THE PRACTICE OF WEB ARCHIVING AT THE DOCUMENT WAREHOUSE

A research report submitted in partial fulfilment of the requirements for the degree of Bachelor of Arts in Records and Archives Management (Honours) to the University of Namibia, Faculty of Humanities and Social Sciences, Department of Information and Communication Studies

BY

LEROY TAKUDZWA KOPERA

(STUDENT NO. 201501887)

2020

SUPERVISOR: PROFESSOR C.T. NENGOMASHA

ABSTRACT

The study investigated the practice of web archiving at the Document Warehouse as a single case study. The objectives of the study were; to establish if the staff were aware of web archiving; to find out the processes of web archiving in the organisations; to determine the barriers and enablers to for successful web archiving; to determine the benefits of web archiving for the organisations; and to recommend ways in which web archiving can be enhanced in the organisation. The study employed a qualitative research approach and two IT staff and a single records employee were the participants selected through purposive sampling. The study was conducted by interviews as data collection methods. All three participants were interviewed and content analysis was used to analyse the collected data.

The study revealed that the Document Warehouse did not systematically practice web archiving. Web archiving was limited to monthly snapshots of the entire server that hosted the content of the web interface and the web records were managed on a content management system called M-Files. This resulted in responses and answers that were not directly reflective of web archiving. The study also showed that some of the challenges were that there was neither a web archiving programme nor a web archiving policy; however, web records were covered in the records management policy with all the other records. The limited minor web archiving practised was linked to M-Files which attracted more clients. The researcher recommends Document Warehouse to get their staff trained in practicing web archiving and to formulate a web archiving programme and web archiving policy.

DEDICATION

I wish to dedicate this research project to my amazing parents, Mrs Fungai Mtandwa Kopera and Mr Rivman Kopera who sacrificed so much for me to get to the point of completing my project and my studies. They are the foundation on which this project is based on and I cannot say enough words to show my gratitude for all they have done for me and this is for them.

ACKNOWLEDGEMENT

I wholeheartedly acknowledge our Almighty Lord for He guided me throughout the process of this research that had so many obstacles because of Covid-19 pandemic. The completion of the research could not have come to fruition if not for God's intervention. Secondly, I would like to show my gratitude to my lovely mother, Mrs Fungai Mtandwa who motivated me to give my best to my research even if we were not in the same country throughout the process of this research and also my father, Mr Rivman Kopera who made it possible for me to complete my research when at times it was difficult but inspired me to pull through and my little brother Delroy, who supported me and gave me a strong push when I needed it the most. I would also like to thank my friends and family who assisted me throughout this whole journey. My deepest gratitude to the Document Warehouse IT and records staff who voluntarily took part in my project and assisted greatly in its success especially Mr Raphael. Last but not least I would like to show my complete appreciation to Professor C.T. Nengomasha, my supervisor who guided me and steered me in the right direction whenever I was lost and made sure I pushed and pumped till I completed my project

DECLARATION

I, Leroy Takudzwa Kopera, solemnly declare that the report titled "The practice of web archiving at the Document Warehouse" has been carried out in every aspect by me under the guidance and supervision of Professor C.T. Nengomasha at the University of Namibia. I also declare that this research report has been submitted strictly for academic purposes and not for any other reason.

Leroy Takudzwa Kopera 26/11/2020

……..... ……

Student's name Date

…….. ……

Supervisor's name Date

TABLE OF CONTENTS

List of tables

List of figures

ABBREVIATIONS AND ACRONYMS

EU	**- European Union**
URL	**- Universal Resource Locator**
EDMS	**- Electronic Document Management System**
HTTP	**- Hypertext Transfer Protocol**
BHO	**- Browser Helper Object**
HTML	**- Hypertext Mark-up Language**
IT	**- Information Technology**
BCP	**- Business Continuity Plan**
InterPARES	**- International Research on Permanent Authentic Records in Electronic System**
ICA	**- International Council of Archives**

CHAPTER ONE

INTRODUCTION

1.1 Introduction

The chapter introduces the research report entitled "The practice of web archiving at the Document Warehouse, Namibia". The Document Warehouse which is an organisation that deals with records and document management. This chapter covers the orientation of the study, statement of the problem, objectives of the study, significance of the study, limitations of the study, summarized methodology, procedures and research ethics.

1.2 Orientation of the study

The rapid technological evolution is now visibly present in the archival profession as archivists are attempting to accommodate both physical and virtual records. Most organisations have websites where potential clients and the public can obtain useful information either about the organization's business functions and structure or basic information about the existence and history of the organization among other interests for the clients. Just as there are physical and tangible documents or records that are permanently stored in an archival repository as archives, there are also intangible records which are obtained electronically on webpages or web sources and permanently stored through a practice that is known as web archiving. The International Internet Preservation Consortium defines web archiving as the process of gathering up web information known as harvesting data that have been published on the World Wide Web ,storing it, ensuring the data is preserved in an archive, and making the collected data available for future research. (Littman et al., 2016). The web archiving process can be viewed as a workflow, whereby web resources are selected, collected, preserved and delivered to users (Brown, 2006).

The inception of archiving websites dates back to 1996 with the first notable web archiving initiative being established and is known as the Internet Archive with its mission statement being 'universal access to all human knowledge' (Brown, 2006). As stated by Brown (2006), the Internet Archive is arguably the biggest web archive in the world and was established with the objective of building a digital library to offer permanent access to historical collections which exist in digital form thereby setting in motion the emergence of international and institutional web archiving programmes from different organisations entities throughout the world. These web archival programmes include the United Kingdom Central Government Web Archive initiated in 2003, The European Digital Archive established in 2004 and the Digital Archive of Chinese Studies established in 2001 (Brown, 2006).

Surveys have shown that web archives are hosted abundantly in developed countries and are almost unknown to developing states specifically on the African continent even though they are quite present throughout the world (Costa, Gomes & Silva, 2017). Organisations such as the European (EU) have operational web archives that archive websites of the EU institutions and most of these are hosted on the europa.eu domain and subdomains and are archived on a regular basis (Publications Office of the European Union, 2019). Ad hoc crawls of websites that will be taken offline or will change substantially can be done on request of the respective EU institution and also be archived like for example, the European Union can archive pages created for certain events (Publications Office of the European Union, 2019). The fact that there is barely enough widespread awareness of web archives is a worrying concern especially with the annual increase of websites belonging to different organisations on the internet since its inception in 1996.

According to Costa et al. (2017), human resources invested in web archiving are scarce and the number of individuals responsible for web archiving programmes is significantly low and there is a potential risk of historical void about organisations' current existence on the web.

As much as information can be stored in other different forms of storage, the web and the internet are a gigantic pool of information that has historical, legal and fiscal value and the most trustworthy method of ensuring their long term preservation is through web archiving. It is essential to collect vital information from the web before it is lost because the internet is not stagnant and web technology evolves gradually. Several organisations especially in the developed countries have begun applying web archiving such as the North Carolina State Archives and State Library of North Carolina which collaborated to develop the North Carolina State Government Website Archives, a collection of captured government websites dating back to the fall of 2005 and available to the public for research (Martin & Eubank 2007). This was monumental in that it was one of the first publicly available comprehensive collections of websites of state government information.

The institution, as any other organization that has presence on the web with active websites, would need to archive their web records. Malicious software programmed by attackers to disrupt computer operation, collect sensitive information, or gain access to private computers is found on web pages and when archiving these pages along with the malware, it poses serious security risks (Adoghe, Kayode, Dike & Olujimi, 2013). Factors which contribute to successful web archiving include keeping all content under one root URL which is the global address of resources and documents on the web and this means that everything in a website can be easily crawled and therefore archived thereby making it easy to identify that the entire site; ensuring that content is not being added and removed between acquisition sessions as this content will not be captured and preserved and this prevents information gaps in the web archives; Information to be captured by web archiving needs to be machine reachable, which means that it can be reached by a web crawler and in that case information that needs a log in, tick box, pick list or search box to access is not machine reachable and so cannot be captured

by a web crawler so the solution is providing alternatives that can be directly downloaded, such as an A-Z list or site map (United Kingdom, The National Archives, 2011).

The Document Warehouse is an offsite document storage and records management company formed in South Africa in 1992 and then later established its Namibia based operations in the year 2006 (The Document Warehouse, n.d.). Their website has information about their services that includes document filing, mailroom services, filerite software which is an electronic document management software (EDMS) to provide effective document management solutions and also archival stationery products. All of the companies aforementioned have accessible websites that are active online.

1.3 Statement of the problem

Web archiving is fast becoming a reliable way of archiving essential virtual records found on websites and the fact that most developing countries especially in Africa who have organisations with active websites do not preserve most of their important web records is troublesome and fatal. Web archiving operations can assist in legislation, preserve historical information that protect legacies of companies, most organisations are not informed about the preservation of websites that holds long term value and how certain legal issues can be averted with practice of web archiving.

Adoghe et al. (2013) state that organisations are however, said to encounter challenges when implementing web archiving, which include economic challenges since web archives require substantial initial investments to cater for technology, research and development, and it must be built to a considerably large scale if it is to save the entire web continuously; limitations of web crawlers whereby web crawlers encounter problems when harvesting contents from database driven web-pages, streamed multimedia files, script code, password protected

content, Java script driven menus and this is referred to as the deep web. This study aimed to investigate these issues to do with web archiving at the Document Warehouse.

1.4 Objectives of the study

The main objective of the study was to investigate the practice of web archiving at the Document Warehouse. The sub-objectives were to:

- Establish if the staff were aware of web archiving
- Find out the processes of web archiving in the organisations
- Determine the barriers and enablers to for successful web archiving
- Determine the benefits of web archiving for the organisations
- Recommend ways in which web archiving can be enhanced in the organisation

1.5 Significance of the study

The findings of this study can contribute to web archiving policy and procedures at the Document Warehouse. The study will also contribute to the body of knowledge on web archiving especially in the Namibian context in the organisations.

