Applying Accessibility Metrics to Measure the Threat Landscape for Users with Disabilities

John Breton
Carleton University
john.breton@carleton.ca

AbdelRahman Abdou
Carleton University
abdou@scs.carleton.ca

***Abstract*—The link between user security and web accessibility is a new but growing field of research. To understand the potential threat landscape for users that require accessibility tools to access the web, we created the WATER framework. WATER measures websites using three security-related base accessibility metrics. Upon analyzing 30,000 websites from three distinct popularity ranges, we discovered that the risk for information leakage and phishing attacks is higher for these users. Over half of the analyzed websites had an accessibility percentage of less than 75%, a statistic that exposes these websites to potential accessibility-related lawsuits. Our data suggests that the current WCAG 2.1 standards may need to be revised to avoid assigning Level AA conformance to websites that undermine the security of users requiring accessibility tools. We make the WATER framework publicly available in the hopes it can be used for future research.**

I. Introduction

As the Internet continues to expand, individuals that require assistance to access the web are increasingly suffering due to a lack of adherence to accessibility best practices. Despite government legislation detailing strict accessibility requirements for websites, the global extent of the Internet exacerbates having these recommendations applied on a large scale [15].

This rapid expansion has also given rise to various attacks perpetrated by bad actors and negligent site operators. Some of the most egregious of these are attacks that target individuals with disabilities. These often succeed due to a lack of adherence to accessibility best practices presented throughout the web [22]. Such attacks normally target the privacy and security of the users involved. As this is a relatively small subset of the global set of web users, little research exists that measures the security risk imposed on users due to websites failing to uphold web accessibility standards [16]. Regardless of the fact that users with disability make up a smaller portion of users on the web, their increased risk of tailored privacy attacks cannot be overlooked.

This paper aims to fill the gap in the literature by measuring the threat landscape for users that require accessibility tools to access websites. Complete assessment of this landscape requires measuring the overall accessibility of websites using metrics that highlight accessibility-related vulnerabilities—those that can be exploited by bad actors if a website is not accessible (i.e., conforming to the Web Content Accessibility Guidelines (WCAG) 2.1 standards).

We focus herein on users of screen readers and alternative means of web page navigation. We analyze the impact of website popularity on its accessibility, and consequently if users are more at risk if they browse sites within a specific popularity range. A long-standing research question proposed by the World Wide Web Consortium (W3C) involves determining if basic accessibility metrics could be used to deem a website accessible [1]. We also aim to provide a data-backed approach to answering that open question. This can pave the path for future research that aims to improve the overall accessibility of websites, which in turn could help reduce the threat landscape for users that require accessibility tools to access the web.

We present our measurement framework, called WATER, and use it to analyze 30k websites from Alexa's top 1M list published on November 11th, 2022.[1] Three runs were performed using WATER, analyzing the top 10k, the middle 10k, and the bottom 10k sites listed in the top 1M. Our analysis suggests that the popularity of websites has no significant correlation with the accessibility of a site. In general, we find that the threat space for users is very large, with over 65% of the sites investigated having the potential to instigate targeted attacks, such as phishing, against these users, and over 80% having the potential to cause information leakage.

We make the following contributions:

1) Construct three basic accessibility metrics related to the ability of a website to minimize threats against users that require accessibility tools to access the web.
2) Develop WATER—a system to assess website conformance to the three base metrics alongside the accessibility percentage of websites across the Internet.[2]
3) Present, upon analysis, data showing that (a) basic accessibility metrics are not enough to determine the accessibility of a website, and (b) the threat landscape for users that require accessibility tools is as large as >80% of the 30,000 domains analyzed.

The rest of this paper is structured as follows. Section II discusses how web accessibility relates to security and pro-

Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb) 2023
3 March 2023, San Diego, CA, USA
ISBN 1-891562-87-8
https://dx.doi.org/10.14722/madweb.2023.23074
www.ndss-symposium.org

[1] This experiment was conducted on November 11th, 2022—prior to the shutdown of the Alexa service. Similarly formatted csv files will work with WATER.

[2] Both the data used in this paper and WATER are available at: https://github.com/john-breton/WATER

vides an overview of the latest web accessibility guidelines. Section III defines the three base metrics being measured. Section IV describes the methodology employed via the use of the WATER framework. Section V presents the collected data. Section VI analyses the results, lists project limitations, and discusses avenues for future research. Section VII touches on some related research in the field of measuring web accessibility. Lastly, Section VIII concludes the paper.

II. Background

In this section, the relationship between web accessibility and user security will be briefly documented, alongside a discussion of the current WCAG 2.1 standards and the notion of accessibility percentage.

A. Web Accessibility and Security

The relationship between web accessibility and security is a recent development. As the literature highlights, the main concerns for users when faced with a lack of web accessibility are privacy-related [11], [18]. Privacy and confidentiality of data are in many cases interconnected. For example, a password leak (security) can lead to divulging a user's mailbox, health records, or bank accounts.

B. WCAG 2.1 and Accessibility Percentage

The *Web Content Accessibility Guidelines (WCAG)* is an exhaustive list of accessibility checks detailed by the W3C [13]. The list itself is divided into four major categories that aim to ensure the web is accessible to all users. These categories are Perceivable, Operable, Understandable, and Robust. The latest revision, WCAG 2.1, was released in 2018.

Conformance is another major section listed in WCAG 2.1. It outlines requirements that must be satisfied if a website wishes to claim they conform to WCAG 2.1. Conformance comes in three levels, Level A, Level AA, and Level AAA. Level AAA is the best type of conformance, but even the WCAG 2.1 document outlines it is not always possible to achieve Level AAA for all types of content. An alternative to the complete conformance requirements laid out in WCAG 2.1 is the notion of an Accessibility Percentage (AP). AP is a quantifiable measure of the overall conformance to all requirements listed in the first four sections of WCAG 2.1. Online tools currently exist to measure the AP of any given web page, such as the online service offered by LevelAccess known as WebAccessibility [3].

It is important to note that AP measures up to Level AA conformance for a website. That means an AP of 100% would indicate a website achieve a conformance level of Level AA according to WCAG 2.1. However, it is not possible to make any claims about the conformance level of a site for anything less than an AP of 100%, as the percentage does not express if all Level A requirements are satisfied or not. Regardless, AP is still a useful measure of a site's overall conformance to the WCAG 2.1 standards and provides a convenient way to compare the conformance of different sites using a quantifiable numerical value compared to the three levels outlined by the WCAG 2.1 document.

