# A Systematic Approach Towards Web Preservation

Muzammil Khan
and Arif Ur Rahman

**ABSTRACT**

*The main purpose of the article is to divide the web preservation process into small explicable stages and design a step-by-step web preservation process that leads to creating a well-organized web archive. A number of research articles are studied about web preservation projects and web archives, and designed a step-by-step systematic approach for web preservation. The proposed comprehensive web preservation process describes and combines strengths of different techniques observed during the study for preserving digital web contents into a digital web archive. For each web preservation step, different approaches and possible implementation techniques have been identified that can be adopted in digital archiving. The potential value of the proposed model is to guide the archivist, related personnel, and organizations to effectively preserved their intellectual digital contents for future use. Moreover, the model can help to initiate a web preservation process and create a well-organized web archive to efficiently manage the archived web contents. A section briefly describes the implementation of the proposed approach in a digital news stories preservation framework for archiving news published online from different sources.*

**INTRODUCTION**

The amount of information generated by institutions is increasing with the passage of time. One of the mediums that uses this information is the World Wide Web (WWW). The WWW has become a tool to share information quickly with everyone regardless of their physical location. The number of web pages is vast. Google and Bing each index approximately 4.8 billion.[1]

Though the WWW is a rapidly growing source of information, it is fragile in nature. According to the available statistics, 80 percent of pages become unavailable after one year and 13 percent of links (mostly web references) in scholarly articles are broken after 27 months.[2] Moreover, 11 percent of posts and comments on websites for various purposes are lost within a year.

According to another study conducted on 10 million web pages collected from the Internet Archive in 2001, the average survival rate of web pages is 1,132.1 days with a standard deviation of 903.5 days. 90.6 percent pages of those web pages are inaccessible today.[3] The information fragility causes this valuable scholarly, cultural, and scientific information to vanish and become inaccessible to future generations.

In recent years, it was realized that the lifespan of digital objects is very short, and rapid technological changes make it more difficult to access these objects. Therefore, there is a need to preserve the information available on the WWW. Digital preservation is performed using the primary methods of emulation and migration, in which emulation provides the preserved digital objects in their original format while migration provide objects in a different format.[4] In the last

**Muzammil Khan** (muzammilkhan86@gmail.com) Assistant Professor, Department of Computer and Software Technology, University of Swat. **Arif Ur Rahman** (badwanpk@gmail.com) Assistant Professor, Department of Computer Science, Bahria University Islamabad.

two decades, a number of institutions worldwide, such as national and international libraries, universities, and companies started to preserve their web resources (resources found at a web server, i.e., web contents and web structure).

The first web archive was initiated in 1996 by Brewster Kahle, named the Internet Archive, and it holds more than 30 petabytes data, which includes 279 billion web pages, 11 million books and texts, and 8 million other digital objects such as audio, video, image files, etc. More than seventy web archive initiatives were started in 33 countries since 1996, which shows the importance of web preservation projects and preservation of web contents. This information era encourages librarians, archivists, and researchers to preserve the information available online for upcoming generations. While digital resources may not replace the information available in physical form, the digital version of these information resources improves access to the available information.[5]

There are different aspects of the preservation process and web archiving, e.g., digital objects' ingestion to the archive during preservation process, digital object's format and storage, archival management, administrative issues, access and security to the archive, and preservation planning. These aspects need to be understood for effective web preservation and will help in addressing the challenges that occur during the preservation process. The Reference Model for Open Archival Information System (OAIS) is an attempt to provide a high-level framework for the development and comparison of digital archives. In web preservation, a challenging task is to identify the starting point of the preservation process and to effectively complete the process which help to proceed further to the other activities. Therefore, the complicated nature of the Web and the complex structure of the web contents make the preservation of the web content even more difficult.

The OAIS reference model helps in achieving the goals of a preservation task in a step-by-step manner. The stakeholders are identified, i.e., producer, management, and consumer, and the packages, i.e., submission information package (SIP), archival information package (AIP) and dissemination information package (DIP), which need to be processed, are clearly defined.[6]

This study aims to design a step-by-step systematic approach for web preservation that helps to understand preservation or archival activities' challenges, especially those that relate to digital information objects at various steps of the preservation process. The systematic approach may lead to an easy way to analyze, design, implement, and evaluate the archive with clarity and different options for an effective preservation process and archival development. An effective preservation process is one that leads to a well-organized, easily managed web archive and accomplishes designated community requirements. This approach may help to address the challenges and risks that confront archivists and analysts during preservation activities.

## STEP-BY-STEP SYSTEMATIC APPROACH

Digital preservation is "the set of processes and activities that ensure long-term, sustained storage of, access to and interpretation of digital information."[7] The growth and decline rates of WWW content and the importance of the information presented on the web make it a key candidate for preservation. Web preservation confronts a number of challenges due to its complex structure, a variety of available formats, and the type of information (purpose) it provides. The overall layout of the web varies domain to domain based on the type of information and its presentation. The websites can be categorized based on two things. First, the type of information (i.e., the web


https://doi.org/10.6017/ital.v38i1.10181

contents) and second, the way this information presented (i.e., the layout or structure of the web page. Examples include educational, personal, news, e-commerce, and social networking websites, which vary a lot in their contents and structure. The variations in the overall layout make it difficult to preserve different web contents in a single web archive. The web preservation activities are summarized in figure 1. The following sections explain the web preservation activities and possible implementation in proposed systematic approach.

### *Defining the Scope of the Web Archive*

The WWW provides an opportunity to share information using various services, such as blogs, social networking websites, e-commerce, wikis, and e-libraries. These websites provide information on a variety of topics and address different communities based on their interest and needs. There are many differences in the way the information is handled and presented on the WWW. In addition, the overall layout of the web changes from one domain to another domain.[8] Therefore, it is not practically feasible to develop a single system to preserve all types of websites for the long term. So, before starting to preserve the web, one (the archivist) should define the scope of the web to be archived. The archive will be either a site-centric, topic-centric, or domain-centric archive.[9]

#### *Site-centric Archive*

A site-centric archive focuses on a particular website for preservation. These types of archives are mostly initiated by the website creator or owner. The site-centric web archives allow access to the old versions of the website.

#### *Topic-centric Archive*

Topic-centric archives are created to preserve information on a particular topic published on the web for future use. For scientific verification, researchers need to refer to the available information while it is difficult to ensure access to these contents due to the ephemeral nature of the web. A number of topic-centric archive projects have been performed including the Archipol archive of Dutch political websites,[10] the Digital Archive for Chinese Studies (DACHS) archive2,[11] Minerva by the Library of Congress,[12] and the French Elections Web archive for archiving the websites related to the French elections.[13]

#### *Domain-centric Archive*

The word “domain” refers to a location, network, or web extension. A domain-centric archive covers websites published with a specific domain name DNS, using either a top-level domain (TLD), e.g., .com, .edu, or .org, or a second-level domain (SLD), e.g., .edu.pk or .edu.fr. An advantage of domain-centric archiving is that it can be created by automatically detecting specific websites. Several projects have a domain-centric scope, e.g., the Portuguese Web Archive (PWA) national websites,[14] the Kulturarw, a Swedish Royal Library web archive collection of.se and .com domain websites,[15] and the UK Government Web Archive collection of UK government websites, e.g., .gov.uk domain websites.

