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

John Breton
Department of Systems and Computer Engineering
Carleton University
Ottawa, Canada
john.breton@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 Internet, the WATER framework was created. WATER measures the Internet using three security-related base accessibility metrics. Using 30,000 websites from three distinct popularity ranges, WATER was able to collect data that demonstrated that these users face an increased risk for phishing attacks and opportunities for information leakage; regardless of website popularity. Further, over 15,000 of the analyzed websites had an accessibility percentage of less than 75%, a statistic that opens these websites to potential accessibility-related lawsuits. Lastly, although the metrics calculated by WATER could not be directly related to the overall accessibility of a website, the collected data suggests that the current WCAG 2.1 standards may need to be revised to avoid assigning Level AA conformance to websites that include the potential for threats to users that require accessibility tools to access the Internet. The WATER framework is made available in the hopes it can be used for future research.**

***Index Terms*—cybersecurity, Internet measurement, accessibility metrics, user security, HTML, Python, WCAG 2.1, W3C**

I. Introduction

As the Internet continues to expand, individuals that require assistance to access the Internet are increasingly suffering due to a lack of available accessibility options. Despite government legislation detailing strict accessibility requirements for websites, the globalization of the Internet ensures that these recommendations are only applicable to a subset of websites currently hosted on the Internet [1].

This rapid expansion has also given rise to various attacks perpetrated by bad actors. Some of the most egregious of these are attacks that prey on individuals that suffer from disabilities; often succeeding due to the lack of consistent accessibility options presented throughout the Internet [2]. These attacks normally target the privacy of the users involved, failing to uphold the confidentiality pillar of the CIA triad. As this is a relatively small subset of the global set of Internet users, little research exists that measures the security risk imposed on users due to websites failing to uphold web accessibility standards [3]. Regardless of the fact that users with disability make up a smaller portion of users on the Internet, 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 a subset of users that require accessibility tools to access the Internet. As disabilities exist in wide varieties, it is impossible to apply the results of such research to all users that require accessibility tools to access the Internet. Instead, this research focuses on users using screen readers and alternative means of web page navigation to access the Internet.

Through active measurements, this paper aims to determine the threat landscape for users that require accessibility tools to access the Internet. This landscape depends on measuring the overall accessibility of websites using basic accessibility metrics that can be linked to attacks bad actors could perform if websites are not conforming to the Web Content Accessibility Guidelines (WCAG) 2.1 standards. Further, it is hypothesized that the popularity of websites could have an impact on their overall accessibility, and as such users may be at more risk if they browse sites within a certain popularity range. This paper wishes to provide clarity on this topic as well. Finally, 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 [4]. This paper aims to provide a data-backed answer to that open research question, as its determination could suggest future research that aims to improve the overall accessibility of the Internet, which in turn could help to reduce the threat landscape for users that require accessibility tools to access the Internet.

To facilitate the questions being posed, a measurement framework known as WATER has been developed. This framework was used to analyze 30k unique domains from the Alexa top 1M list published on November 11th, 2022. In total, three runs were performed using WATER, analyzing the top 10k, the middle 10k, and the bottom 10k sites listed in the target Alexa top 1M list. The data collected suggests that the popularity of websites has no perceivable correlation to the accessibility of a site. Further, while the base metrics used do not appear to be strongly correlated to the overall accessibility of a website, they paint a disturbing picture for users that require accessibility tools to access the Internet. It appears the threat landscape for these users is very large, with over 65% of the sites investigated having the potential to instigate targeted attacks, such as phishing attacks, against these users, and over 80% having the potential to cause information leakage.

This paper presents the following contributions:

1) The formalization of three basic accessibility metrics

that can be directly related to the ability of a domain to prevent specific targeted attacks against users that require accessibility tools to access the Internet
2) A framework known as WATER, that can be used to perform future measurements to assess the changing conformance to the three base metrics alongside the accessibility percentage of websites across the Internet[1]
3) Data to suggest that basic accessibility metrics are not enough to determine the accessibility of a website, as well as presenting the threat landscape for users that require accessibility tools to access the Internet to be as large as >80% of the 30,000 domains analyzed, regardless of website popularity

The rest of this paper is structured as follows. Section II discusses how web accessibility relates to security and provides an overview of the latest web accessibility guidelines. Section III defines the three base metrics being measured. Next, 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 and restates the main takeaways from the analysis.

