“I’m not sure how feasible capture is”: archivability as a dimension of website quality

Brenda Reyes Ayala[1] (Associate Professor)

[1]University of Alberta, Faculty of Education, 11210 87 Ave, Edmonton AB T6G 2G5, Canada

Abstract

This paper presents the results of a study of the quality of archived websites using support tickets from the Internet Archive’s Archive-It (AIT) service, currently the most widely used web archiving service. The study uses grounded theory to create a definition of quality for archived websites that is composed of three dimensions: correspondence, relevance, and archivability. The focus of this paper is on website archivability, which I redefine as the intrinsic properties of a website that make it easier or more difficult to archive. I argue that archivability is a latent construct of information quality expressed by the manifest variables of correspondence and relevance, a finding not previously seen in the literature. The definition is independent of the technology currently in use to create web archives, making it suitable to a wide variety of platforms, preservation contexts, and situations. Furthermore, the paper describes how low levels of website archivability influences the quality of web archives as historical records.

Keywords

web archiving, web archives, archivability, digital preservation, website quality

1. Introduction

Since the 1990s, many cultural heritage institutions, such as museums, archives, and libraries, have undertaken the task of creating a historical record of the web through the practice of web archiving. Web archivists who preserve websites are concerned both with quantity and with quality. Their goal is to create archived websites that are as close as possible in appearance and functionality to the original, live website. Failing to adequately capture a website might mean a flawed or incomplete historical record of it. The importance of creating high-quality web archives was echoed by the results of the 2016 survey on web archiving in the United States, conducted by the National Digital Stewardship Alliance (NDSA). When asked what were their top concerns when developing a web archiving program at their respective institutions, 52% of participants cited quality as a top issue. Quality was the third most-cited concern for respondents, after cost and access and use (60% each) [1].

The most popular web archiving service is the Internet Archive’s Archive-It (AIT), which helps organizations build and manage their own web archives. It currently has over 800 clients *partners* consisting of universities, archives, museums, and libraries in over 24 countries [2]. In 2022, the National Digital Stewardship Alliance (NDSA) conducted another survey of web archiving practices worldwide [3]. The authors found that, of the over 190 institutions that had web archiving programs in place, 71% capture content with Archive-It. Archive-It is popular with many institutions throughout the world, who have entrusted it with the creation and management of their web archives.

Despite the popularity of the Archive-It and the Internet Archive, web archiving is a field with few conceptual tools or theoretical definitions. In a previous work [4] I presented a grounded theory of Information Quality (IQ) for web archives, derived from an analysis of tickets submitted to the Internet Archive’s AIT service by web archivists. Its goal was to create a theory of IQ that is both human-centered and independent of the technology currently in use to create web archives. This theory consists of three

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20-21 2025, Udine, Italy

brenda.reyes@ualberta.ca (B. Reyes Ayala)
https://reyesayala.github.io/ (B. Reyes Ayala)
0000-0002-9342-3832 (B. Reyes Ayala)

dimensions (or core categories) that determine the quality of a web archive: correspondence, relevance, and archivability, along with their subdimensions:

1. Correspondence: degree of similarity, or resemblance, between the original website and the archived website
    - Visual correspondence: similarity in appearance between the original website and the archived website
    - Interactional correspondence: the degree to which a user's interaction with the archived website is similar to that of the original
    - Completeness: the degree to which the archived website contains all of the components of the original
2. Relevance: pertinence of the contents of an archived website to the original website
    - Topic relevance: degree to which an archived website (or a web archive) includes only content that is closely related to that of the original website or the topic of the larger web archive
    - Size relevance: the similarity in size of the archived website to the original website
3. Archivability: degree to which the intrinsic properties of a website make it easier or more difficult to archive

A web archive can be said to have high-quality if it has high correspondence, high relevance, and high archivability. Due to the heavy presence of the dimension of correspondence in the data set, I asserted that it is the most important facet of quality in web archive. This work builds on the theory of IQ for web archives advanced earlier, but turns its attention to archivability, another important dimension. The purpose of this paper is to elaborate on and deepen the original definition of archivability, place it in a human-centred context, and predict how it will affect the future of web archiving. This goal leads to the following research questions:

1. How do people perceive the notion of archivability in web archives?
2. How does website archivability affect web archives and thus the future historical record?