### *Understanding the Web Structure*

After defining the scope of the intended web archive, the archivist will have a better understanding of the interest and expected queries of the intended community based on the resources available or the information provided by the selected domain. The focus in this step is to understand the type of information (contents) provided by the selected domain and how the information has been presented. The web can be understood by two dimensions. The first

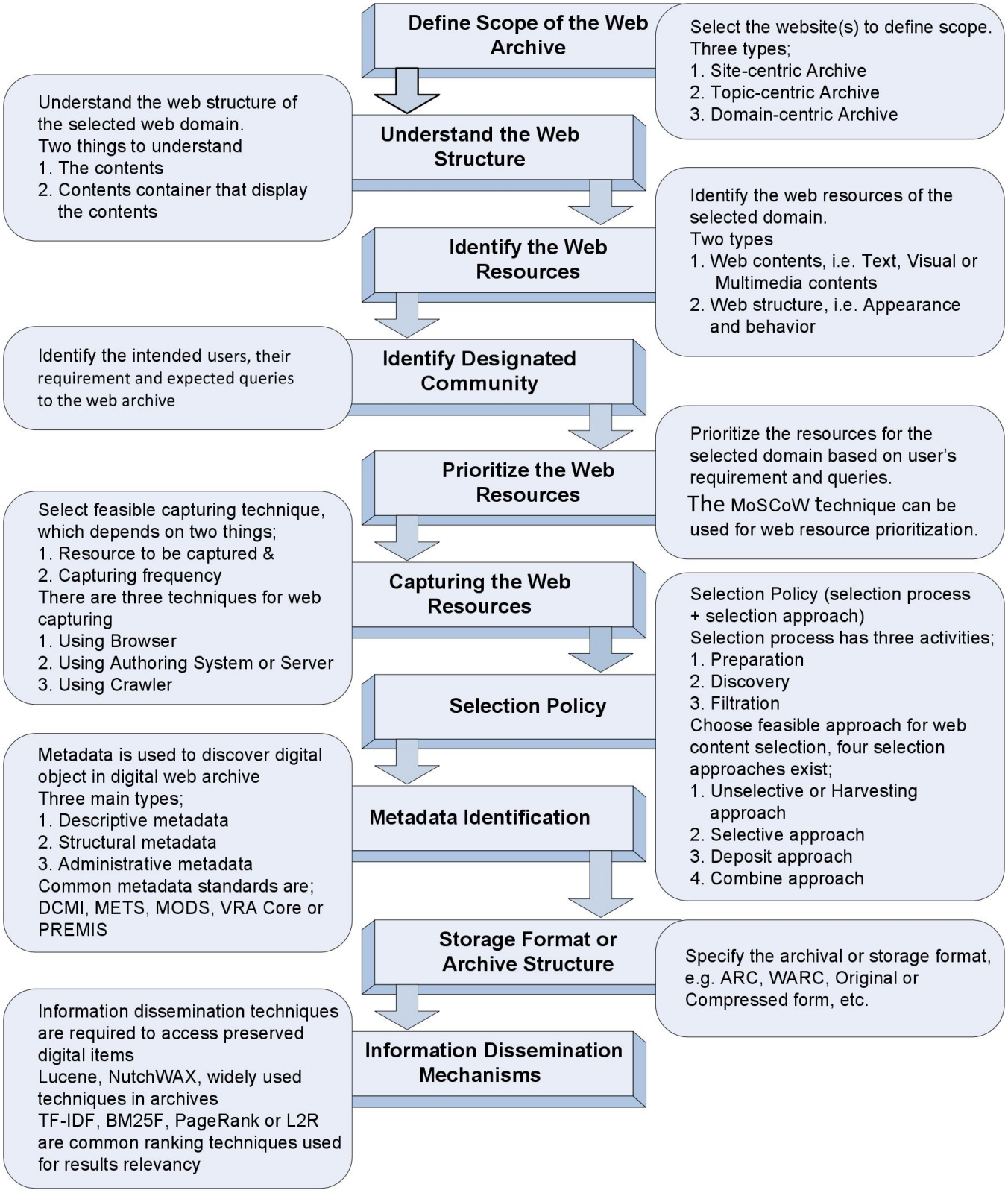


**Figure 1.** Systematic Approach for Web Preservation Process.


https://doi.org/10.6017/ital.v38i1.10181

considers the web as a medium that communicates contents using various protocols, i.e., HTTP, and the second considers the web as a content container, which further presents the contents to the viewers and not simply contents, e.g. the underlying technology used to display the contents.[16] The preservation team should understand such parameters as the technical issues, the future technologies, and the expected inclusion of other related content.

***Identify the Web Resources***

The archivist should understand the contents and the representation of the contents of the selected domain, e.g., blogs, social networking websites, institutional websites, educational institutional websites, newspaper websites, or entertainment websites. All of these websites provide different information and address individual communities that have distinct information needs.

A web page is the combination of two things, i.e., web contents and web structure.[17] The resources which can be preserved are as follows.

*Web Contents*

Web contents or web information can be categorized into the following categories:

- *Textual Contents (Plain Text)*: This category describes textual information that appears on a web page. It does not include links, behaviors, and presentation stylesheets.
- *Visual Contents (Images)*: These contents are the visual forms of information or are a complementary material to the information provided in the textual form.
- *Multimedia Contents*: As another form of information, multimedia contents mainly include audio and video. It may also include animation or even text as a part of a video or a combination of text, audio, and video.

*Web Structure*

Web structure can be categorized in the following categories:

- *Appearance (Web Layout or Presentation)*: This category indicates the overall layout or presentation of a web page. The look and feel of a web page (representation of the contents) are important, which is maintained with different technologies, e.g., HTML or stylesheets, etc.
- *Behavior (Code Navigations)*: Categorized by link navigations, these can be within a website or to other websites, external document links or dynamic and animated features, such as live feed, comments, tagging, or bookmarking.

***Identify Designated Community***

The archivist should identify the designated community of the intended web archive, their functional requirements and expected queries by analyzing them carefully. The designated community means the potential users, such as those who can access the archived web contents for different purposes, i.e., accessing old information that is not available in normal circumstances or referring to an old news article which is not bookmarked properly or retrieving relevant news articles published long ago, etc.