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 [5]–[7]. Of course, privacy falls under maintaining the confidentiality of data and is inherently a security issue, however, the field establishing this link is still evolving [8]. It is possible that the link between web accessibility and security either strengthens or weakens when new research is performed, but for the purposes of this study, it is assumed that there is a definitive connection between web accessibility and security.

B. WCAG 2.1 and Accessibility Percentage

The Web Content Accessibility Guidelines are an exhaustive list of accessibility checks detailed by the W3C [9]. The list itself is divided into four major categories that aim to ensure the Internet 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 [10].

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, three base accessibility metrics are presented that could pose threats to users that access the Internet with accessibility tools such as screen readers and keyboard navigation if they are not satisfied. These metrics make use of conditions that are already used in the calculation of the overall AP of a website, however, they are being recontextualized with a security-focused perspective.

A. Image Tag Alt Adherence (ITAA)

Image Tag Alt Adherence is 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 the Internet with screen readers [2]. 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.

To calculate this metric, the entirety of a web page's <img> tags are collected. Next, a simple check is performed to determine if the following conditions are satisfied:

1) Does the <img> 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, <img> 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

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

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 [9], [11].

The final check is the most subjective. Again, the current discussion suggests avoiding the use of terms such as "image of" and "graphic of" [11]. 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 a simple string check to see if the alt attribute text contains the problem phrases mentioned previously. If this, along with the other two conditions, are all 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:

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

B. Hyperlink Astonishment Minimization (HAM)

Hyperlink Astonishment Minimization is 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 astonishment 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 [8]. This principle is potentially the inspiration for 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 happen upon a hyperlink. The screen reader may read out the text associated with the hyperlink rather than the URL a user will be redirected to in many instances. Naturally, this can allow attackers to redirect these users to malicious sites, as the user is simply hearing the text associated with the hyperlink. 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 to protect users from potential attacks.

To calculate this metric, the entirety of a web page's <a> tags with href attributes are collected. Next, a simple check is performed to determine if the following condition is true:

- Does the text in the <a> tag with 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 that is not login related. Once all hyperlinks have been analyzed, the final HAM score is calculated using the following equation:

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

C. Label Input Mapping (LIM)

Label Input Mapping is a base metric that measures the proportion of <input> tags that an 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 [9], 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. The user has now unintentionally shared private information with users that should not have been made aware of said information.

To calculate this metric, the entirety of a web page's <input> and <label> tags are collected. Next, a simple check is performed to determine if the following condition is true:

- Does id attribute of an <label> tag match the id attribute 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. 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 Internet. A Python framework known as WATER was developed to accomplish this task. WATER can be thought of as three decoupled modules that can run as one cohesive unit. Figure 1 outlines the three modules. In a traditional full execution run, HTML scraping is run first, followed by metric calculations, and ending off with data visualization.

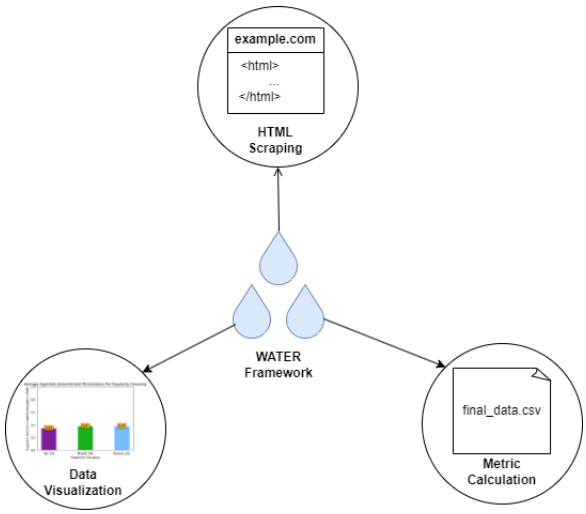

Fig. 1. Overview of the WATER Framework

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 will a csv file containing the domains that are being measured. For each domain in the csv file, 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 for domains were considered. The rationale behind this decision is straightforward. Should a website's homepage be inaccessible, it is unlikely a user that requires accessibility tools to access the Internet would continue to future pages on a website. This makes landing pages a good target for analysis in our research, while allowing for the measurement of additional sub-pages as an area of future work.