III. Base Accessibility Metrics

In this section, we present three security metrics conceptualized using the guidelines posted by W3C in [1]. These metrics make use of conditions that are already used in the calculation of the overall AP of a website, however, they are being re-contextualized with a security-focused perspective. To the best of our knowledge, we are the first to present security-related web accessibility metrics. As the landscape of disabilities is large, we limited the scope of metrics we consider to metrics that could be determined using raw HTML and that were directly applicable to our target group of users (users that use screen readers and alternative means of site navigation). While the creation of additional security-related web accessibility metrics is possible, we do not believe further security-related accessibility metrics could be derived for our specific target group without extra data. We discuss the possibility of adding additional metrics to WATER in Section VI.

We derived the following metrics under a "worst-case" analysis scenario. The attacks that can arise due to a lack of satisfaction with the following three metrics exist in two categories. The first is a purposely designed website owned and operated by a malicious actor, with a target of users that require screen readers and alternative means of navigation to access the web. The second case is a website operator that is not acting in bad faith or with intent to harm users, but their failure to adhere to accessibility best practices exposes this subset of users to potential privacy violations and opportunities for information leakages.

This dual scenario is especially troubling as it demonstrates that users that make use of screen readers or alternative forms of site navigation may be subjected to threats that have not been produced from any malicious intent or any form of gain. As such, when calculating each of the following metrics, we chose to measure the ratio of adherence to best practices as is recommended in [1], as even a single violation could expose these users to threats. We suggest that the more violations that exist for a particular site, the higher the risk is to users that make use of screen readers and alternative means of website navigation. The only way for threats to not exist for these users is to fully adhere to accessibility best practices.

A. Image Tag Alt Adherence (ITAA)

In HTML, `<img>` tags can have an *alt* (short for alternative) attribute, which specifies text that would be displayed in place of the image if the image fails to load, or if screen readers are used. We define the *Image-Tag-Alt-Adherence* as a base metric that measures the proportion of `<img>` HTML tags that have a meaningful *alt* attribute associated with them. It is measured on a scale of 0 to 1, with 1 being the best possible score. A lack of *alt* attributes for `<img>` tags has been previously shown to pose general privacy risks for users that access a website with screen readers [22]. Of specific note, screen readers typically read out the filename of an image when an *alt* attribute is not present, which can lead to users sharing sensitive information due to them being unable to determine which image they are currently selecting. An example of this could occur on social media sites that allow users to upload and store their photos while giving them the option to share photos with friends. Should the *alt* tags for the stored photos

not be meaningful, a user that relies on this information may accidentally share a private photo with their friends when that was not the intent.

To calculate this metric, we collect the entirety of a web page's <img> tags. Next, we check if the following conditions are satisfied:

1) Does the <img> tag have an *alt* attribute?
2) Is the *alt* attribute not empty?
3) Is the *alt* attribute meaningful?

While the first check is straightforward, the next two checks require further explanation. In WCAG 2.1, any <img> tags with an empty *alt* attribute are not considered accessibility violations. The rationale behind that decision is that such images may be simple website decorations. Leaving the *alt* attribute blank fixes the problem of having screen readers read out the full filename of an image, but it does not fully address the potential for privacy violations to occur.

Suppose a website instigates a policy to have a blank *alt* attribute for all images on their web pages. In such situations, the risk that users using screen readers may share private information is still largely relevant. Current talks for best practices suggest that any website decorations that would not require an *alt* attribute be made with CSS instead, yet Section 1.1.1 of WCAG 2.1 does not take this suggestion into account [8], [13]. In an alternative scenario where the website is owned by a bad actor, they could purposefully avoid the use of *alt* attributes on images to promote the chance for these users to expose potentially privacy-sensitive information when using the site. In both cases, the risk for privacy violations of the user remains the same.

The final check is the most subjective. Again, the current discussion suggests avoiding the use of terms such as "image of" and "graphic of" [8]. Ultimately, the *alt* attribute is meant to describe in some detail what is going on in an image, rather than simply stating that it is an image of something. This check is performed using basic string comparison to see if the *alt* attribute text contains any of the above phrases (the ones that should be avoided). If this, along with the other two conditions, are satisfied, an image is considered to pass this check. Once all images have been analyzed, the final ITAA score is calculated using the following equation:

$$\text{ITAA} = \frac{\text{\# of <img> with meaningful } \textit{alt} \text{ attributes}}{\text{total \# of <img> tags}}$$

B. Hyperlink Astonishment Minimization (HAM)

We define the *Hyperlink-Astonishment-Minimization* as a base metric that measures the proportion of <a> tags that contain appropriate textual descriptions for any *href* attributes contained within. It is measured on a scale of 0 to 1, with 1 being the best possible score. *Least-surprise* is a design principle that suggests that actions should be taken to minimize the astonishment a user is subjected to when interacting with a system [19]. This principle is relevant to the *href* recommendations listed in WCAG 2.1. However, it goes beyond design practices in this case.

Consider a scenario where a user is using a screen reader coupled with voice recognition to navigate a website and they arrive to a hyperlink. The screen reader may read out the text associated with the hyperlink, rather than the URL a user will be redirected to. Naturally, this can allow malevolent site owners to redirect these users to malicious sites, as the user is simply hearing the text associated with the hyperlink. In the case of a legitimate site owner, the violation of the *least-surprise* principle [19] can cause users to become disoriented and result in sequence of actions that could threaten the privacy of these users, such as those presented in the definition of ITAA. As such, it is imperative that the text associated with a hyperlink be related in some way to the web page a user will be redirected to upon clicking on the hyperlink, not only to minimize astonishment but also to protect users from potential attacks.

To calculate our HAM metric for a page, we collect the entirety of a web page's <a> tags with *href* attributes, and check if the following condition is true:

- Does the text in the <a> tag with an *href* attribute appear in the hyperlink itself?

