Rethinking the Potentials of Web Crawling Technologies for Digital Library Services in Nigeria

Ahmed Mohammed
Department of Library and Information Science
Bayero University
Kano, Nigeria
E-mail: amkwaru69@gmail.com

Auwalu Dansale Yahaya
Department of Library and Information Science
Bayero University
Kano, Nigeria
E-mail: ad.yahaya.lib@buk.edu.ng

Received: 19-Sep-2022, Manuscript No. IJLIS-22-72877; **Editor assigned:** 22-Sep-2022, PreQC No. IJLIS-22-72877 (PQ); **Reviewed:** 06-Oct-2022, QC No IJLIS-22-72877; **Revised:** 23-Dec-2022, Manuscript No. IJLIS-22-72877 (R); **Published:** 03-Jan-2023, DOI: 10.35248/2231-4911.23.13.843

***Abstract**: The importance of web crawling/web harvesting services in librarianship cannot be over emphasized especially now that libraries around the globe have been mandated to switch to electronic methods so as to satisfy their users effectively. This paper discusses the potential benefits libraries will derive by promoting the use of web crawling technologies for their library services with reference to Nigeria. It further explores the concept and significance of web crawling, web crawlers, approaches to web crawling, and web crawling in libraries. The paper also highlights the experiences of other libraries globally on the integration of crawling technologies in their digital information services. It also justifies the need for librarians in Nigeria to adopt web crawling technologies in their libraries. Challenges that could undermine the successful application of such technologies were equally explored. The paper concludes that, if appropriately used, the content of the chapter will provide valuable additional input to information professionals for the design of the web archiving as a tool for preserving information resources deposited in multiple websites so as to ensure its security by storing it on their local library databases for future use.*

Keywords: Web crawling technologies, Digital library, Nigeria.

Introduction

According to Nielsen web crawling is the method most commonly used to harvest websites on a large scale called macro archiving, or for micro archiving if a researcher wishes to archive a single website including link structures. Web archiving is an emerging concept whose main purpose is to preserve websites with cultural or historical significance. According to Miranda (undated) web archiving systems are useless if one is unable to successfully access the stored data in the next five, ten or more years to come. Crown reported that information published on the web is increasingly becoming the only place where it is available. Because of this, the website is a crucial part of the records and identity of an organization or individual. Web pages should be harvested in their original form and be capable of being delivered as they were on the live web.

As state by Brugger, within a year, 40 percent of the material on the Internet disappears, while another forty percent has been changed, which is why today we can only expect to find twenty

percent of the material that was on the Internet one year ago? Similarly, Kahle corroborated this assertion by saying that web pages only last for about 100 days on average before they change or disappear. This calls for a necessity of the libraries to engage in practices called web archiving services so as to preserve their digital resources for optimum use of their clientele. In response to this, Grotke, opined that many types of organizations archive the web, not just national libraries. Archives, universities, museums, and government organizations are all involved, plus corporations and others preserving their own content. Moreover, Nielsen reported that data can be made less accessible by search engines such as Google by removing them from their indices, so that they can no longer appear in future searches. Thus, according to Grotke many national libraries of today have established web archiving programs, but still others are beginning to consider archiving portions of the web to meet legal deposit requirements or to simply collect the digital output of their own country's citizens.

Literature Review

This write-up is based on a comprehensive review of literature that explains how web crawling is being done. The content of the chapter will provide valuable additional input for the design of the web archiving as a tool for preserving information resources deposited in multiple websites so as to ensure its security by storing it on local databases for future use.

Bragg, Hanna, Donovan, Hukill and Peterson indicated that Web archiving is performed by libraries, archives, companies and other organizations around the world. Anthony, Onasoga, Ike and Ajayi quoted Brewster Kahle, founder of the Internet Archive, once said that "the average life of a web page is 100 days". These authors added that web pages disappear on a daily basis as their owners revise them or servers are brought out of service. Content is lost at an alarming rate, risking not just our digital cultural memory, but also organizational accountability. Therefore, to help preserve web content, websites are captured and archived for long-term access through web archiving. Lyman opined that the average life span of a Web page is only 44 days, and 44 percent of the Web sites found in 1998 could not be found in 1999.

