Towards Real-Time Measurements of Internet Health: Optimizing Large-Scale Web Accessibility Evaluations

Luís P. Carvalho
Northumbria University
United Kingdom, Newcastle upon Tyne
luis.carvalho@northumbria.ac.uk

Tiago Guerreiro
LASIGE, Faculdade de Ciências, Universidade de Lisboa
Lisbon, Portugal
tjvg@di.fc.ul.pt

Shaun Lawson
Northumbria University
Newcastle upon Tyne, United Kingdom
shaun.lawson@northumbria.ac.uk

Kyle Montague
Northumbria University
Newcastle upon Tyne, United Kingdom
kyle.montague@northumbria.ac.uk

ABSTRACT

We rely on large-scale web accessibility evaluations to obtain snapshots of Internet Health and understand trends and behaviours impacting overall web accessibility. Such evaluations are financially and time exhaustive, making the possibility of more real-time measurements of Internet Health infeasible. In this paper, we investigate the impacts of optimising the page selection processes of large-scale web accessibility evaluations. We set out to conduct an automated accessibility evaluation of 1500 websites using the 'Home+' sampling method (for each website, we evaluated the home page and all pages linked belonging to the same domain) as our baseline; then compared the agreement rates of web accessibility evaluations on further sub-sampled datasets. Accessibility data was successfully captured on 987 websites. Our findings demonstrate that a strong accessibility evaluation agreement between the baseline and the sub-sample datasets could be reached with a sub-sample of just 20% of the pages, significantly reducing the effort and resources required to conduct large-scale web accessibility evaluations.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility design and evaluation methods**.

KEYWORDS

Web Accessibility, Large-Scale Analysis, Internet Health

ACM Reference Format:
Luís P. Carvalho, Tiago Guerreiro, Shaun Lawson, and Kyle Montague. 2023. Towards Real-Time Measurements of Internet Health: Optimizing Large-Scale Web Accessibility Evaluations. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23), October 22–25, 2023, New York, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3597638.3608403

ASSETS '23, October 22–25, 2023, New York, NY, USA

ACM ISBN 979-8-4007-0220-4/23/10…$15.00
https://doi.org/10.1145/3597638.3608403

1 INTRODUCTION

Internet Health and our understanding of Web Accessibility relies on regular assessment and auditing of the web pages we interact with and the underlying technologies used to develop and create those websites. The Internet ecosystem is comprised of a number of components, from the browser and front-end HTML; external libraries and web services, down to the low-level communication protocols. It is essential that they work together in harmony to ensure that we have a healthy and accessible web for all. The shifts and changes in the way that we build for the web today have meant that more and more of the sites we use are the result of careful configuration and construction of external libraries and services that provide specific features and functionality, e.g. image galleries, video players or e-commerce carts – oppose to the early days of building everything within the website. As a result, we now have a web that is much more interconnected and small changes or fluctuations to these externally shared libraries and services can have a large impact on our Internet Health and web accessibility.

While it is desirable to have a real-time measurement and understanding of these fluctuations in Internet Health as these ecosystem components evolve and change, it is not feasible to dedicate the resources by way of time and cost to perform large-scale accessibility evaluations on the entire web. As such, prior works have focused on delivering snapshots of particular periods of time [31], or specific domains and categories of websites (e.g. governmental and public sector), which quickly go out of date or don't generalise to the wider behaviours and trends in the web. We also see inconsistencies in the methodologies and sampling approaches of these large-scale evaluations, which in turn could lead to further misrepresentation of the true status of Internet Health.

To that end, in this research, we aimed to explore the impacts of page sampling strategies on web accessibility measurements in hopes of optimising how we do large-scale evaluations.

We conducted an automated web accessibility evaluation on 1500 websites using the 'Home+' [2] as our baseline, resulting in 987 websites after applying our inclusion and exclusion criteria and investigated the effects of sub-sampling on the overall web accessibility evaluation using WCAG 2.1. Our findings demonstrate the need for large-scale evaluations to include more than a website's home page; yet, we found strong agreement and consistency in the evaluation outcomes when sampling just 20% of the website. Moreover, we found that specific WCAG violations were repeated

both within and between websites, suggesting their origins could be through shared templates or libraries. Therefore, if we can optimise large-scale web accessibility evaluations to achieve near real-time measurements of Internet Health, we could tackle these web accessibility vulnerabilities at their source and as they are introduced into the ecosystem to prevent their spread and see a greater impact on overall web accessibility.

2 RELATED WORK

Over the last two decades, the amount of information available on the web and the number of people accessing the internet has grown exponentially [1]. During this time, multiple versions of WCAG have been published, refining the guidelines to make the web more accessible to every internet user. Although guidelines have been available for over 20 years, little progress has been made in developing websites that ensure accessibility. Studies show websites from distinct contexts fail at the most basic level of web accessibility [9, 12, 13, 26]. Primary reasons found for the lack of accessibility on the current web are lack of skill and education on website accessibility [24, 25], unawareness of accessibly standards [11, 18], education not covering web accessibility topics [8, 9], and companies making web accessibility low priority[17, 27, 34]

Richards et al.[31] questioned if some of the improvement of accessibility in the web can be really attributed to an intentional focus on accessibly, suggesting in some cases, the improvement of accessibility was a side effect of new trends and novel technologies (exploitation of new browser capabilities, increased concern with page ranking in search engines, shift toward cross-device and cross-browser content design), revealing that accessibility is not just influenced by human factors but also related to its environment(e.g. browser).

