case study identified several key performance indicators that include speed of the system, user friendly user interface, changing and evolving roles, and responsibilities of staff members. The use of analytical functions to suggest keywords of interest, daily dashboards, etc., on the front-end of Dexter, by default, increases the speed of the system. The team at Partner A includes researchers and managers who rely on Open Cities Lab technical support to assist them with aspects of the advanced technical data when working on existing projects and new areas of media fact-checking of interest to Partner A.

The model was most efficient in exploring the backend construct that represent backend and front development of the big data system. The backend component for the Dexter platform is determined by online services both commercial and open-source solutions. The non-profit organization was able to demonstrate adequate resource management capability looking at Dexter financial obligation and skills expertise needed in the development and production of the Dexter platform. The reliance on cloud computing for hardware; software; outsourced third party services; the work on data structure built in-house; and the use of docker containers that enable mobility in the development process. The frontend development was conducted using the latest web development software ensuring high processing speed, responsive and interactive user interface, use of (metabase) advance reporting tooling and monitoring tooling in both the database side and user interface. Thus, effectively meeting all the quality dimensions outline in the Digital Genesis Data Model. Since the model focuses on the quality of what the big data system outputs, for example the up-to-date data used for reporting analysis. Therefore, this study has added two input quality dimensions to the model. Stage 3 of the model was extended, with additional processes that showed the power of dynamic capability through a collaboration process that involved the technical support team and Partner A. The original model proposed a big data system used by one company with in-house data and a self-contained big data workflow process. In this case study, a company has developed the big data system (Dexter) for Partner A and offers training and guidance support for any new development in the big data system that Partner A wants to generate. One reason for this is that the big data system is complex, and Partner A does not have the technical capability to make advanced use of the system without first consulting and asking for guidance to see what is possible and what is not available, based on the data aspects of the system. Thus, the Digital Genesis Data Model, for the most part, helped answer the

study research questions and the study was able to extend the model, to explore other areas of this big data system.

6.3 Limitations

The study was limited to investigating the big data capability derived from the use of a big data platform as determined collaboratively with users, from the perspective of the developers, data scientists and technological support offered by the creators of the system; and not from the perspective of the users who use the platform daily. The limitation arose because Partner A's company personnel, which included managers and researchers who are the users of the system, were not prepared to be interviewed. A future study could benefit from including the user's feedback.

A limited number of studies report on the full scope of the implementation of a big data platform (Björkman & Franco, 2017; Clark & Rodríguez, 2021; Nakov & Da San Martino, 2020; Prescott, 2016) and this has limited the scope for comparison with other implementations. As the work in this area expands it will be possible to gain more in-depth knowledge and further deconstruct and critically evaluate the individual aspects of the process. The Dexter platform performs descriptive analytics and some analytical functions; but other analytical approaches that may include more advanced techniques, such as machine learning, deep learning, and artificial intelligence, were not considered. However, it was mentioned during the case study analysis that there is a possibility of enhancing Dexter to include performing predictive analytics in the future.

The limitation of Prescott's study (Digital Data Genesis Capability) was that the study only looked at one industry. By conducting this study in another sector, this study will contribute to learning how Digital Data Genesis Capability may impact organisational performance in a broader context. The study focused on descriptive analytics and analytical function and is yet to explore other analytical approaches that may include more advanced machine learning, deep learning, and artificial intelligence, which would allow for predictive modelling, as was mentioned in the case study analysis.

6.4 Future Study

The study was able to extend the literature around the use of big data. However, a considerable gap still exists as this study only looked at one additional sector of the industry (fact-checking). The constant reaction to changes in technology warrants investigation into all industries and sectors of society; and more

should be done to encourage society to be more proactive in integrating impactful technology like big data into their civil society behaviours and habits. This has the potential to ensure that people are guided away from harming each other and themselves, especially in terms of vulnerabilities other industries are experiencing. There is an increase in the number of users joining the internet every day They use smart devices to connect and share information; share and read news commentary; and voice their own opinions. This is all uploaded instantly and even streamed in real time. It is a reality, not a future problem. This study, as with prior studies, should assist in understanding big data use and implementation and more. Similar studies will, in the future, be needed to understand big data use in different sectors. In future work, an investigation into the daily users of a big data system will benefit the study. Other important areas for further studies include other types of analytics, quality dimensions, and the change in job roles caused by big data analytics. Finally, evolving technologies cannot be overlooked. An investigation into big data workflow approaches is a key area of interest to resolve the complexity of big data in the future.