1.6 Limitations of the study

The study was a single case study hence the findings cannot be generalized to other institutions in Namibia. It was initially a multi case study involving three organisations but due to the Covid-19 pandemic the other two organisations excluded themselves from the study for safety reasons hence the findings of the study cannot be generalized. Due to the aforementioned pandemic, face to face interviews were impossible to do so the participants emailed their answers from the interview guides that the researcher gave them.

1.7 Methodology

According to McCombes (2019), a research methodology discusses the methods a researcher used to do research explaining what he/she did and how it was done, allowing readers to evaluate the reliability and validity of the research. The single case study employed the qualitative research approach, applying a case study research design and the records professionals and IT personnel were selected through purposive sampling. The study was carried out with interviews of both IT staff and records professionals from Document warehouse. Those interviews were done through two separate semi structured interview guides, one for IT staff and the other for records professionals.

Validity was guaranteed by response validation which involved participants reviewing the summary of data and findings to check for accuracy of the data collected. Content analysis was applied for analysing the data collected from the interviews of this study. A discussion of the methodology is in chapter 3.

1.8 Procedures

The researcher received a permission letter (Appendix A) from the Department of Information and Communication Studies of the University of Namibia which was used to seek authorisation to conduct the study in the institution. Once authorisation was granted by the form of a non-disclosure agreement (Appendix B), the researcher made appointments with the participants after getting their willingness to participate in the study. All volunteered participants signed a consent form (Appendix C) as no individual was compelled to take part. Participants both in the IT department and records management staff answered questions that were set up in two separate interview guides (Appendix D & E) and the researcher took note as well as recorded the interviews, but only with the permission of the participants.

1.9 Research ethics

Enago Academy (2019) defines research ethics as moral principles that researchers must follow in their respective fields of research. Ethics are the moral principles that a person must follow, irrespective of the place or time and behaving ethically involves doing the right thing at the right time (Enago Academy, 2019).This study was executed well within the limits of research ethics as no participant from the institution was obligated or forced to partake in the study and anonymity was preserved through identification by codes for all participants. To show that participation was voluntary and not imposed on the participants, there were consent forms which were intentionally signed by all participants.

1.10 Chapter summary

The chapter provided the orientation of the study displaying how web archiving is established in developed countries such as United States of America and Australia who have recognized web archiving initiatives and also how web archiving is still yet to be embraced in Africa. The chapter gave the problem statement, objectives, significance of the study and limitation of the study, brief methodology, procedures carried out for the study and research ethics. The next chapter will cover the literature review as well as the theoretical framework.

CHAPTER TWO

Literature review and Theoretical framework

2.1 Introduction

This chapter focuses on the relevant theories and concepts that relate to the subject of research and also reflects on certain gaps in the practice of web archiving. According to McCombes (2020) a literature review is a survey of scholarly sources on a specific topic and it provides an overview of current knowledge, allowing a researcher to identify relevant theories, methods, and gaps in the existing research. McCombes (2020) further highlights that conducting a literature review involves collecting, evaluating and analysing publications such as books and journal articles that relate to your research question and a good literature review does not just summarize sources but it analyses, synthesizes, and critically evaluates to give a clear picture of the state of knowledge on the subject of research. The literature reviewed in this chapter is centred on the practice of web archiving. The literature review is presented under the following topics:

- Broad definition of web archiving
- Techniques and processes of web archiving
- Benefits of practicing web archiving
- Obstacles to web archiving
- Enablers to web archiving

2.2 Web archiving

Websites meet the definition of a public record as they are made or received pursuant to or ordinance in connection with the transaction of public business (Martin & Eubank, 2007). With that realization, websites need to be archived but only the webpages that have existing

value to the organisations. Web archiving enables the capture, preservation and reproduction of valuable content from the live web in an archival setting, so that it can be independently managed and preserved for future generations (Pennock, 2013). Information on websites is volatile due to evolution and updates of technology. Web archiving is the process of collecting websites and the information that they contain from the World Wide Web and preserving these in an archive and it is a similar process to traditional archiving of paper or parchment documents whereby the information is selected, stored, preserved and made available to people (United Kingdom, The National Archives, 2011).

The United Kingdom, The National Archives (2011), further states that access is usually provided to the archived websites, for use by government, businesses, organisations, researchers, historians and the public and in traditional archives, web archives are collected and cared for by archivists, in this case they are known as web archivists. The web contains a vast amount of websites and also information so automation is usually applied to collect the essentially required websites and this is carried out by harvesting websites from their locations on the live web using specially designed software known as crawlers (United Kingdom, The National Archives, 2011). Crawlers travel across the web and within websites, copying and saving the information as they go and the archived websites and the information they contain are made available online as part of web archive collections where they can be viewed, read and navigated as they were on the live web (United Kingdom, The National Archives, 2011).

Web pages have become increasingly more complex over the years, with many loading hundreds or even thousands of images, style sheets, and JavaScript resources which can include advertisements and trackers so since most web pages with JavaScript resources cannot be captured by all web archives (Weigle, 2018). Web archives then make some limited transformations to the original web page by rewriting links and locations of

embedded resources so that they are loaded from the archive instead of the live web and this prevents someone from viewing a web page captured in 2012, for instance, and seeing an advertisement from 2018 embedded in that 2012 web page (Weigle, 2018).

2.3 Web archiving techniques and processes

Technical approaches to web archiving are different according to the scale of operations of an organization. The three main techniques or processes of archiving websites are client-side web archiving, transaction-based web archiving, and server-side web archiving.

2.3.1 Client-side web archiving

This is the main acquisition method both because of its simplicity, scalability and adaptation to a client–server environment and crawlers are adapted to what is the usual way of accessing the web hence allowing archiving of any site that is accessible either freely on the open web, either on intranets or extranets as long as the crawler get the appropriate authorization (Masanes, 2006). Masanes (2006) further states that this method not only adopts the same position as normal web users but also imitates its form of interaction with servers and crawlers starting from seed pages, parsing them, extracting links and fetching the linked document, and furthermore, they reiterate this process with documents fetched and proceed as long as they have links to explore and they find document within the scope defined. Web crawlers make use of the HTTP to collect content responses from the server (Adoghe et al., 2013). Typically such crawlers can capture a wide variety of web material not only documents or text pages, but audio files, images and video, and data files and the success of capture very much depends on how accessible the material is to the crawler web (United Kingdom, The National Archives, 2011).

2.3.2 Transaction based web archiving

This type of approach is operated on the server-side and so requires access to the web server hosting the web content and in this approach, content that is never viewed will never be archived (United Kingdom, The National Archives, 2011). Adoghe et al. (2013) state that this type of web archiving has the advantage of recording exactly what was seen and when however the main limitation of this method is the fact that it requires the use of code on the web-server that is hosting the content, and thus has to be implemented with the collaboration of the server's owner. It is therefore used mainly for internal web archiving by content owners (Adoghe et al., 2013).

2.3.3 Server side web archiving

The last type of acquisition method for web archives is to directly copy files from the server without using the HTTP interface at all and this method, as the transaction based method, can only be used with the collaboration of the site owners, and although it seems to be the most simple, it actually raises serious difficulty to make the content copied usable (Masanes, 2006). It is impossible for crawlers to find path to some documents of a website and files that can only be accessed through a complex interaction like sending a query to a form and this is a portion of the web known as the deep web but in such situations server side archiving can be a solution as it requires active participation of the site administrator (Masanes, 2013)

2.4 Benefits of web archiving

Practising web archiving in organizations such as the Document Warehouse has several benefits such as:

- **Business continuity**: Supporting the current and future activities of organisations just like paper and digital files (United Kingdom, The National Archives, 2011).

Enhancing Resources published on the internet constitute a direct communication between originating organisations and their audience hence improving public relations with existing clients and attracting new clients as well.

- **Preservation of institutional memory:** Archiving websites preserves the historical and cultural value of organizational web archives that is also found and accessed on websites. Institutional memory is the constant image of an organization therefore a preserved website is one component of an organization's history and business record that remains in place even after staff turnover (Ho, 2018). Web information is volatile and it can create a historical gap hence web archiving ensures that information is not lost forever especially vital web information.
- **Information retrieval:** Fast retrieval of archival documents as web archives enable researchers and employees to search in a few seconds millions of documents written from different perspectives hence it is time effective (Miguel & Costa, 2014).

2.5 Obstacles and challenges to web archiving

One of the main barriers to web barriers is the emerging technological advances. When technology changes, websites are affected in that they become obsolete if not updated and for organisations that do not practice web archiving are at the risk of losing important and vital information. On the same issue of technology, frequent updates and corrections are useful features of web-based materials but they can cause confusion and lead to errors or misunderstanding by users if such changes are not recorded (Pennock & Kelly, 2006).

One of the major issues that obstruct the effective operations in web archiving is malware. Malware which is known as malicious software is software programmed by attackers to disrupt computer operation, collect sensitive information, or gain access to private computers and it can appear in the form of scripts, active content, code, and other software and,

furthermore, it include viruses, worms, ransom-ware, trojan horses, key-loggers, root-kits, adware, dialers, spyware, rogue security software, malicious BHOs and other malicious programs (Adoghe et al., 2013). Just like any other business function, web archiving also deals with legal issues and legislation may affect many aspects of web archiving that includes the collection of the content, its preservation and the dissemination of that content to users (Brown, 2006). With the ever-changing state of technology, web archiving is greatly affected in processes of migration which is a preservation action and this affects all forms of records from physical to virtual records as their content and structure is in danger of possible alterations (Brown, 2006). The continual formation of new websites that contain their own web records eventually means that not all the information on websites is archived and there is a selective process as to what requires to be archived according to its value.