Ross et al. [32] developed a conceptual framework based on epidemiology to identify how accessibility barriers could spread through the Android development ecosystem. According to the framework, accessibility barriers can be seen like diseases that infect a host (e.g. the app) through an infectious agent (e.g. UI component) that contains a determinant (the root cause of the accessibility barrier) and that propagates through a transmission mechanism (e.g. a UI Library). This approach looks at the accessibility of an app as the product of its continued interaction with its surrounding environment (e.g. third-party libraries), acknowledging that it can not be understood by just exclusively looking at the app, showing the potential reach and impact of the libraries and their-party software we use to build software.

A study by Hackett et al.[14] revealed that most researchers rely exclusively upon automated testing of the main website page during accessibility evaluations. According to Vigo et al. [33], over 23% to 50% of accessibility barriers can be found by using automated tools, and by using multiple tools, this percentage increases. Conversely, Padure et al.[30] warn caution on using multiple tools since different tools provide different accessibility evaluation results. Although automatic assessment tools to evaluate WCAG compliance are the most used accessibility evaluation method, manual inspection by experts and users testing is required to ensure better accessibility since compliance with WCAG does not mean real web accessibility in all user contexts.

Many studies have been done to evaluate the accessibility of the web, from evaluating websites in specific contexts, like government, higher education, and business, to the general, top most accessed websites. In these studies, the most common result is finding low levels of web accessibility on the evaluated pages. Another similar characteristic of most accessibility evaluation studies is their methodology. Even though they use different tools to evaluate accessibility, different metrics and mixed approaches, most studies focus on evaluating only the home page[8, 16, 19, 20, 22, 23, 28]. The reason for that choice is that home pages tend to be the most developed pages, representing the best possible accessibility for a website[29]. Regarding large-scale evaluations of the web, the WebAIM yearly report of the accessibility of the top one million home pages revealed on the 2023 report that 96.3% of home pages had at least one WCAG 2 failure[3]. Carrico et al. evaluated over 28 million web pages to understand the shape of web accessibility[26].

Over the years, regulations have been implemented to monitor and evaluate the web accessibility of public government-related services. One such example is the European Union's Web Accessibility Directive [7], which requires websites and apps of public institutions to be accessible. Member states must evaluate their websites every three years and create accessibility monitoring reports. Regarding their methodology, they must do both automatic accessibility and manual accessibility evaluations. The number of websites sampled is based on the country's population, while the number of pages per website for the automated analysis is the home page, plus a "number of pages appropriate to the estimated size and complexity of the website". In practice, these made countries have completely distinct methodologies regarding the number of pages selected, with some countries taking the approach of selecting different types of pages (e.g. login page, site map, contact page), while others choosing a hierarchy level selection (e.g. evaluating all pages to 3 levels depth from the main page), displaying the overall lack of universal strategies when it comes to page selection for accessibility evaluations.

A relevant monitoring report is the one created by Portugal[2]. In their report, they conducted a study in trying to determine if 2 distinct approaches to page selection render similar accessibility results. Their approaches were: Home+, meaning the home page plus all pages linked to it; and 2K, meaning collecting a maximum of 2000 pages at multiple levels of depth. The result of those both samples show to follow a similar accessibility trend.

3 EXPLORING THE IMPACT OF PAGE SAMPLING STRATEGIES

Although the research community uses distinct methodologies and sampling approaches on large-scale accessibility evaluations, we do not understand the impact of using such distinct page sampling strategies on our perception of Internet Health. Our goal in this paper is to contribute with an optimal page sampling strategy that is resource-conscious and capable of creating a reliable status of the accessibility of a website. To that end, we conducted an automated accessibility evaluation of 987 websites using the 'Home+'[2] as our baseline and investigated the effects of sub-sampling on the overall web accessibility evaluation using WCAG 2.1.

3.1 Dataset

We started by selecting the top 1 million websites from the DomCop Top 10 Million Websites[1], then we performed a shuffle on the selected websites using the using Fisher-Yates shuffling algorithm[21] and finally selected the first 1500 websites for evaluation. We followed the 'Home+' strategy for each website, meaning we visited each website's home page and all linked pages belonging to the same domain. For each page visited, we captured a detailed accessibility report containing information regarding Web Content Accessibility Guidelines (WCAGs) violations, warnings and passes.

3.2 Accessibility Evaluator

Axe-Core Accessibility Engine[2] was selected as the automatic evaluation tool for this study as it is Open-Source, provides programmatic integration, is capable of running on a local machine without requiring the use of third-party APIs, and has high industry adoption. This engine implements a total of 61 accessibility rules (Axe-Core rules) covering a total of 22 WCAG 2.1 guidelines (multiple Axe-core rules correspond to the same WCAG guidelines but in different contexts), where 16 guidelines correspond to level A conformance, and 5 are level AA conformance[3]. The result of running an Axe-core rule on an HTML page will generate a report containing the accessibility impact of the rule, the corresponding WCAGs and where the rule was applied on the page. Depending on the result of running the rule, it will fall into one of the following categories: Pass, elements on the page where the rule is applicable pass the accessibility check; Violation: Elements on the page where the rule is applicable fail the accessibility check; Warning: Elements on the page where the rule is can not be guaranteed to pass the accessibility check, requiring manual verification; Inapplicable: The rule does not apply to the page. In this paper, whenever it is mentioned in the methodology or results WCAGs accessibility violations, we reference the aggregate of rules that fall within such guidelines. When we mention axe-core rules, we refer to rules as given by the accessibility engine.

3.3 Data Collecting Apparatus

In order to automate the process of crawling the pages of 1500 websites and capturing the accessibility data, we used the software Puppetteer[4]. We distributed the website load to 5 machines working in parallel. To ensure that the automated page crawling was conducted ethically, we ensured that all pages were visited at a reasonable rate by throttling the number of requests per second.