So long as this condition is satisfied, the hyperlink is considered to pass this check. An example of where this is properly implemented is when a hyperlink that redirects to a login page as evidenced by the URL has the associating text: "Click here to login!". Conversely, a failure case would occur if that same text was used to redirect the user to a page not related to the authentic login form. Once all hyperlinks have been analyzed, the final HAM score is calculated using the following equation:

$$\text{HAM} = \frac{\text{\# of <a> with appropriate text for } \textit{hrefs}}{\text{total \# of <a> tags with } \textit{hrefs}}$$

C. Label Input Mapping (LIM)

We define the *Label-Input-Mapping* (LIM) as a base metric that measures the proportion of <input> tags that are unambiguously associated with at least one <label> tag. It is measured on a scale of 0 to 1, with 1 being the best possible score. Mapping an <input> tag to a <label> tag is once again considered in WCAG 2.1 [13], as screen readers can fail if a text input does not have an associated label, leading to ambiguity over what the input field is actually for.

This ambiguity is the main factor that allows unlabeled <input> tags to be used for information leakage attacks. In a scenario where an input field is unlabeled, a screen reader will not explain what the input field is used for. In such situations, users may become confused and submit sensitive information into a field in situations they did not intend to. An example of this could be supplying a password into a field that is used to share a social update with others.

Again, we collect all of a web page's <input> and <label> tags to calculate our LIM metric. We then perform a simple check to determine if the following condition is true:

- Does the ID attribute of a <label> tag match that of an <input> tag?

So long as this condition is satisfied, the input field is considered to pass this check. In these situations, screen readers will read the label whenever a user selects the input to

provide the user with the details surrounding what the input field is expecting from the user. Once all <input> tags have been analyzed, the final LIM score is calculated using the following equation:

$$\text{LIM} = \frac{\text{\# of <input> tags with associating <label> tags}}{\text{total \# of <input> tags}}$$

IV. Measurement Methodology

This section will outline the methodology employed to run active measurements in order to retrieve the data necessary to determine the threat landscape for users that require accessibility tools to access the web. We developed a Python-based system, WATER, to accomplish this task. WATER has three decoupled modules that can run as one cohesive unit. Figure 1 outlines the three modules. Data collection, analysis and visualization (plotting) are fully contained in the WATER system, which facilitates future data collection by other researchers, and is ideal for a longitudinal study that analyzes the evolution of security vulnerabilities arising from improper accessibility configuration. In WATER, the HTML scraping module is run first, followed by metric calculations, and ending off with data visualization.

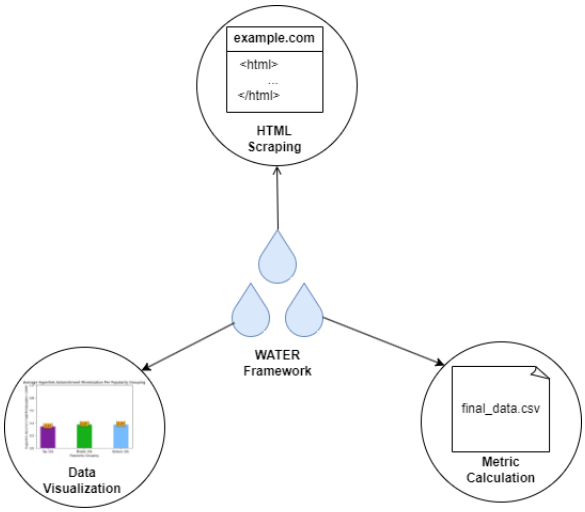

Fig. 1. Overview of the WATER Framework

A. HTML Scraping

The HTML scraping module of WATER is responsible for retrieving the raw HTML of a target web page. This is accomplished by supplying WATER with a csv file containing the domains of the websites to be analyzed. For each domain, WATER will make use of a selenium-powered headless browser to load and execute a simple JavaScript command to scroll to the bottom of the page before the HTML is retrieved. The rationale behind scrolling to the bottom of the page before scraping the HTML is to ensure any images that load as a part of that action have a chance to be collected, as images are needed to properly calculate ITAA.

It should be noted that WATER will only scrape the exact domains specified in the supplied csv file. It will not check for sub-pages, and if those needed to be analyzed they would need to be specified directly within the csv file. For this research, only landing pages (homepages) for domains were considered. While we acknowledge the importance of analyzing other pages of a website [10], we were interested (in this phase of our research) to prioritize breadth over depth—analyze homepages from as many websites as we can. Analyzing additional pages of a website is an area of future work.

WATER supports multiprocessing, and will automatically split the supplied csv between a number of processes that can be specified prior to execution. The operation to fully load a website can take a non-significant amount of time, and having the ability to split the task between processes drastically reduces the execution time of the HTML scraping module of WATER.

Once WATER confirms a website is fully loaded, it will scrape the HTML of the website. In the event a website cannot form a connection within a modifiable timeout period, WATER will move on to the next listed website. If HTML is retrieved, WATER will filter it to only extract the target HTML tags needed for analyzing our metrics. Such filtering then locally stores: all <img> tags, all <a> tags that contain an *href* attribute, all <input> tags, and all <label> tags. Once this filtering process has been completed, the HTML data is written to a temporary JSON file for that specific website that will be used later for metric calculation.

We ran the HTML scraping module three separate times with three csv files. These contained the top 10k, the middle 10k, and the bottom 10k domains listed in the November 11th, 2022 Alexa Top 1M Sites list. This was done to determine if the threat landscape may vary with website popularity. The JSON files for each of these three runs are available alongside the WATER framework, however, the default behaviour of the framework is to discard these files once metric calculations have been completed to maintain space on the user's machine. In total, WATER successfully scraped 8,915 of the top 10k websites, 9,283 of the middle 10k, and 7,325 of the bottom 10k.

B. Metric Calculation

Acting as the main module, the metric calculation module of WATER calculates the three metrics mentioned in Section III. WATER will read through the JSON files created by the HTML scraping module to perform metric calculations. Calculating these metrics does not depend on any queries, and thus this is the fastest component of the WATER framework execution-wise. If a website did not have any data related to a metric (i.e., no <img> tags appeared so ITAA cannot be calculated), it will store the result as 'No Data'. Otherwise, a float between 0.0 and 1.0 is used to represent the score of each individual metric.

While the metric calculation module is mainly responsible for calculating metrics, we extended it to also retrieve the AP of a given website. Unlike the metric calculations, this process is query dependent. As mentioned in Section II, the AP is determined via WebAccessibility, an online web tool provided by Level Access. This process is significantly rate-limited in the WATER framework to avoid overwhelming the service. Queries are submitted using a headless selenium browser, as was the case in the HTML Scraping module. This process does

not occur through an API, and we directly submit the website URL to the WebAccessibility tool using selenium. Should the tool fail to return the AP in a period of fewer than 60 seconds, WATER will record the AP as -1%. Otherwise, WATER will record the exact percentage returned by the WebAccessibility tool.