***Prioritize the Web Resources***

After a comprehensive assessment of the resources of the selected domain and the identification of potential users' requirements and expected queries, the archivist should prioritize the web

resources. The complexity of web resources and their representation cause complications in the digital preservation process. Generally, it may be undesirable or unviable to preserve all web resources; therefore, it is worthwhile to designate the web resources for preservation. The priority should be assigned on the basis of two things: *first*, the potential reuse of the resource and *second*, the frequency with which the resource will be accessed. The resources with no value, little value, or those managed elsewhere can be excluded.

For prioritization of resources, the MoSCoW method can be applied.[18] The acronym MoSCoW can be elaborated as:

***M - MUST have***, the resource must be preserved or resources that must be a part of the archive and preserved. For example, in the Digital News Story Archive (DNSA), the textual news story must be preserved in the archive because the preservation emphasis is on a textual news story.[19] Online news contains textual news stories, and many news stories contain associated images, and a fraction of news stories contain associated audio-video contents.

***S - SHOULD have***, the resource should be preserved if at all possible. Almost all the news stories have associated images; a few news stories have associated audio and video that complement it and should be preserved as a part of the news story in the web archive.

***C - COULD have***, the resource could be preserved if it does not affect anything else or is nice to have. The web structure in DNSA depends on the resources to be used for the preservation of news stories; the layout of the newspaper website could (C) be a part of the preservation process if it does not affect anything, e.g., storage capacity and system efficiency.

***W - WON'T have***, the resource would not be included. Archiving multiple versions of the layout or structure of the online newspaper are not worthwhile and hence would not (W) be preserved.

The prioritization of these resources is very important in the context of web preservation planning because it does not waste time and energy, and it is the best way to handle users' requirements and fulfill their expected queries.

***How to Capture the Resource(s)***

The selection of a feasible capturing technique depends on: first, the resources to be captured and second, the capturing task frequency. There are three web resources capturing techniques, i.e., by browser, web crawler, and authoring system. Each capturing technique has associated advantages and disadvantages.[7]

*Web Capturing Using Browsers*

The intended web content can be captured using browsers after a web page is rendered when the HTTP transaction occurs. This technique is also referred to as a snapshot or post-rendering technique. The method captures those things which are visible to the users; the behavior and other attributes remain invisible. Capturing static contents is one of the disadvantages of web capturing by the browser approach, this approach generally preserved contents in the form of images. It is best for well-organized websites, and commercial tools are available for capturing the web. The following are well-known tools to capture web using browsers.

*WebCapture* (https://web-capture.net/) is a free online web-capturing service. It is a fast web page snapshot tool, which can grab web pages in seven different formats, i.e. JPEG, TIFF, PNG, BMP


https://doi.org/10.6017/ital.v38i1.10181

image formats, PDF, SVG, and postscript files of high quality. It also allows downloading the intended format in a ZIP file and is suitable for long vertical web pages with no distortion in layout.

*A.nnotate* (http://a.nnotate.com/), is an online annotating web snapshot tool to keep track of information gathered from the web efficiently and easily. It allows adding tags and notes to the snapshot and building a personal index of web pages as document index. The annotation feature can be used for multiple purposes, for example, compiling an annotated library of objects for organization, sharing commented web pages, product comparison, etc.

*SnagIt* (https://www.techsmith.com/screen-capture.html) is a well-known snapshot tool for capturing screens with built-in advanced image editing features and screen recording. SnagIt is a commercial and advanced screen capture tool that can capture web pages with images, linked files, source code, and the URL of the web page.

*Acrobat WebCapture* (File > Create > PDF from Web Page...) creates a tagged PDF file from the web page that a user visits while the Adobe PDF toolbar is used for the entire website.[20]

The capture by a browser technique has the following advantages:

- By this technique, the archivist can capture only the displayed contents, and it is an advantage if you need to preserve the displayed contents only.
- It is a relatively simple technique for well-organized websites.
- Commercial tools exist for web capturing using browsers.

In addition, the disadvantages are the following:

- Capturing displayed contents only is a disadvantage if the focus is not on only displayed contents.
- It results in frozen contents and treats contents as if they are publications.
- It loses the web structure, such as appearance, behavior, and other attributes of the web page.

*Web Capturing Using an Authoring System/Server*

The authoring system capturing technique is used for web harvesting directly from the website hosting server. All the contents, e.g., textual information, images, and source code, are collected from the source web server. The authoring system allows the archivist to preserve the different versions of the website. The authoring system depends on the infrastructure of the content management system and is not a good choice for external resources. The system is best for an owned web server and works well for limited internal purposes. The Web Curator Tool (http://webcurator.sourceforge.net/), PANDAS (an old British Library harvesting tool), and NetarchiveSuite (https://sbforge.org/display/NAS/NetarchiveSuite) are known tools use for planning and scheduling web harvesting. They can be used by non-technical personnel for both selection and harvesting web content selection policies. These web archiving tools were developed in a collaboration of the National Library of New Zealand and the British Library and are used for the UK Web Archive (http://www.ariadne.ac.uk/issue50/beresford/). The tools can interface with web crawlers, such as Heritrix (https://sourceforge.net/projects/archive-crawler/). Authoring systems are also referred to as workflow systems or curatorial tools.

The authoring system has the following advantages:

- It is best for web harvesting, which captures everything available.
- It is easy to perform, if you have proper access permission or you own the server or system to access for capturing the resources.
- It works in short to medium term resources and feasible for internal access within organizations.

The disadvantages of web capturing using the authoring system are:

- It captures all available raw information, not only presentations.
- It may be too reliant on the authoring infrastructure or the content management system.
- It is not feasible for large term resources, or for external access from outside organization.

*Web Capturing Using Web Crawlers*

Web crawlers are perhaps the mostly used technique for capturing web contents in systematic and automated manner.[21] Crawler development needs the expertise and experience of different tools, i.e. positive and negative of technologies, and the viability of a tool in a specific scenario. The main advantage of crawlers is that they extract embedded content. Heritrix, HTTrack, Wget, and DeepArc are common examples of web crawlers.

*Heritrix* (https://github.com/internetarchive/heritrix3/wiki) is developed in java, an open source and freely available web crawler, and it was developed by Internet Archive. Heritrix is one of the widely used extensible and web-scale web crawlers in web preservation projects. Initially, the Heritrix was developed for specific purpose crawling of specific websites and now a resourceful or customize web crawler for archiving the web.

*HTTrack* (https://www.httrack.com/) is a freely available configurable browser utility. HTTrack crawls HTML, images, and other files from a server to a local directory and allows offline viewing of the website. The HTTrack crawler downloads a complete website from the web server to a local computer system and makes it available for offline for viewing with all related link-structure and seems like the user is using it online. It also updates the archived websites at the local system from the server and resumes all the interrupted previous extractions. The HTTrack available for both Windows and Linux/Unix operating systems.