6.5 Significance of the study

The study contributes to the fact-checking media sector as well as the strategic planning and practical implementation of big data applications. Furthermore, the study aims to support industries looking into how big data technologies can be designed, implemented, and strategically and operationally used to resolve business and social issues. The research study goes into detail about the complexities of using big data and provides insight into the crucial decisions, procedures and technological tools needed to carry out a successful big data development project.

Additionally, the community in the fields of information systems and data science are provided further understanding of the importance of comprehensive cloud computing solutions, open-source technologies, machine learning, continuous integration, and continuous deployment practices. In addition, the use of the latest content management systems technologies in the development of big data projects are explained. A good decision-making system must also exist, and data quality management plays a significant role and is an embedded part of all decisions and processes.

6.6 Conclusion

By using the Digital Data Genesis Model, this study demonstrated that big data platforms influence the factual quality of digital media in southern Africa, by looking at the process of building and using a big data platform. The model was successful in exploring the capabilities that skilled personnel must possess

to build a successful platform, as well as the process of training users of the system. This includes the assistance required from the technical support team to enable useful analysis of big data to provide data-based intelligence to an audience that has limited technical skills when navigating the big data space.

The technology dealing with big data challenges is always improving. This study provided evidence of the use of Python stack; Heroku; and docker container technologies to handle a big data project; as well as a variety of cloud-based solutions to build, deploy, and manage scaling and hosting a big data project. Big data technologies are continually on an upward trajectory as more companies, institutions, and government and civic society start adopting this form of analysis. Insightful, pioneering data scientists are working to redefine how decision-making is influenced and a lot is expected to change in the future.

> "*Every company has big data in its future and every company will eventually be in the data business" by Thomas H. Davenport (Davenport, 2014)*

References

Adendorff, M. (2017). Large-scale media analysis for driving accountability. https://za.pycon.org/. https://docs.google.com/presentation/d/1akRkw4KmjrrzX5SV3_fyEBZEkDTN5qgU5WFQwtaEuBo/edit#slide=id.g259b9a70bc_1_559

Akter, S., Fosso Wamba, S., Barrett, M., & Biswas, K. (2019). How talent capability can shape service analytics capability in the big data environment? *Journal of Strategic Marketing*, *27*(6), 521-539.

Akter, S., Wamba, S. F., Gunasekaran, A., Dubey, R., & Childe, S. J. (2016). How to improve firm performance using big data analytics capability and business strategy alignment? *International Journal of Production Economics*, *182*, 113-131.

Alexander, D. T. (2019). *Building Big Data Analytics as a Strategic Capability in Industrial Firms: Firm Level Capabilities and Project Level Practices*. Case Western Reserve University.

Alexander, D. T., & Lyytinen, K. (2017). Organizing successfully for big data to transform organizations. *AMCIS 2017 Proceedings.*

Allcott, H., Gentzkow, M., & Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, *6*(2), 2053168019848554.

Amazon. (2021). *Guides and API References*. https://docs.aws.amazon.com/

Anney, V. N. (2014). Ensuring the quality of the findings of qualitative research: Looking at trustworthiness criteria. UDSM Research Repository.

Anspach, N. M., & Carlson, T. N. (2020). What to believe? Social media commentary and belief in misinformation. *Political Behavior*, *42*(3), 697-718.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management*, *17*(1), 99-120.

Baskarada, S. (2014). Qualitative case study guidelines. *The Qualitative Report*, *19*(40), 1-25.

Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). From data quality to big data quality. *Journal of Database Management (JDM)*, *26*(1), 60-82.

Bharadwaj, A., El Sawy, O. A., Pavlou, P. A., & Venkatraman, N. (2013). Digital business strategy: toward a next generation of insights. *MIS Quarterly*, 471-482. Bhattacherjee, A. (2012). Social science research: Principles, methods, and practices. University of South Florida.

Bikakis, N. (2018). Big data visualization tools. *arXiv preprint arXiv:1801.08336*.

Björkman, F., & Franco, S. (2017). How big data analytics affect decision-making: A study of the newspaper industry. Uppsala University.

Bootstrap. (2021). Get started with Bootstrap. Available at: https://getbootstrap.com/docs/5.3/getting-started/introduction/

Boppana, V., & Sandhya, P. (2021). Web crawling based context aware recommender system using optimized deep recurrent neural network. *Journal of Big Data*, *8*(1), 1-24.