Hale (2017) highlights that there are enormous variations in web archival storage especially on the Internet Archive as there is clear bias toward prominent, well- known and highly- rated web pages because smaller, lesser well- known and lower- rated web pages are less likely to be archived. Pennock (2013) highlights the issue of temporal coherence whereby a large website may have changed before a web archive has even finished capturing a copy of it and the temporal integrity of the site is therefore unclear. Web crawlers are a vital part of web archiving to harvest copies of web content however; there is an issue of the limitations of crawlers. This issue arises in the type of content the crawlers are able to capture because there are difficult content such as database or dynamically driven content that is web pages that are generated via a database in response to a request from the user, streamed multimedia files, password-protected content, Javascript-driven menus whereby URLs are generated by dynamic mechanisms which are hard to capture (Pennock, 2013).

Another challenge involves the content accessible only via local site searches therefore crawlers cannot analyse script code and crawling is almost impossible with content on the

deep web which contains content which is difficult for crawlers to see and therefore access, furthermore, the deep web makes up the aforementioned content such as password protected sites and also dynamically driven content (Pennock, 2013). Crawlers have operational limit issues especially when it comes to the crawl memory not having space for the content to be captured as this weakens the progress of crawlers (Pennock, 2013).

Ensuring authenticity, integrity and quality assurance in web archives can be problematic since it is becoming increasingly difficult to identify what comprises an original site and how it should look in the archive and not only do different browsers affect the overall look and feel of a website, but even the content presented to the visitor changes (Pennock, 2013). Determining the originality of the website therefore becomes difficult. Pennock (2013) states that web archiving has technical issues with their long-term preservation which include the sheer complexity of the vast range of formats published on the web as these must not only all be captured, but also a mechanism developed for maintaining access to them. The complex relationships between files that consist of webpages and websites is a challenge as the structural relationships and active links between different files and components of a webpage and website must not only be captured and made to work in an independent archival environment, but those relationships must also be maintained over time (Pennock, 2013). Pennock (2013) further states that this is increasingly difficult if a migration strategy is used and the filenames changed and is even more of a challenge at a domain level.

2.6 Enablers to web archiving

Enablers to web archiving are factors that make the practice possible. A web archiving policy ensures that selection is done according to determined factors outlined in the selection process in the policy. The Web is the main publishing application of the Internet and as such, it consists mainly of the combination of three standards which are the URI(Uniform

Resource Identifier) defining a naming space for object on the Internet, HTTP which defines a client–server interaction protocol using hyperlinks at its core and HTML (Hypertext Mark-up Language) which defines the layout rendering of pages in browsers, therefore these three standards ensure that the web operates effectively and web archiving can also be applied. (Masanes, 2006).

Collaborations in web archiving ensures some sustainability and continuity and this was made possible with State Library of North Carolina and the North Carolina State Archives coming together to develop the North Carolina State Government Website Archives and this was monumental in that the web archives was one of the first publicly available, comprehensive collections of websites of state government information (Martin & Eubank, 2007). A policy guides decision making in any organization which is acceptable by the legislation. Taylor (2017), states that the efficacy of methods of collecting and capturing web records requires guidance and approval from an established policy.

The long term preservation of web records is recommended as they are factors that contribute to the deterioration of web records such as obsolescence. Migration is the process of translating data or digital objects from one computer format to another format in order to ensure users can access the data or digital objects using new or changed computing technologies (Millar, 2009). Even though it is the most common method of transferring records, it is unsafe to migrate web records as it might risk the integrity and their authenticity but can these characteristics of a records can be preserved if done the right way.

For organisations, they can apply techniques such as migration by normalization which involves migrating a digital object from the original software into an open source, standards-based format so that it can be used without having to rely on the original, possibly proprietary, software system used to create it (Millar, 2009). Millar (2009) highlights that

migration by obsolescence is ideal to deal with technology getting out dated but waiting for records to age before migrating them is dangerous, even though the records may have become damaged during a long period in computer storage systems, reducing the quality and integrity of the original at the time of migration, therefore it is recommended to migrate records before they are too out dated.

2.7 Theoretical framework

This study was guided by the web archiving life cycle model. The model is a guideline that displays all the components involved in the management and development of a web archiving programme

2.7.1 Web archiving life cycle model

Bragg and Hanna (2013) define the web archiving life cycle model as an attempt to incorporate the technological and programmatic arms of web archiving into a framework that will be relevant to any organization intending to archive the web and although the model is broken down into individual steps, each action is not discrete as the steps and phases are related, with a significant amount of overlap between them. The diagram of the model is circular in shape to suggest the repetitive nature of the steps in the life cycle as users move through each step eventually finding themselves back at the beginning or repeating certain steps depending on their tasks therefore the model includes circles within circles to suggest these repetitive cycles within the bigger process as shown in the diagram below (Bragg & Hanna, 2013)

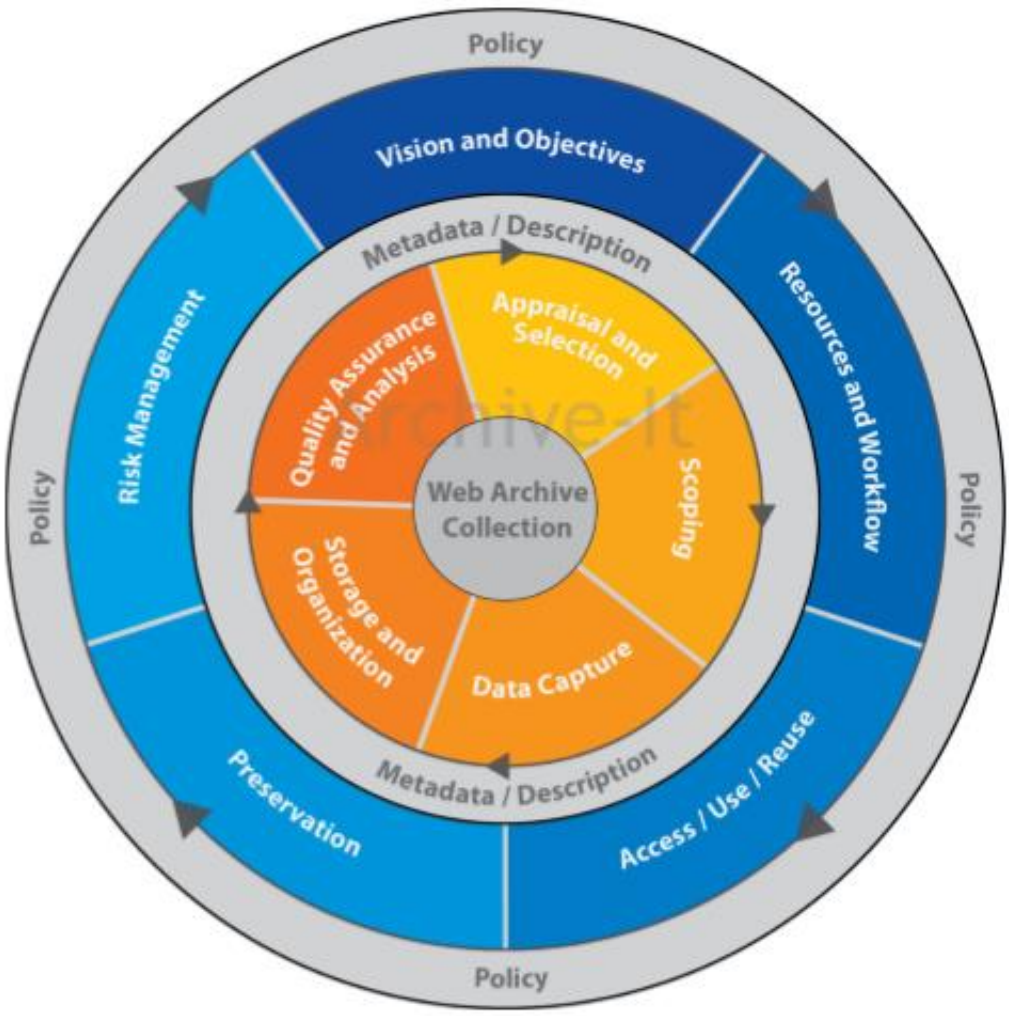

Figure 2.1 Web archiving life cycle model

Source: Bragg and Hanna (2013)

2.7.1.1 Policy

The outmost level of the life cycle shaded in the colour grey is the policy band and almost every aspect of web archiving involves some sort of policy decision whereby these policy decisions may involve developing a new policy specific to web archiving or the adaptation of an existing policy to new collecting activities, and furthermore, by encircling the life cycle steps with a policy band, the model visually represents the ever-present nature of policy making (Bragg & Hanna, 2013). A policy is included in every aspect and component of an

organization as it ensures control and assists in decision making.

2.7.1.2 Metadata and description

In the second band on the diagram also shaded in grey, the model similarly represents metadata and description as metadata was chosen to be assimilated as a band rather than as a segment of the wheel to emphasize that creating, importing and exporting metadata is an on-going process that occurs in tandem with a number of other activities in the life cycle (Bragg & Hanna, 2013). As with most aspects of web archiving, best practices are evolving regarding the use and creation of metadata and descriptive trends for web archives.

The inner blue circle inside the policy band in Figure 2.1 represents the high-level decisions an institution or organisation faces as it sets up and manages its web archiving program and the individual steps are namely vision and objectives; resources and workflow; access, use and reuse; preservation and risk management (Bragg & Hanna, 2013) and they are broadly explained below.