2. Previous Work: "Damaged" web archives and archivability

When deploying crawlers to capture a website, some crawl engineers pay special attention to embedded resources. Embedded resources are files, such as images, videos, or CSS stylesheets, that are present and referenced in a website. A user might not notice their presence, but embedded resources play a key role in ensuring the website looks and operates in the correct way. To this end, crawl engineers might calculate a percentage of missing embedded resources M_m in an archived website and use it to estimate the overall quality of the site. Brunelle, Kelly, SalahEldeen, Weigle, and Nelson [5] showed that M_m is not always consistent with human judgments of the quality of an archived website and was thus not a suitable metric for measuring the damage to an archived website caused by missing embedded resources. Instead, the authors proposed a new metric to assess this damage that is based on three factors: the MIME type, size, and location of the embedded resource [5].

In their iPres paper "CLEAR: A Credible Method to Evaluate Website Archivability", [6] introduced the concept of website archivability. Archivability was defined as the "sum of the attributes that make a website amenable to being archived" [6]. The more easily it was to archive a website, the greater its archivability. The authors introduced a set of facets designed to determine the archivability of a website, termed the Credible Live Evaluation of Archive Readiness, or CLEAR, method. These facets were: standards compliance, performance, cohesion. and metadata usage. Later the authors expanded on their original work by introducing the CLEAR+ method, the incremental evolution of their original CLEAR+ method. According to CLEAR+, The archivability of a website is dependent on the following facets [7]:

- Accessibility (F_A): the ease with which a web crawler can visit a site, traverse its entirety and retrieve it via standard HTTP protocol requests. The website should provide resources so that a web crawler can discover and retrieve its different components (such as individual pages, images, and scripts). This facet also includes performance, or the speed at which a crawler can access the site.
- Standards Compliance (F_S): the website and its individual components conform to common accepted technical standards. For example, its HTML pages, conform to the W3C standards for HTML. It is also important that the website provided content in open file formats, instead of closed, proprietary formats such as QuickTime and Flash.
- Cohesion (F_C): the website does not have components that are dispersed across different locations on the web. For example, images, JavaScript files, and widgets.
- Metadata Usage (F_M): the website contains descriptive metadata such as HTTP headers and HTML META headers. It is important to note that the authors do not commit to a specific metadata model, but recommend using widely-accepted metadata models such as the Dublin Core standards.

Each of these facets has several components, or criteria, each with its own significance. Criteria with high significance are more important to the archivability of a website, and if they are not met, can cause problematic web archiving results or even prevent the website from being archived at all. Medium-significance criteria are not critical but are still important, while low-significance criteria are considered minor issues.

[7] stated that a website's archivability (WA) can be computed by using the sum total of its score for each facet. The value of each facet is the weighted average of its coordinates. The website has a score for each facet, represented as a tuple $(x_1, ..., x_k, ..., x_N)$. The value of x_k is either 0 or 1, which represents a negative or positive answer to a specific criterion. The components of a single facet are not weighted evenly, but are assigned a weight (ω_k) depending on their significance. These weighted scores are then divided to average them. Once the value for each facet has been calculated, the total archivability score for the website can also be calculated using the following equation: $WA = \sum_{\lambda \in \{A,S,C,M\}} w_\lambda F_\lambda$. F_A, F_S, F_C, and F_M represent the value of each facet with respect to accessibility, standards compliance, cohesion, and metadata usage.

[7] also created ArchiveReady, an evaluation system that implements the CLEAR+ model as a web application. ArchiveReady will calculate the websites's archivability and present it to the user in terms of a percentage.

Other researchers have also focused on the notion of archivability and attempted to operationalize it. In their paper "The impact of JavaScript on archivability", [8] defined archivability as the ease with which a website can be archived, which is similar to the concept put forward by [7]. The authors held that the current, live version of a website to be the ideal version. Thus, a perfectly archived website is one that replicates the original, live version in its entirety: "The web page in its live, native environment is the best version possible, and if an archival tool replicates the live web, it has perfectly captured and archived that resource" [8].