Borel, B. (2017). Fact- Checking Won't Save Us from Fake News. *FiveThirtyEight, January*, *4*.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662-679.

Braganza, A., Brooks, L., Nepelski, D., Ali, M., & Moro, R. (2017). Resource management in big data initiatives: Processes and dynamic capabilities. *Journal of Business Research*, *70*, 328-337.

Cheruiyot, D., & Ferrer-Conill, R. (2018). "Fact-Checking Africa" Epistemologies, data and the expansion of journalistic discourse. *Digital Journalism*, *6*(8), 964-975.

Clark, A. M., & Rodríguez, J. (2021). Big Data and Journalism: How American Journalism is Adopting the Use of Big Data. Novum Jus, 15(1), pp.69-89.

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. *Journal of Information Technology*, *30*(1), 44-57.

Corodescu, A.-A., Nikolov, N., Khan, A. Q., Soylu, A., Matskin, M., Payberah, A. H., & Roman, D. (2021). Big Data Workflows: Locality-Aware Orchestration Using Software Containers. *Sensors*, *21*(24), 8212.

Côrte-Real, N., Ruivo, P., & Oliveira, T. (2020). Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value? *Information & Management*, *57*(1), 103141.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications.

Cunliffe-Jones, P. (2020). From Church and Mosque to WhatsApp—Africa Check's Holistic Approach to Countering 'Fake News'. *The Political Quarterly*, *91*(3), 596-599.

Davenport, T. (2014). *Big data at work: dispelling the myths, uncovering the opportunities*. Harvard Business Review Press.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of Big Data based on its essential features. *Library Review*.

De Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V., & Mattos, D. M. (2021). Identifying fake news on social networks based on natural language processing: trends and challenges. *Information*, *12*(1), 38.

Deng, L., Gao, J., & Vuppalapati, C. (2015). Building a big data analytics service framework for mobile advertising and marketing. 2015 IEEE First International Conference on Big Data Computing Service and Applications,

Dering, M. L., & Tucker, C. S. (2017). Generative adversarial networks for increasing the veracity of big data. 2017 IEEE International Conference on Big Data (Big Data),

Dessalk, Y. D., Nikolov, N., Matskin, M., Soylu, A., & Roman, D. (2020). Scalable execution of big data workflows using software containers. Proceedings of the 12th International Conference on Management of Digital EcoSystems,

Docker. (2021). Docker Overview. avaliable at: https://docs.docker.com/get-started/overview/

Dong, J. Q., & Yang, C.-H. (2020). Business value of big data analytics: A systems-theoretic approach and empirical test. *Information & Management*, *57*(1), 103124.

Dosi, G., Nelson, R. R., & Winter, S. G. (2000). *The nature and dynamics of organizational capabilities*. Oxford University Press.

Easterby-Smith, M., & Prieto, I. M. (2008). Dynamic capabilities and knowledge management: an integrative role for learning? *British Journal of Management*, *19*(3), 235-249.

Fadler, M., & Legner, C. (2020). Who Owns Data in the Enterprise? Rethinking Data Ownership in times of Big Data and Analytics. ECIS.

Farquhar, J. D. (2012). *Case study research for business*. Sage.

Fayyad, U., & Hamutcu, H. (2020). Toward foundations for data science and analytics: A knowledge framework for professional standards. Harvard Data Science Review, 2(2).

Fayyad, U., & Hamutcu, H. (2021). How can we train data scientists when we can't agree on who they are. *Harvard Data Science Review, 3 (1)*.

Flask. (2021). User's Guide. Avaliable at: https://flask.palletsprojects.com/en/2.0.x/

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research. Academy of Management Journal, 59(5), pp.1493-1507.

GetSmarter. (2018). *How Big Data is Reshaping Society*. Avaliable at: https://www.getsmarter.com/blog/market-trends/how-big-data-is-reshaping-society/

Ghasemaghaei, M., & Calic, G. (2020). Assessing the impact of big data on firm innovation performance: Big data is not always better data. *Journal of Business Research*, *108*, 147-162.

Gökalp, M. O., Kayabay, K., Zaki, M., Koçyiğit, A., Eren, P. E., & Neely, A. (2019). Open-source big data analytics architecture for businesses. 2019 1st International Informatics and Software Engineering Conference (UBMYK),

Golfarelli, M., & Rizzi, S. (2020). A model-driven approach to automate data visualization in big data analytics. *Information Visualization*, *19*(1), 24-47.

