Universal Access in the Information Society
https://doi.org/10.1007/s10209-022-00956-x

LONG PAPER

Large-scale study of web accessibility metrics

Beatriz Martins[1] · Carlos Duarte[1]

Accepted: 24 November 2022
© The Author(s) 2022

Abstract
Evaluating the accessibility of web resources is usually done by checking the conformance of the resource against a standard or set of guidelines (e.g., the WCAG 2.1). The result of the evaluation will indicate what guidelines are respected (or not) by the resource. While it might hint at the accessibility level of web resources, often it will be complicated to compare the level of accessibility of different resources or of different versions of the same resource from evaluation reports. Web accessibility metrics synthesize the accessibility level of a web resource into a quantifiable value. The fact that there is a wide number of accessibility metrics, makes it challenging to choose which ones to use. In this paper, we explore the relationship between web accessibility metrics. For that purpose, we investigated eleven web accessibility metrics. The metrics were computed from automated accessibility evaluations obtained using QualWeb. A set of around three million web pages were evaluated. By computing the metrics over this sample of nearly three million web pages, it was possible to identify groups of metrics that offer similar results. Our analysis shows that there are metrics that behave similarly, which, when deciding what metrics to use, assists in picking the metric that is less resource intensive or for which it might be easier to collect the inputs.

Keywords Web accessibility metrics · Large-scale accessibility evaluation · Automatic accessibility evaluation · QualWeb

1 Introduction

Web accessibility is defined as the availability and usability of web resources by every single individual, no matter their disabilities. “Web accessibility means that websites, tools, and technologies are designed and developed so that people with disabilities can use them” [1].

Web accessibility can be evaluated by verifying the conformance with standards or guidelines, the most common ones being the Web Content Accessibility Guidelines (WCAG) [2]. Given that the results typically show the accessibility of a page or site in terms of conformance to the set of guidelines being checked, it is not always easy to gauge the accessibility level of the evaluated resource. Approaches such as considering a resource accessible only if it conforms to all the guidelines checked are easy to understand, but do not support understanding how far from being accessible the resource is. Approaches that capture the nuances in the levels of accessibility of web pages or sites could be more useful. This is what web accessibility metrics try to achieve.

Web metrics are defined as procedures “for measuring a property of a web page or website” [3]. For instance, the number of links or the size of an HTML file are two examples of properties that can be computed for a site or web page. As such, accessibility metrics are responsible for measuring the accessibility level of websites or web pages, by synthesizing values of web resources [4].

Web accessibility metrics are formulas that are applied using data provided by accessibility evaluations. This data can be gathered manually, semi-automatically or automatically. For instance, there are metrics that use data collected through expert procedures which, when conducting large-scale evaluations (e.g., comparing multiple sites or monitoring a site with hundreds of pages), make them expensive and impractical choices.

Accessibility metrics are important for multiple purposes and scenarios. The most obvious use is to compare two or more web resources. This can be relevant for practitioners, like web site administrators, that need to compare

Beatriz Martins and Carlos Duarte contributed equally to this work.

✉ Carlos Duarte
caduarte@fc.ul.pt

Beatriz Martins
bmartins@lasige.di.fc.ul.pt

[1] LASIGE, Faculdade de Ciências da Universidade de Lisboa, Campo Grande 016, Lisboa 1749-016, Portugal

Published online: 09 December 2022

subsequent releases of the same website to check for accessibility improvements. But it is also important for researchers, especially when conducting large-scale accessibility evaluations [5], comparing domains of activity [6, 7], geographical areas [8, 9], or user groups [10]. Recently, in the context of the European Web Accessibility Directive[1], the different European member-states reported the results of their accessibility monitoring activities. This large-scale accessibility monitoring exercise was marred by the difficulty in comparing the results reported by the different member-states. The use of a common metric would have mitigated this problem. Other potential benefits from accessibility metrics include support for ranking web pages, which can be relevant for retrieval systems; or being used as a way to provide criteria for adaptations in adaptive hypermedia systems [3].

Given there is a large number of web accessibility metrics available for researchers, auditors or practitioners to choose from, an important question emerges: which one(s) should be used? To help answer this question, it is important to understand how these web accessibility metrics relate to each other and if it is possible to group them according to their similarities and understand the differences between each group.

To identify existing relationships between web accessibility metrics we computed eleven different web accessibility metrics over a set of more than two million web pages. In this article we report the findings of this study. We begin by providing a background about web accessibility metrics and a review of 19 web accessibility metrics that were proposed in the literature. Then, we present the methodology, results and discussion of a study where we compared eleven of the 19 reviewed metrics. Afterwards, we present a second study, where we analyzed the validity of the eleven metrics, by assessing how they rate a set of pages created to demonstrate good and bad accessibility practices. We finish with an analysis of the studies' limitations before concluding.

With this work we contribute the following:

- A review of existing web accessibility metrics, describing a total of 19 metrics applicable at page or website level;
- The results of computing a subset of eleven metrics over a sample of nearly three million web pages;
- An analysis that identifies relationships between metrics and determines groups of metrics that report similar outcomes.

2 Accessibility metrics

According to Vigo, Brajnik and Connor [3], web metrics measure properties of websites or web pages. These metrics can summarize results obtained from a guideline review based evaluation [11]. Additionally, Song, et al., [12] state that web accessibility metrics have the ability to measure the accessibility levels of websites.

Metrics should meet five different aspects [13]. They should:

1. be simple to understand;
2. be precisely defined;
3. be objective;
4. be cost-effective; and
5. give such information so it is possible to have meaningful interpretations.

Freire, et al., also mention that web accessibility metrics are important to understand, control and improve products and processes in companies [13]. Nevertheless, they affirm that it is not possible to define which metric is more effective, since it depends on the project in question and its needs.

Parmanto and Zeng [14] argue that an accessibility metric should be summarized into a quantitative score that provides a continuous range of values so it is possible to understand how accessible and inaccessible the web content is. It is also important to guarantee that the range of the metric's values supports more fine-grained discrimination than accessible and inaccessible. Another property the authors ascribe to high quality metrics is that they should consider the complexity of the websites. It would be convenient if the accessibility metric could be scalable to conduct a large-scale accessibility evaluation.

In conclusion, metrics are useful to process and understand the results obtained from an accessibility evaluation. This approach can also help rank web pages or even explore the accessibility level of web pages or websites. The computation of accessibility metrics can produce, as a result, qualitative or quantitative values.

2.1 Literature review

Before presenting the details of the identified web accessibility metrics, it is important to introduce concepts that help to understand how each metric behaves.

Some metrics use the barrier concept. A barrier is a condition caused by the website or web page that prevents the user to access the web content [15], i.e., a problem found in a certain website or web page that prevents the user to perceive or interact with the web content. Barriers can have different levels of severity.

[1] https://digital-strategy.ec.europa.eu/en/policies/web-accessibility.

Whenever an accessibility evaluation is performed, its outcomes vary according to the compliance with standards. Different outcomes are considered by different metrics, but they can be summarized into: (1) pass, which means that the web content fulfills a certain recommendation; (2) fail, which indicates that the web content does not meet the recommendation; (3) warning, an outcome produced by automated evaluation tools to represent those instances where the tool could not determine the conformance, or lack of conformance, with the recommendation, and the intervention of an human expert is required.

Besides the above aspects, it is important to note that some of the web accessibility metrics that have been reviewed verify the conformance with checkpoints and these checkpoints are grouped into priority levels: priority 1, priority 2 or priority 3. The priority levels in some metrics have associated weights that vary from zero to one. This applies to metrics proposed before the introduction of WCAG 2.0. Metrics proposed after WCAG 2.0 typically verify conformance with success criteria grouped at conformance levels A, AA and AAA.

The score of a metric can be bounded or not. A bounded metric makes it easier to gauge where a score falls within the accessible to inaccessible continuum of values. Unbounded metrics, on the other hand, by not having a defined range of values, can lead to a harder interpretation of whether a resource is accessible or inaccessible.

In the following, we present the metrics we found by searching the existing literature on web accessibility. For each metric, we describe the data it is based on, its output range, and any other considerations regarding its application (e.g., if it is applicable to web pages or web sites).

2.1.1 Failure-rate (FR)

The Failure Rate (FR) was developed by Sullivan and Matson in 2000 [16]. According to Vigo, et al., [17], this metric relates the actual points of failure with the potential points of failure. For instance, if a web page has ten images, all these images are potential barriers if they are not properly defined. If five out of these ten images do not have a proper alternative text, according to the accessibility evaluator, they are actual barriers.

A point of failure can be interpreted in two ways: as an accessibility problem or barrier that occurs on a web page's elements preventing the interaction of a user with the web content; or as the elements that cause accessibility problems. According to the first interpretation, each element can have multiple points of failure, which allows us to count more accessibility problems and better estimate the accessibility level. Therefore, we decided to consider a point of failure as an accessibility problem that occurs on a web page. Consequently, the failure rate can be the ratio between the actual problems that were encountered in a web page and the potential barriers, i.e., all potential problems of a web page that can lead to accessibility issues if they are not properly designed.

Vigo and Brajnik [4] state that the failure rate quantitatively measures the accessibility conformance, having a score from zero to one. A web page with a failure rate of zero is totally accessible, whereas a totally inaccessible web page has a failure rate score of one.

The simplicity of this metric can be explained with the fact that it does not consider the error nature, i.e., "whether checkpoints are automatic errors, warnings or generic problems" [18], or the fact that it does not take into consideration the checkpoints' weights.

$$I_p = \frac{B_p}{P_p} \tag{1}$$

Equation 1 presents the formula for computing the Failure Rate metric, where I_p is the Failure Rate final score, B_p identifies the actual points of failure, and P_p identifies the potential points of failure.

2.1.2 Unified web evaluation methodology (UWEM)

According to Sirithumgul, Suchato and Punyabukkana [10], UWEM 1.0 is an improved version of UWEM 0.5 [19] that was developed in 2006. It is based on user feedback rather than WCAG priority levels [12]. The final value of this metric represents a probability of finding a barrier in a website or web page that could prevent users from completing a certain task [11, 13, 20]. This metric also considers the potential problems and barriers' weights. The UWEM formula is based on the product of the checkpoints' failure rates [20]. Its results are precise and accurate, however, it only takes into consideration 2 priority levels of the WCAG guidelines [21].

The formula can be interpreted as a web page score or a website score. If the website score is wanted, then the UWEM formula will be the sum of the UWEM score of each web page divided by the total number of pages of that website, i.e., the arithmetic mean.

This formula's final score varies between zero and one, where zero means the web page is accessible and one means the web page is inaccessible.

$$UWEM = 1 - \prod 1 - \frac{B_i}{P_i} W_i \tag{2}$$

Equation 2 presents the formula for computing the UWEM metric, where B_i is the total of actual points of failure of a checkpoint i, P_i is the total of potential points of failure of a checkpoint i, and W_i identifies the severity of a certain barrier i (this weight is calculated by simple heuristics, by combining the results of an automatic evaluation and manual testing or by disabled users feedback [22]).