It is obvious that, many organizations create websites as part of their communication with the public and other organizations as they are powerful tools for sharing information. However, from early online content in the early 1990's to around 1997, very little web information survived.
The founding of the International Internet Preservation Consortium (IIPC) in 2003 according to Niu, has greatly facilitated international collaboration in developing standards and open source tools for the creation of web archives. These developments and the growing portion of human culture created and recorded on the web, combine to make it inevitable that more and more libraries and archives will have to face the challenges of web archiving.

Evolution of web crawling services in libraries
The concept of web archiving began at an initial stage in Taiwan. In contrast, internet archive, an unprofitable organization established at San Francisco, has been devoted to collecting all kinds of digital materials for the potential applications of researches since 1996. All kinds of websites are the targets of internet archive. National library of Australia also built PANDORA and joined the web archiving in 1996. The United States of America, the United Kingdom, and Japan executed similar projects consequently.

In the words of Miranda, (undated) the Internet Archive (IA) was one of the first attempts to preserve the web. It was founded in 1996 with the purpose of offering permanent access for researchers, historians, and scholars to historical collections existing in digital format. Although established as a non-profit organization, it collected pages that had been crawled by Alexa

Internet, a commercial company developing web navigation tools. Internet Archive is in close cooperation with institutions like the Library of Congress and the Smithsonian Institute.

Concept of web crawling

According to Crown web archiving is the process of collecting websites and the information they contain from the world wide web, and preserving them in an archive. Information is selected, stored, preserved and made available to people. Access is usually provided to the archived websites, for use by government, businesses, organizations, researchers, historians and the public. Kamba and Yahaya reported that the process of running web archiving software is what is termed as harvesting or crawling. Similarly, Sandvik was quoted by Nielsen viewing web harvesting (crawling) as the process of collecting web material and loading it into a fully brow sable web archive, with working links, media etc.

Pearson, Oury, Goethals, Sierman, and Steinke stated that web archiving is performed by libraries, archives, companies and other organizations around the world and many of these web archives are represented in the international internet preservation consortium. Jacobsen justified that online materials are considered public if anybody can get access to it, whether for free, by paying a fee or by providing personal information to the website (giving your name, e-mail etc.). Websites of associations are similarly considered public, if membership is open to all or segments of the general public. The more simple and static the site is, the more easily it is harvested (Miranda, undated).

Significance of web crawling services

Web archiving is a vital process to ensure that people and organizations can access and re-use knowledge in the long-term, and comply with the needs of retrieving their information. Harvesting with link crawlers has the great advantage of ensuring that entire web pages are collected. Thus, the relations between web objects are preserved, and that the archived material can subsequently be displayed in an interface in which it looks and behaves like the live we. In the words of Anthony, Onasoga, Ike and Ajayi, an ever growing international web archiving community continues to actively develop new tools to improve existing techniques, and to stop the continuous loss of web content caused by the transitory nature of world wide web.
According to Nielsen many good reasons to archive web materials are to:

- Maintain our digital cultural heritage.
- Stabilize and preserve web materials as a research object.
- Be able to document and illustrate claims based on analyses of web materials (whether the web itself is the research object or a source of knowledge about other research objects).

At the time that the internet archive began crawling the web, national libraries and archives around the world also began to see the importance of preserving this global resource. Many countries even sought changes in their laws to allow or mandate them to collect and preserve the output of their nation's creativity.
Chen, Chen and Ting brought that the national Taiwan university library assign metadata for each archived website to provide most useful information to users based on the policies of collection development [1-5].
Jacobsen identified that Denmark began web archiving in 2005 and the experiences are presented with a specific focus on collection-building and issues concerning access. The actual collection requires strategies for harvesting relevant segments of the internet in order to assure as complete coverage as possible.
The national library of France built automated indexing processes to enable fast access to

harvested content. Each file is dated and described to gather only necessary information (original location on the Web, format, size, localization in the archives, etc.). This indexing process makes it possible to then re-play archived websites within their publication environment and browse them by clicking links, just like on the living Web, but in a historical, dated context. In order to be granted access, end users must be over 18 years old and give a proof of their need to access these archives for academic, professional or personal research activities.