3.4 Metrics

For large-scale accessibility, we need ways to talk about accessibility sampling at an aggregate level. As a result, we propose a series of metrics that will accommodate this.

3.4.1 Number of websites per count of violations grouped by WCAG type. We recorded how many unique WCAG violation types occur within the pages of a website and grouped the number of websites by the number of violation types.

3.4.2 Number of websites per WCAG violation type. We recorded how many unique WCAG violation types occur within the pages of a website and counted on how many websites a WCAG violation occurs.

3.4.3 Median number of WCAG violations. We calculated the median number of WCAG violations that occur on the pages of a website. We propose to use this metric to create a ranking displaying the websites with the highest median occurrence of WCAG violations per page. In a large-scale evaluation, this metric could be used to compare variations in the incidence of accessibility barriers over time, providing us with a high-level view of possible new accessibility barrier trends.

3.4.4 Accessibility Severity. To measure the severity of the accessibility barriers found on a website, we used the impact parameter present in each axe-core rule evaluation and defined by the accessibility engine[5]. The axe-core engine defines the impact categories of accessibility barriers as follows:

- **MINOR** - "Considered to be a nuisance or an annoyance bug."
- **MODERATE** - "Results in some difficulty for people with disabilities but will generally not prevent them from accessing fundamental features or content."
- **SERIOUS** - "Results in serious barriers for people with disabilities and will partially or fully prevent them from accessing fundamental features or content."
- **CRITICAL** - "Results in blocked content for people with disabilities and will definitely prevent them from accessing fundamental features or content."

We started by attributing weight to each accessibility impact level where: CRITICAL: 4; SERIOUS: 3; MODERATE: 2; and MINOR: 1. Then, for each axe-core rule violation present in a page, we mapped its accessibility impact to the corresponding weight, summed all the weights and divided them by the total number of rule violations on the page.

$$AS_{page} = \frac{\sum_{i=0}^{TotalViolatedRules} RuleImpactWeight(RuleImpact)}{TotalViolatedRules}$$

Equation 1: Accessibility Severity(AS) - The sum of all violated rule impact weight on the page divided by the total number of violated rules on the page.

To calculate the severity of a website, we use the median severity of the pages belonging to the same website.

3.4.5 Accessibility Home Consistency. We wanted a way to determine how consistent the accessibility barriers are between the home page and the other pages of the same website. To do so, we established the home page as the baseline and compared the types of violated axe-core rules between the home page and every other page. We chose the home page as the baseline because it is the most analysed page in most accessibility studies. We defined the home

[1] https://www.domcop.com/top-10-million-websites
[2] https://github.com/dequelabs/axe-core
[3] https://github.com/dequelabs/axe-core/blob/develop/doc/rule-descriptions.md
[4] https://pptr.dev/
[5] https://github.com/dequelabs/axe-core/blob/develop/doc/issue_impact.md

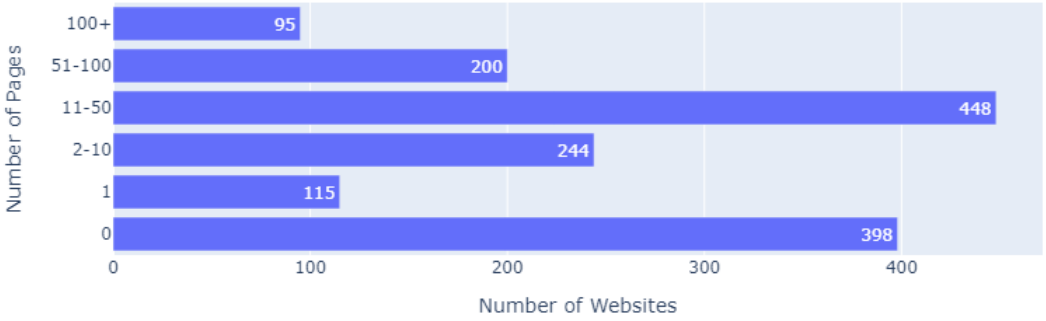


Figure 1: Number of Pages captured per Website

consistency metric as follows:

$$AHC_{page} = \frac{URV_{Page} + URV_{Homepage}}{TotalRules}$$

Equation 2: Accessibility Home Consistency (AHC) - The sum of the number of unique rule violations (URV) on the page and the number of unique rule violations on the homepage divided by the total of rules

3.4.6 Accessibility Inconsistency. We also wanted to measure the overall inconsistency of accessibility barriers existing on a website. To achieve this, we counted the types of violated axe-core rules found on each page and calculated the standard deviation per website. An accessibility inconsistency of zero means full consistency between all pages. This metric will allow us to understand the prevalence of distinct accessibility barriers across the pages of a website.

$$AI_{website} = \sigma_{UniqueRuleViolations}$$

Equation 3: Accessibility Inconsistency (AI) - The standard deviation of the number of unique rule violations found on each page of a website

3.5 Experiments

We ran the metrics and measurements described previously in three different contexts: **1. Home+** On all the pages collected from the 1500 websites, excluding one-page websites. One-page websites were excluded from this research due to the possibility of these websites being single-page applications and our tools not being prepared to analyse such types of pages. **2. Sub-sample without homepage** We selected page samples in increasing increments of 10%(10%-90%) per website, excluding the home page, and repeated the process ten times per sample size. **3. Sub-sample with homepage** We will repeat the process described in 2. However, we added the website's home page for each sample size.