Additionally, 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 the target HTML tags needed for metrics analysis in order to minimize its local memory footprint. These tags are all <img> tags, all <a> 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.

As part of this research, the HTML scraping module was run three separate times with three unique csv files. These files contained the top 10k, the middle 10k, and the bottom 10k domains listed in the November 11th, 2022 Alexa Top 1M Sites list in terms of popularity. 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, 8,915 of the top 10k websites were successfully scraped, 9,283 of the middle 10k websites were successfully scraped, and 7,325 of the bottom 10k websites were successfully scraped.

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, it has also been given the responsibility of retrieving 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 once again made using a headless selenium browser, as was the case in the HTML Scraping module. 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 for 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 terms of popularity. In total, 715 of the 8,915 scraped top 10k websites failed to have their

AP retrieved, 447 of the 9,283 scraped middle 10k websites failed to have their AP retrieved, and 263 of the 7,325 scraped bottom 10k websites failed to have their AP retrieved.

C. Data Visualization

Data visualization is meant to act as the final module as part of the WATER framework. This module will take in a single csv file and produce graphs for various comparisons. It is currently tailored to the research conducted as part of this paper and is presented as a Jupyter notebook. Its existence is to promote replicability and reproducibility of the experiment conducted in this paper.

As this experiment was run three times with websites of differing popularity levels, the module focuses on graphs that compare these three execution runs 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 will be presented in Section V of this paper.

D. Ethical Considerations

As with any Internet measurement study, ethics need to be considered. 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 Internet 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 real risk of overwhelming the WebAccessibility online tool used to determine the AP for a given website. Level Access was contacted by email regarding this research but they have yet to respond. 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.

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. 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.

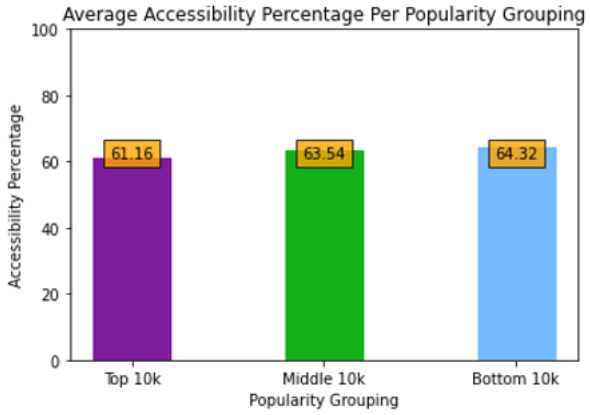

Fig. 2. Average Accessibility Percentage Per Popularity Grouping. The numbers in the orange boxes on top of each bar are the exact averages for each bar

B. Metric Averages Per Popularity Grouping

In order to analyze the most prominent threats to users that require accessibility tools to access the Internet, 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

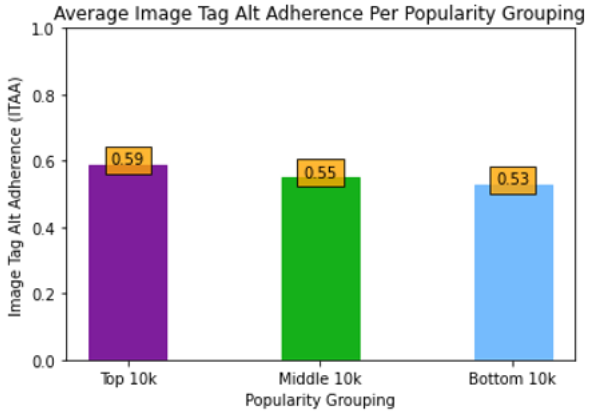

Fig. 3. Average ITAA Per Popularity Grouping. 1.0 is the best possible score while 0.0 is the worst possible score

of the three runs conducted. Here, it can be observed that the top 10k sites, on average, have a higher proportion of images with appropriate alt attributes compared to the lower websites on the popularity list. Importantly this is not large, with the largest difference being only 6%.