Niu remarked that, very large web archives often rely on automatic metadata generation. Some metadata information, such as the timestamp generated when the web resource was harvested, the status code (e.g., 404 for not found or 303 for redirection), the size in bytes, the URI, or the MIME type (e.g., text/html), can be created or captured by crawlers. Metadata information can also be extracted from the Meta tags of HTML pages. Small-scale web archives can afford to create metadata manually.

According to Niu web archive collections have a multilevel hierarchical structure. A web archive collection may include a number of crawling sessions. In each crawling session, a number of websites are crawled. Each website includes many web pages. Each webpage may be composed of many files such as a text file, an image file and a video file. In addition, some metadata for the item level objects, such as file format, size in bytes and date of modification, can be automatically extracted.

The Montana State Library (MSL) web archive seeks to archive state documents, which are now often only available online with an aim of meeting the information needs of state agency employees, provide permanent public access to state publications, support Montana libraries in delivering quality library content and services, work to strengthen Montana public libraries, and provide visually or physically handicapped Montanans access to library resources".

As revealed by Aubry, the web archives in the national library of France are available on 350 computers placed in all research library reading rooms, along with all the library's e-services (information websites, booking facilities, catalogues, internet access) and e-resources (digitized books and images, electronic journals, online databases, CD-ROMs, bookmarks). These computers are available to the public, but users need to make reservations for a place in the Library to be able to use them [6-10].

Crown emphasized that when a website is archived, the context of the information it provides is maintained, as such users can view the information in the context in which it was originally presented.

The advantages of web crawling according to Nelson include:

- The full website can be archived whenever the harvester is set up to follow all links throughout the website, and what is archived usually resembles what was online to a large extent.
- The full length of the individual page is preserved.
- The link structure is preserved, and thus the interrelations between web elements and pages, as well as other websites are also preserved.
- The archived material looks and behaves like the live web (with some important exceptions).
- The archived version is displayed in a browser, and it is clickable so that user can move around by following the links, just like on the live web (except for the temporal issues).
- The html is archived, which means that the archived versions are machine-readable, which provides good possibilities for searching and sorting, and enables links to be clickable.
- It can be performed automatically (in part, as there is a need for ongoing monitoring to evaluate the collected material and deal with any technical difficulties that arise).
- Access to metadata (crawl logs).
- Can be used for big data analysis (e.g. content analysis, network analysis etc.

Beis, Harris and Shreffler remarked that the internet archive, as a non-profit organization based out of California, has been capturing websites since 1996 as part of its mission to provide access to all knowledge. Its automated crawls collect around 1.5 billion pages per week and about 1 million captures per week from individuals using the "save page now" function of its "way back machine".

Enablers of web harvesting

According to Crown the largest web archiving organization crawling the web is the Internet Archive which aims to maintain an archive of the entire world wide web.
Commercial web archiving software and services are also available to organizations that need to archive their own web content for their own business, heritage, regulatory, or legal purposes.
The International Internet Preservation Consortium (IIPC) is an international association seeking the development of efforts for the preservation of Internet content. The IIPC was launched in Paris on July 2003. Its twelve members were the national libraries of France, which is the coordinator of the project, Italy, Finland, Sweden, Norway, Australia, the national and university Library of Iceland, library and archives Canada, the British library, the library of congress (USA), and the internet archive (USA).
Miranda, (undated) brought that The International Internet Preservation Consortium (IIPC) promotes global exchange and international relations, acquiring, preserving and making accessible Internet information for future generations around the world. It aims at enabling the collection of a rich body of worldwide internet content in a way that it can be archived, secured and accessed over time. It fosters the development and use of common tools, techniques and standards that enable the creation of international archives. It encourages and supports national libraries everywhere to address Internet archiving and preservation.
According to Bragg, Hanna, Donovan, Hukill and Peterson Archive-It is a subscription web archiving service that helps partner organizations harvest, build, and manage born digital collections.
Starting in 2010, the University of Alberta began using Archive-It as a broader collection development tool.
Beis, Harris and Shreffler brought that the university of Dayton libraries purchased a subscription to Archive-It in 2014. The university archives and special collections contain records of University departments, student organizations, faculty papers, and other special collection.
Institutions access Archive-It through a web application to harvest, organize, and catalog their collection material. Users can automate the crawls with different frequencies, set parameters about the depth that the crawls will take, and add metadata at the item and collection levels to enable better searching for used.