*Wget* (http://www.gnu.org/software/wget/) is a freely available non-interactive command line tool that can easily be configured with other technologies and different scripts. It can capture files from the web using widely used FTP, FTPS, HTTP and HTTPS protocols, and support cookies as well. It also updates the archived websites and resumes all the interrupted extractions. Wget is available for both Microsoft Windows and Unix operating systems.

The advantages of web crawling:

- Widely used in capturing techniques.
- Can capture specific content or everything.
- Avoids some of the accessing issues, such as: Link rewriting and embedded external content from an archive or live.


https://doi.org/10.6017/ital.v38i1.10181

Disadvantages associated with web crawling:

- Much work is required, as well as tools or development expertise and experience, etc.
- The web crawler does not have the right scope: sometimes, it does not capture everything that it should, and sometimes the crawler captures too much content.

***Web Content Selection Policy***

In the previous steps, the web resources are identified, prioritized based on requirements and expected queries of the designated community, and feasible capturing technique is identified based on capturing frequency. Now, the contents need to be prepared and filtered for selection, and a feasible selection approach needs to be selected based on the contents. A web content selection policy helps to determine and clarify, which web contents are required to be captured based on the priorities, the purpose and the scope of web contents already defined.[22] The decision of the selection policy comprises the description of the context, the intended users, the access mechanisms and the expected uses of the archive. The selection policy may comprise the selection process and selection approach.

The ***selection process*** can be divided into subtasks which, in combination, provide a qualitative selection of web contents to a certain extent, i.e., preparation, discovery, and filtering, as shown in figure 2. The main objective of the **preparation** phase is to determine the targeted information space, the capture technique, capturing tools, extension categorization, granularity level, and the frequency of archiving activity. The best personnel who can provide help in preparation are the domain experts, regardless of the scope of the web archive. The domain experts may be the archivists, researchers, librarians, or any other authentic reference, i.e. a document or a research article. The tools defined in the preparation phase will help to discover intended information in the discovery phase, which can be divided into the following four categories:

1. *Hubs* may be the global directories or topical directories, collection of sites or even a single web page with essential links related to a particular subject or topic.
2. *Search engines* can facilitate discovery by defining a precise query or set of alternative queries related to a topic. The use of specialized search engines can significantly improve the results of discovering related information that can be greatly improved.
3. *Crawlers* can be used to extract web contents such as textual information, images, audio, video and links. Moreover, the overall layout of a web page or a whole website can also be extracted in a well-defined systematic manner.
4. *External Sources* may be non-web sources that may be anything, such as printed material for mailing lists, which can be monitored by the selection team.

The main objective of the **discovery** phase is to determine the source of information to be stored the archive. This determination can be achieved by two ways. First, a manually created entry point list is used to determine the list of entry points (usually links) for crawling the collection manually and updating the list during the crawl. There are two discovery methods, i.e., exogenous and endogenous. Exogenous discovery is used in manual selection and mostly relies on exploitation of an entry point list for hubs, search engines, and on non-web documents. Second, there is an automatically created entry point list to determine the list of entry points by extracting links automatically and obtaining an updated list every time during the crawl. Endogenous discovery is

used in automatic selection and relies on the link extraction using crawlers by exploring the entry point list.

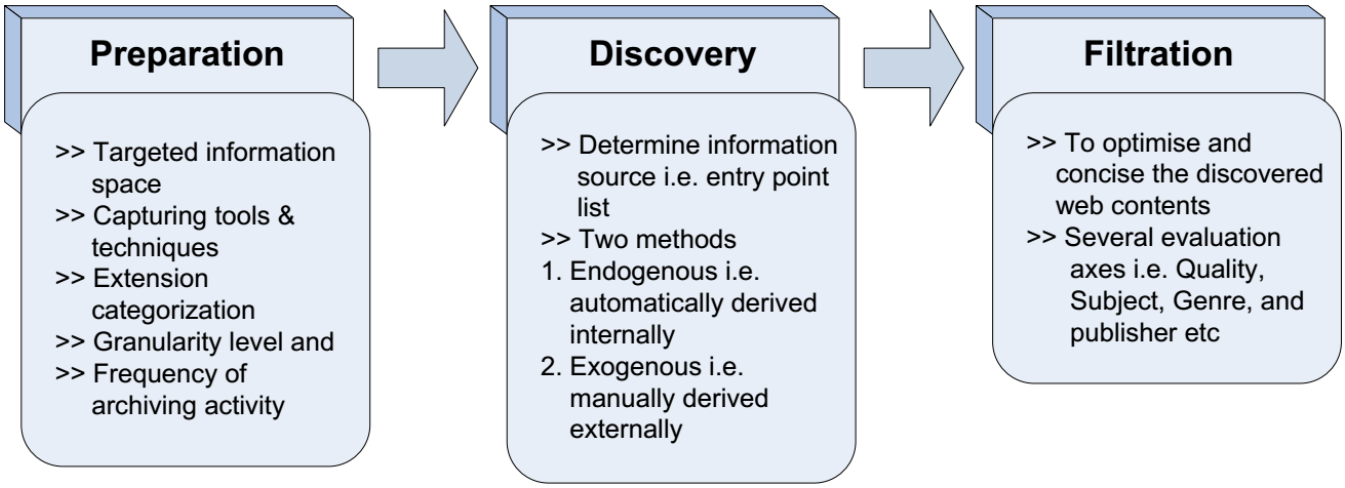


**Figure 2.** Selection Process.

The main objective of the **filtering** phase is to optimize and make concise the discovered web contents (discovery space). Filtering is important in order to collect more specific web content and remove unwanted or duplicated content. Usually, for preservation, an automatic filtering method is used; manual filtering is useful if the robots or automatic tools cannot interpret the web. The discovery and filter phase can be combined practically or logically. Several evaluation axes can be used for the selection policy (e.g., quality, subject, genre, and publisher). In the literature, we have three known techniques for selecting web content.

The ***selection approach*** can be either automatic or manual. Manual content selection is very rare because it is labor intensive: it requires automatic tools for finding the content, and then manual review of that collection to identify the subset that should be captured. Automatic selection policies are used frequently in web preservation projects for web collection, especially for web archives.[23] The selection of the collection approach depends on the frequency with which the web content has been preserved in the archive. There are four different selection approaches for web content collection.

*Unselective Approach*

The unselective approach implies collecting everything possible; by specifically using this approach, the whole website and its related domains and subdomains are downloaded to the archive. It is also referred to as automatic harvesting or selection, bulk selection, and domain selection.[24] The automatic approach is used in a situation where a web crawler usually performs the collection. For example, the collection of websites from a domain, i.e., .edu means all educational institution websites (at domain level) or the collection of all possible contents/pages from a website (harvesting at website level) by extracting the embedded links.