We will use Cohen's weighted kappa coefficient to measure the inter-reliability of the metrics between the data generated on 1. and each sample collected on 2. and 3. We chose Cohen's weighted kappa because the unweighted kappa treats all disagreements equally, and there is a need to preserve scale point differences between ratings of agreement, allowing us to apply a greater penalty to larger disagreements.

4 RESULTS

The crawling of 1500 websites has resulted in a dataset with data from 987 websites(Figure 1). Accessibility data from 115 one-page websites were excluded due to the accessibility tools not being prepared to evaluate single-page apps. No data was from 398 due to website unavailability or problems experienced when crawling. A total of 48,335 pages were visited, with an average of 48.8 (SD = 90.11) pages per website. 95 websites had more than 100 pages, with the maximum number of pages on a website being 1408. The number of pages range with the most websites is 11 to 50, with a total of 448 websites. In total, 1,346,557 axe-core accessibility rule evaluations were run, where 183,075 (13.6%) resulted in accessibility violations, 57,417 (4.3%) resulted in accessibility warnings, and 1,106,065 (82.1%) resulted in accessibility passes.

4.1 Metrics applied to all websites

In this section, we will present the results of each metric when applied to our dataset.

4.1.1 Number of websites per count of violations grouped by WCAG type. Overall we see that websites are failing across a wide range of accessibility violation types (see Table 1), with the most frequent count of accessibility violation types per website being between 7 to 10. the most significant number of websites, 215, having 9 types of accessibility violations. No websites were found to have 15 or more types of violations. The websites with 1, 2 and 14 accessibility violation types have the lowest number of websites, with only 2 websites in each. These results show that websites still contain a high variety of accessibility barriers existing on their pages.

4.1.2 Number of websites per WCAG violation type. The data shows that the most common level of non-compliance is WCAG A, with a total of 4,144 non-compliant pages. The most broken guideline is WCAG 4.1.2 Name, Role, Value, with 922 websites not conforming, followed by WCAG 1.4.3 Contrast Minimum with 883 websites not conforming and 833 websites not conforming to WCAG 2.4.4 Link Purpose. (see Table 2.) These results reflect the persisting absence or incorrect use of syntax required for assistive technology to display

Count of violations grouped by WCAG type	Number of Websites
1	2
2	2
3	36
4	10
5	60
6	60
7	107
8	185
9	215
10	153
11	91
12	51
13	13
14	2

Table 1: Count of violations grouped by their WCAG type

WCAG	WCAG Compliance Level	Websites non-conforming
4.1.2	A	922
1.4.3	AA	883
2.4.4	A	833
1.1.1	A	613
1.4.1	A	586
4.1.1	A	566
1.3.1	A	397
3.1.1	A	288
1.4.4	AA	220
2.1.1	A	94
2.4.2	A	76
2.2.2	A	8
3.1.2	AA	7
1.4.12	AA	3

Table 2: Number of websites per WCAG violation type

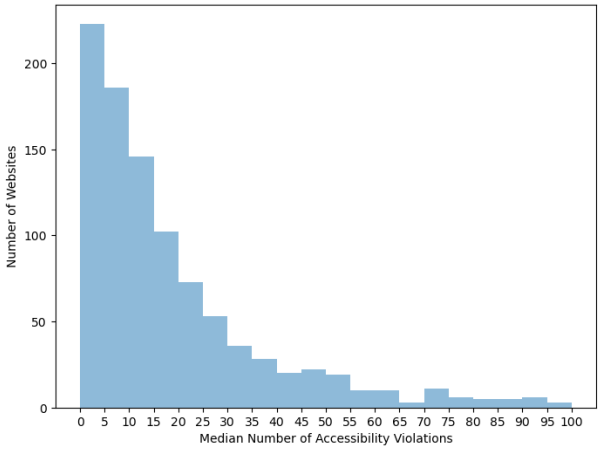


Figure 2: Median number of accessibility violations per website

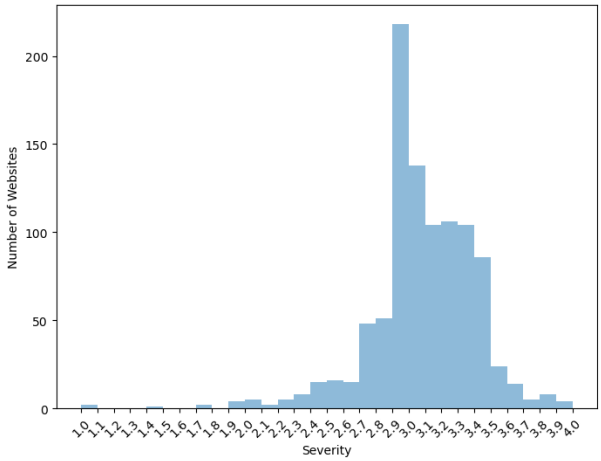


Figure 3: Accessibility Severity

a page correctly and the lack of making visual elements of pages visually accessible. In both cases, it demonstrates the overall lack of importance still given to ensuring the accessibility of pages.

4.1.3 Median number of accessibility violations. The maximum value found was 447 barriers. Most websites (223) contained between 1 and 5 barriers, with an overall tendency to diminish the more significant the number of accessibility barriers (Figure 2). Only 18 websites have more than 100 accessibility barriers.