However, obtaining a perfect copy of the original is an onerous process, made more difficult by the widespread use of the JavaScript programming language. The use of JavaScript, in the form of small pieces of code called *scripts*, has made websites more personalized and interactive. Unfortunately, it has also made websites more difficult to archive. As the authors state, today's archival tools, such as the Heritrix web crawler employed by the Internet Archive, are unable to fully capture and render this complexity [8].

A website that contains JavaScript, such as Google Maps, functions differently from a traditional, HTML-only website. Typically, a web browser requests a website from a server, then proceeds to load the basic elements, such as HTML code and images. After the initial page is loaded, the JavaScript code is executed, This code will then request additional components to be loaded onto the page, such as the panning and zooming functions of an interactive map or geographic location features. [8] define

these type of websites as *deferred representations* because they are not "fully realized and constructed until *after* the client's-side representation is rendered". When attempting to archive such a website, a crawler will usually capture the initial components that are loaded first, but will not capture the other components that are loaded after the JavaScript code is executed. This is because crawlers cannot execute JavaScript code.

To study the impact of JavaScript on archivability, the researchers compiled two sets of archived URLs: some taken from the social media platform Twitter and others from the Internet Archive's Archive-It service [9]. The authors studied the quality of the archived URLs and their use of the JavaScript language, and presented several metrics to measure their archivability. Each URL had a specific number of client-side components (files which execute on the end user's computer, such as JavaScript) and server-side components (files which execute on the server). The authors called these components *parameters* and defined them in Equation 1. The complexity of a single URL was measured as the arithmetic mean of its *depth* (number of levels down from the top-level domain) and the number of client-side and server-side parameters, as shown in Equation 2.

$$F = max(|client - sideparameters|, |server - sideparamters|) \tag{1}$$

$$UC = \frac{|Depth| + F}{2} \tag{2}$$

$$CC = \sum script\ tags \in HTML \tag{3}$$

$$\begin{aligned} Javascript - loaded\ resources = Number\ of\ resources\ loaded \\ -Number\ of\ resources\ in\ HTML\ tags\ and\ CSS \end{aligned} \tag{4}$$

Unlike [7], [8] thought of archivability not as a discrete measurement, but as a dynamic one that changed over time. They found that over half (54.5%) of the URLs in their collection used JavaScript to load embedded resources, an increase of 14.7% between 2005 and 2012. Similarly, JavaScript was responsible for 52.7% of all missing embedded resources during the same time period, an increase of 32.5% [8]. Based on these findings, they concluded that the archivability of websites was being negatively affected by the increasing use of JavaScript, and that in the future, the completeness of archived websites would also decrease as a result.

It is worthwhile to note that the research published by [8] focuses on specific, single URLs, not on an entire website, which can consist of dozens or even thousands of URLs. However, it would be reasonable to assume that, if a single web page becomes less archivable the more JavaScript it contains, the same would apply to a complete website and even an entire web archive. The more JavaScript a website contains, the less archivable it is, and the more JavaScript a collection of websites contains, the less archivable they are as a group.

In 2016, [10] examined how popular open-source tools such as the Heritrix crawler and the Wayback Machine can be used to archive a corporate intranet. They found that the crawler was able to successfully crawl many pages; however, it would sometimes accidentally crawl sensitive information, and could not crawl pages which required private user credentials. Another important issue was the inability to correctly archive any resources that were constructed with JavaScript, such as YouTube, Facebook, and widget dashboards. To improve the archivability of these resources, [10] recommended using a headless browsing client such as PhantomJS, resulting in slower but more complete crawls of websites.

[11] introduced Retrospective Web Archiving (RWA), an approach for filling-in gaps of events which were not archived in real time when they originally took place. They tested their approach by building a retrospective web archive about the 2014 war in Gaza. The authors framed these differences as a matter of "platform archivability', where some platforms contain links that a) suffered from link rot (meaning they no longer functioned on the live web) and others from b) low archival coverage, meaning they had never been archived in the first place. They found that social media websites are particularly vulnerable

to link rot. Shortened URIs on Twitter were the most likely to suffer from both link rot and low archival coverage, while URIs extracted from Google and Facebook were less vulnerable to these conditions.