The need for web crawling services in libraries

According to Nielsen the web is an unpredictable medium. Its pace of change is rapid, and it can be hard to predict what types of new online content and services will become the next hit on the web Anthony, Onasoga, Ike and Ajayi remarked that the internet Web keeps increasing in size, adding several million Web pages daily, and presently consists of billions pages of publicly available content and information. Content on these pages is structured in different ways, comes in of many formats and includes text, videos, and images as well as links between pages and to content in other formats such as PDF or docx. However, web pages are constantly being updated, relocated, or removed and you may not always be able to go back to something you saw before. For example, a video or article you found on a site last month may not be available anymore or the site you referenced in a paper may no longer even exist.

Grotke indicated that with so much global cultural heritage being documented online, librarians, archivists, and others are increasingly becoming aware of the need to preserve this valuable resource for future generations.

Web crawlers (Harvesting Technologies)

Crown stated that as the Web contains a massive amount of websites and information, web archivists typically use automated processes to collect websites. The process involves 'harvesting' websites from their locations on the live Web using specially designed software. This type of software is known as 'a crawler'. Crawlers travel across the Web and within websites, copying and saving the information. The archived websites and the information they contain are made available online as part of web archive collections. Nielsen revealed that when we talk about web archiving, a crawler is often described as a 'harvester'.

Samouelian and Dooley discussed some examples of web crawlers as follows;
Heritrix: In 2002, the Internet Archive released Heritrix, the open source web crawler, which is the software tool that captures content from the World Wide Web. In 2009, the Heritrix crawler's file output, the WARC file, was adopted as an ISO standard for web archiving, demonstrating both the prevalence of active web archiving programs and the importance of the web crawler itself.
A widely used, open source web crawler developed by the internet archive and is one of the principal capture tools used by internet archive and by many others for harvesting websites. It produces web archives but does not allow for the input or generation of additional descriptive metadata within the tool. The acquisition tool which Library of Congress Web Achieves (LCWA) applies is the Heritrix web crawler. All web archives that are members of the International Internet Preservation Consortium (IIPC), including Netarkivet and the Internet Archive, use Heritrix, which is a flexible and scalable harvester.
HT Track: An open-source capture tool that uses an off-line browser utility to download a website to a directory generates a folder hierarchy and saves content that mirrors the original website structure. It also does not allow for input of any descriptive metadata.
Site Story: An open source transactional tool that captures every version of a resource as it is requested by a web browser, thus replicating the user's experience. The resulting archive is effectively representative of a web server's entire history; versions of resources that are never requested by a browser are not included.
Memento group: Open source tooling built on Memento protocols, such as MediaWiki Memento Extension and Memento Time Travel that facilitates access to archived websites through http content negotiation based on capture date.

Web archiving tools
Samouelian and Dooley in their paper discussed the following web archiving tools as follows;
Archive-It: This is a widely used subscription web archiving service from the Internet Archive (IA) that harvests websites using a variety of capture technologies including Internet Archive's open-source web crawler Heritrix. Files are preserved in the IA digital repository and can be downloaded by users for preservation in their own repositories. Archive-It provides 16 Dublin Core metadata fields from which users can choose, as well as the ability to add custom fields that can be added manually at the collection, seed and document levels.
Institutions access Archive-It through a web application to harvest, organize, and catalog their collection material. Users can automate the crawls with different frequencies, set parameters about the depth that the crawls will take, and add metadata at the item and collection levels to enable better searching for users.
Bragg, Hanna, Donovan, Hukill and Peterson identified that the University of Alberta has a very

large network of individuals actively using Archive-It, many of whom are subject specialists.
As reported by Bragg, Hanna, Donovan, Hukill and Peterson, the most active users of Archive-it in Montana State Library are the state publications librarian (who oversees the program), the metadata cataloger, and the library systems programmer/analyst who handles technical issues.