2.1.3 A3

In 2006, Buhler, et al., proposed some changes to the UWEM 0.5 metric [22]. In particular, some probability properties were used as well as some issues related to the complexity of the web page were aggregated. A3 is an improved aggregation formula based on UWEM [11, 13, 20]. Similar to UWEM, A3 also considers the failure rate, i.e., the ratio between the number of barriers (violation of a given checkpoint) and the total number of potential barriers. UWEM and A3 consider the barriers weights coefficients based on the impact on the user of each given barrier [13].

This metric produces a small range of values, that are all between zero and one, where zero means the web page is accessible whereas 1 means the web page is inaccessible.

$$A3 = 1 - \prod_{b} (1 - F_b)^{\frac{B_{pb}}{N_{pb}} + \frac{B_{pb}}{B_p}} \tag{3}$$

Equation 3 presents the formula for computing the A3 metric, where B_{pb} is the total of actual points of failure of a checkpoint b in page p, b is the barrier (checkpoint violation), N_{pb} is the total of potential points of failure of a checkpoint b in page p, and F_b identifies the severity of a certain barrier b (this weight is calculated by simple heuristics, by combining the results of an automatic evaluation and manual testing or by disabled users feedback [22]).

The authors of this metric performed an experimental study to compare the results between A3 and UWEM and understand the differences between them. A checkpoint weight of 0.05 was used for all checkpoints, assuming that all of them would have the same importance. This experiment was conducted with a group of six disabled users that evaluated six web pages. After applying both metrics, the authors concluded that A3 outperformed UWEM in the experiment [11].

2.1.4 Web accessibility barriers (WAB)

The WAB metric was proposed by Hackett, et al., in 2003 [23]. Parmanto and Zeng proposed a new version of the WAB metric in 2005 [14]. It quantitatively measures the accessibility of a web site considering the 25 WCAG 1.0 checkpoints (5 checkpoints in Priority 1, 13 checkpoints in Priority 2, and 7 checkpoints in Priority 3). It applies the concepts of potential problems and weights of the barriers. Barriers' weights are related to the relative importance of a given checkpoint. It takes into consideration the total number of pages of a certain website. The WAB formula is defined as the ratio between the sum of the failure rate of each checkpoint and the priority of that checkpoint [4]. The arithmetic mean of all pages of a website represents the metric score for that website. The Hackett and the Parmanto and Zeng formulas are represented in equations 4 and 5, respectively.

The range of this metric's values is not bounded [18], as there is no limit for this metric's score. The only reference this metric has is the higher its score, the worse the accessibility level of the website. Since this metric takes into consideration 25 WCAG checkpoints out of 65, this metric offers a guideline support of 38%. Nevertheless, according to Brajnik and Vigo [24], WAB is the best individual metric compared to A3, Page Measure (PM) and Web Accessibility Quantitative Metric (WAQM) since it yields an accuracy rate of 96%.

$$WAB = \frac{1}{N_p} \sum_{p} \sum_{c} \frac{fr(p, c)}{\text{priority}_c} \tag{4}$$

Equation 4 presents the formula for computing the WAB by Hackett metric, where $fr(p, c)$ is the failure rate of a certain checkpoint c in web page p, priority_c identifies the priority level of the checkpoint c (1, 2 or 3), and N_p is the total number of web pages of a given website.

$$WAB = \frac{\sum_{j=1}^{T} \sum_{i=1}^{n} (\frac{b_{ij}}{B_{ij}})(W_i)}{T} \tag{5}$$

Equation 5 presents the formula for computing the WAB by Parmanto and Zeng metric, where b_{ij} is the number of actual violations of checkpoint i in page j, B_{ij} is the number of potential violations of checkpoint i in page j, n is the total number of checkpoints, W_i identifies the weight of the checkpoint c, according to its priority level (this weight is calculated from experiments with users with different disabilities [11]), and T is the total number of web pages of a given website.

Parmanto and Zeng [14] weighted the priority levels in the calculation of the WAB score. Priority 1 violations represent a higher weight score since web pages with this level of violations are more difficult to access by people with disabilities.

Ana Belén Martínez, Aquilino A. Juan, Darío Álvarez, and Ma del Carmen Suárez [21] went further and created a quantitative metric based on the WAB metric: WAB∗. The WAB∗ metric is based on WAB and has some UWEM-like extensions. It gets the WAB's precision of the accessibility score and uses more detailed checkpoints, as UWEM does. With all these tools, the authors could build a new metric, namely WAB∗. Martínez, et al. [21], point out the main problems and the main advantages of WAB and UWEM metrics. For instance, WAB performs tests to evaluate checkpoints, yet it is not precise in the way it determines the number of potential violations of each checkpoint. However, it specifies all three priorities' checkpoints. Concerning UWEM, this metric produces more precise results, although it only focuses on priority 1 and 2 checkpoints. Thus, these two metrics are merged into WAB$*$. Consequently, WAB$*$ has more precision in terms of the obtained results. In conclusion, this new metric considers 3 priority levels and has 36 checkpoints (25 WAB checkpoints + 11 UWEM checkpoints). This metric was tested by evaluating 30,600 web pages from banking sector websites. The results show that WAB$*$ outperforms WAB and UWEM.

2.1.5 Overall accessibility metric (OAM)

In 2005, Bailey and Burd [25] proposed OAM. The calculated value considers the number of violations of a checkpoint and the weight of that checkpoint as the confidence level. This confidence level depends on how certain the checkpoint is. There are four confidence levels: certain checkpoints weigh 10, high certainty checkpoints weigh 8, low certainty checkpoints weigh 4 and the most uncertain checkpoints weigh 1. The higher the weight, the more the barrier is penalized.

This metric does not have a bounded range of values. The higher this metric's score, the more inaccessible the web page is.

$$OAM = \sum_{c} \frac{B_c W_c}{N_{\text{attributes}} + N_{\text{elements}}} \tag{6}$$

Equation 6 presents the formula for computing the OAM metric, where B_c is the number of violations of checkpoint c, W_c is the weight of the checkpoint c, $N_{\text{attributes}}$ is the number of HTML attributes on a given web page, and N_{elements} is the number of elements on a given web page.

2.1.6 Page measure (PM)

Later, in 2007, Bailey and Burd [26] proposed Page Measure (PM). This metric "analyzes the correlations between the accessibility of web sites and the policies adopted by software companies regarding usage of CMS or maintenance strategies" [4]. It is similar to OAM (Overall Accessibility Metric), however, instead of using checkpoint weights, the checkpoint priority levels are considered. This metric does not have a bounded range of values. The higher this metric's score, the more inaccessible the web page is.

$$PM = \frac{\sum_{c} \frac{B_c}{\text{priority}_c}}{N_{\text{attributes}} + N_{\text{elements}}} \tag{7}$$

Equation 7 presents the formula for computing the PM metric, where B_c is the number of violations of checkpoint c, $priority_c$ identifies the priority level of the checkpoint c (1, 2 or 3), $N_{attributes}$ is the number of HTML attributes on a given web page, and $N_{elements}$ is the number of elements on a given web page.

2.1.7 SAMBA

Brajnik and Lomuscio proposed SAMBA [27], a semi-automatic method for measuring barriers of accessibility, that combines automatic evaluations with human judgment, and, for this reason, is a semi-automated methodology.

SAMBA is based on WCAG 1.0. This method applies human judgment in the context of a Barrier Walkthrough analysis [27] to estimate aspects related to the automated tool errors and the severity of the barriers. The Barrier Walkthrough method is used for evaluating the web accessibility [28] and it is performed by experts. This manual approach contextualizes the accessibility barriers identified by experts within usage scenarios and these barriers receive a severity score. The severity score of a barrier assumes a value from {0, 1, 2, 3} that corresponds to false positive (FP), minor, major or critical barriers.

This semi-automated approach [27] applies a set of sequential steps. Initially, automatic accessibility tools are used to identify the potential accessibility barriers and the provided results are submitted to human judgment. Then, it is possible to statistically estimate the false positives and the severity of barriers for each website. Finally, barriers are grouped according to disability types and it is possible to derive scores that represent non-accessibility.

This metric computes two accessibility indexes: Raw Accessibility Index (AIr) and Weighted Accessibility Index (AIw). Since AIw is based on confidence intervals manually computed by human experts, its result is represented by an interval $[\underline{AIw}, \overline{AIw}]$. The confidence intervals express the minimum and the maximum percentages of a type of barriers (FP, minor, major or critical) for a specific disability (blind users, deaf users, among others) on a given website. For example, having the interval [6, 12] in column 'critical' and row 'blind' means that, in a given website, there are between 6% and 12% of critical barriers for blind users. The AIw index considers weights that are associated with minor and major severity levels. If both minor and major weights are equal to 1, AIw becomes unweighted (AIu).

SAMBA has a limitation: it cannot cope with false negatives, i.e., problems that were not identified [4]. This means that, although human judgments are used to evaluate and validate the results obtained by the automated tools, they do not deal with the problem of false negatives, since the experts only verify the identified problems. For this reason, the actual issues that were not identified, are not going to be analyzed by the experts, i.e., the problems that are not identified by the evaluation tools are not considered.

$$AI_r = \prod_d (1 - F \cdot \vec{D}_d)^2 \tag{8}$$

$$\underline{AI_w} = \prod_d (1 - F \cdot min\{1, \overline{H_d}\})^2 \tag{9}$$

$$\overline{AI_w} = \prod_d (1 - F \cdot \underline{H_d})^2 \tag{10}$$

$$F = \frac{\text{number of potential barriers}}{\text{number of HTML lines}}, \tag{11}$$

$$\underline{H_d} = \frac{\underline{f}_{d,mnr}}{w_{mnr}} + \frac{\underline{f}_{d,maj}}{w_{maj}} + \underline{f}_{d,cri}, \tag{12}$$

$$\overline{H_d} = \frac{\overline{f}_{d,mnr}}{w_{mnr}} + \frac{\overline{f}_{d,maj}}{w_{maj}} + \overline{f}_{d,cri} \tag{13}$$

In Eq. 8, *F* is the barrier density of a website, *d* is a disability type, and *D* is the disability vector of a website. In Eqs. 9 and 10, H_d is the severity of the barriers of a disability type *d*. Equations 12 and 13 identify *f* as the relative frequency, *mnr* as a minor barrier, *maj* as a major barrier, and *cri* as a critical barrier.

2.1.8 Web accessibility evaluation metric (WAEM)

The Web Accessibility Evaluation Metric Based on Partial User Experience Order [29] was proposed by Song et al. and intends to analyze data from the user experience of people with disabilities. To do so, the authors defined a formula that calculates the weighted accessibility score (Eq. 15), by using the pass rate (Eq. 14), of a certain checkpoint on a website. Besides these formulas, this metric also considers users' experience evaluations through PUEXO pairs. PUEXO (Partial User EXperience Order) defines pairs of websites that establish a comparison in terms of user experience. For instance, the (*a*, *b*) pair indicates that a certain user had a better browsing experience in website *a* compared to website *b*. The PUEXO pairs are then compared to the weighted accessibility scores of the websites in question, by Eq. 16.

Subsequently, the results of Eq. 16 and the users' evaluations are both used to calculate the optimal checkpoint weights (Eq. 17). Equation 17 is not, however, adequate once the user experience is a subjective aspect. For this reason, the authors developed Eq. 18, where they make use of machine learning.