2.7.1.3 Vision and objectives

This is the part where an organization sets its objectives and the reason why they would want to have a web archiving programme. This step in the cycle primarily occurs as institutions initially plan their program, however, institutions do tend to revisit and redefine their web archiving objectives throughout the life of the program as the cycle has a repetitive nature and these parts of reassessment may result from a specific stimulus such as a change of resources or may be an on-going question considered along with and in relation to their other collection policies (Bragg & Hanna, 2013). This is the section where organization decide to practice web archiving according to either their stakeholders' objectives or because they believe that specific web content is at risk of disappearing and therefore needs to be captured and kept accessible--particularly in the case of rapidly changing spontaneous events, like

natural or manmade disasters, political uprisings, and memorials for public figures (Bragg & Hanna, 2013).

2.7.1.4 Resource and workflow

The resources and workflow phase of the life cycle can be considered similarly to policy as they can be applied in multiple areas of the web archiving life cycle model, they can also be considered as general program management terms that can be applied to each of the elements in the model's inner ring which is the web archive collection and by doing so resources and workflow become part of the daily operations (Bragg & Hanna, 2013).

Bragg and Hanna (2013) further highlights that in addition to staffing, the resources and workflow in this model also encompass how institutions manage other resources for example, the decision to collaborate and divide management of the web archiving program between the State Library of North Carolina and the North Carolina State Archives and the two institutions manage a single collection of state government agency websites so in dividing up the day-to-day work, the two agencies have several well-established workflows.

2.7.1.5 Access, use and reuse

Establishing access, use, and reuse policies is vital to a successful web archiving program which means organisations consider whether and how they want to provide open access to their web archives, if and how to promote the collections, as well as how to govern public use of the material hence managing these processes is the primary goal of the access, use and reuse phase of the web archiving life cycle (Bragg & Hanna, 2013). Bragg and Hanna (2013) state that part of the creation of an access policy will include choosing the specific technology or tool to provide access to the archived webpages and furthermore, organisations ensure that their websites are easily accessible especially with visible page history links on the footer of the website which direct visitors to archived versions of the webpage so they can

see how it has changed over time.

2.7.1.6 Preservation

Preservation is an evolving issue for institutions that archive the web, which goes hand in hand with the evolving nature of digital preservation and the development of digital repositories and many organisations and institutes tend to employ several different preservation strategies for example with most organisations relying on the Internet Archive for storage and preservation of their files and associated metadata (Bragg & Hanna, 2013). Preservation is a vital issue because if it is long term then strategies will eventually evolve especially with how rapidly technology changes.

2.7.1.7 Risk management

In developing a web archiving program, many organisations consider the level of risk related to copyright they are willing to accept and how they will manage this risk and whether and how institutions decide to seek permission from site owners before archiving is one of the clearest examples of risk management policy making in action (Bragg & Hanna, 2013). Bragg and Hanna (2013) highlight that the Archive-It service has long used robots.txt which is a web standard as a permissions management tool, which provides an automatic way for site owners to exclude their sites from the archiving process and decisions can be based on discussions with the legal department who give a risk threshold to follow and they ask permission when necessary to stay within this threshold. Risk management decisions can also be seen in the choices institutions make when deciding which sites to archive even though not all organizations ask for permission before capturing content as many organizations are clear that as an archive and/or a library, their organization has the right and the mandate to capture publicly available content on the live web (Bragg & Hanna, 2013). Bragg and Hanna state that eventually risk can be managed and mitigated pre-emptively, and sometimes institutions

may need to address potential issues that come up after archiving of content has taken place.

The remaining stages of the model or those in the deep inner circle describe the day-today activities of managing a web archiving program and they are namely appraisal and selection; scoping; data capture; storage and organization and quality assurance and analysis (Bragg & Hanna, 2013)

2.7.1.8 Appraisal and selection

This involves choosing specific websites for capture and this step comprises of more granular, specific decision points than the broader vision and objectives policy phase of the life cycle and in the appraisal and selection phase institutions choose the specific URLs they will archive, and furthermore, universities that archive the web sometimes take a different approach to site appraisal as they tend to archive the university web presence and/or create collections based on specific themes (Bragg & Hanna, 2013).

2.7.1.9 Scoping

Scoping is done after selecting what needs to be archived and it involves the organisations deciding on whether to archive the entire websites or just portions of the websites and this can be carried out before the first page is captured or after content is harvested as part of the overall collection quality review and also, this part of the life cycle can be quite technical depending on their scoping parameters and the formats of the web content they are capturing (Bragg & Hanna, 2013).

2.7.1.10 Data capture

Bragg and Hanna (2013) highlight that the data capture phase is associated with crawling software and the organisations the frequency and timing of their crawls and when to cut-off long crawls and then they will set their crawls to begin and also given how diverse websites

are in terms of their structure and construction, the data capture step of web archiving can produce a number of surprises. For instance, a site can be much bigger than anticipated and therefore exhaust storage resources and similarly, there are ways for web masters to keep their sites from being archived which can require technological intervention or negotiation between the parties involved (Bragg & Hanna, 2013).

2.7.1.11 Quality assurance and analysis

After completion of data capture, organisations then undertake reviews and assessments of the captured sites' quality and completeness and this can be carried out through reports generated by crawlers or by clicking through the archived websites themselves by way of an access tool like the Wayback software and the process of web archiving can include trial and error (Bragg & Hanna, 2013). Like most aspects of web archiving, no single best practice for quality assurance has emerged among institutions that archive the web and reviewing reports can take time and reviewers need to know what kind of anomalies to look for and sometimes (Bragg & Hanna, 2013). Most organisations have developed their own quality assurance tools to work explicitly with their content and meeting their institutional guidelines which makes the phase of quality assurance and analysis effective in their operations.

The web archiving life cycle model displays the processes involved in ensuring effective web archiving operations. The theory guided the research by reflecting on each and every step taken in archiving websites guided by a policy. It also showed how web archiving can be incorporated into an organization without interfering with other business functions by having its own vision and objectives, risk management, access/use/reuse and the collaborative efforts possible resource and workflow. The web archive life cycle model informed the study on the importance of a policy as the entire model and theory is guided by a policy in every aspect and, furthermore, a policy makes it easy to implement a web archiving initiative in an

organization. The study investigated if web archiving at the Document Warehouse covered the various aspects of the web archiving life cycle.

2.8 Chapter summary

The literature review focused on explaining what web archiving is, as well as the techniques and methods of web archiving that are applied in different organisations. It also covered the benefits that come from practicing web archiving, the barriers that obstruct the success of web archiving and also enablers which make it possible for web archiving to be an operational function in organisations and institutes. The literature review also showed that in the existence of web archiving, it has recurring issues such as technological advancement, obsolescence and malware issues.

The theoretical framework of the study was based on the web archive life cycle model, which assists organisations that intend to establish a web archival programme in their operations and it displays that most steps can be repeated at certain points in the life cycle of web archives without too many inconveniences on the other sections of web archiving, which ensures that any addition and subtractions can be done. The next chapter is on research methodology.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

According to McCombes (2020), research methodology is an explanation of what you did and how you did it allowing readers to evaluate the reliability and validity of the research and it should include the type of research you did, how you collected your data and how you analysed the data supported with the rationale for choosing the methods used. The chapter presents the method utilized to conduct the research. It highlights the research approach used, data collection methods applied, the population, sample, research methods, reliability and validity, procedure, data analysis and research ethics.

3.2 Research design

The single case study of the Document Warehouse employed the qualitative research approach. Qualitative research is primarily concerned with understanding human beings' experiences in a humanistic, interpretive approach and good quality studies applies standards of trustworthiness such as member-checking, stepwise replication, and audit trails, each of which seeks to verify the substance of what participants said so that interpretations are not subjective iterations of the researcher's own belief system (Jackson II, Drummond & Camara, 2007).

A case study research design is defined as an empirical inquiry that investigates a contemporary phenomenon within its real-life context; when the boundaries between phenomenon and context are not clearly evident; and in which multiple sources of evidence are used (Yin, 2014). A case study is also described as an intensive, systematic investigation of a single individual, group, community or some other unit in which the researcher examines

in-depth data relating to several variables and examines complex phenomena in the natural setting to increase understanding of them (Heale & Twycross, 2018).

3.3 Data collection methods

The study was carried out using interviews. Interviews are ideal in order to gather in-depth information on web archiving practices from Document Warehouse.

Adams (2015) states that the advantages of carrying out semi-structured interview include:

- The opportunity to know the independent thoughts of each individual in a group thereby getting separate thoughts and answers that can assist in the findings.
- Participants will be more candid and open in semi-structured interviews than in a focus group interview
- The interviewer can conduct a formative programme evaluation and one-on-one interviews with the participants
- The interviewer can examine uncharted territory with unknown but potential momentous issues where the interviewer need maximum latitude to spot useful leads and pursue them.

3.3.1 Interviews

An interview is an important data gathering technique involving verbal communication between the researcher and the subject and they are commonly used in survey designs. (Mathers, Hunn & Fox, 2000). In exploratory and descriptive studies, there are also a range of approaches to interviewing, from completely unstructured in which the subject is allowed to talk freely about whatever they wish, to highly structures in which the subject responses are limited to answering direct questions (Mathers, Hunn & Fox, 2000).

Mathers, Hunn and Fox (2000) further highlights that the quality of the data collected in an interview will depend on both the interview design and on the skill of the interviewer, for example, a poorly designed interview may include leading questions or questions that are not understood by the subject and that interviewer may consciously or unconsciously influence the responses that the subject makes. Interviews can be structured, semi-structured or unstructured depending on the type of research being carried out. In interviews there is an advantage of close-up communication between interviewer and interviewees which contributes to better understanding of what is required in the interviewing process.