Netarchive Suite: This is open-source software which is used to plan, schedule and run web crawls. Net archive suite consists of several modules, including a harvester module (which uses Heritrix) for defining, scheduling and running crawls; an archive module that serves as a preservation repository for harvested material; and an access module for viewing harvested material.

Social feed manager: This is an open-source web application that allows users to create collections of data from social media platforms, including Twitter, Tumblr and Flickr. The application harvests the data, including images and web pages linked from or embedded in the social media content, and use Heritrix *via* public APIs. Consistent descriptive metadata such as creator and date are inherently embedded in social media posts and are automatically generated. Other descriptive metadata pertaining to selection of a collection (including title and a description) or set of collections must be manually generated.
Littman et al. advocated that social media archiving practices and tools should align with those of web archiving.

Way back: Released by the internet archive in 2005 is software that replays the already archived webpages. It consists of a group of interrelated applications that index and retrieves captured web content, and display the content in a web based user interface. For example, the Beta Way back Machine, which released in October 2016, expands its search capabilities, including the ability to perform a keyword search against an index of terms contained within the archive. This tool allows anyone to capture an individual website page and add it permanently to the internet archive's collection of websites.

Web archive discovery: This is an open source tool that enables full-text search within web archives by indexing the archived files. The only descriptive metadata captured are crawl date, crawl year and Way back date.

Web curator tool: This is an open-source workflow management tool designed to enable non-technical users to manage the web archiving process. It can be customized to include a variety of specialized tasks to support web acquisition and description, including permission/authorization, selection/scoping/scheduling, harvesting, quality review and archiving. Harvest date is automatically captured, while all other descriptive metadata must be added by the user.

Web recorder: This is an open source tool which preserves user's experience. It captures the exact sequence of navigation through a series of web pages or digital objects. The recording produces the archived files with a high level of detail that can then be replayed using the Web recorder. It captures descriptive metadata elements of the resources such as creator, title, capture date/time, archive file format, title of collection (if included in the metadata) and all URLs the user visited during a recording session.

Web crawling in libraries

Library of North Carolina archive social media feeds generated by state agencies on Facebook, Twitter and Flickr because they see these feeds as extensions of their official web based records.
Anthony, Onasoga, Ike and Ajayi remarked that the largest web archiving organization crawling the web is the internet archive, an organization which aims to maintain an archive of the entire world wide web. The internet archive, is a non-profit organization which aims to build a digital library of Internet sites, it has been archiving websites and web content since 1999 with the purpose of preserving virtually "everything" and currently crawls portions of the web every few months. It is home to the largest collection of web content, with more than 2 petabytes of compressed data and over 150 billion websites captures. Internet Archive archives websites by the topics, such as World War II, the US election, and so on. To allow libraries, archives, or

content creators to create collections and archive selective web content, the Internet Archive created Archive-It, a subscription-based web archiving service. Web pages captured by the Internet Archive or through Archive-It are full-text and are searchable through their respective services. In addition to its own search and browse tools, the Internet Archive uses the Way-back Machine interface to allow users to search for specific URLs that may have been archived Anthony, Onasoga, Ike and Ajayi.

Discussion

A consortium of universities, government agencies, professional associations, and other entities with an interest in digital preservation called in USA national digital stewardship alliance, has conducted four surveys of institutions that have web archiving programs to learn more about the current state of the field from 2011-2017. Their findings reported that the most recent survey from 2017 found that 60.5% of organizations engaging in web archiving were colleges or universities. Public libraries made up 12.6% of the total. Other institutions with web archiving programs include federal and state governments (13.5% combined); and historical societies, commercial organizations, consortia, K-12 schools, museums, and other 13.4% combined.

Experiences from other countries

In 1996, the Swedish Royal library started one of the pioneer web archiving projects called Kulturarw. It used an active collection policy based on the development of a harvesting robot by the Royal Library. At first, public access was not part of the plan, but in 2002 the Royal Library allowed public access to the collection. Kulturarw uses a harvesting crawlers running twice a year to collect the Sweden sites. Its archive is around 4 to 5 TB and is growing at an average rate of 2-3 TB per year (Miranda, undated).