As seen in [29], "results demonstrate that WAEM really can better match the accessibility evaluation results with the user experience of people with disabilities on Web accessibility". Nevertheless, the user experience is a subjective problem and varies according to the user. This means that it is complicated to confirm a relationship between user experience and web accessibility, since different users can have different user experiences [29].

When using WAEM, the higher the weighted accessibility score, the more accessible the website is.

$$p = \frac{s}{h} \tag{14}$$

Equation 14 presents the formula for computing the Pass Rate, where *p* is the pass rate of a checkpoint, *s* is the number of pages of a website a checkpoint passed, and *h* is the total number of web pages of a website.

$$q_i = P_i w = \sum_{j=1}^{m} P_{i,j} w_j \tag{15}$$

Equation 15 presents the formula for computing the Weighted Accessibility Score, where q_i is the weighted accessibility score of a website i, $P_{i,j}$ is the pass rate of a checkpoint j on a website *i*, *m* is the number of checkpoints, and w_j is the weight of a checkpoint *j*, according to its priority level.

$$f((a, b), w, P) = \begin{cases} 1 : P_a w > P_b w \\ 0 : otherwise \end{cases} \tag{16}$$

Equation 16 presents the formula for computing the function *f*, where (*a*, *b*) is a PUEXO pair that represents an order identified by disabled users, *w* is the set of checkpoints' weights, and *P* is the matrix of the pass rates of all websites.

$$\begin{aligned} argmax_w &= \sum_{i=1}^{k} f(L_i, w, P) \\ s.t. \sum_{j=1}^{m} w_j &= 1; \quad \forall i, w_i > 0 \end{aligned} \tag{17}$$

Equation 17 presents the formula for computing the optimal checkpoint weight vector *w*, where *w* is the set of checkpoints' weights, *L* is the matrix that contains all pairs of websites, *i* is the website, *j* is the checkpoint, *m* is the number of checkpoints, and *P* is the matrix of pass rates.

$$\begin{aligned} argmin_w &= \sum_{i=1}^{k} e_i \\ s.t. \sum_{j=1}^{m} w_j &= 1; \quad \forall, e_i \geq 0, w_i > 0, P_{L_{i,1}} w + e_i > P_{L_{i,2}} w \end{aligned} \tag{18}$$

Equation 18 presents the formula for computing the optimal checkpoint weight vector *w*, where *i* is the website, *e* is the error tolerance vector, *P* is the matrix of pass rates, *m* is the number of checkpoints, and L is the matrix that contains all pairs of websites.

2.1.9 Reliability aware web accessibility experience metric (RA-WAEM)

RA-WAEM is a metric that assesses the severity of accessibility barriers by considering the user experience of disabled people [12]. The authors of this metric wanted to overcome the limitation of only using checkpoint weights, by reflecting the user experience of people with disabilities.

This metric's approach is similar to WAEM's approach. RA-WAEM is also aligned with PUEXO, which represents a pair of ordered websites, according to user experience. As RA-WAEM is similar to WAEM, its process is identical to WAEM's formulas 14, 15 and 16. From Eq. 16, RA-WAEM exhibits a different behavior. This metric also aims to calculate the optimal checkpoint weights as shown in Eq. 19. However, this last equation is not continuous. For this reason, Eq. 20 emerged. Yet, the fact that user experience is subjective and influenced by users' expertise level and objectivity, led to a reliability aware model (Eq. 21) where they introduce the reliability level. This new formula is the main difference between RA-WAEM and WAEM.

The results shown by Song, et al., in their study [12], assert that RA-WAEM outperforms WAEM, since it is more stable and reliable concerning the user experience of disabled people. One limitation of both RA-WAEM and WAEM metrics is that the users that are picked to evaluate the accessibility of a set of web pages, may not have a certain expertise level, ending up compromising the final metric results. For instance, users with low expertise would probably have more difficulty, considering a website as inaccessible [12]. Whenever the experience of more volunteers is considered, the performance of both metrics decreases. Nevertheless, results indicate that RA-WAEM is significantly less affected than WAEM [12].

With RA-WAEM, the higher the weighted accessibility score, the more accessible the website is.

$$\begin{aligned} argmax_w &= \sum(a,b,u) \in L \quad f(a,b,w,P) \\ s.t. \sum_{j=1}^{m} w_j &= 1; \quad \forall 1 \leq j \leq m, w_j > 0 \end{aligned} \tag{19}$$

Equation 19 presents the formula for computing the optimal checkpoint weight vector w, where w is the set of checkpoints' weights, L is the matrix that contains all pairs of websites a and b ordered by disabled user u, i is the website, j is the checkpoint, m is the number of checkpoints, and P is the matrix of pass rates.

$$\begin{aligned} argmin_w &= \sum_{i=1}^{k} e_i \\ s.t. \sum_{j=1}^{m} w_j &= 1; \quad \forall 1 \leq j \leq m, w_j > 0; \\ \forall (a,b,u) &\in L, e_{a,b} \geq 0, P_a w + e_{a,b} > P_b w \end{aligned} \tag{20}$$

Equation 20 presents the formula that corrects 19, since it is not continuous, where (a, b, u) is a tuple containing the PUEXO pair of websites a and b that were evaluated by the disabled user u, e is the error tolerance, P is the matrix of pass rates, m is the number of checkpoints, j is the checkpoint, w is the checkpoints' weights, and L is the matrix that contains all pairs of websites a and b ordered by disabled user u.

$$\begin{aligned} argmin_w &= \sum(a,b,u) \in L e_{a,b} r_u \\ s.t. \sum_{j=1}^{m} w_j &= 1; \quad \forall 1 \leq j \leq m, w_j > 0; \\ \forall (a,b,u) &\in L, e_{a,b} \geq 0, P_a w + e_{a,b} > P_b w \end{aligned} \tag{21}$$

Equation 21 presents the formula for computing the reliability aware model, where (a, b, u) is a tuple containing the PUEXO pair of websites a and b that were evaluated by the disabled user u, e is the error tolerance, r is the reliability level vector, P is the matrix of pass rates, m is the number of checkpoints, j is the checkpoint, w is the checkpoints' weights, and L is the matrix that contains all pairs of websites a and b ordered by disabled user u.

2.1.10 Barrier impact factor (BIF)

BIF is the barrier impact factor. According to Battistelli, et al. [30], this metric analyzes each accessibility error with respect to the way it affects disabled users' browsing by means of assistive technologies. It evaluates the accessibility, against the WCAG guidelines, using a list of assistive technologies or disabilities affected by each error. Each error represents a success criterion failure that was detected by the accessibility evaluation tool. It is necessary to define a barrier-error association table in advance that represents a list of assistive technologies affected by each error.

The main goal is to understand the impact factor of each barrier on a specific assistive technology or disability (for example, a screen reader). The result score refers to the amount of detected errors that were identified for each assistive technology and it also considers the weight of that assistive technology. This weight's value varies according to the success criterion conformance level: level A errors weigh 3, level AA weigh 2 and level AAA weigh 1.

This metric's range of values is not defined. Nevertheless, the minimum score it can have is 0, which represents the absence of barriers. The higher this metric's score, the higher the impact of a certain barrier on a specific type of assistive technology/disability.

$$BIF(i) = \sum_{error} error(i) \times weight(i) \tag{22}$$

Equation 22 presents the formula for computing the BIF metric, where $BIF(i)$ is the barrier impact factor of an assistive technology i, $error(i)$ is the number of detected errors that affect the assistive technology i, and $weight(i)$ is the weight of assistive technology i (1, 2 or 3).

2.1.11 Web accessibility quantitative metric (WAQM)

WAQM was proposed by Vigo, et al. [18], and overcomes some limitations of previous measures (i.e., lack of score normalization and consideration of manual tests). It considers the WCAG guidelines classified according to the 4 principles: Perceivable, Operable, Understandable and Robust [13]. This metric measures the conformance using percentages [31], and it produces one score for each WCAG guideline in addition to an overall score. It considers the severity of checkpoint violations according to WCAG priorities and it provides normalized results.

Unlike other metrics, WAQM also takes into account the problems that are identified as warnings by the accessibility evaluation tools [13]. It not only considers automatic tests but also manual tests.

According to Vigo, Arrue, Brajnik, Abascal and Lomuscio [18], this metric was proposed to overcome the drawbacks of the WAB and FR metrics as they do not focus on specific user groups, cover less guidelines and do not consider expert manual evaluation results.

This metric is based on the sum of failure rates for groups of checkpoints which are grouped according to their priority levels and their WCAG 2.0 principles (Perceivable, Operable, Understandable, Robust) [20]. The authors defined weights for each priority level: W1 = 0.8, W2 = 0.16 and W3 = 0.04 for checkpoints with priorities 1, 2 and 3, respectively.

Since WAQM was considered to be tool dependent, there was the need to see if it was possible to prove the opposite [18]. Therefore, Vigo, et al., in their study [18], wanted to have similar outcomes, regardless of the evaluation tool being used. For this matter, the authors proposed a method to reach independence of the tools for every possible scenario. A total of 1363 web pages from 15 websites were evaluated against the WCAG guidelines, using the automated evaluation tools EvalAccess and LIFT. They used 2 different tools to understand the behavior of the WAQM metric when the accessibility is measured by different tools. So, they tuned two WAQM parameters (a and b) to obtain independence. However, WAQM proved to be tool independent when conducting large scale accessibility evaluations with more than 1400 web pages [4].

WAQM's normalized values range from zero to one hundred, where the latter corresponds to the maximum accessibility level.

$$WAQM = \frac{1}{N} \sum_{x \in \{p,o,u,r\}} N_x \sum_{y \in \{e,w\}} \frac{N_{x,y} \sum_{z \in \{1,2,3\}} W_z A(x,y,z)}{N_x} \tag{23}$$

$$A(x,y,z) = \begin{cases} \frac{-100}{b} \frac{B_{x,y,z}}{P_{x,y,z}} + 100, if \frac{B_{x,y,z}}{P_{x,y,z}} < \frac{a-100}{a-100/b} \\ -a\left(\frac{B_{x,y,z}}{P_{x,y,z}}\right) + a, otherwise \end{cases} \tag{24}$$

Equations 23 and 24 present the formulas for computing the WAQM metric, where N is total number of checkpoints, N_x is the number of checkpoints from a specific principle x ($x \in$ {Perceivable, Operable, Understandable, Robust}), $N_{x,y}$ is the number of checkpoints from a principle x and type of test y ($y \in$ {automatic, manual)}, W_z is the weight of the checkpoint, according to its priority level z, $B_{x,y,z}$ is the number of accessibility errors of a checkpoint of priority level z, type of test y and principle x, $P_{x,y,z}$ is the number of test cases of a checkpoint of priority level z, type of test y and principle x, a is a variable that varies between 0 and 100, and b is a variable that varies between 0 and 1.

2.1.12 Navigability and listenability

Fukuda et al. [32] proposed two different web metrics. These metrics are responsible for evaluating the usability for blind users.

Navigability is responsible for evaluating the structure of the web page elements. It evaluates headings, intra-page links, labels, among other HTML elements of a certain web page. Listenability takes into consideration the alternative texts and denotes how properly built they are.