3.4 Population

Walliman (2011) defines population as a collective term used to describe and explain the total quantity of cases of the type which are the subject of the study and it can consist of objects, people and even events. It is important to find out how much of a representative is the sample of the whole population, in other words, how similar are characteristics of the small group of cases that are chosen for the survey to those of all of the cases in the whole group which is the population (Walliman, 2011). For this study, the population comprised of seven IT staff and three records keeping staff from the Document Warehouse. This population was ideal for the study because it covers both aspects of web archiving which comprises of records management and administration and also web preservation.

3.5 Sample

Sampling occurs when researchers observe a portion or part of a larger group of potential participants and use the results to make statements that apply to this broader group or population (Salkind, 2010). Salkind (2010) highlights that the extent to which the research findings can be generalized or applied to the larger group or population is an indication of the external validity of the research design and the process of choosing or selecting a sample is

an integral part of designing sound research, and furthermore in theory, a sound sampling method will result in a sample that is free from bias meaning that each individual in the population has an equal chance of being selected and is reliable.

Purposive sampling also known as judgmental sampling is a strategy in which particular settings persons or events are selected intentionally so as to provide important information that cannot be obtained from other choices (Maxwell, as cited in Taherdoost, 2016). Purposive or judgmental sampling was used to select the records keeping and IT professionals who took part in the study. In this case the records keeping and IT personnel were selected due to their responsibilities of managing records and provision of IT services respectively. According to their vast knowledge and expertise, one records keeping staff and two IT staff from the institution was selected to participate in the study.

3.6 Research instruments

This section of the methodology describes the types of data collection instruments which were utilized for the research. A research instrument is a tool used to collect, measure, and analyse data related to your subject and research instruments can be tests, surveys, scales, questionnaires, or even checklists (Duquesne University, 2020). The instruments utilized for this study were semi-structured interview guides.

3.6.1 Semi-structured interview guide

A semi structured interview guide is a list of open-ended questions and topics that need to be covered during the conversation or interview, usually in a particular order even though the interviewer can add extra questions about an unexpected but relevant area that emerges and sections that do not apply to the participant can be negated (Robert Wood Johnson Foundation, 2008). Two separate semi-structured interview guides were used for this study so

as to have an advantage of proficiency in both fields. There was one interview guide for records management personnel and a different one for the IT personnel.

The major advantage of a semi-structured interview guide is that the interviewer will have a list of themes and areas to be covered and there may be some standardised questions, but the interviewer may omit or add to some of these questions or areas, depending on the situation and the flow of the conversation (Neville, 2007).

3.7 Reliability and validity

Reliability and validity are crucial features of any research. Reliability measures where the researcher must prove that the process and the results have replicable outcomes and also the qualitative research must be consistent (Thakur & Chetty, 2020). Walliman (2011) states that in relation to human perception and intellect, reliability is the power of memory and reasoning to organize data and ideas in order to promote understanding. Reliability was ensured in this study by meticulous documentation of the research process and honest reporting of the findings and, furthermore, the researcher guaranteed consistency in what was recommended of the participants by conducting the study himself.

Validity relates to the appropriateness of any research value, tools and techniques, and processes, including data collection and validation and it also establishes the soundness of the methodology, sampling process, data analysis process, and the conclusion of the study (Thakur & Chetty, 2020). Validity is the outcome goal of research and is based on trustworthiness and external reviews and verification is the first step in achieving validity of a research project (Cresswell, 2007). Validity was ensured in this qualitative research through response validation or member checking (Cresswell, 2012). This was secured by presenting the summary of data and findings to all the research participants to check for accuracy of the data and also to see if the findings concurred with their perspectives.

3.8 Data analysis

In qualitative research of study typically you gather a text database, so the data analysis of text consists of dividing it into groups of sentences called text segments and determining the meaning of each group of sentences so, rather than using statistics, you analyse words or pictures to describe the central phenomenon under study and thereby the result may be a description of individual people or places (Cresswell, 2012). Cresswell (2012) highlights that in qualitative studies in which you both describe individuals and identify themes, a rich, complex picture emerge and from that complex picture, a researcher can make an interpretation of the meaning of the data by reflecting on how the findings relate to existing research and rather than relying on statistical procedures, the qualitative researcher analyses the words to group them into larger meanings of understanding such as codes, categories or themes.

Content analysis refers to the process of categorizing verbal or behavioural data to classify, summarize and tabulate the data and it discovers correlations and patterns in how concepts are communicated (Luo, 2020.). Content analysis was utilized for analysing the data collected from the interviews of this study and it was applied by way of identifying a term that appears consistently in all the interviews and the word that appeared next to it thereby analysing the meaning of the term that linked all the interview participants which assisted in understanding the concepts surrounding the practice of web archiving in the institution that is part of the case study. The analysed data was presented in both a descriptive and narrative form

3.9 Chapter summary

This chapter discussed the research methodology and. A qualitative research approach was applied for this case study of the Document Warehouse. The population of the study as well as sample were also covered. The data collection methods were interviews using two separate

interview guides for the IT staff and the records management personnel. Every research needs to be trustworthy and consistent thereby, reliability and validity was ensured as explained in this chapter. Data was analysed through content analysis. The following chapter is on data analysis and presentation.

Chapter 4

Data analysis and presentation

4.1 Introduction

This chapter presents the findings of the study titled "The practice of web archiving at the Document Warehouse" whose data was collected through interviews using two separate semi-structured interview guides, one for records keeping and the other for IT staff. In qualitative research typically you gather a text database so the data analysis of text consists of dividing it into groups of sentences, called text segments, and determining the meaning of each group of sentences (Cresswell, 2012). Instead of using statistics, a researcher analyses words or pictures to describe the central phenomenon under study (Cresswell, 2007). The motive for analysing data is to combine all the gathered and collected data and then interpreting that data so as to form a conclusion for the findings.

The chapter consists of under mentioned sections that are based on the interview guides and other thematic areas that emerged from the content analysis.

- General information of participants of the study
- Web archiving practice
- The application of a web archiving technique
- Benefits of web archiving
- Challenges of web archiving
- Enablers to web archiving
- Appraisal and selection
- Access to and use of web archives
- Quality assurance

4.2 General information on participants of the study

At the Document Warehouse data was gathered from the Systems Administrator, the System Developer and the Records Administrator.

***Table* 4.1 Participants**

Respondents	**Codes**
IT Systems Developer	A
IT Systems Administrator	B
Records Administrator	C

The aforementioned respondents in Table 4.1 were the participants involved in the study and they were all identified by codes A-C. The interviews were with the IT Systems Developer (A), the Systems Administrator (B) and the Records Administrator (C).

The study sought to determine the tasks of participants that worked with the web and also records management. The System Developer (A) ensured that the clients were assisted in developing software as a service and also implementing solutions in IT that contributed greatly to hosting data for the organisation's clients. The System Administrator (B) installed and configured software and hardware for the organisation and worked hand in hand with participant (A) for all IT related issues of the Document Warehouse. The Records Administrator (C) dealt with the record-keeping of vital documents and records and also created, managed and maintained an effective filling system.

The study also sought to elicit the working experience of the participants in their field so as to ensure reliability of their response to the interview questions hence, they were asked the

number they had been working in both IT and records management. The findings are shown below in Table 4.2

***Table* 4.2 Number of years in current job**

Respondents	**Number of years in the existing job title**
IT Systems developer	8
IT Systems administrator	3
Records administrator	6

The table above shows that the respondents who had expertise in their field for more than half a decade were the IT System Developer and Records Administrator with 8 and 6 years of work experience, respectively whilst the IT System Administrator had 3 years of working experience. It can, therefore be established that all the participants had enough experience to give sufficient and reliable information on the study.

4.3 Web archiving practice

One of the objectives of the study was to find out if the staff were familiar with and aware of web archiving hence it was raised as a question to the participants and the findings showed that all three participants were not aware of the term ‘‘web archiving’’ but could only relate to the two terms web and archiving separately. All the participants had to ask for clarification for the meaning and practice of web archiving before they proceeded with the interview.

4.4 The application of a web archiving technique

One of the aims of the study was to see what processes of web archiving the institution applied in their business operations in which Participant A stated that they took monthly snapshots of the entire server which hosted the content of the web interface. This shows that

the institution applied and practiced some web archiving even though they did not recognize it as such because taking snapshots of web pages is a method of capturing web records and it forms part of remote harvesting or Client-Side Web Archiving..

Participant A also mentioned that most of the data including scanned documents and monthly snapshots were stored on M-files which is a content management tool that also retains web records and this is the institution's storage which also manages all records in it. Participant B stated that the Document Warehouse take snapshots of the clients' data on the servers and also offer training services so that their clients could do it themselves if they possess the software.

4.5 Benefits of web archiving

One of the goals of the study was to establish the benefits of web archiving to institutions such as the Document Warehouse and the question was therefore raised to establish if the organisation was realising any benefits from web archiving. Participant C responded by indicating that with taking monthly snapshots of the web interface, web data was stored on M-files and could be virtually made available to them if requested. Both Participant A and B indicated that applying M-files was cost effective and economical to the institution whilst also saving time during retrievals that would have been utilized in other costly digital preservation. M-files as a content management tool enabled the usability and management of web records.

Participant C stated, "More clients have been drawn to require the Document Warehouses' services because of their effective content management system that is effective in records management. The application of a content management tool has attracted clients which have contributed to the growth of the business".

4.6 Challenges to web archiving

Web records are volatile in nature hence one of the aims of the study was to look at challenges and obstacles to web archiving. Participant B indicated that since one of the services of Document Warehouse was to teach clients how they can host or store their own data and web records, the process could be exhaustive especially with clients who are not fully familiar with the task and software hence it could be time consuming and there was no web archiving programme. In rare cases of losing certain records stemmed from the web interface, Participant C stated that there are cost implications to recreating deleted or lost records that were vital to either the institution or the clients.