Google. (2021). Welcome to Google Analytics. Available at: https://analytics.google.com/analytics/web/provision/#/provision

Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information & Management*, *53*(8), 1049-1064.

Hafiz, K. (2008). Case study ecmple. *The Qualitative Report*, *13*(4), 544-559.

Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2010). Multivariate data analysis: A global perspective (Vol. 7). Pearson.

Hallikainen, H., Savimäki, E., & Laukkanen, T. (2020). Fostering B2B sales with customer big data analytics. *Industrial Marketing Management*, *86*, 90-98.

Henke, J., Leissner, L., & Möhring, W. (2020). How can journalists promote news credibility? Effects of evidences on trust and credibility. *Journalism Practice*, *14*(3), 299-318.

Heroku. (2021). Learn about building, deploying, and managing your apps on Heroku. Available at: https://devcenter.heroku.com/

Hu, H., Wen, Y., Chua, T.-S., & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *IEEE access*, *2*, 652-687.

Hyett, N., Kenny, A., & Dickson-Swift, V. (2014). Methodology or method? A critical review of qualitative case study reports. *International Journal of Qualitative Studies on Health and Well-being*, *9*(1), 23606.

Irizarry, R. A. (2020). The role of academia in data science education. *Harvard Data Science Review*, *2*(1).

Jaca, C., Rodríguez, M. Z., Tecnun, E. V. V., & Álvarez, M. J. (2016). Exploring information capability and its role in innovation. *Journal of Globalization, Competitiveness & Governability/Revista de Globalización, Competitividad y Gobernabilidad/Revista de Globalização, Competitividade e Governabilidade*, *10*(1), 66-81.

Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors influencing big data decision-making quality. *Journal of Business Research*, *70*, 338-345.

Johnson. (2021, 10/09/2021). *Global digital population as of January 2021*. https://www.statista.com/. https://www.statista.com/statistics/617136/digital-population-worldwide/

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research–Moving away from the "What" towards the "Why". *International Journal of Information Management*, *54*, 102205.

Karimi, J., Somers, T. M., & Bhattacherjee, A. (2007). The role of information systems resources in ERP capability building and business process outcomes. *Journal of Management Information Systems*, *24*(2), 221-260.

King, K. K., & Wang, B. (2021). Diffusion of real versus misinformation during a crisis event: A big data-driven approach. *International Journal of Information Management*, 102390.

Kroeze, J. H. (2012). Interpretivism in IS–a postmodernist (or postpositivist?) knowledge theory. Americas Conference on Information Systems AMCIS2012 Seattle.

Kumar, N. (2021). *Big Data Using Hadoop and Hive*. Stylus Publishing, LLC.

Lab, O. C. (2021). Open Cities Labs. Available at: https://opencitieslab.org/odd/home

Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. *Journal of Big Data*, *2*(1), 1-36.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6*(70), 1.

Lighthouse. (2021). Lighthouse. Available at: https://github.com/GoogleChrome/lighthouse

Lyko, K., Nitzschke, M., & Ngomo, A.-C. N. (2016). Big data acquisition. In *New Horizons for a Data-Driven Economy* (pp. 39-61). Springer, Cham.

Mani, M., & Fei, S. (2017). Effective big data visualization. Proceedings of the 21st International Database Engineering & Applications Symposium,

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10), 60-68.

McElhiney, P. R. (2018). *Scalable Web Service Development with Amazon Web Services* University of New Hampshire.

Mehmood, N. Q., Culmone, R., & Mostarda, L. (2017). Modeling temporal aspects of sensor data for MongoDB NoSQL database. *Journal of Big Data*, *4*(1), 1-35.

Menczer, F., & Hills, T. (2020). Information overload helps fake news spread, and Social Media Knows It. *Scientific American*, *323*(6), 54-61.

Menditto, A., Patriarca, M., & Magnusson, B. (2007). Understanding the meaning of accuracy, trueness and precision. *Accreditation and Quality Assurance*, *12*(1), 45-47.

Metabase. (2021). Metabase documentation. Available at: https://www.metabase.com/

Michalsons. (2020). POPI Commencement Date. Available at: https://popia.co.za/

Mutendadzmera, T. (2015). Big Brands Need Big Data: Investigating How Major Retail Brands in South Africa Use Consumer Data to Shape Consumer-brand Engagement. The IIE, 2015.