In [12], I examined how clients of Archive-It form mental models of websites and web archives that are often at odds with the realities of both web archiving and the web itself. these misconceptions are often not addressed until "something goes wrong", that is, a problem has occurred and they need to contact an AIT employee for help and support. Additionally, these misconceptions can create false expectations about what current web archiving technologies are capable of. If they are not addressed, web archivists run the risk of assuming that everything can be preserved, when this is not the reality.

3. Methodology

This paper focuses on the specific dimension of relevance in web archives, and was part of a larger project to build a comprehensive theory of IQ in web archives. The methodology described in this section is the same as the one employed in my previous work "Correspondence as the Primary Measure of Quality for Web Archives: A Grounded Theory Study" [4]. Though I describe it here, additional details can be found in that work.

Barney Glaser and Anselm Strauss created the methodology of Grounded Theory (GT), which they defined as "the discovery of theory from data - systematically obtained and analysed in social research' [13]. For the authors, theory was not a perfected product that explains all facets of a phenomenon, but a process, an ever-developing entity. GT is an inductive methodology; working closely from the data, the researcher begins the work of generating a theory.

3.1. Data Gathering and Processing

The Internet Archive's Archive-It (AIT) is a subscription-based web archiving service that helps organisations build and manage their own web archives. Archive-It is currently the most popular web archiving service, with over 600 clients (called "partners') consisting of universities, state libraries and archives, museums, and national libraries in several countries [14]. The accounts of Archive-It clients are managed by a team of partner specialists. When a client encounters a problem with Archive-It, she first opens a support ticket using Zendesk, a popular customer-service platform. The ticket is received by a partner specialist, who is then responsible for addressing the issue.

AIT support tickets are a rich source of information regarding quality problems in web archives. They contain the opinions and views of individuals who are experienced creators of web archives, well-versed in web archiving processes, and familiar with institutional web preservation goals, whether they be clients or the partner specialists themselves. They contain rich descriptions of how quality problems are detected, analysed, and addressed, and are thus an ideal dataset for studying quality in all its dimensions.

The first step was to obtain Archive-It support tickets in order to analyse them. Since these tickets belonged to the Internet Archive, I negotiated a researcher agreement with the organisation to obtain support tickets from the years 2012 through 2016. The tickets received comprised a wide variety of institutions reflecting AIT's client base, from national libraries, to private organisations, to universities and museums from Europe, North America, and Asia. After the tickets were cleaned, I randomly selected the same amount of tickets for each year from 2013 through 2016. This randomisation approach was taken to minimise the selection bias that might have occurred if I had manually chosen which tickets to analyse. The final dataset of 645 tickets was then imported into the NVivo software package, a popular program for performing qualitative data analysis [15].

Among other conditions, the research agreement stipulates that the researcher anonymise any personal or institutional information present in the tickets, as well as any other potentially identifying information. In order to comply with the terms of this agreement, all the information presented in this paper has been anonymised: identifying elements such as personal names, names of institutions, and website addresses have been removed or changed.

3.2. Data Analysis

The tickets collected were Level 1 support tickets that had been submitted by AIT client. They included the initial question submitted by the client, the response given by the AIT partner specialist, and any subsequent communication between the two. It is important to note that not all the AIT tickets deal with issues of quality in a web archive. Quite a few deal with collection management issues, such as how to manage user accounts for a collection of web archives, storage limitations, and questions about the privacy or public access to archived content. This research focuses on tickets in which the client discusses a perceived flaw in an individual archived website or an entire web archive. From prior experiences, I had seen that these types of tickets are the most likely to deal with issues of quality.

Support tickets not pertaining to quality issues were classified as such and separated from the main data of interest. Each ticket analysed consisted of the original ticket submitted by the client, the response sent by the AIT partner specialist, and any subsequent interactions between them. Tickets could be quite brief, consisting of three interactions (the original client ticket, the partner specialist's response, and the client's response), or they could have many interactions over time, spanning weeks or even months. A total of 305 tickets and 2544 interactions were analyzed.

These support tickets were analysed using the GT techniques of open coding and theoretical memos to identify the main concepts and categories present in the data. According to the precepts of GT, after several rounds of coding, the researcher will reach *saturation*, a state when nothing new is being extracted from the data. Per the guidelines of Grounded Theory, only the core categories (that is, the ones that explain most of the variation in quality) are part of the final theory. In order to increase the quality and rigour of the study, I engaged in purposeful peer review. University professors were periodically invited to audit the entire research project, including the codebook, preliminary findings, and core categories. In addition to peers, employees of the Internet Archive were also invited to see the findings and comment on them.