Once all three metrics and the AP for a website have been determined, a csv entry is created. This entry lists the website URL, ITAA, HAM, LIM, and finally the AP. Once all JSON files generated by the HTML scraper module in a directory have been analyzed, the results are saved to a final csv file whose path can be specified by the user. This csv file can then be used by the data visualization module for analysis purposes. In total, three csv files were generated as part of this research, separating between the top 10k, the middle 10k, and the bottom 10k websites. In total, we failed to retrieve the AP for 715 of the 8,915 scraped websites in the top 10k, 447 of the middle 9,283, and 263 of the bottom 7,325.

C. Data Visualization

Data visualization is the final module of the WATER framework. This module takes in a single csv file and produces graphs for various comparisons. It is currently tailored to the research conducted as part of this paper and is presented as a Jupyter notebook [4].

As this experiment was run three times with websites of differing popularity levels, the module focuses on graphs that compare these three executions against one another. Whenever data involving AP is considered, data points where the AP was not successfully retrieved are excluded from the graph. Likewise, in cases where metric scores are being considered but the website did not have the required tags to calculate a score for the metric, those data points are also excluded from the constructed graphs. The graphs generated from the data used for this experiment are presented in Section V of this paper.

D. Ethical Considerations

When it comes to raw data, WATER pulls only the front-facing HTML served by a website to a typical browser-based user, alongside its publicly available AP. As both sources of data are public and not privacy sensitive, the data itself is not considered of ethical significance. This also allows the data visualization module to be excluded from ethical considerations, as it only visualized the public data collected by the HTML scraper and the metric calculation modules.

However, both the HTML scraper and metric calculation modules must be considered. Both modules make use of active measurement techniques to retrieve data that could impact computational resources. In the case of the HTML scraper module, WATER visits each target website a single time and scrapes the HTML that is displayed. This is unlikely to have any impact on the availability of any specific website, thus it is unlikely that the HTML scraper module performs any operation that can be deemed of ethical significance.

As for the metric calculation module, there is a risk of overwhelming the WebAccessibility online tool used to determine the AP for a given website.[3] Regardless, to combat the potential impact during the runs performed for this paper, the multiprocessing capabilities of WATER are disabled. Instead, each check is performed sequentially and 30 seconds apart. This process resulted in roughly 7 days of execution time to fully gather the data used in this research, but it was important not to overwhelm the service and potentially impact its operation. At the time of writing, no notice has been received from Level Access regarding the use of its service and it is deemed unlikely their service was impacted as part of this research.

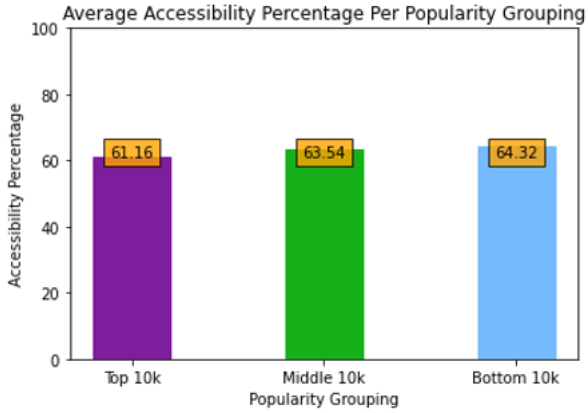

Fig. 2. Average Accessibility Percentage Per Popularity Grouping. The numbers in the orange boxes on top of each bar are the exact averages for each bar

V. Results

A. AP Averages Per Popularity Grouping

To determine whether the threat landscapes changed alongside the popularity of a website, the average AP for each of the three runs was compared. In total, 24,019 of our 25,523 successfully scraped websites had their AP successfully returned. The remaining 1,504 are excluded from the following results. Figure 2 displays the averages for each of the three runs conducted. Interestingly, accessibility does not appear to be significantly different between the top 10K websites and the bottom 10K websites, sitting at a roughly 62% average across the board.

B. Metric Averages Per Popularity Grouping

In order to analyze the most prominent threats to users that require accessibility tools to access the web, the averages for each of the three measured metrics were examined. These were once again separated by popularity grouping to determine if the threat landscape was based on the popularity range of given websites. Figure 3 displays the average ITAA for each of the three runs conducted. Of the 25,523 websites that were successfully scraped by WATER, 22,492 had data that allowed for the calculation of ITAA (7,665 from the top 10k, 8,231 from the middle 10k, and 6,596 from the bottom 10k). Here, it can be observed that the top 10k sites, on average, have a higher proportion of images with appropriate

[3]We emailed Level Access regarding this research, but haven't gotten a response as of this writing.

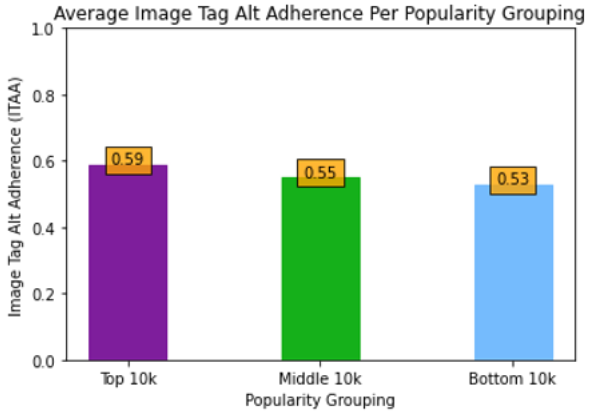

Fig. 3. Average ITAA Per Popularity Grouping. 1.0 is the best possible score while 0.0 is the worst possible score

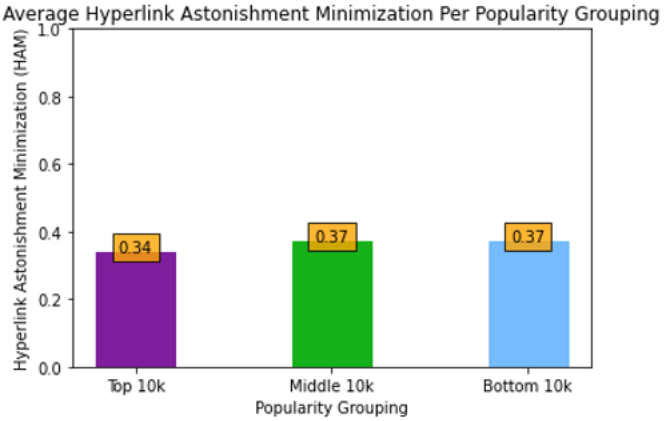

Fig. 4. Average HAM Per Popularity Grouping. 1.0 is the best possible score while 0.0 is the worst possible score

alt attributes compared to the lower websites on the popularity list. Importantly this is not large, with the largest difference being only 6%.