Nakov, P., & Da San Martino, G. (2020). Fact-Checking, Fake News, Propaganda, and Media Bias: Truth Seeking in the Post-Truth Era. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts,

Nelson, R. R., Todd, P. A., & Wixom, B. H. (2005). Antecedents of information and system quality: an empirical examination within the context of data warehousing. *Journal of Management Information Systems*, *21*(4), 199-235.

Newstools. (2021). A beginners Guide to "Fake News", dodgy News and News Credibility. Available at: https://newstools.co.za/

Ponterotto, J. G. (2006). Brief note on the origins, evolution, and meaning of the qualitative research concept thick description. *The Qualitative Report*, *11*(3), 538-549.

Prescott, M. E. (2014). Big data and competitive advantage at Nielsen. *Management Decision*.

Prescott, M. E. (2016). Big data: Innovation and competitive advantage in an information media analytics company. *Journal of Innovation Management*, *4*(1), 92-113.

Pröllochs, N. (2021). Community-Based Fact-Checking on Twitter's Birdwatch Platform. *arXiv preprint arXiv:2104.07175*.

Ramasamy, A., & Chowdhury, S. (2020). Big Data Quality Dimensions: A Systematic Literature Review. *JISTEM-Journal of Information Systems and Technology Management*, *17*.

React. (2021). Getting Started. Avaliable at: https://reactjs.org/docs/getting-started.html

Sain, S., & Wilde, S. (2014). Review of soft skills within knowledge management. In *Customer Knowledge Management* (pp. 7-55). Springer.

Saling, L. L., Mallal, D., Scholer, F., Skelton, R., & Spina, D. (2021). No one is immune to misinformation: An investigation of misinformation sharing by subscribers to a fact-checking newsletter. *Plos One*, *16*(8), e0255702.

Sekaran, U., & Bougie, R. (2019). *Research methods for business: A skill building approach*. John Wiley & Sons.

Shozi, N. A., & Mtsweni, J. (2017). Big data privacy in social media sites. 2017 IST-Africa Week Conference (IST-Africa),

Silverman, C. (2016). This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed News*, *16*.

Sivarajah, U., Irani, Z., Gupta, S., & Mahroof, K. (2020). Role of big data and social media analytics for business to business sustainability: A participatory web context. *Industrial Marketing Management*, *86*, 163-179.

Song, I. Y., & Zhu, Y. (2016). Big data and data science: what should we teach? *Expert Systems*, *33*(4), 364-373.

SQLAlchemy. (2022). The Python SQL Toolkit and Object Relational Mapper. Available at: https://www.sqlalchemy.org/

Stančin, I., & Jović, A. (2019). An overview and comparison of free Python libraries for data mining and big data analysis. 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),

Starman, A. B. (2013). The case study as a type of qualitative research. *Journal of Contemporary Educational Studies/Sodobna Pedagogika*, *64*(1).

Stencel, M., & Luther, J. (2021). *Fact-checking census shows slower growth*. Duke.

Strohbach, M., Daubert, J., Ravkin, H., & Lischka, M. (2016). Big data storage. In *New horizons for a data-driven economy* (pp. 119-141). Springer, Cham.

Suleh-Yusuf, M. (2020). Big Data, Internet Privacy and the Vulnerabilities of the African Regulatory Landscape. ResearchGate 10.7176/EJBM/12-17-14.

Sulla, V., Zikhali, P., & Cuevas, P. F. (2022). Inequality in Southern Africa: An Assessment of the Southern African Customs Union. World Bank. Available at: https://doi.org/10.1596/37283

Sun, E. W., Chen, Y.-T., & Yu, M.-T. (2015). Generalized optimal wavelet decomposing algorithm for big financial data. *International Journal of Production Economics*, *165*, 194-214.

Surbakti, F. P. S., Wang, W., Indulska, M., & Sadiq, S. (2020). Factors influencing effective use of big data: A research framework. *Information & Management*, *57*(1), 103146.

Tambe, P. (2014). Big data investment, skills, and firm value. *Management Science*, *60*(6), 1452-1469.

Teece, D. J. (2007). Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal*, *28*(13), 1319-1350.

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal*, *18*(7), 509-533.

Torabi Asr, F., & Taboada, M. (2019). Big data and quality data for fake news and misinformation detection. *Big Data & Society*, *6*(1), 2053951719843310.