4. Findings

As previously discussed, the notion of archivability has already received some attention from academic researchers. It was defined by [8] as the ease with which a website can be archived. I redefine archivability as the intrinsic properties of a website that make it easier or more difficult to archive. Archivability is highly dependent on the technology being employed to do web archiving. As technology evolves over time, web components that were previously thought to be unarchivable might become archivable, and viceversa. Archivability proved to be a prominent dimension, as it appears 101 times in 78 tickets. The data showed several factors that greatly affect the archivability of a website. According to the data presented here, archivality problems occur because a website:

1. has changed the way the content is delivered to the user.
2. is media-heavy or contains much dynamic content.
3. renders content in a unique, "non-standard' way.

Table 1 presents examples of the first situation. The marker *C:* refers to the original query presented by the AIT client, while the marker *AIT:* refers to the answer given by the web archivist. Many websites routinely change the way the content is delivered to the user, thus a website can go from being easily archivable to practically unarchivable fairly quickly. As one AIT employee said: "The web, and specifically social networking sites can be a moving target.' When websites change their internal functionality, it can result in the archived website looking different from the original (tickets 08 and 129) and missing content (tickets 08 and tickets 258).

Cases where archivability was negatively impacted by the heavy presence of dynamic content are shown in Table 2. Generally, sites that utilize technologies such as JavaScript, Flash, and streaming audio and video are difficult to capture and render like the original. This finding is consistent with the work of [6] and [8]. A special case of this situation is seen with websites that are database and form or

Table 1
Examples of archivability problems caused by websites changing how it delivers content to users.

Ticket No	Text of the Ticket
8	*C:* Our Facebook page didn't get archived. When I viewed what was crawled, all the came up was basically a blank Facebook page. *AIT:* Both Facebook and Twitter have made some changes recently to the way they set up their sites, which requires a little bit of work on our end to catch up. 1. For Facebook, your site was archived, there is just an issue that is keeping the archived page from displaying normally. Our engineers are working on this and it should be fixed this week. I will let you know as soon as I have further information. 2. For Twitter, they recently removed the "more" button from twitter feeds and instead users access older tweets by scrolling down the page. The way this feature is set up makes it difficult for our crawlers to access the older content that is not displayed automatically.
129	*C:* I am getting an error on the following Facebook crawl. *AIT:* Facebook made a change to the settings for their stylesheets
258	*C:* in the "Township of ___" collection I am trying to capture this facebook site: http://www.facebook.com/pages/township *AIT:* We are still generally able to capture the initial content on a Facebook timeline; however the most recent change from Facebook has made it one again difficult to capture dynamically loading content as a user scrolls down through the page

search-driven, such as library catalogs, web forms, or search engines. As the AIT employee explains, these are elements that depend on a myriad of complex, dynamic interactions that cannot be replicated in an archived website.

Sometimes websites will have unique or unusual ways of rendering content, which can negatively affect archivability, as seen in Table 3. For example, some content management systems can create endlessly repeating directory structures (such as http://somesite.com/news, http://somesite.com/news/news, and http://somesite.com/news/news/news). The presence of these will cause the crawler to go into infinite loops (crawler traps) in an attempt to capture all levels of the website. This can lead to poor-quality archived websites, stalled or incomplete crawls, and large amounts of unnecessary data.

5. Discussion

5.1. How do people conceptualise the notion of archivability?

Archivability is not a dimension of quality that is directly perceived by most AIT clients. AIT employees, who have a deep knowledge of and experience with the technical process of archiving websites, were much more likely to perceive a quality problem as an archivability problem. Instead, AIT clients framed archivability issues as correspondence or relevance problems that negatively affected the quality of an archived website or of an entire web archive. For example, in Table 1, ticket 8 describes a situation where the AIT client is writing because she sees a blank Facebook page, while it is the AIT employee that is able to determine that the blank page is being caused by Facebook changing how it delivers its content. That is, an archivability problem is being framed in terms of a problem with visual correspondence, since the archived website does not look like the original. Similarly, in Table 2, in ticket 100, the AIT client frames the problem as one of interactional correspondence, that is, the archived site does not "flip through the slides" in a slideshow, and so does not behave as the original website does. Table 4 lists the support tickets analysed, as well as the IQ dimension in which they are framed.