Each of these two metrics executes a set of calculations using the aDesigner (Accessibility Designer) engine. This approach is responsible for the visualization of the Web's usability for blind users through colors and graduations [32].

2.1.13 Web interaction environments (WIE)

Lopes and Carriço proposed, in 2008, a metric that quantifies Web accessibility [33]. It calculates the proportion of checkpoints that are violated on a web page [4]. To do so, it considers a set of checkpoints and, for each of them, it verifies if a checkpoint c is successfully evaluated or if it fails [33]. If it is successfully evaluated, then $v_c = 1$, otherwise $v_c = 0$.

This metric's values have a limited range from zero to one, where one means the web page in question is totally accessible and all checkpoints that were evaluated in that web page have passed.

$$WIE(p) = \frac{\sum v_c}{n} \tag{25}$$

Equation 25 presents the formula for computing the WIE metric, where $WIE(p)$ is this metric's final score for a page p, v_c is a variable that assumes 1 if a checkpoint c passes, otherwise is 0, and n is the number of checkpoints.

2.1.14 Conservative, strict and optimistic

Conservative, Strict and Optimistic are the three web accessibility metrics defined by Lopes, Gomes and Carriço in 2010 [5]. These metrics are based on the results of a checkpoint evaluation of an HTML element: PASS, FAIL or WARN. For each checkpoint, a PASS result indicates that an HTML document compliance is verified; a FAIL result specifies an HTML document compliance that is not verified; and a WARN result specifies it is impossible to verify the HTML document compliance. The main difference between these three metrics resides in the way they consider WARN results. They all contemplate the number of PASS results and the number of applicable elements to evaluate the accessibility results of an automatic accessibility evaluation tool. The conservative metric considers WARN results as failures, the optimistic metric considers them as passes, and the strict metric does not consider them at all.

These three metrics' scores range from zero to one, where one means the web page in question is totally accessible.

$$rate_{\text{conservative}} = \frac{\text{passed+warned}}{\text{applicable}} \tag{26}$$

$$rate_{\text{optimistic}} = \frac{\text{passed}}{\text{applicable}} \tag{27}$$

$$rate_{\text{strict}} = \frac{\text{passed}}{\text{applicable-warned}} \tag{28}$$

Equations 26, 27 and 28 present the formulas for computing the Conservative, Optimistic and Strict metrics, respectively, where *passed* is the number of passes, *applicable* is the number of applicable elements, *warned* is the number of warnings.

2.1.15 eXaminator

According to Benavidez [34], this metric classifies specific situations that can be positive or negative. eXaminator presents a quantitative index that measures the accessibility of a web page. It uses the WCAG 2.0 as a reference. It has two different modes to calculate the qualifications:

- Standard: eXaminator applies all tests. Some of the tests identify errors, while others are responsible to qualify good practices;
- Strict: eXaminator applies only the set of most secure tests, i.e. the tests that have less possibilities of creating false positives or false negatives.

The author considers that not all tests have the same importance, i.e., they need different weights. This means that it is necessary to first weight the tests to make sure their relative weight reflects their differences from each other. The weight calculation is reflected in Eq. 29, and it is the multiplication between the Confidence of the test and the Value. Both Value and Confidence vary between 0 and 1, meaning that the weight will always be a value between 0 and 1. The Value variable depends on the WCAG conformance levels: level A: $V = 0.9$; level AA: $V = 0.5$; level AAA: $V = 0.1$. The Confidence variable verifies, for each test, what procedures are applicable when running it and, for each procedure that cannot be verified, the confidence decreases by 0.1.

eXaminator uses a matrix with information about each test, in particular the Element (E), Situation (S), Note (N), Tolerance (T) and Fraction (F). The Element identifies one or a set of HTML elements and the test is only applied if the element is present in the web page or if the element is *all*. The Situation identifies one or a set of HTML elements that fulfills a certain condition. The Note is the initial qualification of the test that was applied to the first detected situation. It is an absolute value that varies between 1 and 10, where 10 means the test classification result is excellent. The Tolerance is the error tolerance threshold, i.e., indicates the maximum number of errors that are allowed to happen in a specific situation. If the number of errors exceeds the Tolerance, the final test classification decreases by 1 point. Finally, the Fraction variable represents the quantity of errors that decrease the initial note by 1.

The final score of a web page is the ratio between the sum of all tests by the sum of their respective weights. This metric's result uses a scale from 1 to 10, where 1 represents a very bad accessibility level and 10 means otherwise.

$$P = C * V \tag{29}$$

Equation 29 presents the formula for computing the Test Weight, where P is the final weight score, C is the confidence of the test, and V is the value of the test.

Afterwards, there are three different tests that can be applied: True/False tests, Test of proportional type and Test of decreasing type [34].

$$R = N * P \tag{30}$$

Equation 30 presents the formula for computing the True/False tests, where R is the result of the test, N is the Note, and P is the weight of the test.

$$R = N * (1 - S/E) * P \tag{31}$$

Equation 31 presents the formula for computing the Proportional Type tests, where R is the result of the test, N is the Note, S is the Situation, E is the Element, and P is the test weight.

$$R = (N - (S - T)/F) * P \tag{32}$$

Equation 32 presents the formula for computing the Decreasing Type tests, where R is the result of the test, N is the Note, S is the Situation, T is the Tolerance, P is the test weight, and F is the Fraction.

2.1.16 Web accessibility barrier severity (WABS)

Instead of being concerned with conformance to priority levels, this metric focuses on the barriers that limit the accessibility based on their severity. It ranks each web accessibility barrier found in all web pages of a data set of websites [15]. Each barrier has an associated numerical weight according to its severity and impact on the accessibility level. The authors define the barriers' weights as suggested in [18].

The final result of this metric represents a value for each barrier based on the priority class it violates and based on the web page that is being assessed, i.e. the final score will relate to a specific barrier in a certain web page of a website.

It was possible to emphasize two aspects when using this metric. First, each barrier is unique and has different properties, even if it belongs to the same priority level as other barriers, and, second, barriers that belong to priority level one are not necessarily more severe compared to the others.

This metric's formula covers three different measurements [15]:

1. the importance of a barrier to the remaining barriers that violate the same priority level;
2. the importance that barrier has to the web page;
3. the importance of the barrier to all the remaining barriers in all websites.

The formula considers the frequency of a given barrier in a certain web page, the total number of barriers that violate the same priority level in a specific web page, the total number of web pages where a given barrier appears in, the total number of web pages and the total number of barriers that appear in a given web page.

The score result is bounded from zero to one. The closer the result is to zero, the less severe the barrier is.

$$WABS = \frac{\sqrt{\sum_{d=1}^{k} freq(bi)^2}}{\sqrt{\sum_{d=1}^{k} b(pc)^2}} * \frac{n(bi)}{N} * \frac{Pc}{\sqrt{\sum_{d=1}^{k} (b)^2}} \tag{33}$$

Equation 33 presents the formula for computing the WABS metric, where $freq(bi)$ is the frequency of a barrier bi, d is the web page that is being checked, k is the last web page to be tested, $b(pc)$ is the total number of barriers that violate the same priority level pc, $n(bi)$ is the number of web pages the barrier bi appears in, N is the total number of web pages, Pc is the weight of the priority level of a checkpoint, and b is the total number of barriers of that web page d.

2.1.17 Summary

To present an overview about all the 19 web accessibility metrics that were studied and reviewed from the literature, Table 1 gathers important aspects about each metric.

We can also analyze the complexity of the metrics from the perspective of collecting data for their application and resources needed for their computation.

Almost all metrics use the number of barriers or passes found through an accessibility evaluation. The single exception is the WIE that relies on checkpoint compliance, without using more detailed information at the element level, therefore being the easiest to compute in what concerns data collection.

A significant number of metrics also rely on identifying the potential barriers. This is needed for the Failure Rate metric, and for all the other metrics that use the Failure Rate or a similar metric: UWEM, A3, WAB, SAMBA, WAQM and WABS. The Conservative, Optimistic and Strict also require potential barriers (to compute the applicable elements), even though they do not use the Failure Rate, and two of them (the Conservative and the Strict) additionally require the number of warnings. This type of data is traditionally made available from accessibility evaluations, either from automated tools or manual evaluations, with potential barriers being identified from applicable elements that have been marked as pass or warnings.

OAM and PM require the number of attributes and elements. SAMBA requires the number of lines in the HTML document. Even though these might not be standard data resulting from accessibility evaluations, they are not hard to obtain from an HTML parser.

The metrics that potentially raise issues when considering data collection include:

- SAMBA - requires an expert evaluation of the issues identified by an automated tool;

Table 1 19 studied web accessibility metrics

Metric	References	Year of publication	Applicability	Range of values
Failure Rate (FR)	[16]	2000	Page level	0-1, where 0 is totally accessible
Unified Web Evaluation Methodology (UWEM 1.0)	[19]	2006	Page/website level	0-1, where 0 is totally accessible
A3	[22]	2006	Page level	0-1, where 0 is totally accessible
Web Accessibility Barriers (WAB) by Hackett	[23]	2004	Website level	the higher its score, the worse the accessibility level
Web Accessibility Barriers (WAB) by Parmanto and Zeng	[14]	2005	Website level	the higher its score, the worse the accessibility level
Web Accessibility Evaluation Metric (WAEM)	[29]	2017	Website level	N/A
Web Accessibility Quantitative Metric (WAQM)	[18]	2007	Page level	0-100, where 100 is totally accessible
Web Interaction Environments (WIE)	[33]	2008	Page level	0-1, where 1 is totally accessible
Conservative	[5]	2010	Page level	0-1, where 1 is totally accessible
Strict	[5]	2010	Page level	0-1, where 1 is totally accessible
Optimistic	[5]	2010	Page level	0-1, where 1 is totally accessible
Overall Accessibility Metric (OAM)	[25]	2005	Page level	the higher its score, the worse the accessibility level
Page Measure (PM)	[26]	2007	Page level	the higher its score, the worse the accessibility level
SAMBA	[27]	2007	Page level	N/A
Reliability Aware Web Accessibility Experience Metric (RA-WAEM)	[12]	2018	Website level	N/A
Barrier Impact Factor (BIF)	[35]	2011	Assistive technologies/Disability types	the higher its score, the higher the impact of the barrier on an assistive technology/disability
Navigability and Listenability	[32]	2005	Page level	N/A
Web Accessibility Barrier Severity (WABS)	[15]	2017	Accessibility barriers	0-1, where the closer the score is to 0, the less severe the barrier is
eXaminator	[34]	2012	Page level	1-10, where 1 is totally inaccessible

- WAEM and RA-WAEM - require user experience evaluations by users with disabilities for obtaining the PUEXO pairs; and
- BIF - requires classification of the barriers (errors in BIF) by assistive technology;

In summary, from a data collection perspective, the data required by most metrics should be easily accessible from accessibility evaluations. The exception is data that requires human intervention, be it from experts that classify outcomes of evaluations, or from user tests.

A different type of issues, not related to data collection, is raised by eXaminator. Tests in eXaminator need the definition of multiple parameters. Even though this needs to be done only once, these parameters are not available when data is collected through different tools or methodologies.