The University of Nebraska-Lincoln libraries are now a member of Archive-It, the Internet Archive's subscription web archiving service, allowing the institution to identify, archive, and preserve web content of historic importance to the university.

The university of Dayton Libraries began using Archive-It to capture websites relevant to their collecting policies in 2015. However, the collections were only made available to users from the university of Dayton page on the Archive-It website.

National Library of Australia in 1996 built PANDORA Digital Archiving System (PANDAS). Now, it is built in collaboration with nine other Australian libraries and cultural collecting organizations. It focuses on collecting the websites in Australia and categorizes the websites into eighteen subjects, including arts, education, politics, sciences, history, and so on. Their users could easily browse archived websites by the alphabetical order. The web crawler adopted by PANDORA is HT Track (Chen, Chen and Ting, undated)

In June, 2006 the UK Web Archiving Consortium (UKWAC) was originated by six institutions with British Library as the Leading partner. Others are the National Archives, National Library of Wales, National Library of Scotland, JISC, and Wellcome Trust. This Consortium aims at collecting websites related to UK. It allows the participating institutions to have their own policies for archiving. For example, Wellcome Library concentrates on the medical websites while British Library focuses on the critical cultural, historical or political issues. UKWAC engaged in web archiving services after reaching the agreement with the owners of websites (Chen, Chen and Ting, undated)

The national diet library of Japan proposed Web Archiving Project (WARP) was matured in 2006. It has collected more than 3,000 websites of governments, universities, and special topics and unofficial institutions. WARP harvests each website at least once per year and also features keyword search function.

Bragg, Hanna, Donovan, Hukill and Peterson discovered that Montana's state library created a

portal on their own website that provides access to their Archive-It collections in addition to providing access to data collected using the Archive-It service, Montana State Library extracted older webpages dating back to 1996 from the internet archive's general web archive. These webpages are accessible from the portal along with their Archive-It data, which dates back to 2006. They also found other innovative ways to draw attention to their web archives. All of their webpages contain a "page history" link in the footer. These links direct visitors to archived versions of the webpage so that they can see how it has changed over time.

Webs inform all was set up by the internet laboratory of Peking University. The growth of their archived websites is about forty-five millions of web pages per month and the total number of web pages is three billions as at April, (Chen, Chen and Ting, undated).

Similarly, the Marian Library Archivist captured a blog that is related to their significant collection. They also capture the organizational website of the Mariological Society of America, a theological organization whose paper materials are maintained at the University of Dayton.

Approaches to web harvesting

According to Chen, Chen and Ting (undated), there are three options for setting up schedule of harvesting websites. The first option is to harvest target websites regularly. The second option is to harvest target websites immediately. The third option is to harvest target websites between two pre-setup dates.

As brought by Crown there are 3 main technical methods for archiving web content which are:

- **Client-side web archiving**: This is the most popular method employed, because of its relative simplicity and its scalability. This method allows the archiving of any site that is freely available on the open web, making it attractive to institutions with an interest in preserving websites owned and managed by other organizations or individuals. 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.
- **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. It is much less frequently employed as a methodology and records the transactions between the users of a site and the server. The main constraint is the need for access to the server, which will require agreement and collaboration with the server's owner. The primary advantage here is that it is possible to record exactly what was seen and when, so this approach is particularly attractive as a method for internal corporate and institutional archiving, where legal accountability or compliance is important.
- **Server-side web archiving**: This method involves directly copying files from the server. This can only be employed with the consent and collaboration of the server owner. The main benefits of this approach rest with the possibility of being able to archive parts of the site which are inaccessible to client side crawlers.

Crown summarized that client side archiving is the most popular method and can be carried out remotely and on a large scale. Transaction-based and server side approaches require active collaboration with the server owners and need to be implemented on a case by case basis.

Broad crawl also called bulk harvesting/snapshot harvesting is a form of broad harvesting that attempts to harvest more or less everything, or at least as much as possible. This type of harvesting usually takes a long time-up to several months. Examples of such programmers that attempt to harvest all relevant domains are Netarkivet which bases its activities on two types of lists of URLs. The first is a list of all Danish domains registered with the Danish national domain name registrar DK Hostmaster, *i.e.* all domains on the top-level domain .dk. The second list is called Danica, *i.e.* records on Denmark and the Danes. This list is maintained by Netarkivet's curators, who are constantly on the look-out for new, relevant domains to be added to the list. On this website, users can suggest domains for the Danica list, which will then be harvested if the

curators find that they are relevant.