Website archivability can thus be seen as a *latent* dimension, because it is hidden from most people, and framed in terms of other quality problems. **The archivability of a website can only be perceived**

***after* its archived counterpart exhibits a quality problem.**

In Social Sciences research, the term *manifest variables* is used to describe variables that we can directly observe [16]. Manifest variables stand in contrast to *hypothetical constructs*, also known as *latent constructs*, which cannot be directly observed [17]. In order to detect the presence of a latent construct or measure its impact, manifest variables are used as operational definitions. "We therefore assume that the presence and strength of the manifest variable reflects, albeit imperfectly, the presence and strength of the hypothetical construct" [16].

I advance that website archivability is one such latent construct, because it cannot be directly measured until the website is actually archived. Any proposed archivability measurement that is taken before the website is actually archived, such as those presented by [7], is a *probability* measure and is at best an estimate of the likelihood that a website will be preserved. The actual archivability of a website can only be seen after archival.

5.2. How does website archivability affect web archives and thus the future historical record?

As can be seen in the findings, a website with low archivability can negatively impact the quality of its archived version by causing correspondence or relevance problems. An archived website with low correspondence will look different from the original, behave in a different, degraded manner, and have missing elements. An archived website with low relevance will have content that is unrelated to that of the original website, or will be much larger in size than the original. *Low archivability in a website leads to low-quality web archives, and low-quality web archives lead to low-quality historical records.*

The degree of archivability of a website can be estimated *a priori* by calculating how much of it is composed of dynamic content, such as JavaScript. However, its true archivability of a website can only be determined *a posteriori* by detecting correspondence or quality problems. This points to the nature of web archiving as a reactive practice instead of a proactive one. **Any substantial change in web technologies, standards, or platforms necessitates a change in web archiving practice in order to "catch up" and create high-quality web archives that result in a high-quality historical record.**

Many parts of the web have always been out of reach: websites that are database-, form-, or search-driven have always been impossible to capture. The increasing reliance on dynamic, client-side technologies such as JavaScript has also done much to decrease website archivability [8]. Given these findings, it is safe to conclude that the web is no longer archivable, and will become increasingly so as time passes. Current web archiving technologies cannot adequately capture the web as it is now, yielding, at worst, highly-degraded versions of the original.

6. Conclusion

This paper makes the following contributions:

1. It presents a detailed, comprehensive definition of website archivability, one of the most important dimensions of information quality in web archives.
2. This definition of archivability is human-centred and grounded in how web archivists perceive quality in web archives.
3. The definition is independent of the technology currently in use to create web archives, making it suitable to a wide variety of platforms, preservation contexts, and situations.
4. It characterizes archivability as a latent construct of information quality, expressed by the manifest variables of correspondence and relevance, a finding not previously seen in the literature.
5. It describes how low website archivability influences the quality of web archives as historical records.

As historian Ian Milligan stated "web archives do not provide a perfect representation of the past...but neither do traditional archives, which have had to be very selective with what they select, appraise, and preserve"[18]. Though this paper paints a sobering picture of decreasing website archivability and its implications for the historical record of the web, perhaps it is not web archivists, but historians, who will be the most able to deal with the errors and omissions of the past web.

References

[1] J. Bailey, A. Grotke, E. McCain, C. Moffatt, N. Taylor, Web Archiving in the United States: A 2016 Survey, Research Report, 2016. URL: http://ndsa.org/publications/.

[2] Archive-It, Learn more, https://archive-it.org/learn-more, 2021.

[3] S. Abrams, Z. Collier, E. Colón-Marrero, keondra bills freemyn, N. Krabbenhoeft, M. E. Wertheimer, A. Wickner, 2022 web archiving survey results, 2023. URL: http://ndsa.org/publications/.

[4] B. Reyes Ayala, Correspondence as the primary measure of quality for web archives: A grounded theory study, in: M. Hall, T. Merčun, T. Risse, F. Duchateau (Eds.), Digital Libraries for Open Knowledge, Springer International Publishing, Cham, 2020, pp. 73–86.