In what concerns the complexity of computing the metric, only one metric group stands out. WAEM and RA-WAEM, by running a vector optimization procedure, require more computing resources than the other metrics. The time complexity of the other metrics is linear. From these, more computing resources are required by A3 and WABS, for computing exponentiation and square root operations, respectively.

3 Comparing accessibility metrics

To compare a subset of the metrics presented in the previous section, we planned a study based on a large-scale evaluation of web pages.

3.1 Methodology

In this section we present the methodology followed in the study. We introduce the automated evaluation tool we used and the data set that was evaluated. We then describe what metrics we were able to compare, based on the constraints of running a large-scale study, which prevented us from comparing metrics that rely on human judgment. We also describe how the metrics were implemented, based on the results provided by the used tool. We conclude this section with a description of how we analyzed the data.

3.1.1 QualWeb and evaluation data set

To run a study based on a large number of accessibility evaluations of web pages, the only viable option is to conduct automated accessibility assessments [36]. To run those evaluations, we used QualWeb[2] [37].

QualWeb is an automated web accessibility engine. It performs a set of tests on a web page that check conformance with ACT-Rules[3] and WCAG Techniques 2.1[4]. For each web page evaluated, we extracted from the QualWeb report the number of elements that passed, the number of elements that failed and the number of warnings for each test. We also collected information about the test being applicable or not to the web page. This information is useful since we want to consider only applicable tests when computing the metrics. When an applicable test passes, it means that it has no failures nor warnings. If an applicable test returns no errors, but has at least one warning, the test outcome is “warning”. If the test has at least one element that fails, the test fails.

As previously mentioned, QualWeb has two types of tests: ACT-rules tests, which test a web page against a set community approved checks; and WCAG techniques, which test a web page against the tool developer’s interpretation of specific WCAG techniques. To ensure that only checks that correspond to consensual interpretation of the WCAG are used and increase the validity of the results, we used only the outcomes of the ACT-Rules tests in this study. In our study we used the 0.6.1 version of QualWeb, which tested a total of 72 ACT-Rules.

QualWeb was used to evaluate a total of 2,884,498 web pages. The pages were obtained from CommonCrawl[5]. CommonCrawl is an open corpus of web crawl data, including metadata and source of billions of web pages since 2013. We used the most recent crawls to obtain the URLs of the pages. The pages were evaluated in the period from March 2021 to September 2021. The 2,884,498 pages correspond to a total of 166,311 websites, averaging 21 pages per website. The distribution of pages and websites per top level domain is presented in Table 2.

Table 2 Number of web pages by top-level domain

Top-level Domain	Number of web pages
.asia	322,208
.au	174,371
.com	166,388
.org	154,154
.pt	139,807
.gov	130,689
.info	125,233
.uk	122,543
.es	120,085
.it	120,085
.fr	113,513
.de	109,135
.us	106,432
.news	105,196
.eu	102,995
.net	102,671
.br	100,353
.edu	93,956

The evaluation found a total of 86,644,426 errors, averaging 30 errors per page and 521 errors per website. The highest number of errors on a single web page was 15,645 and the lowest was zero. The highest number of errors on a website was 878,776 and the lowest zero. The ACT-Rules violated in most pages were: ACT-R76 (Text has enhanced contrast), ACT-R37 (Text has minimum contrast) and ACT-R12 (Link has accessible name). The ACT-Rules with the highest number of errors were ACT-R76 (Text has enhanced contrast) having 33,109,298 errors.

3.1.2 Applicable metrics

The analysis of the accessibility metrics shows that not all metrics can be studied with this data set. For instance, metrics that require human judgment cannot be considered, since it is not viable to produce expert judgment over such a large set of pages. Therefore, we needed to identify the ones that could be computed with our data. From the 19 presented metrics, we found that 11 metrics could be computed with our dataset composed from the ACT-Rules evaluation results. Most of the applicable metrics use WCAG 1.0 which considers checkpoints rather than success criteria. Since we used WCAG 2.1 in our accessibility evaluation, when computing the accessibility metrics, we will refer to the checkpoints as success criteria. Each ACT-Rule has corresponding success criteria. Each success criterion has an associated principle and conformance level. Through these success criteria, it was possible to define which principle(s) and conformance level(s) characterize each test. As one test can have more than one success criterion, it can also have more than one principle or priority level. This information is required by some of the metrics.

[2] http://qualweb.di.fc.ul.pt/evaluator/.

[3] https://act-rules.github.io/rules/.

[4] https://www.w3.org/TR/WCAG21/.

[5] https://commoncrawl.org/.

Table 3 Constants used to compute the WAQM metric

Constants	Description	Value
Nall	Number of all tests	72
N	Number of applicable tests	51
Np	Number of Perceivable tests	28
No	Number of Operable tests	11
Nu	Number of understandable tests	7
Nr	Number of Robust tests	11
Npe	Number of automatic perceivable tests	26
Noe	Number of automatic operable tests	11
Nue	Number of automatic understandable tests	6
Nre	Number of automatic robust tests	10
Npw	Number of manual perceivable tests	5
Now	Number of manual operable tests	4
Nuw	Number of manual understandable tests	1
Nrw	Number of manual robust tests	1
a	Constant	20
b	Constant	0,3

The FR metric is the simplest metric to compute. It requires the number of potential and actual problems. For each page, the sum of all elements that failed a test and the sum of all elements applicable to the test are computed. Having both totals for all the tests, it is possible to calculate the failure rate of the page. It is important to highlight that some tests might evaluate the same elements of the page. However, they evaluate different aspects and so they cannot be counted only once, since we are considering the total number of failures and not the total number of elements that failed. For the remaining accessibility metrics that utilize the number of potential and actual points of failure for success criteria, the same logic was applied.

The WAQM metric is the most complex to compute. It considers the priority level and its weight, the type of the test, i.e., if it is manual or automatic, and the principle(s) of each test. WAQM is computed for each test and its computation relies on a number of parameters. Table 3 presents the parameters we used. Parameters *a* and *b* are constants because "the tuning was not necessary because WAQM proved to be independent of the tool when conducting large-scale evaluations (approx. 1400 pages)" [4]. The other parameters were tuned to the QualWeb tool and the ACT-Rules it tests.

It was also possible to compute the UWEM and A3 metrics for each web page, since they both rely on the FR of each checkpoint of that page. Since both metrics are computed using a weight that is obtained from user feedback, we had to determine this weight according to the priority levels, due to time and resources constraints. UWEM already calculates a score for each website, by calculating its web pages' average score. Besides applying the UWEM metric to websites as Vigo, et al. define [17], we decided to additionally use another procedure to convert this metric into a website metric, as will be described further on.

WIE, Conservative, Optimistic and Strict are four simple metrics that can be easily applied with our data, as they only require the number of applied success criteria, the number of elements, the number of warnings, the number of fails and the number of passes.

In what concerns metrics that are applicable to websites, instead of web pages, we considered WAB by Parmanto and Zeng and WAB by Hackett. The two WAB formulas were applied as one requires the priority level and the other one the weight of the priority level. Both formulas also calculate the failure rate and consider the total number of web pages a website contains.

Other metrics like SAMBA or eXaminator were not considered, either because of the lack of information in our data or the fact that the metric is semi-automated, which means that it needs manual intervention. For instance, the indexes that are computed in SAMBA concern the disability type; and eXaminator considers information about HTML elements that are evaluated in each page. Yet, we could partially use the WAEM/RA-WAEM metric. Since both WAEM and RA-WAEM require users' intervention as they evaluate pairs of websites, i.e. PUEXO pairs, to be compared to the results of the weighted accessibility score computation (Eq. 16), we could only consider the weighted accessibility score (Eq. 15) that can be automatically computed. This score is used in the WAEM and RA-WAEM metrics' process to classify a website and to compare the results with user classifications, and it considers the number of pages a success criteria passes in that website.

We did not consider the OAM nor the Page Metric metrics since they both consider the number of HTML attributes in their formulas. We do not have that information from the QualWeb reports. Also, we did not consider the two metrics by Fukuda et al. [32] since we do not have information regarding the aspects both formulas take into consideration (alt attributes, reaching time of a given element, page size).

We could not apply BIF since it needs a table that relates the errors that were identified by the accessibility evaluation tool with the assistive technologies that are affected by these errors. For this reason, it is not viable to attend to all the errors of our 2.8 million web pages sample and identify which assistive technologies are affected by them.

WABS was not considered since it classifies the accessibility barriers based on their severity, which means that it refers to the severity of a barrier that was identified in a set of websites and their respective web pages. Thus, this metric is focused on a specific problem that hinders the user's interaction. The final result of this metric would be a list of barriers that were found in our web pages data set and their respective severity scores. For this reason, it is

Table 4 Descriptive statistics for web page metrics

Metric	Average	Standard deviation	Best score	Worst score	First quartile	Third quartile
FR	0.0673	0.0780	0	1	0.0201	0.0856
A3	0.6657	0.3203	0	1	0.4077	0.9443
UWEM	0.3842	0.3212	0	0.9997	0.1010	0.800
WAQM	82.8626	19.1529	100	0	76.3289	95.6360
WIE	0.5545	0.1561	1	0	0.4375	0.6667
Conservative	0.3936	0.2316	1	0	0.2018	0.5556
Optimistic	0.6015	0.2314	1	0	0.4453	0.7852
Strict	0.4973	0.2692	1	0	0.2658	0.7209

not possible to correlate metrics that evaluate the accessibility of web pages with metrics that evaluate accessibility barriers.

The following list summarizes what metrics were analyzed in our study.

- Web page metrics:
  - Failure-rate (FR);
  - Unified Web Evaluation Metric (UWEM);
  - A3;
  - Web Accessibility Quantitative Metric (WAQM);
  - Web Interaction Environments (WIE);
  - Conservative;
  - Optimistic;
  - Strict.
- Website metrics:
  - Web Accessibility Barriers (WAB) by Hackett;
  - Web Accessibility Barriers (WAB) by Parmanto & Zeng;
  - Web Accessibility Evaluation Metric (WAEM).

3.1.3 Metrics comparison and analysis

With our goal being to understand the similarities between different accessibility metrics, we computed their correlation pairs. We tested the normality of the data using the Shapiro-Wilk and Kolmogorov-Smirnov tests. We found that our data did not follow a normal distribution. Therefore, we used the Spearman correlation in our analysis. Following the recommendations from Statstutor [38], absolute correlation values above 0.4 represent moderate or stronger correlation. In our analysis we considered two metrics to have similar results when they are at least moderately correlated.

It is important to take into consideration the fact that some metrics are applicable to websites whereas others are applicable to web pages. To be able to compare all metrics, the web page metrics were converted to web site metrics via two different approaches:

1. Computing the metric score based on the sum of the evaluation results of all the pages of the website; and
2. Calculating the average of the metric score for all web pages of the web site, similar to the UWEM strategy.

Besides analyzing the pairwise similarity obtained from the correlation, we used this information to cluster the correlation scores and find if groups of metrics present similar behaviors in our data set. For this analysis we used hierarchical clustering [39].