Ünal, R., & Çiçeklioğlu, A. Ş. (2019). The function and importance of fact-checking organizations in the era of fake news: Teyit. org, an example from turkey. *Media Studies*, *10*(19), 140-160.

Vargo, C. J., Guo, L., & Amazeen, M. A. (2018). The agenda-setting power of fake news: A big data analysis of the online media landscape from 2014 to 2016. *New Media & Society*, *20*(5), 2028-2049.

Vennam, S. (2020). *Cloud Computing*. https://www.ibm.com. https://www.ibm.com/cloud/learn/cloud-computing

Wahyuningsih, T. (2020). Problems, Challenges, and Opportunities Visualization on Big Data. *Journal of Applied Data Sciences*, *1*(1), 20-28.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J.-f., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, *70*, 356-365.

Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*.

Warren, J., & Marz, N. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Simon and Schuster.

White, C. (2011). Using big data for smarter decision making. *BI Research*, 1-10.

Xu, P., & Kim, J. (2014). Achieving Dynamic Capabilities with Business Intelligence. PACIS,

Yin, R. K. (2009). *Case study research: Design and methods* (Vol. 5). Sage.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. *arXiv preprint arXiv:1905.12616*.

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., & Metcalf, J. (2017). Ten simple rules for responsible big data research. PLoS Computational Biology, 13(3), p.e1005399.

Appendix A: Ethical Clearance Approval

UNIVERSITY OF KWAZULU-NATAL
INYUVESI YAKWAZULU-NATALI

29 October 2020

Mr Andile Samkeliso Dlomo (210533333)
School of Management, IT & Governance
Westville Campus

Dear Mr Dlomo,

Protocol reference number: HSSREC/00001979/2020
Project title: The influence of Big Data on monitoring the factual quality of digital media in Southern Africa
Degree: Masters

Approval Notification – Expedited Application

This letter serves to notify you that your application received on 14 September 2020 in connection with the above, was reviewed by the Humanities and Social Sciences Research Ethics Committee (HSSREC) and the protocol has been granted **FULL APPROVAL**

Any alteration/s to the approved research protocol i.e. Questionnaire/Interview Schedule, Informed Consent Form, Title of the Project, Location of the Study, Research Approach and Methods must be reviewed and approved through the amendment/modification prior to its implementation. In case you have further queries, please quote the above reference number. PLEASE NOTE: Research data should be securely stored in the discipline/department for a period of 5 years.

This approval is valid until 31 October 2021.
To ensure uninterrupted approval of this study beyond the approval expiry date, a progress report must be submitted to the Research Office on the appropriate form 2 - 3 months before the expiry date. A close-out report to be submitted when study is finished.

All research conducted during the COVID-19 period must adhere to the national and UKZN guidelines.

HSSREC is registered with the South African National Research Ethics Council (REC-040414-040).

Yours sincerely,

[redacted]

Professor Dipane Hlalele (Chair)

/ms

Humanities & Social Sciences Research Ethics Committee
UKZN Research Ethics Office Westville Campus, Govan Mbeki Building
Postal Address: Private Bag X54001, Durban 4000
Tel: +27 31 260 8350 / 4557 / 3587
Website: http://research.ukzn.ac.za/Research-Ethics/

Appendix B: Interview Schedule for Open Cities Lab team

The opening

Hi <Interviewee name>, I am Andile Dlomo a UKZN masters research student. As an academic researcher I have been given the opportunity to conduct the study with the Open Cities Lab team. Matt thought it would be a good idea to have an interview with you so that I could gain a better understanding of the process you went through while working on the Dexter big data platform.

I'd like to inquire about your background, education, and experience, as well as your participation with the Dexter big data platform.

I'm hoping to use this information to better understand how big data analytics is being used in the decision-making process for generating high-quality digital content. It should take about 30-45 minutes to complete the interview. Are you available to answer some inquiries right now?

The body

Thank you. There are 3 aspects to understanding the platform according to the model the study will be using. There are the back-end development and the front-end development and the user interaction with the big data system. I will start by asking questions about your background.