A section of the data preservation community believes that technically it is a relatively cheaper, quicker collection approach and yields a comprehensive picture of the web as a whole. In contrast, its significant drawbacks are that it generates huge unsorted, duplicated, and potentially useless data, consuming too many resources.


https://doi.org/10.6017/ital.v38i1.10181

The Swedish Royal Library's project Kulturarw3 harvests websites at domain level, i.e., collecting websites from a .se domain which is a physically located website in Sweden and one of the first projects to adopt this approach.[25] Usually, national-based web archive initiatives adopt the unselective approach, most notably NEDLIB, a Helsinki University Library harvester, and AOLA, an Austrian Online Archive.[26]

*Selective Approach*

The selective approach was adopted by the National Library of Australia (NLA) in the PANDAS project in 1997. In this approach, a website is included for archiving based on certain predefined strategies and on the access and information provided by the archive. The Library of Congress' project Minerva and the British Library project "Britain on the Web" are the other known projects that have adopted the selective approach. According to NLA, the selected websites are archived based on NLA guidelines after negotiation with the owners.[27]

The inclusion decision could be taken at one of the following levels:

- *Website level*: which websites should be included from a selected domain, e.g., to archive all educational websites from high level domain ".pk".
- *Web page level*: which web pages should be included from a selected website, e.g., to archive the homepages of all educational websites.
- *Web content level*: which type of web contents should be preserved, e.g., to archive all the images from the homepages of educational websites.

A selective approach is best if the numbers of websites to be archived are very large or the archiving process is targeting the entire WWW and wants to narrow down the scope by identifying the resources in which the archivists are more interested. This approach performs implicit or explicit assumptions about the web contents that are not to be selected for preservation. It may be very helpful to initiate a pilot preservation project, which identifies: What is possible? What can be managed? In addition, some tangible results may be obtained easily and quickly in order to enhance the scope of the project in a broader perspective. The selective approach may be based on a predefined criterion or based on an event.

**Selective approach based on criteria** involves selecting web resources based on various pre-defined sets of criteria. NLA's guidance characterizes the criteria-based selective approach as the "most narrowly defined method," and described it as "thematic selection." A simple or a complex content-selection criteria can be defined, which depends on the overall goal of preservation. For example, all resources owned by an organization, all resources of one genre, i.e., all programming blogs, resources contributed to a common subject, resources addressing a specific community within an institution, i.e., students or staff, all publications belonging to an individual organization or group of organizations, all resources that may benefit external users or an external user's community, e.g., historians, or alumni.

**Selective approach based on event** involves selecting web resources or websites based on various time-based events. The archivists may focus on websites that address national or international important events, e.g., disasters, elections, and the football world cup, etc. Event-based websites have two characteristics: (1) very frequent updates and (2) website content is lost after a short time, e.g., a few weeks or a few months. For example, the start and end of a term or

academic year, the duration of an activity, e.g., research project, appointment, or departure of a new senior official.

*Deposit Approach*

In the deposit collection approach, the information package is submitted by the administrator or owner of the website which includes a copy of the website with related files that can be accessed through different hyperlinks. The archival information package is applicable to the small collection (of a few websites), or the owner of the website can initiate the preservation project, e.g. a company can initiate a project for preserving their website. The deposit collection approach was adopted by the National Archives and Records Administration (NARA) for the collection of US federal agency websites in 2001 and by Die Deutsche Bibliothek (DDB, http://deposit.ddb.de/) for the collection of dissertations and some online publications. New digital initiatives are heavily dependent on administrator or owner support and provide an easy way to deposit new content to the repository, e.g., in the MacEwan University's institutional repository, the librarians leading the project tried to offer an easy and effective way to deposit their archival contents.[28]

*Combined Approach*

There are advantages and disadvantages associated with each collection approach. The ongoing debate is which approach is best in a given situation. For example, the deposit approach should be an inexpensive agreement with the depositors. The emphasis is to use the combination of automatic harvesting and selective approaches as these two approaches are cheaper as compared to other selection approaches because a few staff personnel are required and cope with technological challenges. This initiative was taken by the Bibliothque Nationale de France (BnF) in 2006. The BnF automatically crawls information regarding the updated web pages and stores it in an XML-based "site delta" and uses page relevancy and importance, similar to how Google ranks pages, to evaluate individual pages.[29] The BnF used a selective approach for the deep web (that is, web pages or websites that are behind a password or are otherwise not generally accessible to search engines), referred to as "deposit track."

***Metadata Identification***

Cataloging is required to discover a specific item from the digital collection. An identifier or set of identifiers is required to retrieve a digital record in digital repositories or an archive. For digital documents, this catalog or registration or identifier is referred to as metadata.[30] Metadata are structured information concerning resources that describe, locate (discover or place), manage, easily retrieve (access) and use digital information resources. Metadata are often referred to as "data about data" or "information about information", but it may be more helpful and informative to describe these data as "descriptive and technical documentation."[31]

Metadata can be divided into the following three categories:

1. *Descriptive metadata* describes a resource for discovery and identification purposes. It may consist of elements for a document such as title, author(s), abstract, and keywords, etc.
2. *Structural metadata* describes how compound objects are put together, for example, how sections are ordered to form chapters.


https://doi.org/10.6017/ital.v38i1.10181

3. *Administrative metadata* imparts information to facilitate resource management, such as when and how a file was created, who can access the file, its type, and other technical information. Administrative metadata is classified into two types: (1) rights management metadata addresses intellectual property rights and (2) preservation metadata contains information needed to archive and preserve a resource.[32]

Due to new information technologies, digital repositories, especially web-based repositories, have grown rapidly over the last two decades. This interest prompts the digital libraries communities to devise metadata strategies to manage the immense amount of data stored in digital libraries.[33] Metadata play a vital role in the long-term preservation of digital objects and important to identify the metadata which may help to retrieve a specific object from the archive after preservation. According to Duff et al., "the right metadata is the key to preserving digital objects."[34]

There are hundreds of metadata standards developed over the years for different user environments, disciplines, and for different purposes; many of them are in their second, third, or nth edition.[35] Digital preservation and archiving requires metadata standards to trace and ensure its access to the digital objects. Several of the common standards are briefly discussed below.