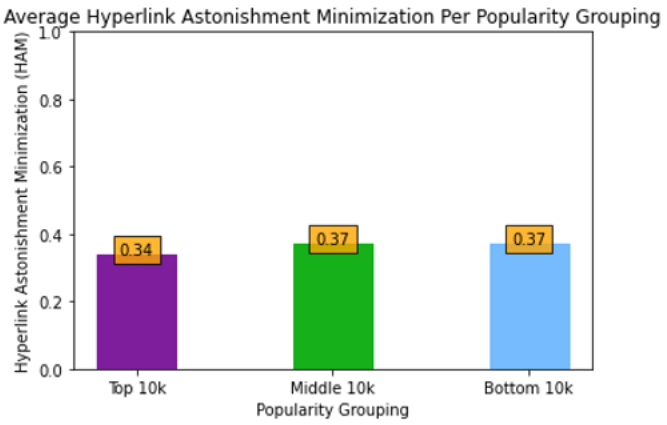

Fig. 4. Average HAM Per Popularity Grouping. 1.0 is the best possible score while 0.0 is the worst possible score

Figure 4 displays the average HAM for each of the three runs conducted. As can be seen, the astonishment percentage for hyperlinks is less than 50% across the board, demonstrating that this metric appears to be more than likely unsatisfied than satisfied. Interestingly, the top 10k sites are 3% worse compared to the middle 10k and bottom 10k sites.

Figure 6 displays the average LIM for each of the three runs conducted. Clearly, very few sites are correctly implementing label and input mapping for all <input> tags, with less than 20% satisfaction across the board. This seems 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 5 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. Additionally, it appears no websites were registered as having an AP of 10%.

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 5. 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 would seem to imply no correlation between the base metrics being calculated as part of this research and the AP of a website, which raises alarms for the current WCAG 2.1 standards that will be discussed further in Section VI.

D. All Websites with a Combined 3.0 Metric Score

Out of all of the websites analyzed, only 8 had the data necessary to calculate scores for all three metrics being investigated and achieved a 3.0 combined total metric score, as seen in Figure 9. Surprisingly, not a single website from the bottom 10k managed to achieve a combined metric score of 3.0. Even more interestingly, the average accessibility percentage is higher for websites from the middle 10k compared to websites in the top 10k. Again, the fact that all of these are not 100% should make it demonstrably clear that these metrics are likely not correlated with the accessibility percentage of a website.

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 can be seen in Figure 8, roughly two-thirds of all websites have an AP that is greater than 50%. However, if we increase the threshold to 75%, only 25.9% of websites have an AP of greater than 75%. This is concerning, as an AP of 75% has been shown to be grounds for accessibility lawsuits [12]. This could also be indicative of other factors that threaten the security of users that require accessibility tools to access the Internet beyond the three base metrics observed in this research. 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 user security in order to determine the full extent of the threat landscape for users that require accessibility tools to access the Internet.

VI. Discussion

In this section, the threat landscape for users that require accessibility tools to access the Internet, as revealed by the results of this paper, is presented. Limitations of the WATER framework are also presented, alongside a discussion of avenues of future work that could make use of the created 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 Internet. In particular, the results for both HAM and LIM imply that users that make use of screen readers or keyboard navigation could be at risk for phishing attacks and potential information leakages, regardless of the popularity range a website falls under.

It is important to note that failure to satisfy the metrics presented in this paper is not indicative of a security threat to users. However, as discussed in Section III, there exist scenarios where failing to satisfy these metrics can result in attacks against users. The data suggests that there is a large