Part A: Development of Dexter platform

1. Kindly provide a brief description of your job title and responsibility?

2. 2.1 How did you and your team choose the hardware and software for the Dexter platform?
   2.2 How did you and your team choose the information architecture for the Dexter platform?
   2.3 Which option for acquiring the necessary data structure components was used from the following for the Dexter platform: inhouse, outsource, co-sourcing?
   2.4 And why did you choose that option for acquiring the necessary data structure components?
3. The Dexter platform provides descriptive media analytics. Is this correct, or are there other types of media analysis I should be made aware of? i.e. predictive analytic, prescriptive analytic, outcome analytic, etc.
   3.2 How was the Dexter platform planned on descriptive media analytics?
4. How did you and your team go about the information architecture of the Dexter platform ?

   4.1 Was the data captured elsewhere and then streamed?
   4.2 Is the data stored locally?
   4.3 How was the big data handling (data checking) managed?

4.4 From data on the Dexter platform a monitor curator has been mentioned. What role does a monitor curator have in the big data handling?

4.5 Who has access to the security of the Dexter platform ? (No names just job positions)?

4.6 How was the security of the Dexter platform secured in terms of backup or archiving of data?

5. How flexible is the design of the Dexter platform in terms of allowing resetting of the data itself?
6. How adjustable is the dexter platform design in allowing reconfiguration of ways of accessing the data?

Part B: Use and performance of the Dexter platform

1. Is the Dexter platform only found on a web-based virtual environment?

1.1 and why not deploy it on an app /desktop virtual environment ?

2. How easy is it to use the Dexter platform user interface?
3. How understandable is the information and operations of the dexter platform user interface?
4. How powerfully built is the dexter platform user interface?

4.1 How do you do performance metrics on the Dexter platform user interface ?

4.2 How do you ensure availability of the Dexter platform user interface ?

4.3 How do you check the loading speed/downtime of the Dexter platform user interface ?

5. How close is the Dexter platform to what it had been desired to do initially when the project started?
6. Were all client specifications on the Dexter platform met or were some modified?
7. Does the media analysis from the Dexter Platform allow up-to-date reporting when requested i.e. is the data dynamically updated in real-time?
8. How do you ensure up-to-date media analysis reporting of the Dexter platform?
9. How was the Dexter platform querying of information setup for access by decision-makers?
8. What standard reports are available for decision-makers?

9. How customisable is the standard report provided by the Dexter platform ?
10. How accessible is the option for decision-makers to generate customized query reports from the dexter Platform ?
11. How did the media analysis from Dexter platform provide a complete picture of what was queried from your client's feedback?
12. What measures were used for calculating performance?
13. Were there benchmarks applied to the measures?

14. How are the performance measures bench marked?
15. At the end of the service cycle how do Open Cities evaluate the performance of the Dexter platform ?
16. At the end of the service cycle how do client(s) evaluate the performance of the service ?

Transition - Well, it's been interesting learning more about you. Let me summarize the information I gathered throughout our interview in a few words.

Finally, I'd like to express my gratitude for the time you spent on our interview. Is there anything else you think I should know before the session concludes? I should be able to get all of the information I require. Would it be okay if I called you again in the future if I had any additional questions? Thank you one more.

Appendix C: Clarification Interview Schedule

Back-end Questions

You did mention that a lot of work had been done into the addition of a far greater set of indices for the dexter system to do its own scraping/crawlers, having that being done on the Newstools site.

1. What were the criteria for choosing the Newstools system ?
2. Which of the following 5 V's (volume, velocity, value, variety, veracity) aspects of the nature of the input data were considerations when you considered the use of the Newstools platform?
3. How much does the Newstools play a part in alleviating those data challenges 5V's (volume, velocity, value, variety, veracity)?
4. What do you feel is still a limitation to the Dexter platform in terms of the data input being only textual data?

Kindly help me draw a picture of the current data flow architecture.

5. How does the framework, object relational mapping, tasks, natural language processing, scraping, templating and analysis flow from one role to another for the dexter platform?
6. How has it changed from the one presented at PyCon talk?
7. Where does the Newstools system and amazon web service and other new tools feature in the data flow architecture?
8. Also what tools have been added on the architecture you mention google tools, react, python 3, what other tools are there and in what phase in the data flow and UI do we see them being utilized?

Front-end/ User interface part

The studies say in some cases the success of a big data initiative is dependent on the quality of its source. Some believe that using known quality IO standards is best.

9. What makes Dexter news data sources reliable and not prone to being found as fake news ?
10. How do you ensure trustworthy data in the 109 news sources before crawlers begin?
11. What measures are taken to ensure credible data of the 109 news sources before crawlers begin ?