National or regional domains: Harvesting that focuses on selected top level domains, such as .dk or .uk. The data collection is usually done through broad crawls.

Selective harvesting: Harvesting of specific domains. Here one might for instance focus on harvesting in even more depth than is usually done in broad crawls. Selective crawls as the name indicates are targeted at selected domains. It has been reported that In Netarkivet's selective crawls, around 80-100 highly dynamic domains are selected, *i.e.* domains with a high level of activity that are considered to be particularly important, such as news sites and "frequently-visited websites belonging to the authorities, the commercial sector and civil society".

Miranda (undated) remarked that the Preserving and Accessing Networked Documentary Resources of Australia (PANDORA) is an initiative of the national library of Australia and was established in 1996 and regarding its limited resources, the national library decided to take a selective approach. This is because, collecting digital publications is a time consuming and expensive task, so the national library of Australia decided to focus on publications considered valuable for research in the present and in the future. PANDORA is growing at a rate of 26 GB per month and around 518.932 files per month.

Event harvesting: This is the type of harvesting in which an attempt is made to harvest all websites that are relevant in relation to a particular event. Some event harvesting may be planned in advance, e.g. harvesting all websites in relation to a general election or the Olympics, while other events, such as natural disasters or terrorist attacks, are unpredictable, in which case harvesting can only be initiated when we become aware that the event is something that ought to be preserved for posterity. Giving access to these Harvested archives is not the same as giving access to documents physically located on the Library premises. This is because, harvested websites are not recorded in the Library catalogue as the collection is too large and too heterogeneous; it would be impossible to establish an exhaustive list of archived websites, to know their titles and their detailed content. Event crawling collect websites containing content about significant events, including planned events, such as elections, major sporting events, and unpredictable events such as natural disasters or man-made crises. But it depends on which events are considered relevant to collect.

Thematic harvesting: Similar to selective harvesting, but centered on a particular theme or subject area, which is considered particularly important to preserve. This can be similar to event harvesting, but is not necessarily restricted to the time limit of an event.

Special harvestings

In addition to these types of harvests, Netarkivet also undertakes so-called special harvestings, usually for short periods, or on just a single occasion. This might be done if websites are due to be closed, or if a researcher has specific archiving requests in connection with a research project. Some of these special harvestings aim to collect videos (which are not covered by broad or selective crawls). There are for example several special harvestings that include YouTube videos. Special harvestings can also be used to test new technologies for the collection of streamed content, or other content that requires special techniques in order to be collected.

The crawler made snapshots of as much of the web as it could find, downloaded the content, and eventually made it accessible. After selecting the sites, an agreement is secured with the site owners. Then the material is collected using web harvesters or directly provided by the owners (Miranda, undated).

According to Niu in the case where a library decides to use a bibliographic approach and create only a single level description, it should choose the unit of description based on the scale of its archives and the resources available. A unit of description on a higher level such as a whole

website means less detailed description and then fewer metadata records will be created. The library of congress and the Harvard university web archive create one MARC record for its web archive collection that includes many websites. This MARC record is searchable through their library catalog. A unit of description on a lower level such as the page level results in more detailed description and more metadata records will be created. In addition to the MARC records for a web archive collection, the library of congress web archive also creates MODS records for websites. These MODS records are searchable in the web archive but not accessible through the library catalog.

Niu, opined that existing web archives demonstrate a variety of methods and approaches to selecting, acquiring, organizing, storing, describing and providing access. These approaches are affected by external factors, such as the legal environment and the relationships between web resource producers and the web archive, as well as internal factors, such as the nature of archived web content, the nature of the archiving organization, the scale of the web archive, and the technical and financial capacity of the archiving organization.

Bragg, Hanna, Donovan, Hukill and Peterson identified that Columbia university contacts site owners directly and formally asks permission to archive websites before they begin their harvests.

Not all organizations ask for permission before capturing content; 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.