4.1.4 Accessibility severity per website. Figure 3 shows that most websites have a severity rating between 2.9 and 3.4, ranging between the *Serious* to *Critical* categories of severity. The severity range 2.9 to 3.0 has the highest count of websites, totalling 218. The minimum severity being recorded is 1 (Minor), and the maximum is 4 (Critical). These results show how impactful the accessibility barriers are on the user experience of people with disabilities. Since most websites score three or above, this means people with disabilities are either facing *Serious* barriers that fully or partially prevent them from accessing fundamental features or are being completely blocked from accessing content. It is also worth noting that while the number of accessibility violations found in this dataset is relatively low, with only 13% of all accessibility rule evaluations being violations, the severity of the identified barriers is greater. Demonstrating that even a small count of barriers can have a substantial impact on the overall accessibility of a website. An example of an accessibility barrier with *Serious* severity is ARIA commands not having an accessible name. For screen reader users, this means that they are not able to discern the purpose of elements with role="link", role="button", or role="menuitem" thus making it impossible to use the functionality of such elements.

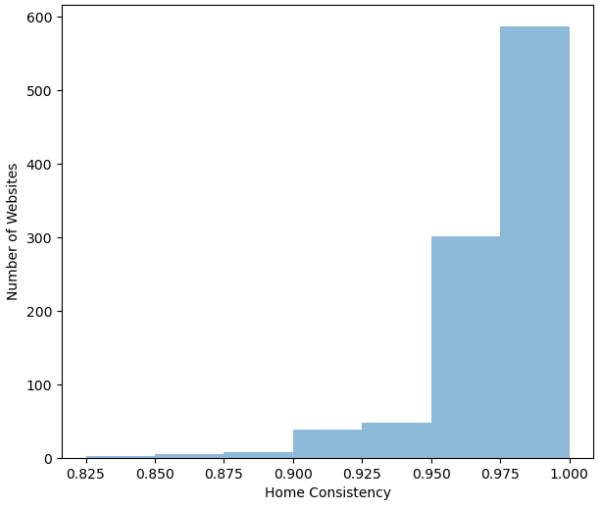


Figure 4: Home accessibility consistency

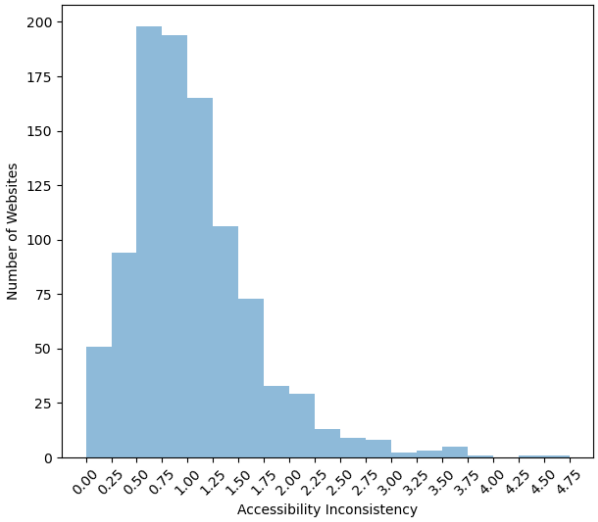


Figure 5: Accessibility Inconsistency

4.1.5 Home accessibility consistency. Regarding accessibility barrier consistency, 57 websites did not contain accessibility barriers on their home pages, thus excluded from this analysis (Figure 4). The overall accessibility consistency between home pages and other pages is high. The minimal consistency level is 83.6%, while 279 websites have full consistency between the homepage and other pages. As shown earlier in the paper, only 13% of all rules evaluated were violations. As such, having a high level of consistency of 90% to 100% was expected, with the accessibility rules passed impacting this metric the most and demonstrating that the differences between *home* and other pages are not extensive.

4.1.6 Accessibility Inconsistency. Regarding accessibility consistency (illustrated in Figure 5), only 39 websites had a rating of 0, meaning full consistency between all pages belonging to those websites. Overall the highest incidence of inconsistency can be found in the range from 0.5 to 0.75, with a total of 201 websites. Above this inconsistency range, the number of websites displays a downward trend, with the maximum inconsistency registered being one website with 4.76. The result of this metric shows that there is a prevalence of pages within a website with distinct types of accessibility barriers that are not present in all website pages. Possible factors for this could be the use of distinct components, libraries or templates on different pages within the website.

4.2 Page Sampling

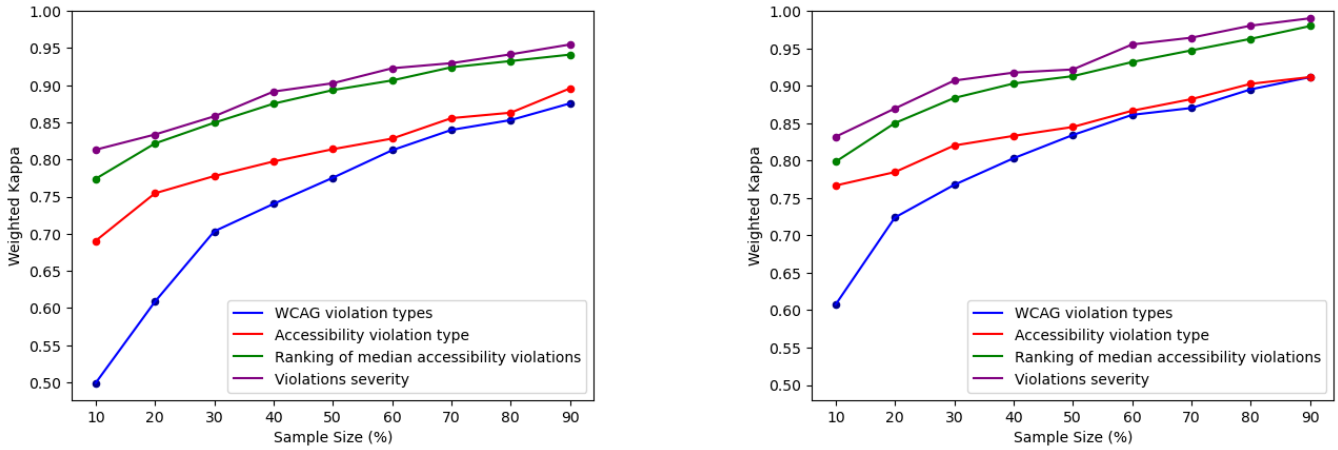
In this section, we will present the result of calculating Cohen's Weighted Kappa between the dataset of all website pages excluding one-page websites and homepage, with a random sample of pages for each website. Figure 6a presents a linear plot of the metrics "WCAG violation types by website", "Number of websites per WCAG violation type", "Median number of accessibility violations", and "Accessibility severity per website", of the random percentage of the page selected per website versus the Kappa agreement between whole dataset and a set percentage of randomly selected pages per website. The overall metrics displayed an upward trend, where the bigger the sample size, the higher the level of agreement. The metric "WCAG violation types by website" displayed an overall lower level of agreement, starting with an agreement of 0.5 when the sample size is 10% and finishing at an agreement level of 0.9 when the sample size is 90%. The metric with the overall highest level of agreement is the "Accessibility severity per website", with an agreement level of 0.81 at a sample of 10% and finishing at an agreement of 0.95 at a sample size of 90%.