4.7 Enablers to web archiving

This section aimed to question what made web archiving possible or easier to apply in the organisation. Participant A stated that the collaboration between records personnel and IT staff enabled web archiving to be more effective and efficient as they also need to collaborate when a digital or web record was lost. Participant C indicated, "There was risk management policy for web records or digital preservation and it was part of their Business Continuity Plan (BCP) and is updated yearly". This BCP contains the risk assessment and plans to mitigate risks for both digital and hard copy archiving.

4.8 Appraisal and selection

The question aimed to look at the appraisal and selection of web archives of the institution. When it came to which web records or digital content to archive, Participant A stated that they had a blanket approach which covered all content that was stored for clients and for internal information. Participant B further added by stating that the reason for this was that they dealt with client data or web content that was not defined as important or not, and they had a policy in place that did not allow them to know what the data was hence they treated it

all as important. When it came to who had the authority to choose what web content or digital content to be archived, Participant A and B indicated that clients were identified by the companies' authority as to who had access to what content on the web and the Document Warehouse had no authority on such matters unless their clients permitted them to do so. This was defined at the beginning of a project or acquiring of a new client.

4.9 Access to and use of web archives

The objective of the question of access and use was to find out about restrictions and security of web records and how clients could access their own web records. In terms of access and use of the digital or web content, Participant B stated that the clients could access and use their own data with the assistance of an employee from Document Warehouse for easy retrieval and avoiding unrestricted access. Participant B further added, "There are measures put in place for restricting unauthorised access to clients' information and also the Document Warehouse data".

4.10 Quality assurance

Participant B indicated that assessment on archived web records so as to ensure authenticity was done by their systems such as the M-files which had built in security that managed secured and reliable transfers of data from hardcopy to soft copy. Participant C stated, "To ensure quality assurance for all digital content, every document which eventually becomes a digital record undergoes a series of quality checks before and after it is made digital" .This means that authenticity was defined within the system that managed the content and kept it secure behind firewalls and multiple security protocols

4.11 Summary

This chapter presented the data based on the responses from participants A, B and C. The study established that all the participants were not familiar with the term web archiving; however, it was showed that the Document Warehouse practice some web archiving in the form of monthly snapshots and the use of a content management tool that captured web and digital records. The findings showed that all the participants did not recognise the aforementioned method and technique as part of web archiving. The study showed that the clients had access to their own data without compromising others clients as guided by the risk management policy. The Document Warehouse also backed up its data including snapshots to their server and all data was treated as important unless the client advised otherwise. Assessment of digital content and the website was carried out within the security in the systems; and authenticity was defined within that same system. The next chapter discussed the findings, presents conclusions and recommendations.

Chapter 5

Discussion of Findings, Summary, Conclusions and Recommendations

5.1 Introduction

This chapter discusses the findings of the study and compares it to previous literature whilst interpreting those findings. Moreover, this chapter summarizes the findings, provides conclusions of the findings and give recommendations.

5.2 Discussion of findings

According to Azar (2020) the discussion of findings is putting your research in context limiting the discussion to the essential points and going back to the literature and grappling with what your findings mean, including how they fit in with previous studies and if they differ from previous literature and findings, they should be an explanation as to why. While the amount of discussion required in a thesis may vary according to discipline, all disciplines expect some interpretation of the findings that make those connections (Monash University, 2020).

The discussions of findings are presented under the following subheadings:

- Participants' awareness of web archiving
- The application of web archiving
- Benefits of web archiving
- Challenges of web archiving
- Enablers to web archiving
- Appraisal and selection
- Access to and use of web archives
- Quality assurance

5.2.1 Participants' awareness of web archiving

The study revealed that all three participants (A, B and C) were not familiar with web archiving and its importance to an organization's records management function. They could only understand the two words "web" and "archiving" independently and not as one term. All of them needed some clarification and more explanation before they could begin with the interview but still did not fully understand what web archiving was all about. The study showed that the participants' unfamiliarity to web archiving proved that there was no web archiving programme in their business functions and their activities related to web archiving were done on an ad hoc basis. Any organization with a functioning website needs to be preserving vital web records and the staff should be made aware of the importance of web records. Both IT and records employees need to know and understand the practice of web archiving because archiving websites gives organisations the chance to provide access to legacy information that they may not necessarily want to keep on their 'live' website (United Kingdom, The National Archives, 2011).

5.2.2 The application of web archiving

The study revealed that the Document Warehouse applied a web archiving capturing method in their operations. Participant A stated that they took monthly snapshots of the entire server which hosted the content of the web interface and this method of capturing web records is known as remote harvesting (InterPARES/ICA, 2012). It is part of a web archiving approach called Client-Side Web Archiving (United Kingdom, The National Archives, 2011). The snapshots were digitally stored in an enterprise content management tool called M-files which is utilized for digital preservation and also for information and document management for organisation (Google Play, 2020). The study also revealed that Document Warehouse applied the web archiving approach even though the employees were not aware of what it is

called. The above findings are similar to what is reported by The United Kingdom, The National Archives (2011) which states that archived websites can be viewed, read and navigated as they were on the live web, but are preserved as snapshots of the information at particular points in time. At the moment there is only a single method of capturing web records at the Document Warehouse, and if the employees were to be made fully aware of how a content management tool could be effective and cost effective when it concerns capturing web records, it would be beneficial to both the clients and Document Warehouse.

5.2.3 Benefits of web archiving

The study discovered that with the applying of web archiving, the Documents Warehouse has attracted more clients to hire their services and this has contributed to the growth of the organisation. Use of M-files allowed web records management to be possible hence records personnel could manage selected records captured from websites. Adoghe et al. (2013) states that long term research is a usefully benefit from archiving the web, as users have access to web content over time. This assisted clients of Document Warehouse who would be able to access and use their web records over a period time whereby their records could have been already deleted and lost if not for web archiving.

5.2.4 Challenges to web archiving

The study revealed that since Document Warehouse also taught their clients who wanted to host their own web records and data, it was time consuming as some clients needed extended time periods to fully grasp hosting their own web records. Due to the fact that the Document Warehouse did not have either a web archiving policy or web archiving programme, challenges to web archiving were not clearly stated since it was carried out on an ad hoc basis. Web records would eventually be compromised because of the absence of a web archiving programme. The study also discovered that losing or deleting records accidentally

was costly as well as time consuming to recreate those records and also there was no web archiving policy specified for solely web records and this could be because web archiving is not included in the Archives Act 12 of 1992. Similar findings relating to legislation were reported by Pennock (2013) stating that legislative challenges that not only inhibit collection but also limit access remain one of the greatest issues for collecting institutions. Pennock (2013) states that the long-term preservation of web archives is affected greatly by the rapidity of technology changes. Other literature state that web archives are typically accessed through a search interface, although some permit browsing but it is currently not possible to cross-search the publicly available web archives (Ball, 2010).

Repeated crawling of large, frequently updated websites causes temporal incoherency and it may take several days or even longer to crawl a large website, during which time the web sites are undergoing changes and this could lead to harvesting a website that never really existed (Niu, 2012)

5.2.5 Enablers to web archiving

The study revealed that in terms of what enables web archiving in Document Warehouse; it was the collaboration between the IT staff and records personnel. Web archiving requires individuals with skills in digital and electronic management systems and also in archives and records management to make it possible and effective in an organisation. Collaboration in web archiving practices are instrumental in any organisation as explained by Bragg and Hanna (2013) that, "Collaboration and division of management of the web archiving programme between the State Library of North Carolina and the North Carolina State Archive have established proper workflows and the two agencies alternate responsibility for conducting the crawls, and both institutions perform quality control of the data harvested".

The study also revealed that the Document Warehouse had a risk management policy that covered both physical and virtual records. Web records were also included in that same policy as there was no specific web archiving policy just for web records. The risk management policy was a permanent part of the organisation's functions as it is included in business continuity plan which is updated annually. Bragg and Hanna (2013) state that in developing a web archiving programme, many institutions consider the level of risk related to copyright they are willing to accept and how they will manage this risk and, whether and how institutions decide to seek permission from site owners before archiving is one of the clearest examples of risk management policy making in action.

5.2.6 Appraisal and selection

The study discovered that the Document Warehouse had a blanket approach which covered all content that was stored for clients and for internal information including web records. This was done since they dealt with client data or web content that was not defined as important or not, and they had a policy in place that did not allow them to know what the data was hence they treated it all as important and that was the appraisal and selection. It is stated in Bragg and Hanna (2013) that the appraisal and selection phase of web archiving involves choosing specific websites for capture, however, institutions choose the specific URLs they will archive and organisations and institutions make these choices in different ways.

It was also revealed by the study that the Document Warehouse had no authority as to selecting which web records to archive as this was only authorised by the owners of those records who are the clients. The clients inform the Document Warehouse on which data or web records to archive. Existing web archiving efforts use the following selection criteria to determine what to preserve: domain such as .gov or .edu, topic or event, media type and genre and theoretically, selection based on objective criteria can be easily automated (Niu,

2012). On a technical level, it is easy for software to decide the media type whether it is audio, video or textual and domain like e.g., .gov, or .au of web resources and similarly, it should not be very hard to differentiate between such genres as online journals or blogs, or notice the differences between blog posts and comments (Niu, 2012). Macro appraisal entails appraising and selecting web resources based on aggregates of web pages rather than individual web pages and it reduces the size of the problem and makes the appraisal process more efficient and also the aggregates can be decided on different levels (Niu, 2012). Relevant literature also states that selection criteria, such as domain or media type, can be associated with either a value-based selection or a representative sampling method (Chen et al., as cited in Niu, 2012).

5.2.7 Access to and use of web archives

The study revealed that the clients could access and use their own records with the assistance of a Document Warehouse employee for faster retrieval. This prevents any authorised access from occurring as all clients will only have access to the web records that belong to them and this further emphasised by the measures put in place by the Document Warehouse to restrict unauthorised access to the clients' information.