Figure 4 displays the average HAM for each of the three runs conducted. Of the 25,523 websites that were successfully scraped by WATER, 23,099 had data that allowed for the calculation of HAM (7,897 from the top 10k, 8,447 from the middle 10k, and 6,755 from the bottom 10k). As can be seen, the astonishment percentage for hyperlinks is less than 50% across the board, demonstrating that this metric appears to be likely unsatisfied. Interestingly, the top 10k sites are 3% worse compared to the middle 10k and bottom 10k sites. While we cannot be sure why this result occurred, it should be noted that on average the top 10k sites had 229 hyperlinks on average, which is nearly double the average number of hyperlinks in the middle 10k (139) and the bottom 10k sites (129). These extra hyperlinks may be the reason for a lower overall HAM score average for the top 10k sites, but further research would be required to confirm this. Figure 5 displays the average LIM for each of the three runs conducted. Of the 25,523 websites that were successfully scraped by WATER, 18,222 had data that allowed for the calculation of HAM (6,525 from the top 10k, 6,595 from the middle 10k, and 5,102 from the bottom 10k). Clearly, very few sites are correctly implementing label

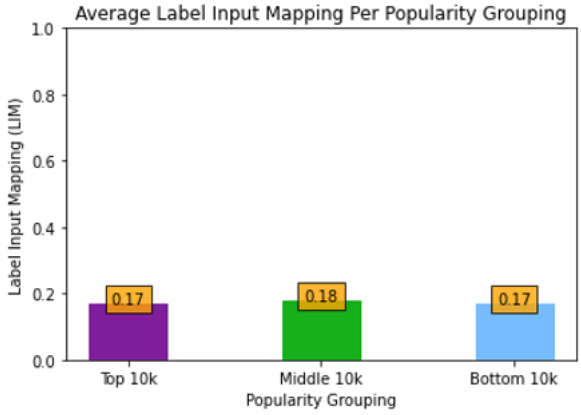

Fig. 5. Average LIM Per Popularity Grouping. 1.0 is the best possible score while 0.0 is the worst possible score

and input mapping for all `<input>` tags, with less than 20% satisfaction across the board. This appears to be the biggest area for improvement from the metrics being measured.

C. Metric Score And AP Correlation

To determine if base metrics could be used to successfully predict the accessibility of a given website, the combined metric score of the three base metrics being investigated was plotted against the AP of a given website, once again separated by website popularity ranges. Figure 6 displays the total metric score versus AP for each of the three data runs. From the graphs, it is apparent there is a deeper concentration of data points toward their relative centers.

To further look for a possible correlation between the accessibility of a website and base metrics, regardless of website popularity, the data from the three previous runs were combined. Figure 7 combines the graphs seen in Figure 6. Unsurprisingly, no trends appear following this data combination. Once again, the data is concentrated toward the center of the graph. The most common data points appear to be websites with a total metric score between 1.0 and 1.5, and their respective accessibility scores are most likely to be either around 40% or 60%. Again, it should be noted that there are websites that scored 100% AP but did not achieve a total metric score of 3.0. This highlights the importance of the metrics we create herein, especially in assessing accessibility properties related to security and privacy vulnerabilities. This raises an alarm for the current WCAG 2.1 standards, which we discuss further in Section VI.

D. All Websites with a Combined 3.0 Metric Score

Out of the 25,523 websites that were successfully scraped by WATER, 17,818 had enough data to calculate all 3 metrics and successfully returned an AP (6,358 from the top 10k, 6,452 from the middle 10k, and 5,008 from the bottom 10k). Of these 17,818 sites, only 8 achieved a 3.0 combined total metric score, as seen in Figure 8. Surprisingly, not a single website from the bottom 10k managed to achieve a combined metric score of 3.0. Even more interestingly, the average AP is higher for websites from the middle 10k compared to websites in the top 10k. Note that while we have previously established that a

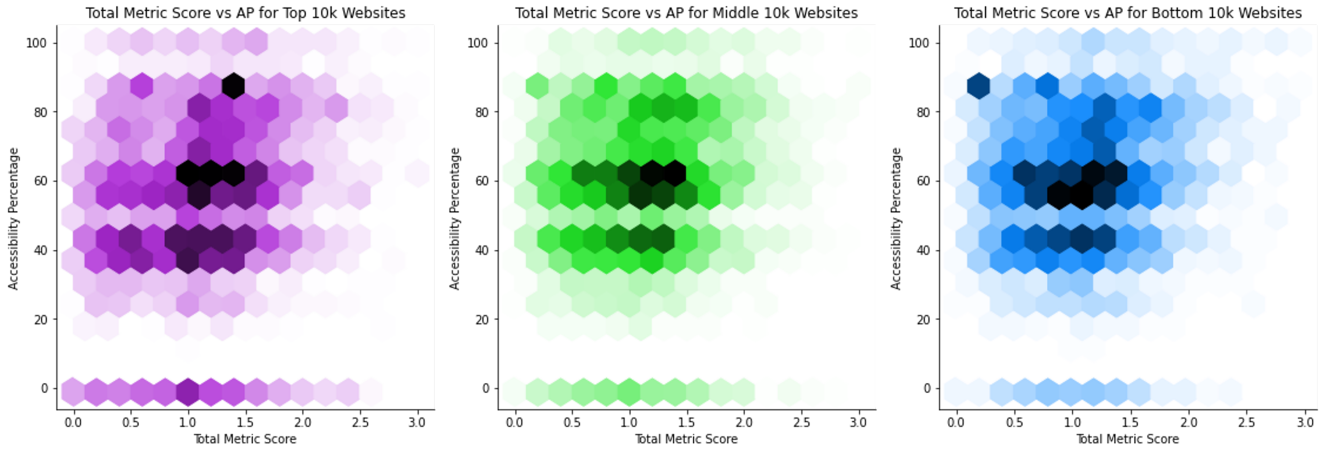

Fig. 6. Total Metric Score vs AP separated by website popularity range. The deeper the colour, the more data points appear within that specific hexagon. 3.0 represents the best possible combined metric score, while 0.0 represents the worst