Figure 6b present the same metrics and axis as above, but now we selected not only a percentage of random pages per website but also the home page of every website. The overall trend of the metrics remains upward, with the bigger difference being a higher level of agreement at smaller sample sizes. The lowest kappa value remains associated with the metric "WCAG violation types by website", but the level of agreement at 10% sample is now 0.66. In Figure 7, we present the impact of adding the home page to the page sampling on the kappa agreement. Overall the variation displays a downward trend, where the smaller the sample of pages, the biggest the impact for the agreement when considering the home page. On the metric "Ranking of median accessibility violations," the variance remains mostly constant for every sample size. We recommend a page sample size of 20% (including the homepage) for a 'substantial' level of agreement or a 50% sample for an 'almost perfect agreement' (including the homepage).

5 DISCUSSION

5.1 The heterogeneity of web page accessibility within websites

The data shows high consistency when comparing the types of accessibility barriers present on the home with the rest of the web pages, but only around 30% of pages achieve full consistency. This supports the need for using page sampling methods like 'Home+' and not just evaluating the home pages as many large-scale assessments have in the past. This reveals that using just the home page as a way to measure the accessibility of a website might be creating



(a) Excluding homepage **(b) Including the home page**

Figure 6: Weighted Cohen's Kappa of all pages vs random sample of pages

an exceedingly optimistic view of a website's accessibility and missing the key barriers. Moreover, large-scale assessments in the past have not examined websites from the perspective of accessibility consistency found between pages of a website; where instead, all scores were based exclusively on the individual assessment of the accessibility of a page. The creation of an understanding of the accessibility consistency between pages could lead to finding the causes of accessibility barriers on a website, where if an accessibility barrier is present across most of the pages of a website, it could mean its root cause is present on the template or external library used by the website.

5.2 Optimising the selection of pages

Our results demonstrate that using 20% of the 'Home+' pages provides good coverage and consistent representation of the wider website accessibility on most websites – significantly reducing the volume of pages that would need to be scanned regularly. In this study, we randomly selected 20% of pages of each website, but further exploration is required to understand which 20% to evaluate as we believe the page selection should consider a page's components and complexity. Hambley et al.[15] provides an example of how pages could be chosen using DBSCAN, which aggregates pages into clusters of similarity and selects a representative page, thus reducing the number of pages for an auditor to evaluate. In trying to apply this solution in the context of large-scale assessment for Internet Health, a major barrier is having to examine most pages of a website in order to establish a representative sample. As such, in the context of large-scale assessments, there is a need for sampling methods that do not require an extensive pre-analysis of the pages in order to establish a sample population and instead use low-resource methods that use easily accessible data from a website. One such approach would be to use the paths from the URLs for sample creation, where for example, we assume the use of page templates, and a page could be selected per unique URL path found on a website. Future research on page selection optimisation must also consider the following factors: 1) *access to the web through mobile devices* – the use of smartphone imposes distinct accessibility requirements that impact web accessibility; and many websites provide unique versions to desktop and mobile 2) *Large-size websites* – The use of 'Home+' in this study led to the inclusion of 95 websites with 100+ pages. We believe in most of these websites, many pages are templated, and pages minimally differ in regard to accessibility, revealing the possibility of requiring to analyse fewer pages to obtain a representative sample; 3) *Website purpose or context* – We envision that depending on the purpose of the website (e.g. e-commerce, education, social network), the minimal percentage of the pages required to be sampled could be further optimised, in part due to the shared technologies (e.g., e-commerce cart and checkout) used to build such websites.