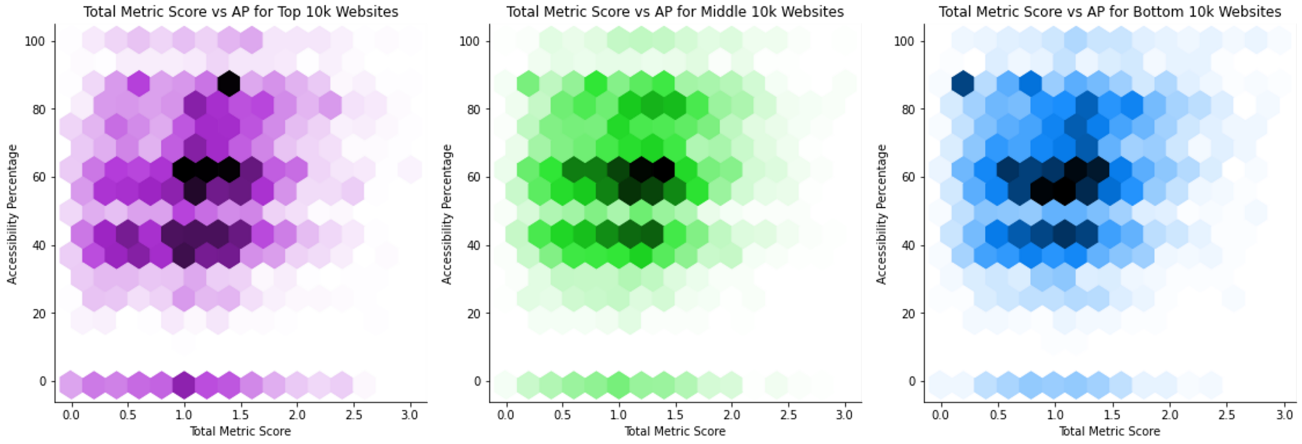

Fig. 5. Total Metric Score vs AP separated by website popularity range. The deeper the colour, the more data points appear within that specific hexagon. 3.0 represents the best possible combined metric score, while 0.0 represents the worst

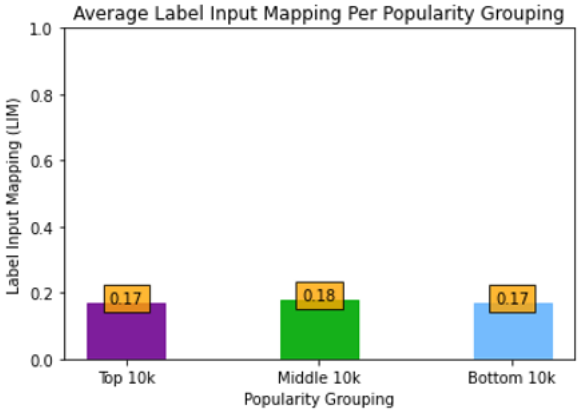

Fig. 6. Average LIM Per Popularity Grouping. 1.0 is the best possible score while 0.0 is the worst possible score

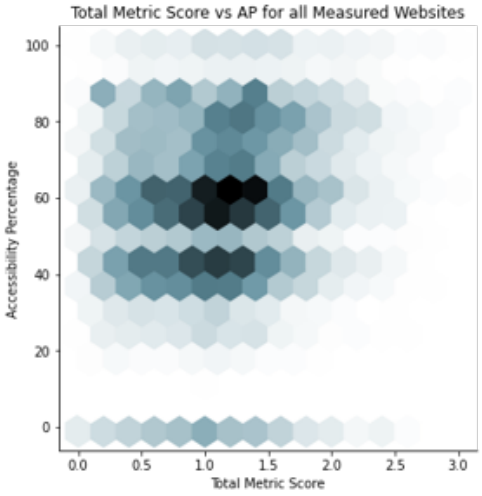

Fig. 7. Total Metric Score vs AP. The deeper the colour, the more data points appear within that specific hexagon. 3.0 represents the best possible combined metric score, while 0.0 represents the worst

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 seems to not be tied to the overall accessibility of websites. Moreover, the base metrics calculated did not vary much between popularity ranges, with the largest variation being for ITAA with a 6% difference between the top 10k websites in terms of popularity and the bottom 10k websites in terms of popularity. This consistency with the data suggests that users that require accessibility tools to access the Internet have a wide-spanning threat landscape, and even the most popular sites cannot be considered excluded from the list of potentially threatening websites for these users.

A secondary goal of this research was to determine if base metrics could be used to determine 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 base metrics cannot be used to determine the accessibility of a website. This is not entirely surprising, as the WCAG 2.1 standards are an exhaustive list of guidelines and it seemed unlikely that cherry-picking a few of these metrics could be enough to correlate with the overall accessibility of the entire website measured against all accessibility checks [4], [9].