3.2 Results

This section presents the results of the metrics comparison study. We begin by presenting an overview of the outcomes of each metric in the full set of web pages evaluated. We then examine the similarity between metrics across different contexts: metrics over web pages, and metrics over websites, exploring both ways previously introduced to compute a website metric from web page metrics.

3.2.1 Descriptive statistics

Table 4 presents descriptive statistics of the scores for all metrics that are applicable at page level.

Regarding the descriptive statistics for web page metrics, we can observe some worthwhile points. The FR metric average indicates a very optimistic perspective on the accessibility of the evaluated web content. Additionally, the standard deviation is very small, indicating that the web pages scores do not vary much from the average. Also, WAQM presents a positive perspective about the accessibility of the Web, as the average is approximately 83, which is close to the score that expresses the highest accessibility level. The UWEM metric is slightly positive concerning the accessibility, having an average of 0.38 and given the fact that the lower the score, the more accessible the web page is. The WIE, Optimistic and Strict metrics present an intermediate accessibility level average. In contrast, A3 and Conservative metrics report a negative perspective about the accessibility

Table 5 Descriptive statistics for website metrics, adding the evaluation results of all website pages

Metric	Average	Standard deviation	Best score	Worst score	First quartile	Third quartile
FR	0.0832	0.0836	0	1	0.0296	0.1080
A3	0.8390	0.2713	0	1	0.8301	1.0000
UWEM	0.4728	0.3294	0	0.9997	0.1715	0.8131
WAQM	79.4362	21.7414	100	0	71.7592	94.8497
WIE	0.5176	0.1662	1	0.04	0.400	0.6250
Conservative	0.4327	0.2191	1	0.0006	0.2640	0.5799
Optimistic	0.6366	0.1994	1	0.0006	0.5065	0.7857
Strict	0.5390	0.2410	1	0.0006	0.3515	0.7273
WAB-H	0.4742	0.6927	0	5.8333	0.0263	0.6875
WAB-PZ	0.3053	0.4799	0	4.2	0.0133	0.400
WAEM	4.1765	1.3273	8.68	0.0072	3.3353	5.1446

Table 6 Descriptive statistics for website metrics, considering the average of the website pages' metric scores

Metric	Average	Standard deviation	Best score	Worst score	First quartile	Third quartile
FR	0.0859	0.0849	0	1	0.0312	0.1111
A3	0.6744	0.3007	0	1	0.4569	0.9255
UWEM	0.4433	0.3208	0	0.9997	0.1562	0.800
WAQM	77.9130	21.3167	100	0	70.1056	93.1619
WIE	0.5715	0.1562	1	0.0588	0.4645	0.6797
Conservative	0.4457	0.2182	1	0.00055	0.2777	0.5955
Optimistic	0.6470	0.1970	1	0.00055	0.5178	0.7968
Strict	0.5527	0.2384	1	0.00055	0.3688	0.7402
WAB-H	0.4742	0.6927	0	5.8333	0.0263	0.6875
WAB-PZ	0.3053	0.4799	0	4.2	0.0133	0.400
WAEM	4.1765	1.3273	8.68	0.0072	3.3353	5.1446

of the evaluated web pages, as their values are closer to the inaccessible reference.

Tables 5 and 6 present descriptive statistics of the scores of all metrics at the website level. Table 5 shows results where scores for page level metrics were calculated by adding the evaluation results for all pages of the same website. Table 6 results were calculated averaging the page metric results for all pages of the same website.

In relation to the descriptive statistics for website metrics, some metrics reported consistent behavior while for others some differences to the web page metrics could be observed. The FR average indicates that the accessibility of the evaluated web content is very optimistic, as it was observed in this metric's web page version. Also, WAQM still presents a more positive perspective about the accessibility of the Web, as it was stated in the descriptive analysis of the web page metrics. The UWEM metric has slightly increased its average when compared with this metric applied to web pages. Nevertheless, it still provides a positive perspective about accessibility. The Optimistic and Strict website metrics' average also increased, yet they do not show a clear difference compared to the web pages metrics. The average of the website scores using the WIE metric decreased when the evaluation results for all pages of the same website was considered, compared to this metric's web page version. The same did not happen when considering the average of the page metric results for all pages of the same website, as it shows an increase on its average. Conservative, as a website metric, also has a similar negative perspective about web accessibility compared to the same metric applied to web pages. The A3 metric for websites, in particular, considering the average of the website pages' scores, did not show a noticeable difference in the average result, compared to the A3 metric for web pages (around 0.67). Nevertheless, a considerable difference between these last two approaches for A3 metric was detected in the website level, concerning a website as a web page, having an average of approximately 0.84. Interestingly, the WAB-H and WAB-PZ metrics reveal differences in their averages that might be justified from the worst scores. WAB-H evaluated a website that had a score of 5.8333, which represents the most inaccessible website. Yet, the WAB-PZ worst score was 4.2, which indicates that the accessibility issues are less weighted compared to WAB-H. Since WAB-PZ, WAB-H and WAEM do not provide a limited range of values, it is more complicated to define the accessibility level by their scores.

Table 7 Spearman correlation scores for web page metrics (moderate, strong and very strong correlation scores are displayed in bold)

	FR	A3	UWEM	WAQM	WIE	Conservative	Optimistic	Strict
FR	1							
A3	0.0008	1						
UWEM	0.0008	**0.8375**	1					
WAQM	0.0002	−0.0173	-0.0175	1				
WIE	−0.0006	**−0.6342**	**−0.4963**	0.0099	1			
Conservative	0.0001	−0.0178	0.0226	0.0061	0.0403	1		
Optimistic	0.0004	−0.0209	0.0240	0.0070	0.0456	**0.9042**	1	
Strict	0.0003	−0.0193	0.0231	0.0068	0.0427	**0.9706**	**0.9733**	1

bold represents scores that have moderate or higher correlation

Table 8 Spearman correlation scores for website metrics (moderate, strong and very strong correlation scores are displayed in bold)

	WAEM	WAB-H	WAB-PZ
WAEM	1		
WAB-H	−0.1183	1	
WAB-PZ	−0.1850	**0.9858**	1

bold represents scores that have moderate or higher correlation

As can be observed from Table 1, different metrics have scores ranging from 0 to 1, others from 0 to 100, and yet others are unbounded. Bounded ranges support easier to interpret results, by allowing to compare a score with the limits of the range. For example, one intuitively expects that an UWEM score close to zero represents an accessible page, while a score close to one represents an inaccessible page. With unbounded ranges, since there is only one limit, this comparison is not always possible. For example, a WAB score close to zero represents an accessible page, but what about a score of 1? Or 5? The data collected in this study, and presented in Tables 4, 5 and 6, provides not only a reference for the unbounded ranges (the extreme value for WAB-H was 5.83, for WAB-PZ was 4.2, and for WAEM was 8.68) but also gives an indication of how metric scores are distributed. This allows us to interpret values from the metrics more precisely. For example, a UWEM score of 0.5 probably represents a web page that is less accessible than an A3 score of 0.5. The first quartile for A3 is 0.4077 (considering web page metrics), which is much closer to 0.5 than the first quartile for UWEM (0.101).

3.2.2 Web page metrics

Table 7 presents the Spearman correlation coefficients between every pair of page metrics.

The highest correlation score obtained was between the Strict and Optimistic metrics ($\rho = 0.9733$), followed by Conservative and Strict metrics ($\rho = 0.9706$) and Optimistic and Conservative metrics ($\rho = 0.9042$). They seem to have a very strong positive correlation, since they are all based on the ratio of passed tests over applicable tests, differing only on how warnings are considered. Possibly, the number of warnings classified by QualWeb was not high enough to ensure clear differences between the results of each of these three metrics.

A3 shows a strong positive correlation with UWEM ($\rho = 0.8375$), which is expected since these two metrics are similar. Also, A3 has a strong negative correlation with WIE ($\rho = -0.6342$). Since WIE considers the number of elements that pass, the higher the WIE score, the higher the accessibility level of the web page. A3 shows an opposite behavior. Similarly to A3, UWEM shows a moderate negative correlation with WIE ($\rho = -0.4963$). This behavior might be explained from the fact that both UWEM and A3 consider the number of elements that failed while WIE considers the number of success criteria that passed in a page. If a page that fails all the success criteria that are tested, also fails one element per test, the number of failed elements will be similar to the number of failed success criteria.

Interestingly, no other pairs of metrics are correlated. This means that FR and WAQM are not correlated to any other metric.

3.2.3 Website metrics

In the present study we considered 3 accessibility metrics that are exclusively applied at website level: WAEM, WAB-H and WAB-PZ. Table 8 presents the Spearman correlation coefficients between all pairs of metrics.

The WAB metrics show a very strong positive correlation as expected ($\rho = 0.9858$), since they share the same formula, with just one little difference: WAB by Hackett considers the priority level (1, 2 or 3) of the checkpoint (success criterion, in our case), whereas WAB by Parmanto and Zeng considers the weight of the checkpoint priority level ($P1 = 0.8$, $P2 = 0.16$ and $P3 = 0.04$). Both WAB metrics are not correlated with WAEM.

Domain as a web page

Table 9 Spearman correlation scores for website metrics, considering a domain as a web page (moderate, strong and very strong correlation scores are displayed in bold)

	FR	A3	UWEM	WAQM	WIE	Conservative	Optimistic	Strict	WAB-H	WAB-PZ	WAEM
FR	1										
A3	0.1779	1									
UWEM	**0.4442**	**0.4131**	1								
WAQM	**−0.5423**	−0.2140	**−0.7285**	1							
WIE	−0.1154	**−0.5942**	−0.3649	0.2485	1						
Conservative	−0.0188	−0.3528	−0.0967	0.0333	**0.4838**	1					
Optimistic	−0.1336	−0.3519	−0.1329	0.0775	0.4523	**0.8759**	1				
Strict	−0.0984	−0.3730	−0.1283	0.0673	**0.4898**	**0.9573**	**0.9704**	1			
WAB-H	**0.4217**	−0.2910	**0.5698**	**−0.5675**	−0.0233	−0.0024	−0.0556	−0.0378	1		
WAB-PZ	**0.4071**	−0.2222	**0.6457**	**−0.6249**	−0.0525	−0.0098	−0.0591	−0.0434	**0.9858**	1	
WAEM	−0.3125	**−0.5252**	**−0.5057**	**0.4902**	**0.6566**	0.2870	0.2764	0.3044	−0.1183	−0.1850	1

bold represents scores that have moderate or higher correlation

Table 10 Spearman correlation scores for website metrics, considering the average of the web pages' scores (moderate, strong and very strong correlation scores are displayed in bold)

	FR	A3	UWEM	WAQM	WIE	Conservative	Optimistic	Strict	WAB-H	WAB-PZ	WAEM
FR	1										
A3	**0.4568**	1									
UWEM	**0.4914**	**0.8612**	1								
WAQM	**−0.5606**	**−0.7167**	**−0.7917**	1							
WIE	−0.2018	**−0.6053**	**−0.4411**	**0.3914**	1						
Conservative	0.0014	−0.3129	−0.0916	0.0177	**0.5604**	1					
Optimistic	−0.1283	−0.3499	−0.1392	0.0748	**0.5253**	**0.8740**	1				
Strict	−0.0910	−0.3567	−0.1332	0.0630	**0.5685**	**0.9536**	**0.9718**	1			
WAB-H	0.4083	**0.5310**	**0.6161**	**−0.5036**	−0.2607	−0.0262	−0.0813	−0.0646	1		
WAB-PZ	0.3992	**0.5957**	**0.6846**	**−0.5665**	−0.2798	−0.0291	−0.0816	−0.0666	**0.9858**	1	
WAEM	−0.3390	**−0.5563**	**−0.5077**	**0.5369**	**0.5968**	0.2577	0.2583	0.2839	−0.1183	−0.1850	1

bold represents scores that have moderate or higher correlation

Table 9 presents the Spearman correlation coefficients for all metrics, with the metrics being computed by considering the domain as a page with the evaluation results of all the pages of the domain.