InterPARES/ICA (2012) states that ensuring accessibility to web-based materials over time raises the same accessibility issues as surround other electronic records hence there are steps that can be taken to mitigate these issues including ensuring materials are carefully managed including maintaining the trustworthiness of web records and identifying and mitigating the management risks, planning for obsolescence, the use of widely supported standards, implementing security measures to protect against either deliberate or accidental alteration, and ensuring environmental control and monitoring. The search capability of different web

archives depends on the richness of metadata and the search and indexing tools used (Niu, 2012).

5.2.8 Quality assurance

As per quality assurance, the study revealed that assessment on archived web records so as to ensure authenticity was done by their systems such as the M-files which had inbuilt security systems meaning that authenticity was guaranteed within the system that managed the content and kept it secure behind firewalls and multiple security protocols. Organisations carry out quality assurance based on their content hence they are various ways of doing quality assurance. Bragg and Hanna (2013) explain that quality assurance can be done through reports generated by crawlers or by clicking through the archived websites themselves by way of an access tool like the Wayback software, however, there is no single best practice for quality assurance.

5.3 Summary of findings

Cresswell (2012) defines summary as a statement that reviews the major conclusions to each of the research questions or hypotheses and it represents general, rather than specific conclusions. The summary was organized according to issues under which data was presented in Chapter 4.

5.3.1 Web archiving practice

The summarized findings revealed that the Document Warehouse practiced web archiving but not as a separate planned programme. This was gathered from IT and records management staff who even though they practised some web records capturing technique, they were not aware it was web archiving.

5.3.2 The application of a web archiving technique

The findings showed that the Document Warehouse used snapshots as a method to capture web records and these were stored and managed in an enterprise and content management system called M-files. This showed that they practised some web archiving.

5.3.3 Benefits and challenges of web archiving

All the participants could not give clear and direct responses of benefits and challenges to web archiving because there was no web archiving programme in the Document Warehouse hence the findings were the benefits and challenges of the little web archiving carried out when necessary. The use of M-files allowed the managing and use of web records even after their initial source or web page had been updated hence web records management was possible. The findings indicated that obstacles to web archiving in the Document Warehouse included cost implications due to accidental loss or deletion of records and the exhaustive and time consuming teaching of clients who want to host their own data and web records.

5.3.4 Enablers to web archiving

Findings showed that a risk management policy for all records in all formats with web archives included made it possible to apply web archiving at the institution since there was no web archiving policy. There was collaboration between records personnel and IT staff in web archiving at the Document Warehouse.

5.3.5 Appraisal and selection

The findings showed that all data and web records were treated as important and the client selected the web records they required to be archived by the institution. The Document Warehouse had no authority as to which web records must be archived.

5.3.6 Access and use of web archives

The findings revealed that clients accessed and used their own web records with the help of a Document Warehouse employee so as for better retrieval of web records.

5.3.7 Quality assurance

The findings showed that quality assurance was carried out in M-files. The M-Files had a built in security system which also guaranteed authenticity.

5.4 Conclusions

Conclusions are presented and drawn from the analysed findings to determine if they met the objectives of the study (Cresswell, 2012). The conclusions of this study were drawn from the objective and sub-objectives.

5.4.1 To investigate the practice of web archiving at the Document Warehouse

Findings revealed that the Document Warehouse did not practice web archiving. There was neither a web archiving programme nor a web archiving policy. There however, was a presence of brief web archiving practice but on ad hoc basis when it is needed and necessary and not as an organization function.

5.4.2 To establish if the staff were aware of web archiving

The findings showed that all the research participants were not aware of web archiving and could only understand the terms ‘web’ and ‘archiving’ independently.

5.4.3 To find out the processes of web archiving in the organisations

The findings revealed that since there was neither a web archiving programme nor a web archiving policy, there were no official and established processes of web archiving in Document Warehouse. Findings also showed that they had a process of web archiving which

was taking monthly snapshots of the entire server which hosted the content of the web interface even though it was recognised as a web archiving process by the organisation. These snapshots were managed in a content management system called M-files.

5.4.4 To determine the barriers and enablers to for successful web archiving

The study showed that the barriers and enablers to web archiving could not be fully recognised or clearly stated due to the absence of a web archiving programme which is a challenge within itself. The study revealed that the loss or deletion of records accidentally was costly to the Document Warehouse as well as time consuming to recreate those records. It was also revealed by the study that web records were covered in the risk management policy of the organisation and there was collaboration between the IT and records staff.

5.4.5 To determine the benefits of web archiving to records management

The study revealed that all the participants could not give clear benefits of web archiving for the organisation. The study showed that due to their use of M-files, records personnel were enabled to maintain the usability and management of web records as they could be stored and managed longer than they were on the web.

5.4.6 To recommend ways in which web archiving can be enhanced in the organisation.

One of the objectives of the study was to come up with recommendations on how web archiving can be enhanced in the organisation. Based on the findings of the study, the following recommendations can be made:

1. The Document Warehouse should send its employees to training on web archiving
2. The Document should apply a second method of capturing web records such as automated capture.

3. The Document Warehouse should have specific employees assigned to teaching clients who wish to host their own data and web records so that it does not impede other activities of the institution.
4. The Document Warehouse should have a web archiving programme and web archiving policy for all web records.
5. The Document should create an access policy for web records which would contribute in giving access of web records to its clients through an approved policy.
6. The Document Warehouse should have its own quality assurance tool for web archives other than the one that came with M-Files so as to further strengthen authenticity.

5.6 Area for further research

This study focused on a private company so it is essential that future research focuses on public companies so as to find out their position on web archiving.

5.7 Final conclusions

This study focused on the practice of web archiving using the Document Warehouse as a single case study and a qualitative research approach was used with interviews as data collection methods. Through purposive sampling, three out of the ten employees in records management & IT were selected for the study. The study revealed that all the participants were unaware of what web archiving was, however, the Document Warehouse practiced web archiving without labelling it as such and viewed it as a form of digital preservation. Management of web records was ensured by capture using monthly snapshots which were managed in M-files, an enterprise content management system which also had an inbuilt security system that guaranteed authenticity. Cost implications due to the loss or deletion of web records created challenges for the Document Warehouses in their archiving of websites.

They did not have a web archiving policy but had a risk management policy which was part of their business continuity plan which included web archiving. An advantage to the Document Warehouse for practicing web archiving was that it attracted more clients.

The main recommendations were that the Document Warehouse should send their employees to training on the practice of web archiving especially the IT and record staff. They should also have a web archiving policy and programme to in order to carry web archiving activities efficiently. The study's findings add to the body of knowledge on the practice of web archiving and can be used to guide both public and private organisations on web archiving. The study can open a discussion on implementing web archiving in Namibian organisations' main functional activities and display the major role records management play in the effective practice of web archiving uopn collaboration with the IT personnel so as to ensure that the content, context and structure of web records is maintained.

References

Adams, W. C. (2015). Conducting semi-structured inteviews. In K.E. Newcomer, H.P. Patty & J.S. Wholey, *Handbook of practical program evaluation* (pp.492-505). San Francisco, CA: Jossy Bass Whiley.

Adoghe, A., Kayode, O., Dike, U. I., & Olujimi, A. (2013). Web archiving: techniques, challenges and solutions. *International Journal of Management and Information Technology*, *5*(3), 598-603.

Alex Ball. (2010). *Web archiving* (version 1.1). Edinburgh, UK: Digital Curation Centre.

Azar, B. (2020). *Discussing your findings.* Retrieved from https://www.apa.org/gradpsych/2006/01/findings#

Brown, A. (2006). *Archiving websites: A practical guide for information management professionals*. London, UK: Facet Publishing.

Bragg, M., & Hanna, K. (2013). *Web archiving life cycle model. The Archive-It Team, Internet Archive*. Retrieved from http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf

Cresswell, J. W. (2012). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (4th ed.). Boston, MA: Pearson Education

Cresswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage Publications Inc.

Costa, M., Gomes, D., & Silva, M. J. (2017). The evolution of web archiving. *International Journal on Digital Libraries*, *18*(3), 191–205.

Duquesne University. (2020). *What are research instruments.* Retrieved from

https://guides.library.duq.edu/researchinstruments

Enago Academy. (2019). *Research ethics & misconduct: What researchers need to know.* Retrieved from https://www.enago.com/academy/principles-of-ethical-research/

Gomes, D., & Costa, M. (2014). The importance of web archives for humanities. *Imternational Journal of Humanities and Arts Computing, 8*(1), 106-123.

Google Play. (2020). *M-Files.* Retrieved from https://play.google.com/store/apps/details?id=com.mfiles.mobile&hl=en&gl=US

Hale, A. S., Blank, G., & Alexander, V. D. (2017). Live versus archive: Comparing a web archive to a population of web pages. In N. Brügger & R. Schroeder (Ed.), *The web as history: Using web archives to understand the past and the present* (pp. 45-61). London, UK: UCL Press

Heale, R & Twycross, A. (2018). What is a case study? *Evid Based Nursery*, *21*(1), 7-8.

Ho, J. (2018). *The Chicago Community Trust: Web Archiving: Keep Your Organization from Losing Its Memory.* Retrieved from https://www.cct.org/2017/08/web-archiving-keeping-your-organization-from-losing-its-memory/

InterPARES/ICA. (2012). *Digital records pathways: Topics in digital oreservation: Module 7: Management and preservation of records in web environments.* Draft July 2012. Retrieved from http://interpares.org/ip3/display_file.cfm?doc=ip3_canada_gs12_module_7_july-2012_DRAFT.pdf

Jackson II, R. L., Drummond, D. K., & Camara, S. (2018). What Is Qualitative Research? *Qualitative Research Reports in Communication, 8*(1), 21-28.