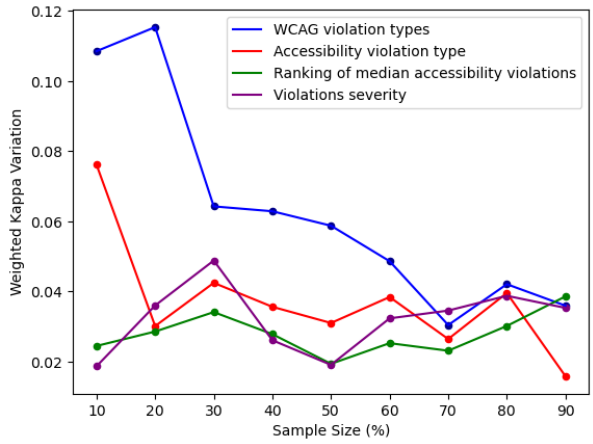


Figure 7: Variation of Kappa between the random sample of pages with and without home page

5.3 Accessibility barrier repetition

Most accessibility violations are repeated errors likely due to the result of templates, frameworks and reused libraries or code within the website. Given the findings of Richards et al[31]. and Ross et al[32], we suspect that these issues are also shared between websites, and we could be using large-scale approaches to identify those issues and their root causes. Moving to real-time monitoring of the internet health, would mean that we could directly measure and observe the impacts or changes to web accessibility as the result of updates or changes to their shared resources and libraries. This is very similar to the way the software vulnerability field works, where there is constant monitoring and testing of software to find possible bugs and vulnerabilities. In software vulnerability analysis, one of the main components since the area's inception has been vulnerability databases[5, 6], where integrating these databases has become almost seamless in the current developer ecosystem. An example is GitHub; through vulnerability scanning of code repositories[4] and a bot connected to a vulnerability database[10], developers receive alerts and continuous updates if any of their code or used external libraries have security issues. We envision that similar approaches could be applied to the web accessibility field as a way of diminishing accessibility barriers caused by external libraries.

6 CONCLUSION

Regular monitoring and assessment of websites to understand internet health are required to ensure the creation of a web that is accessible to all. Still, such attempts to understand the overall accessibility of the web are scarce, superficial and vary greatly in methodologies, partly due to the cost, resources and time required for thorough automatic large-scale accessibility evaluations of the web. In this paper, we explored new strategies for large-scale accessibility evaluations and explored effective strategies for selecting suitable page subsets representative of the overall accessibility of a website. We visited 987 websites and collected accessibility evaluations from all home pages and linked pages, and used Cohen's Kappa to determine a level of page sampling representative of the website's overall accessibility. Results suggest that 20% to 30% of pages should be selected for representative accessibility evaluation.

REFERENCES

[1] 2021. Internet users, UK - Office for National Statistics - Date accessed: 04/05/2023. https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/internetusers/2020

[2] 2022. Report on the 2020/21 Monitoring Period - acessibilidade.gov.pt - Date accessed: 04/05/2023. https://www.acessibilidade.gov.pt/publicacao/monitoring1/

[3] 2022. WebAIM: The WebAIM Million - The 2023 report on the accessibility of the top 1,000,000 home pages - Date accessed: 09/01/2023. https://webaim.org/projects/million/

[4] 2023. Github Features · Security - Date accessed: 30/07/2023. https://github.com/features/security

[5] 2023. NVD - Home - Date accessed: 30/07/2023. https://nvd.nist.gov/

[6] 2023. Vulnerability Database - OSV - Date accessed: 30/07/2023. https://osv.dev/list

[7] 2023. Web Accessibility | Shaping Europe's digital future - Date accessed: 04/05/2023. https://digital-strategy.ec.europa.eu/en/policies/web-accessibility

[8] Patricia Acosta-Vargas, Sergio Luján-Mora, Tania Acosta, and Luis Salvador-Ullauri. 2018. Toward a combined method for evaluation of web accessibility. *Advances in Intelligent Systems and Computing* 721 (2018), 602–613. https://doi.org/10.1007/978-3-319-73450-7{_}57/COVER

[9] Patricia Acosta-Vargas, Sergio Lujan-Mora, and Luis Salvador-Ullauri. 2016. Evaluation of the web accessibility of higher-education websites. *2016 15th International Conference on Information Technology Based Higher Education and Training, ITHET 2016* (11 2016). https://doi.org/10.1109/ITHET.2016.7760703

[10] Mahmoud Alfadel, Diego Elias Costa, Emad Shihab, and Mouafak Mkhallalati. 2021. On the use of dependabot security pull requests. *Proceedings - 2021 IEEE/ACM 18th International Conference on Mining Software Repositories, MSR 2021* (5 2021), 254–265. https://doi.org/10.1109/MSR52588.2021.00037

[11] Humberto Lidio Antonelli, Sandra Souza Rodrigues, Willian Massami Watanabe, and Renata Pontin De Mattos Fortes. 2018. A survey on accessibility awareness of brazilian web developers. *ACM International Conference Proceeding Series* (6 2018), 71–79. https://doi.org/10.1145/3218585.3218598

[12] Barbara Rita Barricelli, Pierlauro Sciarelli, Stefano Valtolina, and Alessandro Rizzi. 2017. Web accessibility legislation in Italy: a survey 10 years after the Stanca Act. *Universal Access in the Information Society 2017 17:1* 17, 1 (2 2017), 211–222. https://doi.org/10.1007/S10209-017-0526-Z

[13] Ramiro Gonçalves, José Martins, Jorge Pereira, Manuel Au-Yong Oliveira, and João José P. Ferreira. 2012. Enterprise Web Accessibility Levels Amongst the Forbes 250: Where Art Thou O Virtuous Leader? *Journal of Business Ethics 2012 113:2* 113, 2 (4 2012), 363–375. https://doi.org/10.1007/S10551-012-1309-3

[14] Stephanie Hackett and Bambang Parmanto. 2009. Homepage not enough when evaluating web site accessibility. *Internet Research* 19, 1 (2009), 78–87. https://doi.org/10.1108/10662240910927830