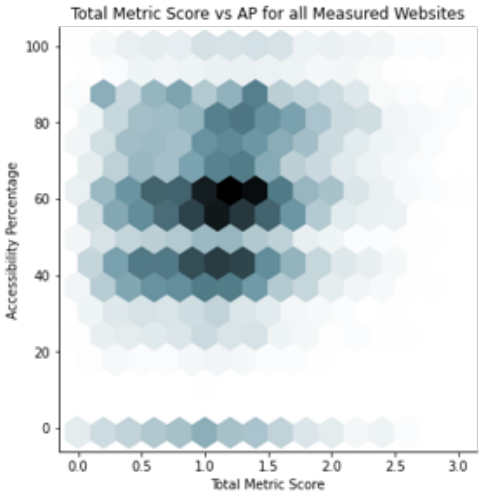

Fig. 7. Total Metric Score vs AP. The deeper the colour, the more data points appear within that specific hexagon. 3.0 represents the best possible combined metric score, while 0.0 represents the worst

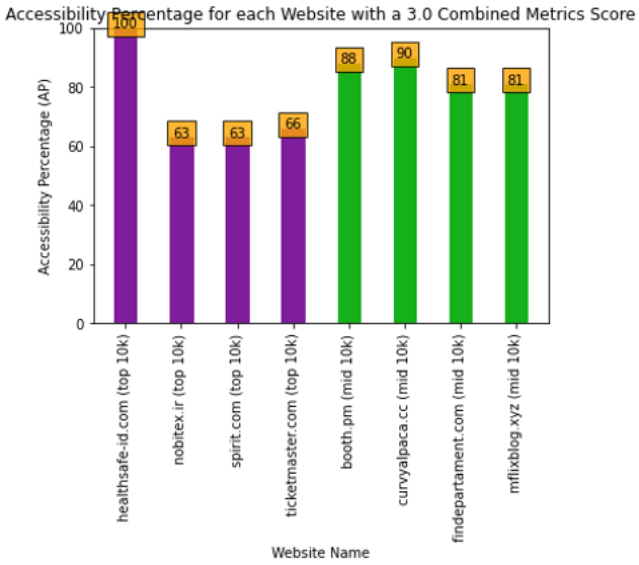

Fig. 8. Accessibility Percentage for each Website with a 3.0 Combined Metrics Score. 100% implies Level AA compliance with WCAG 2.1 standards

website with a 100% AP does not necessarily satisfy 3.0 in our metrics, the results in Figure 8 demonstrate that the opposite is also true—having a 3.0 in our metrics does not imply 100% AP.

E. Overall Analysis of AP for all Analyzed Websites

Lastly, an analysis of the AP for all websites measured in this study is presented. As previously mentioned, 24,019 of our 25,523 websites had their AP returned by the WebAccessbility tool provided by Level Access and only those 24,019 websites are considered in the following analysis. As can be seen in Figure 9, roughly two-thirds of all websites have an AP $> 50\%$, and only 25.9% have an AP $> 75\%$. This is concerning, as an AP of 75% or less might be grounds for accessibility lawsuits [2]. This could also be indicative of other factors that threaten the security of users requiring accessibility tools, beyond the three metrics constructed herein. While accessibility violations do not necessarily imply security violations, these results should prompt further investigation into potential additional base metrics that can be linked back to users' security in order to determine the full extent of the threat landscape for users requiring accessibility tools to access the web.

VI. Discussion

Following up on the above results, we now discuss the threat landscape for users of accessibility tools. We also shed some light on the limitations of the WATER framework. Finally, we entertain research avenues to extend this, which can make use of the WATER framework.

A. Current Threat Landscape

The results from this work paint a concerning picture of the current threat landscape plaguing users that require accessibility tools to access the web. In particular, the results for both HAM and LIM imply that users that make use of

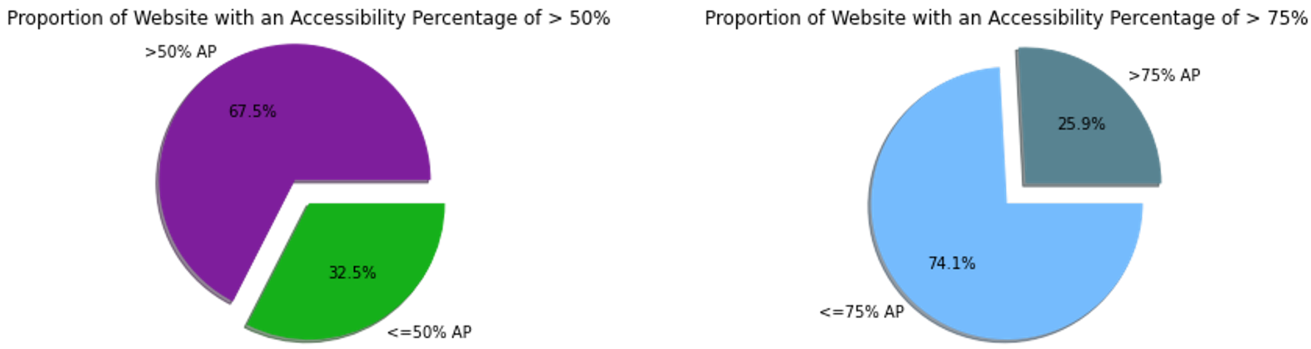

Fig. 9. Left: Proportion of Website with an Accessibility Percentage of $> 50\%$. Right: Proportion of Website with an Accessibility Percentage of $> 75\%$

screen readers or keyboard navigation could be at risk for phishing attacks and potential information leaks. This appears to be the case for popular and unpopular websites alike.

It is important to note that failure to satisfy the metrics presented herein does not necessarily imply a security threat. However, as discussed in Section III, there exist scenarios where failing to satisfy these metrics can result in attacks. The data suggests that there is a large proportion of websites where these attacks could happen, and that in itself is cause for concern.