Littman, J., Chudnov, D., Kerchner, D., Peterson, C., Tan, Y., Trent, R., … Wrubel, L. (2018). API-based social media collecting as a form of web archiving. *International Journal on Digital Libraries*, *19*(1), 21–38.

Luo, A. (2020). *Scribber: What is content analysis and how can you use it in your research?* Retrieved from https://www.scribbr.com/methodology/content-analysis/#:~:text=%20How%20to%20conduct%20content%20analysis%20%201,the%20units%20of%20meaning%20into%20the...%20More%20

Mathers, N., Fox, N, J., & Hunn, A. (2000). Using interviews in a research project. In Wilson, A., Williams, M., & Hancock, B. (Eds.), *Research approaches in primary care* (pp. 113-134). Oxford: Radcliffe Medical Press.

Martin, K. E., & Eubank, K. (2007). The North Carolina State government website archives: A case study of an American government web archiving project. *New review of Hypermedia and Multimedia*, *13*(1), 7–26.

Masanes, J. (Ed.). (2006). *Web Archiving.* New York, NY, USA: Springer

McCombes, S. (2019). *How to write a literature review.* Retrieved from https://www.scribbr.com/dissertation/literature-review/

McCombes, S. (2019). *How to write a research methodology.* Retrieved from https://www.scribbr.com/dissertation/methodology/

Monash University. (2020). *Reporting and discussing your findings.* Retrieved from https://www.monash.edu/rlo/graduate-research-writing/write-the-thesis/writing-the-thesis-chapters/reporting-and-discussing-your-findings#discuss-your-findings

Neville, C. (2007). *Effective learning service: Introduction to research and research methods.* West Yorkshire, UK: University of Bradford.

Niu, J. (2012). *An overview of web archiving. D-Lib Magazine*, 18 (3/4). Retrieved from http://www.dlib.org/dlib/march12/niu/03niu1.html

Pennock, M. (2006). Web archiving. *DPC Technology watch report 13/03/2013*. Retrieved from https://www.dpconline.org/docs/technology-watch-reports/865-dpctw13-01-pdf/file

Pennock, M., & Kelly, B. (2006). Archiving web site resources: A records management view. *Proceedings of the 15th International Conference on World Wide Web*, (June), 987–988.

Publications Office of the European Union. (2020). *About the web archive of the European Union institutions*. Retrieved from https://op.europa.eu/en/web/web-tools/euwebarchive

Qualitative Research Guideline Project. (n.d.). *Semi structured interviews.* Retrieved from http://www.qualres.org/HomeSemi-3629.html

Robert Wood Johnson Foundation. (2008). *Qualitative research guidelines project: Semi-structured interviews.* Retrieved from http://www.qualres.org/HomeSemi-3629.html

Roper, M. (Ed.). (2009). *Module 4, Preserving electronic records: Training in Electronic Records Management*. London, UK: IRMT.

Taylor, N. (2017). Introduction to the special issue on web archiving. *Journal of Western Archives, 8*(2), 1-6.

Thakur, S., & Chetty, P. (2020). *Project Guru: How to establish the validity and reliability of qualitative research?* Retrieved from https://www.projectguru.in/how-to-establish-the-validity-and-reliability-of-qualitative-research/

Taherdoost, H. (2018). Sampling methods in research methodology; How to choose a sampling technique for research. *SSRN Electronic Journal*, *5*(September), 18–27.

The Document Warehouse. (n.d.). *Company history.* Retrieved from https://thedocumentwarehouse.com/about-tdw/company-history/

The Document Warehouse. (n.d.). *Services*. Retrieved from https://thedocumentwarehouse.com/services/

Salkind, N. J. (2010). *Encyclopedia of research design.* Thousand Oaks, CA: SAGE Publications Inc.

United Kingdom (UK), The National Archives. (2011). *Web archiving guidance*. Retrieved from http://www.nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf

Walliman, N. (2011). *Research methods: The basics.* Milton Park, Abingdon: Routledge, Taylor & Francis Group.

Weigle, M. C. (2018). *Insights from the social sciences: On the importance of web archiving.* Retrieved from https://items.ssrc.org/parameters/on-the-importance-of-web-archiving/

Yin, R. K. (2014). *Case study research design and methods* (5th ed.). Thousand Oaks, CA: Sage.

APPENDIX A

NON-DISCLOSURE AGREEMENT (NDA)

Purpose of the Non-Disclosure Agreement

This Non-disclosure Agreement is entered into by The Document Warehouse and Mr Leroy Kopera, a fourth-year student conducting research in partial fulfilment of the B.A. in Records and Archives Management (Hons) at the University of Namibia.

The purpose of this non-disclosure agreement is to clarify that the information The Document Warehouse provides during this interview will never be used for anything other than academic purposes and that any person you disclose the information to will use this information for academic reasons only.

Signing this agreement means you have read and understood the agreement.

Signed at Windhoek, on the 13 day of November 2020.

(Signature)_____ Witness _____

On behalf of "The Document Warehouse (Pty) Ltd"

Signed at 10:52hrs on the 14th day of July 2020.

(Signature)_____ Witness Rivman Kopera

On behalf of "the Student"

APPENDIX B

PERMISSION LETTER

University of Namibia, Private Bag 13301, Windhoek, Namibia
340 Mandume Ndemufayo Avenue, Pioneerspark
☎ +264 61 206 3111; URL.: http//www.unam.edu.na

UNAM
UNIVERSITY OF NAMIBIA

8 June 2020

To Whom It May Concern
Document Warehouse

Re: Request for Permission to Conduct Research

We wish to introduce to you Mr Leroy Kopera, a fourth year student conducting research in partial fulfillment of the B.A. in Records and Archives Management (Hons). His research project is titled *"The practice of Web Archiving in Namibia"*.

We are requesting your assistance by granting the student permission to conduct the study in your organisation. We rely on the support of our stakeholders for the success of our programmes.

His contact details are +264 81 626 3020 and koperaleroy@yahoo.com.

Thank you in advance for your support.

Yours Sincerely

Prof C.T. Nengomasha

Supervisor, Department of Information and Communication Studies

Cell: 0812787617; Office: 2063641; email: cnengomasha@unam.na

APPENDIX C

INFORMED CONSENT FORM

Title of Research Project: "The practice of web archiving in three Namibian institutions"

Researcher: Kopera Leroy T.
Student number- 201501887
Supervisor: Prof. C. T. Nengomasha - 061 206 3641
cnengomasha@unam.na

Information

This research explores the practice of archiving web content in the Namibian state by using case studies of three Namibian institutions.

All the information collected as part of this study is confidential. No name or identity will be published in the write up of the findings and confidentiality and anonymity will be adhered to at all times.

Participation is voluntary and you may choose not to participate or withdraw from participation at any time. I however, appeal to you to assist in the success of this research through your participation.

If you voluntarily agree to participate in this study, please indicate your consent by signing this form.

Signature: _____

Date : _____

APPENDIX D

Interview guide for the records personnel

1. GENERAL INFORMATION

1.1 What is your job/title in the organization?

1.2 How long have you been working with records?

2. WEB ARCHIVING

2.1 Web archiving practice

2.1.1 Are you familiar with the practice of web archiving?

2.1.2 Does the organization practice web archiving?

2.2 Web archiving technique

2.2.1 Are you directly involved in the web archiving process?

2.2.2 If yes, what type of web archiving technique do you use?

2.2.3 If not, are there any provisions for preserving vital web content?

2.3 Benefits to web archiving

2.3.1 What are the benefits of web archiving to you as records personnel?

2.3.2 How has the organisation benefited from applying web archiving?

2.4 Challenges to web archiving

2.4.1 What are some of the challenges to web archiving have you encountered?

2.4.2 Are there any measures that can be put in place to limit those challenges?

2.5 Enablers to web archiving

2.5.1 Is there a policy that guides web archiving in the organisation?

2.5.2 If not, are there any protocols that guide decision in relation to web archiving?

2.5.3 Do you collaborate with the IT staff when it comes to web archiving?

2.6 Risk management

2.6.1 Is there a risk management policy for web records?

2.7 Access to and use of web archives.

2.7.1 Who has access to the web archives?

2.7.2 Is the public allowed access and use of the organisation's web archives?

APPENDIX E

Interview guide for the IT personnel

1. GENERAL INFORMATION

1.1 What is your position/job title in the organization?

1.2 How long have you been working in IT?

2. WEB ARCHIVING

2.1 Web archiving practice

2.1.1 Are you familiar with the practice of web archiving?

2.1.2 Does the organization practice web archiving?

2.1.3 Who has the authority on what web content is to be archived?

2.2 Web archiving technique

2.2.1 What type of web archiving technique or process do you use?

2.2.2 If there is none, are there any provisions for digital or web records preservation?

2.3 Benefits of web archiving

2.3.1 What are the benefits of web archiving to you as IT personnel?

2.3.2 How has the organization benefited from utilizing web archiving in their business functions?

2.4 Obstacles to web archiving

2.4.1 What obstacles to web archiving have you faced in the organisation?

2.4.2 Do you think those obstacles can be dealt with so as to improve web archiving?

2.5 Enablers to web archiving

2.5.1 Is there a policy that guides web archiving?

2.5.2 If not, are there any protocols that guide decision making when it comes to web archiving?

2.5.3 Do you collaborate with the records personnel in archiving websites?

2.6 Appraisal and selection

2.6.1 How does the organisation choose which web records to archive?

2.6.2 If not, does the organisation archive all the web records?

2.6.3 Who has the authority on what web content should be archived?

2.7 Quality assurance

2.7.1 How do you assess the archived web records so as to ensure authenticity?