[15] Alexander Hambley, Yeliz Yesilada, Markel Vigo, and Simon Harper. 2023. OPTIMAL-EM: Optimised Population Sourcing for Web Accessibility Evaluation. In *20th International Web for All Conference*. ACM, New York, NY, USA, 171–172. https://doi.org/10.1145/3587281.3587962

[16] Kelly A. Harper and Jamie DeWaters. 2008. A Quest for website accessibility in higher education institutions. *The Internet and Higher Education* 11, 3-4 (1 2008), 160–164. https://doi.org/10.1016/J.IHEDUC.2008.06.007

[17] Soon G. Hong, Silvana Trimi, Dong W. Kim, and Joon H. Hyun. 2015. A Delphi Study of Factors Hindering Web Accessibility for Persons with Disabilities. *https://doi.org/10.1080/08874417.2015.11645784* 55, 4 (6 2015), 28–34. https://doi.org/10.1080/08874417.2015.11645784

[18] Yavuz Inal, Frode Guribye, Dorina Rajanen, Mikko Rajanen, and Mattias Rost. 2020. Perspectives and Practices of Digital Accessibility: A Survey of User Experience Professionals in Nordic Countries. *ACM International Conference Proceeding Series* 20 (10 2020). https://doi.org/10.1145/3419249.3420119

[19] Abid Ismail and K. S. Kuppusamy. 2018. Accessibility of Indian universities' homepages: An exploratory study. *Journal of King Saud University - Computer and Information Sciences* 30, 2 (4 2018), 268–278. https://doi.org/10.1016/J.JKSUCI.2016.06.006

[20] Shaun K. Kane, Jessie A. Shulman, Timothy J. Shockley, and Richard E. Ladner. 2007. A web accessibility report card for top international university web sites. *ACM International Conference Proceeding Series* 225 (2007), 148–156. https://doi.org/10.1145/1243441.1243472

[21] Donald E. Knuth. 1969. *Seminumerical algorithms. The Art of Computer Programming*. Vol. 2. Addison-Wesley, Reading, MA. 139–140 pages.

[22] K. S. Kuppusamy and V. Balaji. 2021. Evaluating web accessibility of educational institutions websites using a variable magnitude approach. *Universal Access in the Information Society* 22, 1 (3 2021), 241–250. https://doi.org/10.1007/S10209-021-00812-4/FIGURES/5

[23] Serhat Kurt. 2011. The accessibility of university web sites: The case of Turkish universities. *Universal Access in the Information Society* 10, 1 (3 2011), 101–110. https://doi.org/10.1007/S10209-010-0190-Z/TABLES/6

[24] Jonathan Lazar, Alfreda Dudley-Sponaugle, and Kisha Dawn Greenidge. 2004. Improving web accessibility: a study of webmaster perceptions. *Computers in Human Behavior* 20, 2 (3 2004), 269–288. https://doi.org/10.1016/J.CHB.2003.10.018

[25] Humberto Lidio Antonelli, Sandra Souza Rodrigues, Willian Massami Watanabe, and Renata Pontin de Mattos Fortes. [n. d.]. A survey on accessibility awareness of Brazilian web developers ACM Reference format. *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion* ([n. d.]). https://doi.org/10.1145/3218585

[26] Rui Lopes, Daniel Gomes, and Luís Carriço. 2010. Web not for all: A large scale study of web accessibility. *W4A 2010 - International Cross Disciplinary Conference on Web Accessibility Raleigh 2010* (2010). https://doi.org/10.1145/1805986.1806001

[27] VellemanEric M., NahuisInge, and GeestThea. 2017. Factors explaining adoption and implementation processes for web accessibility standards within eGovernment systems and organizations. *Universal Access in the Information Society* 16, 1 (3 2017), 173–190. https://doi.org/10.1007/S10209-015-0449-5

[28] Carlos Máñez-Carvajal, Jose Francisco Cervera-Mérida, and Rocío Fernández-Piqueras. 2019. Web accessibility evaluation of top-ranking university Web sites in Spain, Chile and Mexico. *Universal Access in the Information Society* 20, 1 (3 2019), 179–184. https://doi.org/10.1007/S10209-019-00702-W

[29] J. Nielsen. 2000. Designing web usability. (2000). https://doi.org/10.3/JQUERY-UI.JS

[30] Marian PADURE and Costin PRIBEANU. 2020. Comparing Six Free Accessibility Evaluation Tools. *Informatica Economica* 24, 1/2020 (3 2020), 15–25. https://doi.org/10.24818/ISSN14531305/24.1.2020.02

[31] John T. Richards, Kyle Montague, and Vicki L. Hanson. 2012. Web accessibility as a side effect. *ASSETS'12 - Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (2012), 79–86. https://doi.org/10.1145/2384916.2384931

[32] Anne Spencer Ross, Xiaoyi Zhang, James Fogarty, and Jacob O. Wobbrock. 2020. An Epidemiology-inspired Large-scale Analysis of Android App Accessibility. *ACM Transactions on Accessible Computing (TACCESS)* 13, 1 (4 2020). https://doi.org/10.1145/3348797

[33] Markel Vigo, Justin Brown, and Vivienne Conway. 2013. Benchmarking web accessibility evaluation tools: Measuring the harm of sole reliance on automated tests. *W4A 2013 - International Cross-Disciplinary Conference on Web Accessibility* (2013). https://doi.org/10.1145/2461121.2461124

[34] Yeliz Yesilada, Giorgio Brajnik, Markel Vigo, and Simon Harper. 2013. Exploring perceptions of web accessibility: a survey approach. *http://dx.doi.org/10.1080/0144929X.2013.848238* 34, 2 (2 2013), 119–134. https://doi.org/10.1080/0144929X.2013.848238