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 5. In fact, some websites achieved an AP of 100% with a combined score of less than 1.0. This is likely due to the variations in checks included as part of the WATER framework compared to the formal WCAG 2.1 standards. 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

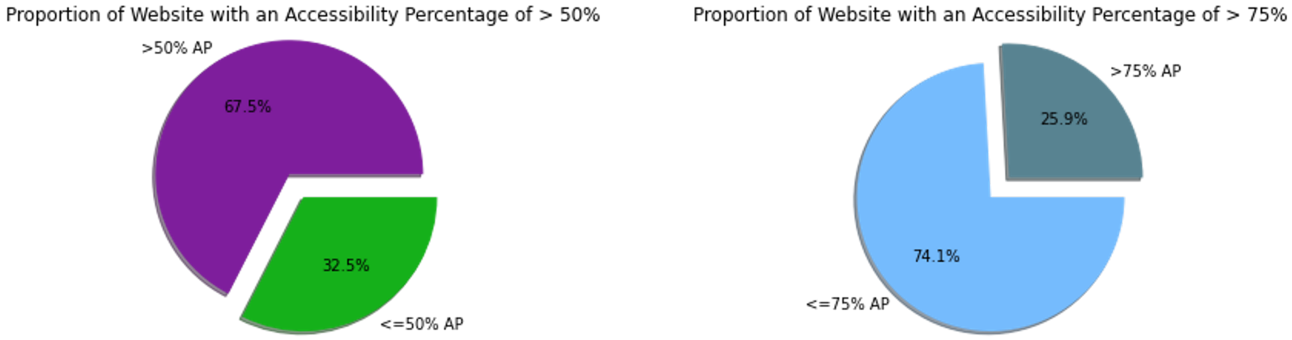

Fig. 8. Left: Proportion of Website with an Accessibility Percentage of ¿ 50%. Right: Proportion of Website with an Accessibility Percentage of ¿ 75%

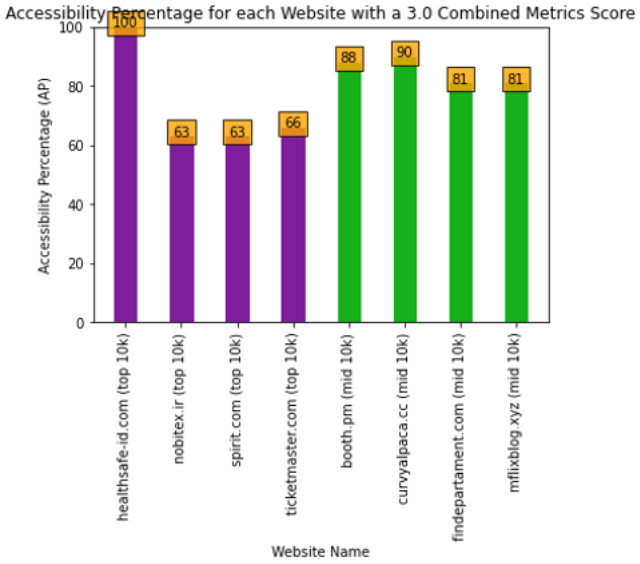

Fig. 9. Accessibility Percentage for each Website with a 3.0 Combined Metrics Score. 100% implies Level AA compliance with WCAG 2.1 standards

AP of 100% demonstrates that security is not being properly considered as part of the current WCAG 2.1 standards and that it may be beneficial to reevaluate the criteria to achieve Level AA conformance to include checks that ensure the security of users are taken into account and measures are taken to minimize the threats they may face while navigating through the Internet.

B. Limitations

The experiment conducted as part of this research is far from free of limitations. To begin, the scope of the measurements spans only a very small portion of the IPv4 space. Due to this, it is impossible to definitively conclude the accessibility of the entire Internet. Yet this small subset adopted for this research does promote certain trends. Regardless, a formal analysis should make an effort to scan the entire IPv4 space, and the WATER framework has been designed to accommodate such work.

Next, 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. While efforts were made to compare the analysis provided by the WebAccessibility tool to other online tools, the analyses were inherently flawed as the WebAccessibility tool provided by Level Access was the only service that made use of the latest WCAG 2.1 standards and provided results for free. As such, the AP results could be skewed, which seems likely given the fact that no website was reported as having an AP of 10%. Unfortunately, as tools that calculate AP are difficult to come by, the compromise to use the WebAccessibility tool was made despite the previously highlighted issues.