As it would be expected, and matching to the web pages correlation scores, the Conservative metric has a very strong and positive correlation with Optimistic and Strict: $\rho = 0.8759$ and $\rho = 0.9573$, respectively. Also, as observed in the web pages scores, Strict and Optimistic metrics still have the same strong positive correlation ($\rho = 0.9704$). WAEM appears to have a strong positive correlation with WIE ($\rho = 0.6566$). A3 has a moderate positive correlation with UWEM ($\rho = 0.4131$) and it is negatively correlated with WAEM ($\rho = -0.5252$) and WIE ($\rho = -0.5942$). UWEM has a moderate negative correlation with WAEM ($\rho = -0.5057$), while WAQM has a moderate positive correlation with WAEM ($\rho = 0.4902$). UWEM shares a strong negative correlation with WAQM ($\rho = -0.7285$) and positive correlation with both WAB-PZ ($\rho = 0.6457$) and WAB-H ($\rho = 0.5698$). In contrast to UWEM, WAQM has negative correlations with WAB-PZ ($\rho = -0.6249$) and WAB-H ($\rho = -0.5675$). FR shows a positive moderate correlation with UWEM ($\rho = 0.4442$), with WAB by Hackett ($\rho = 0.4217$) and with WAB-PZ ($\rho = 0.4071$). It presents a moderate negative correlation with WAQM ($\rho = -0.5423$).

Average of the web pages' scores

Table 10 presents the Spearman correlation coefficients for all metrics, with the domain metric being computed by averaging the metrics of the pages belonging to the domain.

As expected, the Conservative metric has a very strong and positive correlation with Optimistic and Strict: $\rho = 0.8740$ and $\rho = 0.9536$, respectively. These values are very similar to the ones obtained when considering a domain as a web page. Also, as observed in the web pages scores, Strict and Optimistic metrics have the same strong positive correlation ($\rho = 0.9718$). Unlike results obtained from considering a domain as a web page, A3 has a very strong positive correlation with UWEM ($\rho = 0.8612$). Since A3 and UWEM are very similar, having a strong correlation, WAQM also shares similar correlations with both metrics: $\rho = -0.7167$ and $\rho = -0.7917$, respectively. The same happens with WAB-PZ and WAB-H. Since these two metrics are strongly correlated, their correlations with WAQM are also very similar: $\rho = -0.5665$ and $\rho = -0.5036$, respectively.

UWEM also has a strong positive correlation with WAB-PZ ($\rho = 0.6846$) and WAB-H ($\rho = 0.6161$). A3 has a similar correlation score with WAB-H ($\rho = 0.5310$) and WAB-PZ ($\rho = 0.5957$). However, it has a negative correlation with the remaining metrics: moderate with WAEM ($\rho = -0.5563$), and strong with WIE ($\rho = -0.6053$) and WAQM ($\rho = -0.7167$).

UWEM and WAQM have moderate correlation with WAEM: $\rho = -0.5077$ and $\rho = 0.5369$, respectively. They both share a moderate to almost moderate correlation with WIE: $\rho = -0.4411$ and $\rho = 0.3914$ WAEM appears to have a moderate positive correlation with WIE ($\rho = 0.5968$). WIE correlation scores with Conservative, Optimistic and Strict metrics vary between 0,52 and 0,57.

FR shows a positive correlation, although moderate, with UWEM ($\rho = 0.4914$) and A3 ($\rho = 0.4568$). However, FR presents a negative correlation with WAQM ($\rho = -0.5606$).

3.3 Discussion

To identify the groups of metrics, we used hierarchical clustering [39] that groups similar metrics into clusters according to the correlation matrices. To define the number of clusters, we had to cut the clustering tree in order to define the different clusters. To obtain relevant clusters, we analyzed the dendrograms to determine the best cluster distance where we would cut the clustering tree. With respect to web page clusters, we decided to cut the clustering tree where the cluster distance is 1. Concerning the website metrics clusters, we cut the clustering tree where the cluster distance is approximately 0.7. With these choices, we aimed at having a reasonable number of clusters, while avoiding clusters with elements that are too far apart. The cuts are represented in the dendrograms by the red or yellow horizontal lines.

With regards to the web accessibility metrics' results, and concerning the web page metrics, we could find four groups: Conservative, Strict and Optimistic; WIE, A3 and UWEM; FR; and WAQM. The clusters are illustrated in Figure 1.

The Conservative, Strict and Optimistic metrics have similar formulas based on the number of passed tests over applicable tests, differing on whether warnings are considered passes, fails or not applicable. For this reason, it is expected that they produce similar outcomes and are clustered together.

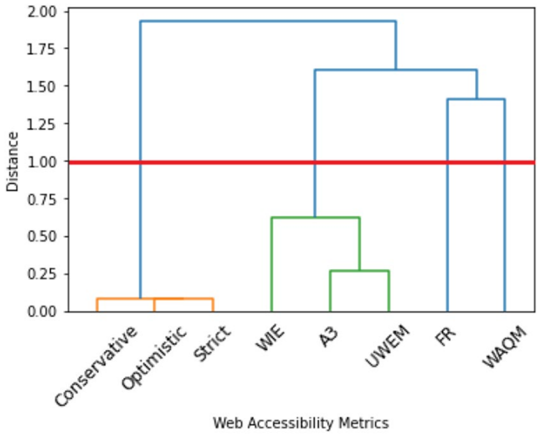

Fig. 1 Clusters of web page metrics

A3 and UWEM also have very similar formulas, which justifies the high correlation between their outcomes. While A3 and UWEM are based on the ratio of actual and potential barriers, WIE considers pass rates of checkpoints. Even though those perspectives of measuring the accessibility of web pages are different, the fact is that they seem to be correlated. This might be relevant information when deciding on using one of these metrics, since WIE requires information that is easier to obtain and less resources to compute than A3 or UWEM.

WAQM did not form a group with FR, as their distance is higher than 1. The distance between these two metrics may be significant, since their correlation is almost null.

With respect to website metrics, in particular interpreting a website as a web page, the following five clusters were identified, as represented in Figure 2:

- Conservative, Strict and Optimistic;
- A3, WIE and WAEM;
- FR;
- WAB-H, WAB-PZ;
- UWEM and WAQM.

From these groups, only one cluster is similar to the web page metrics' groups. In fact, the web pages cluster WIE, A3 and UWEM is similar to one cluster of the website metrics (A3, WIE and WAEM), except the fact that the UWEM metric is not included into the A3 and WIE group.

The main difference is the fact that it now includes the correlations with the website metrics (WAB-H, WAB-PZ and WAEM). Besides this inclusion, the WAQM seems to be closer to UWEM, compared to the web page metrics' results. This could be happening because all the web pages data are grouped together to provide the website final score,

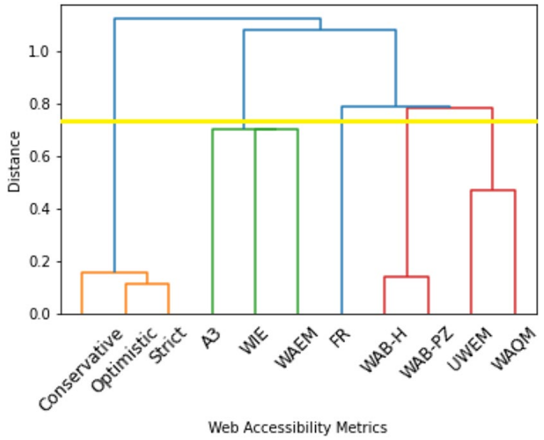

Fig. 2 Clusters of website metrics, interpreting a website as a web page

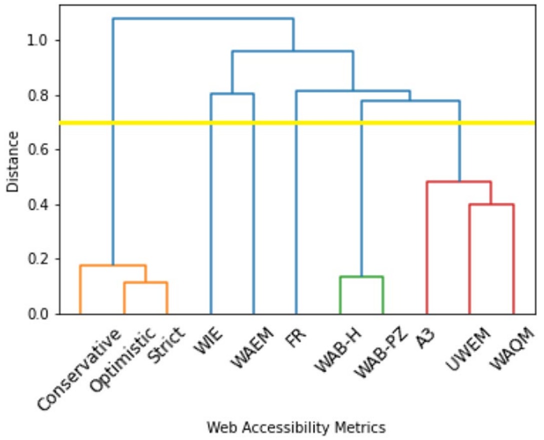

Fig. 3 Clusters of website metrics, calculating the average of the web pages' scores

modifying the metrics' behavior. The FR is still distant from the remaining metrics. WAQM and UWEM calculate the failure rate, which might justify the cluster they are grouped in. WAB-PZ and WAB-H had a very strong correlation, so it was expected they would be part of the same cluster. They share similar formulas that only differ in the way they consider the success criterion weight: as the priority level or the weight of the priority level. WIE may relate with WAEM, since they both consider when a checkpoint passes: WIE increments one every time a certain checkpoint passes on a website, while WAEM counts the number of pages where a checkpoint passes. For instance, if a certain checkpoint passes on a website that only has one web page, it will count as one for WIE and also for WAEM. The main difference between them is that WIE considers the total number of checkpoints, while WAEM not only considers the number of website pages but also the weight of the checkpoint.

Regarding the average of the web pages' scores, we identified the following 6 clusters (Figure 3):

- Conservative, Strict and Optimistic;
- FR;
- A3, UWEM and WAQM;
- WIE;
- WAEM;
- WAB-H and WAB-PZ.

The above groups show that Conservative, Strict and Optimistic metrics belong to the same cluster, as seen before. However, there are two main differences when comparing with the other website metrics approach: (1) A3 is now part of the UWEM and WAQM cluster; (2) WAEM and WIE do not belong to the same cluster, as their cluster distance is higher than the previously defined threshold.

Another interesting aspect is the fact that Conservative, Optimistic and Strict are always in the same independent group, in all the three approaches we have mentioned. Perhaps because the number of warnings of the considered domains and web pages is not that significant to the point of changing these metrics' results, since the only difference between these three metrics' formulas is the way the warnings are considered.

In all metrics' clusters, the FR metric does not form a group with any other metric, even though some of them incorporate the failure rate in their formulas, indicating that their results do not correlate with the FR metric results. Thus, we can recognize that the metrics that integrate the FR, consider other important information in their scope that makes them different from the FR. For instance, WAQM is more complex than FR, taking into account the principles, the type and the priority levels of success criteria. For this reason, when opting for one of these metrics, FR seems a more straightforward and easy choice, but it should be kept in mind that the other metrics may give more relevant information.