*Dublin Core Metadata Initiative* (DCMI, http://dublincore.org/) was initiated at the 2nd World Wide Web conference in 1994 and was standardized by ANSI/NISO Z39.85 in 2001 and ISO 15386 in 2003.[36] The main purpose of the DCMI was to define an element set for representing web resources; initially, thirteen core elements were defined which later increased to a fifteen-element set. The elements are optional, repeatable, can be followed in any order, and expressed in XML.[37]

*Metadata Encoding and Transmission Standard* (METS, http://www.loc.gov/standards/mets/) is an XML metadata standard intended to represent information of the complex digital objects. METS elements evolved from the early project Making of America II "MOA2" in 2001, supported by the Library of Congress and sponsored by the Digital Library Federation "DLF" and registered with National Information Standards Organization "NISO" in 2004. A METS document contains seven major sections in which each contains different aspects of metadata.[38]

*Metadata Object Description Schema* (MODS, http://www.loc.gov/standards/mods/) was initiated by the MARC21 maintenance agency at the Library of Congress in 2002. MODS elements are richer then DCMI, simpler then MARC21 bibliographic format and expressed in XML.[39] The MODS identified the widest facets or features of an object and presented nineteen high-level optional elements.[40]

*Visual Resources Association Core Strategies* (VRA Core, http://www.loc.gov/standards/vracore/) was developed in 1996, and the current version 4.0 was released in 2007. The VRA core is a widely used standard for art, libraries, and archives for such objects as paintings, drawings, sculpture, architecture, and photographs, as well as books and decorative and performance art.[41] The VRA core contains nineteen elements and nine sub-elements.[42]

*Preservation Metadata Implementation Strategies* (PREMIS, http://www.loc.gov/standards/premis/) was developed in 2005, sponsored by the Online Computer Library Center (OCLC) and the Research Libraries Group (RLG), includes a data dictionary and some information about metadata. PREMIS defined a set of five interactive core semantic units or entities and XML schema for endorsing digital preservation activities. It is not

concerned with discovery and access but with common metadata, and for descriptive metadata, other standards (Dublin Core, METS or MODS) need to be used. The PREMIS data model contains intellectual entities (contents that can be described as a unit, e.g., books, articles, databases), objects (discrete units of information in digital form, which can be files, bitstreams, or any representation), agents (people, organization, or software), events (actions that involve an object and an agent known to the system) and rights (assertion of rights and permission).[43]

It is indisputable that good metadata improves access to the digital object in the digital repository. Therefore, the creation and selection of appropriate metadata make the web archive accessible to the archive user. Structure metadata helps to manage the archival collection internally, as well as the related services, but may not always help to discover the primary source of the digital object.[44] Currently, there are many semi-automated metadata generation tools. The use of these semi-automatic tools for generating metadata is crucial for the future, considering the operation's complexity and cost of manual metadata origination.[45]

***Archival Format***

The web archive initiatives select websites for archiving based on relevance of contents and the intended audience of the archived information. The size of the web archives varies significantly depending on their scope and the type of content they are preserving, e.g., web pages, PDF documents, images, audio, or video files.[46] To preserve these contents, a web archive uses different storage formats containing metadata and utilizes data compression techniques. The Internet Archive defined the ARC format (http://archive.org/web/researcher/ArcFileFormat.php), later used as a defacto standard. In 2009, the Internet Organization for Standardization (ISO) established the WARC format (https://goo.gl/0RBWSN) as an official standard for web archiving. Approximately 54 percent of web archive initiatives applied ARC and WARC formats for archiving. The use of standard formats helps the archivists to facilitate the creation of collaborative tools, such as search engines and UI utilities to efficiently manipulate the archived data.[47]

***Information Dissemination Mechanisms***

A well-defined preservation process can lead to a well-organized web archive that is easy to maintain and easy to retrieve a specific digital object from the collection using information dissemination techniques. Poor search results are one of the main problems in information dissemination of web archives. The users of a web archive expend excessive time to retrieve intended documents or information to satisfy the user's query. Archivists are more concerned with "ofness," "what collections are made up of," although archive users are concerned with aboutness, "what collections are about."[48]

To use the full potential of web archives a usable interface is needed to help the user to search the archive for specific digital object. Full text and keyword search are the dominant ways to search the unstructured information repository, evidently observed from the online search engines. The sophistication of search results against user queries is based on the ranking tools.[49] The access tools and techniques are getting the attention of researchers, and approximately 82 percent of European web archives concentrate on such tools, which makes these web archives easily accessible.[50] The Lucene full-text search engine and its extension NutchWAX is widely used in web archiving. Moreover, for the combination of semantic descriptions that already rely on or are implicit within their descriptive metadata, reasoning-based or semantic searching of the archival


https://doi.org/10.6017/ital.v38i1.10181

collection can enable the system to produce novel possibilities for the archival content retrieval and browsing.[51] Even in the current era of digital archives, mobile services are adopted in digital libraries, e.g., access to e-books, libraries databases, catalogs, and text messaging are common mobile services offered in university libraries.[52] In a massive repository, a user query retrieves millions of documents, which makes it difficult for users to identify the most relevant information. The ranking model estimates the results relevancy based on user's queries using specified criteria to overcome this problem and sorts the results by placing the most relevant result at the top.[53] There are a number of ranking models that exist in the literature, e.g., conventional ranking models, e.g., TF-IDF, BM25F, temporal ranking models, e.g., PageRank, and learning to rank models, e.g., L2R.

The findings of the systematic approach for web preservation are used to automate the process of the digital news-story preservation. The steps of the proposed model are carefully adopted to develop a tool that is able to add contextual information to the stories to be preserved.

## DIGITAL NEWS STORIES PRESERVATION FRAMEWORK

The advancement of web technologies and maturation of the internet attracts news readers to access news online that is provided by multiple sources and to obtain the desired information comprehensively. The amount of news published online has grown rapidly, and for an individual, it is cumbersome to browse through all online sources for relevant news articles. The news generation in the digital environment is no longer a periodic process with a fixed single output, such as printed newspapers. The news is instantly generated and updated online in a continuous fashion. However, because of different reasons, such as the short lifespan of digital information and the speed of generation of information, it has become vital to preserve digital news for the long term. Digital preservation includes various actions to ensure that digital information remains accessible and usable, as long as they are considered important.[54] Libraries and archives preserve by carefully digitizing newspapers considering as a good source of knowing the history. Many approaches have been developed to preserve digital information for the long term. The lifespan of news stories published online varies from one newspaper to another, i.e., from one day to a month. However, a newspaper may be backed up and archived by the news publisher or national archives; in the future, it will be difficult to access particular information published in various newspapers regarding the same news story. The issues become even more complicated if a story is to be tracked through an archive of many newspapers, which requires different access technologies.

The Digital News Story Preservation (DNSP) framework was introduced to preserve digital news articles published online from multiple sources.[55] The DNSP framework is planned based on adopting the proposed step-by-step systematic approach for web preservation to develop a well-organized web archive. Initially, the main objectives defined for the DNSP framework are:

- To initiate a well-organized national level digital news archive of multiple news sources.
- To normalize news articles during preservation to a common format for future use.
- To extract explicit and implicit metadata, which would be helpful in ingesting stories to the archive and browsing through the archive in the future.
- To introduce content-based similarity measures to link digital news articles during preservation.