Beis, Harris and Shreffler reported that the collections Librarian/Archivist of the U.S. Catholic collection created a collection of 10 catholic blogs, with their creators' permission. It is crawled quarterly for archiving in Archive-It.

Archive-It partners sometimes seek out website owners to get written permission before beginning to harvest. The Archive-It service has long used robots.txt (a web standard) as a permissions management tool, which provides an automatic way for site owners to exclude their sites from the archiving process.

Challenges associated with successful web crawling services

Grotke emphasized that for institutions that are just starting web Archiving programs a number of factors are worth considering, as well as various social, technological and legal challenges that organizations are typically faced with. He further stated that Web archiving is technically challenging, and many of the initial challenges can be in integrating web archiving into the traditional library processes and training staff on how to work with this material. In addition to determining what to preserve, other considerations should include: staffing and organization, legal situations or mandates, types of approaches to crawling and the tools used, and access.

A technical problem exist which has to do with the incoherence of collected web pages, this occurs if during the harvesting process of the web crawler (which could take several weeks), parts of the website have been updated and web content from the top of the seed URL no longer matches those at the lower levels, thus the resulting collection is not a coherent representation of the website.

As part of the challenges, Dooley revealed that:

- Most tools built for web archives focus on capturing and storing technical metadata for accurate transmission and re-creation but capture minimal descriptive metadata, in part because so little exists in the captured files. Descriptive metadata therefore must be created manually, either within the tool or externally.
- The title of a site (as recorded in its metadata) and the date of harvesting are routinely captured, but it may not be possible to extract them automatically.
- Not all tools define descriptive metadata in the same way.
- The hope for auto-generation of descriptive metadata may be fruitless unless or until

creators of textual web pages routinely embed more metadata that can be available for capture.

- Complex web content that has been archived is sometimes presented in a way that exceeds the limits of users' technical knowledge, constituting a widespread barrier to use.
- A need for user support services derives from the complexity of accessing and using web archives.

Disadvantages of web crawling as brought by Nelson include:
The archived version does not necessarily look exactly what was online as some objects cannot be archived, such as videos and streamed content, as well as applications that use Flash, JavaScript, etc.

1. Content that requires user interaction cannot be archived.
2. Difficult to spatially delimit.
3. Temporal inconsistencies.
4. Risk of the harvester getting caught in 'bot traps'. Bot trap is a 'trap' intended or unintended that generates links and creates an endless loop of requests, causing the harvester to go in circles or crash.

Conclusion

The paper concludes that, if appropriately used, the content of the chapter will provide valuable additional input to information professionals for the design of the web archiving as a tool for preserving information resources deposited in multiple websites so as to ensure its security by storing it on their local library databases for future use.

References

1. Anthony, A., et al. "Web archiving: Techniques, Challenges, and Solutions". *International Journal of Information Technology* 5.3 (2013):598-603.
2. Aubry, S. "Introducing Web Archives as a New Library Service: The Experience of the National Library of France". *Liber Quarterly* 20.2 (2010):179–199.
3. Jacobsen, G. "Web Archiving: Issues and Problems in Collection Building and Access". *Liber Quarterly* 18.4 (2008):366–376.
4. Kahle, B. "Locking the Web Open, a Call for a Distributed Web Internet Archive Blogs". *Blog Archive Org* (2015).
5. Kamba, M. A and Yahaya, A. D. "Application of Web Harvesting Services for Digital Collaboration among Libraries in Nigeria". Bayero *Journal of Education in Africa* 8 (2020).
6. Thelwall M and Stuart D. "Web crawling ethics revisited: Cost, privacy, and denial of service". *The Journal of the Association for Information Science and Technology* 57.13 (2006):1771-1779.
7. Wang, X. et al. "Hidden web crawling for SQL injection detection". In 2010 3rd IEEE *International Conference on Broadband Network and Multimedia Technology*. 26 (2010):14-18.
8. Hwang, J. "Development of training image database using web crawling for vision-based site monitoring". *Automation in Construction* 135 (2022):104141.
9. Hu, H., Ge, Y., and Hou, D. "Using web crawler technology for geo-events analysis: A case study of the Huangyan Island incident". *Sustainability* 6.4 (2014):1896-912.