As made evident by the results, the popularity of a website appears disconnected from its overall accessibility. Moreover, the base metrics calculated did not vary much between popularity ranges, with the largest variation being for ITAA (6% difference between the top 10k and the bottom 10k). Such consistency suggests that users that require accessibility tools have a wide-spanning threat landscape. And even the most popular sites do undermine the privacy and security of users of accessibility tools. In fact, the bottom 10k sites at times produced better results than the top 10k sites, such as the case with the average HAM scores in Figure 4. While WATER does not currently collect data beyond raw HTML, we discuss possible expansions to the framework that could provide insight into this result in Section VI-C.

A secondary goal of this research was to determine if our base metrics could be used to assess the accessibility of a website. As made evident by the lack of linear trends between the total metric scores and the AP of websites, the data suggests that our base metrics should not be used to determine the overall website accessibility. This is not entirely surprising, as the WCAG 2.1 standards are an exhaustive list of guidelines [1], [13], whereas our base metrics are focused on security vulnerabilities arising specifically from the lack of proper implementation of accessibility-related parameters.

A final important highlight of this research deals with the fact that websites can achieve an AP of 100% without satisfying the three base metrics targeted in this research, as seen in Figure 6. In fact, some websites achieved an AP of 100% while having a combined score of less than 1.0 (out of 3.0) in our base metrics. As mentioned in Section III, certain checks for the metrics are more strict compared to the official guidelines. This practice was adopted in order to accurately measure the potential security consequences of these metrics failing to be satisfied by a website. The fact that websites can successfully achieve an AP of 100% suggests that security may not be sufficiently considered as part of the current WCAG 2.1 standards. It may be beneficial to revise the criteria for achieving Level AA conformance to include checks that ensure the security of users is taken into account, and measures are taken to minimize the threats they may face while navigating the web.

B. Limitations

While our list of analyzed websites could have been larger, the adopted sample does highlight certain trends. A more complete analysis can be facilitated by the WATER framework, which has been designed to accommodate such work. The reliance on the WebAccessibility tool provided by Level Access to determine the AP of a website is an inherent limitation, as the tool is closed-source. It is difficult to determine the accuracy of the tool with regard to its analysis. Upon exploring the viability of other tools, we found that the WebAccessibility tool provided by Level Access was the only service accessible to us that made use of the latest WCAG 2.1 standards. We acknowledge that our detection methods are simplistic and can be easily evaded by bad actors. For example, a malicious website operator may modify a hyperlink to include the text contained within a `<a>` tag while still redirecting the user to a phishing link. Similar techniques can be used to evade ITAA and LIM violations. However, we ultimately chose to proceed with this simplistic detection method to demonstrate that even if bad actors are employing these techniques, detection for these metrics remains significant and users remain at risk with not only possible phishing attacks (in the case of HAM) but also *least-surprise* violations. The use of simplistic checks more readily ensures that the other source of threats (innocent but negligent site owners) were caught by the execution of WATER. Regardless, WATER's metric calculation module will continue to evolve as more insight into the evasion tactics employed by bad actors is discovered.

Lastly, this research dealt with only a small subset of base metrics. The entire catalogue of disabilities is expansive and difficult to fully measure. We focused on analyzing the effect of improper accessibility configurations on users that require screen readers and comparable means of web page navigation. Our results are thus relevant specifically to these tools.

C. Future Work

The WATER framework was designed to be easily adaptable for future measurements due in part to its modular design and its open-source availability. As such, there are numerous avenues for future work that could make use of the created WATER framework. One such avenue would be to run the same experiment with a larger data set. While we only analyzed landing pages, calculating the metric scores for all pages of a website could highlight different results from our initial investigation into the topic. To properly draw conclusions about web trends, it is can be valuable to measure as much of the web as possible.

Another potential avenue of research includes the expansion of WATER to check for additional base metrics. Many additional metrics can be devised and incorporated into WATER to measure web accessibility given the raw HTML data that WATER currently collects. Certain non-security-related metrics that could be added to WATER, such as text comprehensibility, are discussed in [1]. While these will not be strictly security related, it could be of interest to compare the results between security-related accessibility metrics and non-security-related ones.

It is also possible to expand WATER to collect additional data from sites, such as included JavaScript files and their contents, the CSS/XSS files, among other data sources. This additional data would provide a more complete picture of factors that could impact users with disabilities when visiting a website. For example, users with limited motor functions may benefit from a metric to determine if popups provide adequate timeouts prior to auto-accepting or closing. This could provide valuable insight into the potential privacy concerns that occur as a result of this auto-acceptance or premature closing, especially if these popups deal with the acceptance of a privacy policy or a site's terms and conditions. These additional metrics could help inform a clearer picture of the threats facing users with disabilities and the different measures that can be employed to protect them.

A fourth point that could be explored is sustained measurements over a prolonged period of time. This would require running WATER constantly over a measurement period to analyze how metric scores and AP varies over time. This could provide insight into the work some websites may be making towards improving their overall accessibility to conform with WCAG 2.1 standards.

VII. Related Work

Surprisingly, there is little literature with studies that actually measure web accessibility. Wille et al. [23] presented the notion of a measurement such as AP back in 2016 when they performed one of the first Internet-wide measurement studies based on the most recent WCAG standards at the time, WCAG 2.0. Beyond this, Johari et al. [12] attempted a questionnaire-based measurement approach to understand the impact that a lack of accessibility meant for Persons with Disabilities (PWDs). Neither of these studies focuses on the threat landscape that arises due to a lack of web accessibility.

Some additional user-focused studies have been conducted that measure the security implications of accessing the Internet for users that are visually impaired. Lau et al. [14] put forth and executed a suggested research methodology that can be applied to studies involving users that are visually impaired. Abdolrahmani et al. [5] investigated how users with visual disabilities determine the credibility of sites and the information contained within to determine effective ways to communicate credible information for users which can be extended to best security practices. In 2021, Napoli et al. [17] observed the behaviours of users with visual disabilities and discovered significant usability issues that prevented them from being able to identify risks.