The Digital News Story Extractor (DNSE) is a tool developed to facilitate the extraction of news stories from the online newspapers and to migrate to a normalized format for preservation. The normalized format also includes a step to add metadata in the Digital News Stories Archive (DNSA) for future use.[56] To facilitate the accessibility of news articles preserved from multiple sources, some mechanisms need to be adopted for linking the archived digital news articles. An effective term-based approach "Common Ratio Measure for Stories (CRMS)" for linking digital news articles in DNSA is introduced that links similar news articles during the preservation process.[57] The approach is empirically analyzed, and the results of the proposed approach are compared to get conclusive arguments. The initial results computed automatically using a common ratio measure for stories are encouraging and are compared with the similarity of news articles based on human judgment. The results are generalized by defining a threshold value based on multiple experimental results using the proposed approach.

Currently, there is ongoing work to extend the scope of DNSA to dual languages, i.e., Urdu and English, as well as content-based similarity measures to link news articles published in Urdu-English. Moreover, research is underway to develop tools for exploiting the linkage created among stories during the preservation process for search and retrieval tasks.

## SUMMARY

Effective strategic planning is critical in creating web archives; hence, it requires a well-understood and a well-planned preservation process. The process should result in a well-organized web archive that includes not only the content to be preserved but also the contextual information required to interpret the content.

The study attempts to answer many questions by guiding the archivists and related personnel, such as: How to lead the web preservation process effectively? How to initiate the preservation process? How to proceed through different steps? What are the possible techniques that may help to create a well-organized web archive? How can the archived information can be used to its greatest potential?

To answer these questions, the study resulted in an appropriate step-by-step process for web preservation and a well-organized web archive. The targeted goal of each step is identified by researching the existing approaches that can be adopted. The possible techniques for those approaches are discussed in detail for each step.

## REFERENCES

[1] "World Wide Web Size," The size of the World Wide Web, visited on Jan 31, 2019, http://www.worldwidewebsize.com/.

[2] Brian F. Lavoie, "The Open Archival Information System Reference Model: Introductory Guide," *Microform & Imaging Review* 33, no. 2 (2004): 68-81; Alexandros Ntoulas, Junghoo Cho, and Christopher Olston, "What's New on the Web? The Evolution of The Web from a Search Engine Perspective," in *Proceedings of the 13th International Conference on World Wide Web-04* (New York, NY**:** ACM, 2004), 1-12.


https://doi.org/10.6017/ital.v38i1.10181

[3] Teru Agata et al., "Life Span of Web Pages: A Survey of 10 million Pages Collected in 2001," *IEEE/ACM Joint Conference on Digital Libraries*, (IEEE, 2014), 463-64, https://doi.org/10.1109/JCDL.2014.6970226.

[4] Timothy Robert Hart and Denise de Vries, "Metadata Provenance and Vulnerability," *Information Technology and Libraries* 36, no. 4 (Dec. 2017): 24-33, https://doi.org/10.6017/ital.v36i4.10146.

[5] Claire Warwick et al., "Library and Information Resources and Users of Digital Resources in the Humanities," *Program* 42, no. 1 (2008): 5-27, https://doi.org/10.1108/00330330810851555.

[6] Lavoie, "Open Archival Information System Reference Model."

[7] Susan Farrell, K. Ashley, and R. Davis, "A Guide to Web Preservation," *Practical Advice for Web and Records Managers Based on Best Practices from the JISC-Funded PoWR Project* (2010), https://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf.

[8] Lavoie, "Open Archival Information System Reference Model;" Farrell, Ashley, and Davis, "Guide to Web Preservation."

[9] Peter Lyman, "Archiving the World Wide Web," Washington, Library of Congress (2002), https://www.clir.org/pubs/reports/pub106/web/.

[10] Diomidis Spinellis, "The Decay and Failures of Web References," *Communications of the ACM* 46, no. 1 (2003): 71-77, https://dl.acm.org/citation.cfm?doid=602421.602422.

[11] Digital Archive for Chinese Studies (DACHS) Archive2 https://www.zo.uni-heidelberg.de/boa/digital_resources/dachs/index_en.html, visited on Jan 31, 2019.

[12] Julien Masanès, "Web Archiving Methods and Approaches: A Comparative Study," *Library Trends* 54, no. 1 (2005): 72-90, https://doi.org/10.1353/lib.2006.0005.

[13] Hanno Lecher, "Small Scale Academic Web Archiving: DACHS," in *Web Archiving* (Berlin/Heidelberg: Springer, 2006), 213-25, https://doi.org/10.1007/978-3-540-46332-0_10.

[14] Daniel Gomes et al., "Introducing the Portuguese Web Archive Initiative," in *8th international Web Archiving Workshop* (Berlin/Heidelberg: Springer, 2009).

[15] Gerrit Voerman et al., "Archiving the Web: Political Party Web Sites in the Netherlands," *European Political Science* 2, no. 1 (2002): 68-75, https://doi.org/10.1057/eps.2002.51.

[16] Sonja Gabriel, "Public Sector Records Management: A Practical Guide," *Records Management Journal* 18, no. 2 (2008), https://doi.org/10.1108/00242530810911914.

[17] Farrell, Ashley, and Davis, "Guide to Web Preservation."

[18] Jung-ran Park and Andrew Brenza, "Evaluation of Semi-Automatic Metadata Generation Tools: A Survey of the Current State of the Art," *Information Technology and Libraries* 34, no. 3 (Sept, 2015): 22-42, https://doi.org/10.6017/ital.v34i3.5889.

[19] Muzammil Khan and Arif Ur Rahman, "Digital News Story Preservation Framework," in *Digital Libraries: Providing Quality Information: 17th International Conference on Asia-Pacific Digital Libraries, ICADL 2015* Seoul, Korea, December 9-12, 2015 (*Proceedings*, vol. 9469, Springer, 2015), 350-52, https://doi.org/10.1007/978-3-319-27974-9; Muzammil Khan, "Using Text Processing Techniques for Linking News Stories for Digital Preservation," PhD Thesis, Faculty of Computer Science, Preston University Kohat, Islamabad Campus, HEC Pakistan, 2018.

[20] Dennis Dimick, "Adobe Acrobat Captures the Web," *Washington Apple Pi Journal* (1999): 23-25.

[21] Trupti Udapure, Ravindra D. Kale, and Rajesh C. Dharmik, "Study of Web Crawler and Its Different Types," *IOSR Journal of Computer Engineering (IOSR-JCE)* 16, no. 1 (2014): 01-05, https://doi.org/10.9790/0661-16160105.

[22] Dora Biblarz et al., "Guidelines for a Collection Development Policy Using the Conspectus Model," *International Federation of Library Associations and Institutions, Section on Acquisition and Collection Development* (2001).

[23] Farrell, Ashley, and Davis, "Guide to Web Preservation;" E. Pinsent et al., "PoWR: The Preservation of Web Resources Handbook," http://jisc.ac.uk/publications/programmerelated/2008/powrhandbook.aspx (2010); Michael Day, "Preserving the Fabric of Our Lives: A Survey of Web Preservation Initiatives," *Lecture Notes in Computer Science* (Berlin/Heidelberg: Springer, 2003): 461-72, https://doi.org/10.1007/978-3-540-45175-4_42.