Lastly, this research dealt with only a small subset of base metrics. The entire catalogue of disabilities is expansive and difficult to fully measure. This research focused on users that suffer from disabilities that necessitate them to use screen readers and alternative means of web page navigation to access the Internet, but this inherently limits the overall landscape of threats present to this subset of Internet users. As such, the threat landscape provided by the research is likely very conservative in nature. Further, it is possible that additional base metrics could be used to determine the accessibility of a website, however further research would be needed to back up such a claim.

C. Future Work

The WATER framework was designed to be easily adaptable for future measurements. 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. To properly draw conclusions about Internet trends, it is vital to measure as much of the Internet that is possible to be measured. As such, rerunning the same three-metric experiment over the entire IPv4 space to determine if the trends presented in this paper hold would be an interesting research opportunity.

Another potential avenue of research includes the expansion of WATER to check for additional base metrics. There are

numerous additional metrics that exist to measure web accessibility. These are not strictly security related, however, it could be novel to compare results between security-related accessibility metrics and non-security-related accessibility metrics. It may be determined that non-security-related accessibility metrics are more tightly coupled to the AP of a website versus security-related accessibility metrics, which would further promote the need to revise current standards to take security issues into account when determining the accessibility of a website.

A third point that could be explored is sustained measurements over a prolonged period of time. This would require running WATER on a daily basis 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. 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 [13]. Beyond this, Johari et al. attempted a questionnaire-based measurement approach to understand the impact that a lack of accessibility meant for Persons with Disabilities (PWDs) [14]. Neither of these studies focuses on the threat landscape that arises due to a lack of web accessibility.

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. 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 [15]. Freire et al. 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. [15], [16]. Vigo et al. 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 [17]. More recently, Alsaeedi 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 [18]. Ultimately, the field is still evolving but its adoption of security as part of their evaluation remains to be seen.

VIII. Conclusion

In conclusion, the threat landscape for users that require accessibility tools to access the Internet is of great concern. Using the created WATER framework, it was 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 websites on the Internet. The overall accessibility of websites is trailing behind recommended standards, with over 15,500 websites having an AP of less 75%. It also appears that while base metrics are not enough to predict the AP of a website, the data collected suggests that the current WCAG 2.1 standards may need to be revised to prevent websites that fail to properly satisfy the security concerns of users from achieving Level AA conformance. It is hoped that the WATER framework and data presented in this paper serves as a starting point for future research into the field of the link of accessibility and security, alongside web accessibility measurements.

References

[1] 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.

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

[3] 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.

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

[5] 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.

[6] K. Renaud and L. Coles-Kemp, "Accessible and inclusive cyber security: a nuanced and complex challenge," *SN Computer Science*, 2022.

[7] D. Napoli, "Accessible and usable security: Exploring visually impaired users' online security and privacy strategies," Ph.D. dissertation, Carleton University, 2018.

[8] W. Stallings, L. Brown, M. D. Bauer, and M. Howard, *Computer security: principles and practice*. Pearson Upper Saddle River, 2012, vol. 2.

[9] 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/

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

[11] 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

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

[13] K. Wille, R. R. Dumke, and C. Wille, "Measuring the accessability based on web content accessibility guidelines," in *2016 Joint Conference of the International Workshop on Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA)*. IEEE, 2016, pp. 164–169.

[14] K. Johari and A. Kaur, "Measuring web accessibility for persons with disabilities," in *2012 Fourth International Conference on Computational Intelligence and Communication Networks*. IEEE, 2012, pp. 963–967.

[15] M. Vigo, M. Arrue, G. Brajnik, R. Lomuscio, and J. Abascal, "Quantitative metrics for measuring web accessibility," in *Proceedings of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*, 2007, pp. 99–107.

[16] A. P. Freire, R. P. Fortes, M. A. Turine, and D. M. Paiva, "An evaluation of web accessibility metrics based on their attributes," in *Proceedings of the 26th annual ACM international conference on Design of communication*, 2008, pp. 73–80.

[17] M. Vigo, J. Brown, and V. Conway, "Benchmarking web accessibility evaluation tools: measuring the harm of sole reliance on automated tests," in *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, 2013, pp. 1–10.

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