While few studies focused on measuring the extent of web accessibility, there exists numerous papers that evaluate existing tools and metrics that measure web accessibility. Vigo et al. [20] highlighted the lack of quantitative accessibility metrics and presented three use cases where their existence would be vital, including QA, web accessibility monitoring, and information retrieval. Freire et al. [9] collected many web accessibility metrics as part of a literature review to demonstrate there were still gaps in the field of quantitative metrics as previously highlighted by Vigo et al. [20]. Vigo et al. [21] would later evaluate developments in the field of web accessibility evaluation tools, where the problem with reliance on a single testing service is highlighted prominently. More recently, Alsaeedi [7] provided a comparison between two web accessibility tools known as Wave and SiteImprove to determine their effectiveness as frameworks for site owners to improve the accessibility of their websites based on the WCAG 2.0 standards. Ultimately, the field is still evolving but sufficient adoption of security as part of evaluation remains to be seen.

There exist studies that focus on establishing the connection between privacy/security concerns with accessibility. Wang et al. [22] established that people with disabilities face additional challenges when it comes to managing their privacy and there is a need to define new guidelines that can be used to design accessible tools for these users to reduce the number of threats they face. This result was previously observed by Ahmed et al. [6], although the focus was on visually impaired individuals and revealed unique privacy concerns that were not being satisfied for these users by current technology at the time.

To the best of our knowledge, we are the first to estimate the threat landscape for users that require accessibility tools to access the web. We hope that the WATER framework can be reused for future research into this important but understudied topic.

VIII. Conclusion

In conclusion, the threat landscape for users that require accessibility tools to navigate the web is of great concern. Using the created WATER framework, we found that users that use screen readers and alternative means of website navigation could be at risk for targeted phishing attacks and potential information leakages for more than 60% of the analyzed websites. The overall accessibility of websites is trailing behind recommended standards, with over 15,500 websites found to have an AP of less than 75%. We determined that the scores of the base metrics could not be used to estimate

the AP of a website. However, our analysis suggests that the current WCAG 2.1 standards may need to be revised to disallow websites that expose users of accessibility tools to security threats from achieving Level AA conformance. We hope that the WATER framework and data presented in this paper serve as a starting point for future research exploring the accessibility–security relationship.

REFERENCES

[1] "Benchmarking web accessibility metrics." [Online]. Available: https://www.w3.org/WAI/RD/wiki/Benchmarking_Web_Accessibility_Metrics

[2] "Ada & wcag compliance (free scan)," Oct 2022. [Online]. Available: https://www.accessibilitychecker.org/

[3] "Test your site for accessibility," Aug 2022. [Online]. Available: https://www.webaccessibility.com/

[4] "Project jupyter," Jan 2023. [Online]. Available: https://jupyter.org/

[5] A. Abdolrahmani and R. Kuber, "Should i trust it when i cannot see it? credibility assessment for blind web users," in *Proceedings of the 18th international acm sigaccess conference on computers and accessibility*, 2016, pp. 191–199.

[6] T. Ahmed, R. Hoyle, K. Connelly, D. Crandall, and A. Kapadia, "Privacy concerns and behaviors of people with visual impairments," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3523–3532.

[7] A. Alsaeedi, "Comparing web accessibility evaluation tools and evaluating the accessibility of webpages: proposed frameworks," *Information*, vol. 11, no. 1, p. 40, 2020.

[8] M. Contributors, "Html: A good basis for accessibility - learn web development: Mdn," Nov 2022. [Online]. Available: https://developer.mozilla.org/en-US/docs/Learn/Accessibility/HTML#good_semantics

[9] A. P. Freire, R. P. Fortes, M. A. Turine, and D. M. Paiva, "An evaluation of web accessibility metrics based on their attributes," in *ACM conference on Design of Communication*, 2008.

[10] S. Hackett and B. Parmanto, "Homepage not enough when evaluating web site accessibility," *Internet Research*, 2009.

[11] R. Ismailova, "Web site accessibility, usability and security: a survey of government web sites in kyrgyz republic," *Universal Access in the Information Society*, vol. 16, no. 1, pp. 257–264, 2017.

[12] K. Johari and A. Kaur, "Measuring web accessibility for persons with disabilities," in *IEEE International Conference on Computational Intelligence and Communication Networks (CICN)*, 2012.

[13] A. Kirkpatrick, J. O Connor, A. Campbell, and M. Cooper, "Web Content Accessibility Guidelines (WCAG) 2.1," Jun. 2018, edTechHub.ItemAlsoKnownAs: 2405685:8TX6IQZM. [Online]. Available: https://www.w3.org/TR/WCAG21/

[14] E. Lau and Z. Peterson, "A research framework and initial study of browser security for the visually impaired," in *Symposium on Usable Privacy and Security (SOUPS)*, 2015.

[15] E. T. Loiacono and S. Djamasbi, "Corporate website accessibility: does legislation matter?" *Universal access in the information society*, vol. 12, no. 1, pp. 115–124, 2013.

[16] Y. Lu and L. Da Xu, "Internet of Things (IoT) cybersecurity research: A review of current research topics," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2103–2115, 2018.

[17] D. Napoli, K. Baig, S. Maqsood, and S. Chiasson, ""i'm literally just hoping this will work:" obstacles blocking the online security and privacy of users with visual disabilities," in *Symposium on Usable Privacy and Security (SOUPS)*, 2021.

[18] K. Renaud and L. Coles-Kemp, "Accessible and inclusive cyber security: a nuanced and complex challenge," *Springer SN Computer Science*, vol. 3, no. 5, pp. 1–14, 2022.

[19] P. C. Van Oorschot, *Computer Security and the Internet: Tools and Jewels from Malware to Bitcoin*. Springer, 2021.

[20] M. Vigo, M. Arrue, G. Brajnik, R. Lomuscio, and J. Abascal, "Quantitative metrics for measuring web accessibility," in *International cross-Disciplinary conference on Web accessibility (W4A)*, 2007.

[21] M. Vigo, J. Brown, and V. Conway, "Benchmarking web accessibility evaluation tools: measuring the harm of sole reliance on automated tests," in *International Cross-Disciplinary Conference on Web Accessibility (W4A)*, 2013.

[22] Y. Wang and C. E. Price, "Accessible privacy," in *Modern Socio-Technical Perspectives on Privacy*. Springer, Cham, 2022, pp. 293–313.

[23] K. Wille, R. R. Dumke, and C. Wille, "Measuring the accessability based on web content accessibility guidelines," in *IEEE Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA)*, 2016.