[24] Pinsent et al., "PoWR:"; Day, "Preserving the Fabric."

[25] Allan Arvidson, "The Royal Swedish Web Archive: A Complete Collection of Web Pages," *International Preservation News* (2001): 10-12.

[26] Andreas Rauber, Andreas Aschenbrenner, and Oliver Witvoet, "Austrian Online Archive Processing: Analyzing Archives of the World Wide Web," *Research and Advanced Technology for Digital Libraries* (2002): ECDL 2002. Lecture Notes in Computer Science, vol 2458, (Berlin/Heidelberg: Springer, 2002), 16-31, https://doi.org/10.1007/3-540-45747-X_2.

[27] William Arms, "Collecting and Preserving the Web: The Minerva Prototype," *RLG DigiNews* 5, no. 2 (2001).

[28] Sonya Betz and Robyn Hall, "Self-Archiving with Ease in an Institutional Repository: Micro Interactions and the User Experience," *Information Technology and Libraries* 34, no. 3 (Sept. 2015): 43-58, https://doi.org/10.6017/ital.v34i3.5900.

[29] Serge Abiteboul et al., "A First Experience in Archiving the French Web," in *International Conference on Theory and Practice of Digital Libraries*, (Berlin/Heidelberg: Springer, 2002), 1-15, https://doi.org/10.1007/3-540-45747-X_1; Sergey Brin and Lawrence Page, "Reprint of:

The Anatomy of a Large-Scale Hypertextual Web Search Engine," *Computer Networks* 56, no. 18 (2012): 3825-33, https://doi.org/10.1016/j.comnet.2012.10.007.

[30] Masanès, "Web Archiving."

[31] NISO-Press, "Understanding Metadata," National Information Standards (2004), http://www.niso.org/publications/understanding-metadata.

[32] Ibid.

[33] Jane Greenberg, "Understanding Metadata and Metadata Schemes," *Cataloging & Classification Quarterly* 40, no. 3-4 (2009): 17-36, https://doi.org/10.1300/J104v40n03_02.

[34] Michael Day, "Preservation Metadata Initiatives: Practicality, Sustainability, and Interoperability," *Publishers: Archivschule Marburg* (2004): 91-117.

[35] Jenn Riley, *Glossary of Metadata Standards* (2010).

[36] Corey Harper, "Dublin Core Metadata Initiative: Beyond the Element Set," *Information Standards Quarterly* 22, no. 1 (2010): 20-31.

[37] Jane Greenberg, "Dublin Core: History, Key Concepts, and Evolving Context (Part One)," in *Slide Presentation on dc-2010 International Conference on Dublin Core and Metadata Applications Pittsburgh, PA* (2010).

[38] Cundiff V. Morgan, "An Introduction to the Metadata Encoding and Transmission Standard (METS)," *Library Hi Tech* 22, no. 1 (2004): 52-64, https://doi.org/10.1108/07378830410524495; Leta Negandhi, "Metadata Encoding and Transmission Standard (METS),"*In Texas Conference on Digital Libraries*, TCDL-2012 (2012).

[39] Sally H. McCallum, "An Introduction to the Metadata Object Description Schema (MODS)," *Library Hi Tech* 22, no. 1 (2004): 82-88, https://doi.org/10.1108/07378830410524521.

[40] R. Gartner, "MODE: Metadata Object Description Schema," *JISC Techwatch Report TSW* (2003): 03-06. www.loc.gov/standards/mods/.

[41] VRA-Core, "An Introduction of VRA Core," http://www.loc.gov/standards/vracore/VRA Core4 Intro.pdf, Created: Oct 2014.

[42] VRA-Core, "VRA Core Element Outline," http://www.loc.gov/standards/vracore/VRA Core4 Outline.pdf, Created: Feb 2007.

[43] Priscilla Caplan, "Understanding PREMIS," *Washington DC, USA: Library of Congress*, (2009), https://www.loc.gov/standards/premis/understanding-premis.pdf; J. Relay, "An Introduction to PREMIS," *Singapore IPRESS Tutorial*, (2011), http://www.loc.gov/standards/premis/premistutorial iPRES2011 singapore.pdf.

[44] Jennifer Schaffner, "The Metadata is the Interface: Better Description for Better Discovery of Archives and Special Collections, Synthesized from User Studies," *Making Archival and Special Collections More Accessible,* 85 (2015).

[45] Joao Miranda and Daniel Gomes, "Trends in Web Characteristics," in *Web Congress, 2009. LA-WEB'09. Latin American*, (IEEE, 2009), 146-53, https://doi.org/10.1109/LA-WEB.2009.28.

[46] Daniel Gomes, João Miranda, and Miguel Costa, "A Survey on Web Archiving Initiatives," *Research and Advanced Technology for Digital Libraries* (2011): 408-20, https://doi.org/10.1007/978-3-642-24469-8_41.

[47] Ibid.

[48] Schaffner, "Metadata is the Interface."

[49] Miguel Costa and Mário J. Silva, "Evaluating Web Archive Search Systems," in *International Conference on Web Information Systems Engineering* (Berlin/Heidelberg: Springer, 2012), 440-454. https://doi.org/10.1007/978-3-642-35063-4_32.

[50] Foundation, I, "Web Archiving in Europe," technical report, *CommerceNet* Labs (2010).

[51] Georgia Solomou and Dimitrios Koutsomitropoulos, "Towards an Evaluation of Semantic Searching in Digital Repositories: A DSpace Case-Study," *Program* 49, no. 1 (2015): 63-90, https://doi.org/10.1108/PROG-07-2013-0037.

[52] Liu Yan Quan and Sarah Briggs, "A Library in the Palm of Your Hand: Mobile Services in Top 100 University Libraries," *Information Technology and Libraries* 34, no. 2 (June 2015): 133, https://doi.org/10.6017/ital.v34i2.5650.

[53] Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval* 463. (New York: ACM Pr., 1999).

[54] Daniel Burda and Frank Teuteberg, "Sustaining Accessibility of Information through Digital Preservation: A Literature Review," *Journal of Information Science*, 39, no. 4 (2013): 442-58, https://doi.org/10.1177/0165551513480107.

[55] Muzammil Khan et al., "Normalizing Digital News-Stories for Preservation," in *Digital Information Management (ICDIM), 2016 Eleventh International Conference on* (IEEE, 2016), 85-90, https://doi.org/10.1109/ICDIM.2016.7829785.

[56] Khan, et al., "Normalizing Digital News."

[57] Muzammil Khan, Arif Ur Rahman, and M. Daud Awan, "Term-Based Approach for Linking Digital News Stories," in *Italian Research Conference on Digital Libraries* (Cham, Switzerland: Springer, 2018), 127-38, https://doi.org/10.1007/978-3-319-73165-0_13.