[5] J. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle, M. L. Nelson, Not all mementos are created equal: measuring the impact of missing resources, International Journal on Digital Libraries (2015) 1–19. doi:10.1007/s00799-015-0150-6.

[6] V. Banos, Y. Kim, S. Ross, Y. Manolopoulos, CLEAR: A credible method to evaluate website archivability, Presented at the 10th International Conference on Preservation of Digital Objects (iPRES 2013), 2013. URL: http://www.academia.edu/10967309/CLEAR_a_credible_method_to_evaluate_website_archivability.

[7] V. Banos, Y. Manolopoulos, A quantitative approach to evaluate website archivability using the CLEAR+ method, International Journal on Digital Libraries (2015) 1–23. doi:10.1007/s00799-015-0144-4.

[8] J. Brunelle, M. Kelly, M. Weigle, M. L. Nelson, The impact of JavaScript on archivability, International Journal on Digital Libraries (2015) 1–23. doi:10.1007/s00799-015-0140-8.

[9] Archive-It, Learn more, 2014. URL: https://archive-it.org/learn-more.

[10] J. F. Brunelle, K. Ferrante, E. Wilczek, M. C. Weigle, M. L. Nelson, Leveraging heritrix and the wayback machine on a corporate intranet: A case study on improving corporate archives, D-Lib Magazine 22 (2016). URL: http://www.dlib.org/dlib/january16/brunelle/01brunelle.html. doi:10.1045/january2016-brunelle.

[11] A. Ben-David, 2014 not found: a cross-platform approach to retrospective web archiving, Internet Histories 3 (2019) 316–342. URL: https://doi.org/10.1080/24701475.2019.1654290. doi:10.1080/24701475.2019.1654290. arXiv:https://doi.org/10.1080/24701475.2019.1654290.

[12] B. Reyes Ayala, When expectations meet reality: common misconceptions about web archives and challenges for scholars, International Journal of Digital Humanities (2021). doi:https://doi.org/10.1007/s42803-021-00034-3.

[13] B. Glaser, A. Strauss, The Discovery of Grounded Theory: Strategies for Qualitative Research, Aldine Transaction, 2009. URL: http://amazon.com/o/ASIN/0202302601/.

[14] Archive-It, Learn more, 2020. URL: https://archive-it.org/learn-more.

[15] QSR International, Nvivo product range, 2016. URL: http://www.qsrinternational.com/nvivo-product.

[16] M. E. Kite, J. Whitley, Bernard E., Principles of Research in Behavioral Science., Routledge, New York, NY, 2018.

[17] S. El-Den, C. Schneider, A. Mirzaei, S. Carter, How to measure a latent construct: Psychometric principles for the development and validation of measurement instruments, International Journal of Pharmacy Practice 28 (2020) 326–336. URL: https:

//onlinelibrary.wiley.com/doi/abs/10.1111/ijpp.12600. doi:https://doi.org/10.1111/ijpp.12600. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijpp.12600.

[18] I. Milligan, Historiography and the Web, Sage Publications Ltd, Los Angeles, CA, USA, 2018, pp. 16–29.

Table 2
Examples of archivability problems caused by websites with dynamic content.