Strategic maker and Outcome

In the video coverage PyCon year 2017 called : Large-scale media analysis for driving accountability I gathered that a level analysis focus had been presented for instance, children's media diversity rating, investment and journalist + media house at the time.

12. You mentioned analytical framework previously, what were the criterion for designing and implementing analytical framework for dexter?
13. Who is involved in the process of finding analysis focus areas?
14. How were the questions defined in the different analysis focus areas?
15. What formal technical support is available for your partner in terms of the Dexter platform?

Appendix D: Alignment Matrix

Main RQ/ research sub-questions	Variables in RQ	Measurement of Variable	Reference	Interview Questions for Open cities lab team:
1) How does the Dexter project's big data Capability influence the big data strategy used to safeguard the factual quality of digital media?	Big data Capability	Choosing IT		1.Kindly provide a brief description of your job title and responsibility?
				2.2.1 How did you and your team choose the hardware & software for the dexter platform?
				2.3 Which option for acquiring the necessary data structure components was used from the following for the dexter platform: inhouse, outsource, co-sourcing?
				2.4 And why did you choose that option for acquiring the necessary data structure components?
		Integration IT		The Dexter platform provides descriptive media analytics. Is this correct, or are there other type media analysis I should be made aware of? I.e predictive analytic ,prescriptive analytic, outcome analytic, etc..
				3.1 If answer is true to previous question, how was the Dexter platform planned for all media analysis types as a whole?
				3.2 Or were there different strategies used for different media analysis types?
				Alternative Question 3.3 if answer not true, how was the Dexter platform planned on the descriptive media analytics?
				4 How did you and your team go about the information architecture of the dexter platform?
				4.1 Was the data captured elsewhere and then streamed?
				4.2 Is the data stored locally?
		Managing Big data		4.3 How was the big data handling (data checking) managed?
				4.4 What role does a monitor curator have in the big data handling?
				4.5 Who has access to the security of the dexter platform? (No names just job positions)?
				4.6 How was the security of the dexter platform secured in terms of backup or archiving of data?
		Reconfigurability		5.How flexible is the design of the dexter platform in terms of allowing resetting of the data itself?
				6.How adjustable is the dexter platform design in allowing reconfiguration of ways of accessing the data?

Main RQ/ research sub-questions	Variables in RQ	Measurement of Variable	Reference	Interview Questions for Open cities lab team:
2. How the Dexter project is monitoring output used by strategy makers to safeguard the factual quality ofdigital media?	apabilty Outputs (IT Vie	Accessibility		1 Is the dexter platformonly found on a web-based virtual environment? or 1.1 and why not deploy it on app /desktop virtual environment ?
	operable			2 How easy is to use the dexter platformuser interface?
	understandable,			3 How understandable is the information and operations ofdexter platform user interface?
	Robust			4 How robust is the dexter platformuser interface?
				4.1 How do you do performance metrics on the dexter platform user interface ?
				4.2 how do you ensure avaliability ofthe dexter platformuser interface ?
				4.3 How do you check the loading speed/downtime ofthe dexter platformuser interface ?
		Accuracy		5. How close is the dexter platform to what it had been desired to do initially when the project started?
				6. Were all clients specification met or were some modified?
		Completeness		7 Does the media analysis fromthe Dexter Platformallow up-to-date reporting when requested i .e is the data dynamically updated in real -time?
		Currency		8 How do you ensure up-to-date media analysis reporting ofthe Dexter platform?
	Strategy maker	Action		9 How was the Dexter platformquerying ofinformation setup for access by decision-makers?
				10.What standard reports are available for decision-makers?
				11. How customizable is the standard report provided by the Dexter platform ?
				12. How accessible is the option for decision-makers to generate customized query reports fromthe dexter Platform ?
		knowledge		13. how did the media analysis fromDexter platformprovide a complete picture ofwhat was queried fromyour client's feedback?
		competency		

Main RQ/ research sub-questions	Variables in RQ	Measurement of Variable	Reference	Interview Questions for Open cities lab team:
3. What is the assessed performance of the Dexter project, in terms of safeguarding the factual quality of digital media?	factual quality of digital	Performance		14. What measures were used for calculating performance?
				15. Were there benchmark applied to the measures?
				16. How are the performance measures bench marked?
				17. At the end of the service cycle how do Open Cities evaluate the performance of the dexter platform?
				18. At the end of the service cycle how do client(s) evaluate the performance of the dexter platform?