Ticket No	Text of the Ticket
100	*C:* In reviewing our crawls, I have noticed a few of our pages do not display correctly. The problem pages either utilize flash or javascript. I know javascript can be problematic. The following pages are the not displaying correctly: • http://www.___.edu/ Video does not display/does not flip through slides • http://www.___edu/y79.xml Does not flip through slides • www.___.edu/y55333.xml Does not display additional photographs • http://www.___.edu/y213.xml Does not display photograph slides • http://www.___.edu/BBQ.xml Does not display photograph slides • http://www.___.edu/toc.xml Does not display photograph slides *AIT:* flash and Javascript can be difficult to capture or display sometimes
76	*C:* From the crawl report, it looks like a reasonable number of urls were captured, but also a good number show up in the "out of scope" column. Then, when I checked the Wayback version I got a message that the archived site cannot be displayed within a frame (see screenshot). I'm not sure what this means or if it is possible to adjust the crawl to fix it. Do you have any suggestions for the best way to capture these videos? *AIT:* streaming video can be difficult to archive sometimes
369	*C:* The athletics department has their game day programs online. I see to be able to view the sections but can't see a way to capture printer-friendly formats from their link. Is this possible? *AIT:* It looks like the site uses a fair bit of javascript to generate those "printer friendly' pages, but I'm not sure how feasible capture is
2884	*C:* under the About Us tab, under Press Room, the tabs other than News Releases (___ in the news, Annual report, Media Kit, and Social Media) do not work: *AIT:* Regarding the tabs on the Press Room URL, I am not sure if we will be able to capture this content due to the dynamic way in which these links are generated
	Special case: Websites that are database and form or search-driven
30	*C:* Much of this content is located in databases, so, in general I'm curious about how Archive-It will handle these databases. Here are two examples: • http://www.ourhistory.org/search.asp#index The results of a search come up on a site with a URL with the same seed, but will Archive-It crawl this database? • http://www.ourhistory.org/iq/register/welcome.asp Same as the site above, the results remain on the same seed, will Archive-It crawl the database? *AIT:* if database driven parts of sites have direct links to the content, the crawler will capture those, however the crawler can't enter search terms or interact with forms, so if that is the only way to access the database content, the crawler likely will not automatically be able to access that content
3481	*C:* This site has a database backend and is queried via text input fields on the phonebook homepage...Could I get some help on how to do this successfully? *AIT:* Because of their interactive nature, search boxes cannot operate in an archived website in the same way as they would on the live web
3458	*C:* I would like to know if there is any way I can capture the search feature of the website *AIT:* Search boxes are something that will not behave in an archived site like they do on the live web. We can archive content that would be returned by using the search function (as you noticed with the "Browse All Projects' button) however, the crawler is not able to archive the database or search engine that the live site search runs off of

Table 3
Examples of archivability problems caused by websites rendering content in unique ways.

Ticket No	Text of the Ticket
464	*C:* When you're navigating through the catalogs themselves, you also come across the same issue of not being able to get from 1 page to 2 page or 3 or 4 or 5. However, all of the lots are captured, but you have no way of accessing them through the site *AIT:* The way that this site does it's navigation is significantly more complicated than your average site due to the form based dropdowns that you notice to the right of the pagination at the top of the list. The "Sort" and "per page" options are actually forms, so instead of simply clicking on links to subsequent or previous pages (the way that most sites do pagination), the crawler would actually have to select an option from the dropdown and submit a form each time, in order to get content back. These are types of interactive behavior the crawler does not perform by default, so it will require additional development...Because this site is so uniquely complicated in the way it has implemented pagination, any work our engineers put into developing a new crawling feature to capture it would be very specific to this site and likely not transferrable to other examples
3423	*C:* I've done a test crawl on all ".stateu.edu" while it has captured thousands of pages it also seems to determine many "stateu.edu" pages to be "out of scope". These pages are not blocked by a robots.txt. Why would that be happening? *AIT:* We do see these types of repetitive URLs from time to time, and they appear to be generated by code in certain implementations of content management systems like Drupal
3001	*C:* We're having some trouble limiting the URLs on one of crawls to a reasonable number. We're actually getting the content we'd like, but we're also getting a ton of extraneous URLs that are either bad content or aliases to content we're already capturing with another URL. *AIT:* After taking a look at the queued URLs for this host, it appears that the crawler is running into a trap that we see from time to time on some websites (including some Drupal sites) where the site generates links with repeating directories
86	*C:* We ran another test crawl on this site and now we seemed to have opened a can of worms for the main site we wanted to crawl. It looks like we have the Flickr URL under control and may just put a limit on the number captured. The main URL that we want, captured 83, 600 with over 1 million in the que *AIT:* The issue with your http://www.pl.gov/tef/ site is one that we see from time to time, where something in the way the site is put together creates urls with repeating directories that all point back to the same page

Table 4
Archivability problems and the IQ dimensions and sub-dimensions in which they are framed.

Information Quality Dimension and Sub-Dimensions	Ticket No
Correspondence, Visual	8, 129, 100, 76
Correspondence, Interactional	2884, 464
Correspondence, Completeness	258, 369, 30, 3481, 3458
Relevance, Topic	3423
Relevance, Size	3001, 86, 76