4 Metric validity

Our analysis allows detecting what metrics produce similar outcomes, but it does not reflect the validity of the outcomes. This is a result of the fact that the metrics were computed from a set of automated evaluation results. Automated evaluation tools are only capable of identifying a subset of the real accessibility problems in a web page. Therefore, a page that gets a good outcome on an automated accessibility evaluation might have undetected accessibility problems. This means that a metric computed

Table 11 Accessible and inaccessible pages

Web pages	
Accessible web pages	Inaccessible web pages
https://wsnet2.colostate.edu/cwis24/acns/web-accessibility/Example	https://wsnet2.colostate.edu/cwis24/acns/web-accessibility/Example/index-inaccessible.html
https://www.w3.org/WAI/demos/bad/after/home.html	https://www.w3.org/WAI/demos/bad/before/home.html
https://www.washington.edu/accesscomputing/AU/after.html	https://www.washington.edu/accesscomputing/AU/before.html
https://www.w3.org/WAI/demos/bad/after/template.html	https://www.w3.org/WAI/demos/bad/before/template.html

Table 12 Web page metrics scores for assessing the metrics' validity (scores that do not reflect the accessibility of the web page are displayed in bold)

	FR	A3	UWEM	WIE	Conservative	Optimistic	Strict	WAQM
Accessible								
https://wsnet2.colostate.edu/cwis24/acns/web-accessibility/Example	0,00218	0	0	0,875	0,8297	0,9454	0,93827	98,823
https://www.w3.org/WAI/demos/bad/after/home.html	0,01954	0,02192	0,0057	0,6	**0,3909**	0,70684	0,5714	89,005
https://www.washington.edu/accesscomputing/AU/after.html	0,00998	0	0	0,7333	0,525	0,6367	0,5910	90,667
https://www.w3.org/WAI/demos/bad/after/template.html	0,0185	0,0192	0,0057	0,6	**0,384**	0,7159	0,5746	89,005
Inaccessible								
https://wsnet2.colostate.edu/cwis24/acns/web-accessibility/Example/index-inaccessible.html	**0,271**	0,997	0,967	0,308	0,1050	0,2514	0,1230	33,2218
https://www.w3.org/WAI/demos/bad/before/home.html	**0,1453**	0,9998	0,939	0,267	0,1738	0,444	0,238	38,6422
https://www.washington.edu/accesscomputing/AU/before.html	**0,0704**	0,995	0,915	**0,6154**	**0,7605**	**0,901**	**0,885**	51,7320
https://www.w3.org/WAI/demos/bad/before/template.html	**0,1448**	0,999	0,9518	0,2667	0,1785	0,468	0,251	35,420

bold represents scores that do not reflect the accessibility

on that evaluation might indicate an accessibility level that is better than the reality.

4.1 Methodology

To investigate what metrics might better reflect the actual accessibility level of web pages, we conducted a further analysis. Since it was not feasible to conduct manual assessments of the accessibility of the large data set, we compiled a small data set composed of web pages that are published online with the purpose of demonstrating good and bad accessibility practices. Table 11 presents the web pages we considered in our analysis. Two pairs of web pages were developed by universities while the other two pairs are part of the Before and After Demonstration (BAD)[6] published by the W3C Web Accessibility Initiative. They have been created mostly for educational purposes, providing instructors with ready to access examples of how web content should be designed to be accessible, but also how it could be designed in an inaccessible way. One limitation common to all the pages is their age. All of them were developed prior to the publication of WCAG 2.1. Therefore, some of the criteria introduced in the WCAG 2.1 will not have been explicitly explored in

Table 13 Website metrics scores for assessing the metrics' validity (scores that do not reflect the accessibility of the website are displayed in bold)

	WAEM	WAB-PZ	WAB-H
Accessible Domains			
wsnet2.colostate.edu	11.06	0.0008	0.007
www.w3.org	**4.430**	0.001	0.006
www.washington.edu	8.860	0.001	0.011
Inaccessible Domains			
wsnet2.colostate.edu	**5.440**	0.83	1.10
www.w3.org	1.920	0.406	0.521
www.washington.edu	3.840	0.811	1.042

bold represents scores that do not reflect the accessibility

[6] https://www.w3.org/WAI/demos/bad/Overview.html.

these examples. In our analysis, we computed the metrics outcome for all the pages in Table 11, and analyzed the accessibility level they reported the pages to have.

4.2 Results

Table 12 presents the scores of each page level metric for each of the pages used to assess the validity of the metrics. Table 13 presents the scores for website metrics. To avoid the canceling effect in these metrics of having a website with the same number of accessible and inaccessible pages we split each website in two websites: a good website with the accessible pages and a bad website with the inaccessible pages.

4.2.1 Web page metrics

The results of this experiment show that the FR metric produces similar scores when evaluating accessible and inaccessible web pages. Since 1 means that the web page is completely inaccessible and 0 means otherwise, we were expecting to have values close to 1 for the inaccessible web pages and close to 0 for the accessible web pages. The FR scores for all the accessible web pages seem to be coherent and close to 0. However, all the inaccessible web pages also have low values, indicating a positive accessibility level.

WIE, Conservative, Optimistic and Strict metrics exhibit a score close to 1 for the same inaccessible web page, which means that this page is close to be completely accessible. Also, these metrics' scores for this particular inaccessible page are higher than some accessible pages' scores. This means that the inaccessible page is more accessible than some accessible pages, according to WIE, Conservative, Optimistic and Strict metrics' results. The remaining scores for these metrics seem to be coherent, except for the Conservative metric that classifies two accessible web pages as inaccessible, by showing scores close to some of the inaccessible pages' scores.

A3, UWEM and WAQM are the only three metrics that demonstrated coherent scores for all accessible and inaccessible web pages. The WAQM metric shows values close to 100 for all the accessible web pages. For the inaccessible pages, this metric varies from around 33 to 51, which is not close to 0, but still lower than the scores of the accessible pages. A3 and UWEM exhibit the correct behavior as all the accessible pages scores are close to 0 and the inaccessible pages scores are close to 1. Nevertheless, A3 metric scores for inaccessible web pages are closer to 1 compared to UWEM metric scores for the same web pages.

4.2.2 Website metrics

Since the three website accessibility metrics do not have a range of scores limited by two values, the level of accessibility of a certain website becomes uncertain. The main conclusion we can take from the WAB metric is that the higher the score, the more inaccessible the website is. Nevertheless, it is also possible to detect that the accessible domains' scores are really close to 0, which indicates the domains are more accessible. In addition, and in contrast to the accessible domains, the inaccessible domains' scores are higher and close to 1.

By observing Tables 5 or 6, the accessible scores are in the first quartile, while the inaccessible scores belong to the third quartile, indicating that WAB-PZ and WAB-H may have an appropriate representation of the accessibility. Nevertheless, the authors of these two WAB metrics [14, 23] performed a study [40] where they refer to the meaning of the WAB scores accessibility level. For instance, for those websites with a WAB score of 5.5 or less, the website is more accessible as it "has better conformance to the WCAG" [40]. Therefore WAB scores higher than 5.5 indicate more accessibility barriers, and so, a worse accessibility level. Comparing this information to the obtained scores of our study, we can see that all websites' WAB scores vary between 0.0069 and 1.0998. These scores indicate that all websites (including the inaccessible domains) tend to have a small number of accessibility barriers.

As for the WAEM, the higher this metric's score, the more accessible the website is. Consequently, we cannot define whether a website is accessible or inaccessible. This metric seems to be the only one with incoherent results as the www.w3.org accessible domain presents a lower score compared to the wsnet2.colostate.edu inaccessible domain.

4.3 Discussion

To define which metric is the most suitable option, it is important to analyze their results regarding the accessible and inaccessible web pages' and domains' evaluations.

FR seems to have coherent scores for all the accessible web pages. This means that all the accessible web pages have expected results. However, all the inaccessible web pages have low scores, indicating that these web pages are accessible when they are not.

WIE, Conservative, Strict and the Optimistic metrics always fail to assess the accessibility level of the inaccessible web page https://www.washington.edu/accesscomputing/AU/before.html, showing high scores that indicate the web page is accessible. Also, Conservative assigns a score below 0.5 to the https://www.w3.org/WAI/demos/bad/after/home.html accessible web page, which means that this page is not accessible.

Regarding the website metrics, it was possible to state that WAB-PZ and WAB-H seem to have an optimistic behavior for inaccessible pages, considering these metrics' ranges defined by Hackett and Parmanto [40].

A3, UWEM and WAQM seem to have the expected behavior. Interestingly, these three metrics form a cluster in the analysis of domain accessibility based on the average of the scores of the web pages belonging to the domain. Hence, whenever there is the need to measure the accessibility level of a website, one of these three metrics can be considered, as they are all correlated in this specific approach. Regardless of the available resources, the UWEM metric is the least resource intensive.

Nevertheless, if we investigate deeper into the validity analysis results of each of these metrics, we can detect two important aspects that will clarify which metric seems to have the best performance: (1) WAQM metric scores vary between 0 and 100 where 0 means the resource is totally inaccessible, and the scores of the inaccessible pages are not close to 0. Instead, they are above 33, which indicates that this metric is not that discriminating regarding those inaccessible web pages; (2) UWEM and A3 have both inaccessible and accessible scores close to 1 and 0, respectively, which indicates a correct and consistent behavior. Still, A3 metric scores for inaccessible pages can be more discriminating compared to UWEM, as they are all closer to 1.

In conclusion, although UWEM is less resource intensive, A3 provides more discriminating scores, being the most valid metric in this study.

5 Limitations

We acknowledge the accessibility evaluation reports are the result of an automated tool and that this type of tools is limited in the scope of the accessibility problems they can test [36, 41]. Given that our main objective is to compare web accessibility metrics, and that all metrics compared were applied to the same dataset, we believe the impact of this limitation to not be significant. However, for the part of the study that checks the validity of the metrics, this limitation can be significant, since it is probable that several accessibility problems in the web pages have not been identified.

Furthermore, in what relates to the validity study, we acknowledge the sample size is limited and the results presented are just indicative. A study with further web pages is needed to assess the generalizability of these findings.

6 Conclusion

This article compared eleven accessibility metrics by computing them over a dataset of nearly three million web pages evaluated by the QualWeb automated accessibility evaluation tool. The main goal was to understand if these metrics correlate with each other. The studied web accessibility metrics included FR, A3, UWEM, WAQM, WIE, Conservative, Optimistic, Strict, WAB-PZ, WAB-H and WAEM.

By analyzing the pairwise correlations we were able to identify groups of metrics. When considering the subset of metrics that are applicable at page level, we identified four clusters of distinct metrics. By looking at the full set of metrics applicable at site level, we identified a larger number of groups. This information is relevant when a decision between using one metric over another is needed. By knowing that the outcomes of two metrics are similar, it becomes possible to choose the one that is less resource intensive, or from which it is easier to obtain the data required to compute the metric, for instance.

Additionally, we ran an experiment with a small number of web pages with known levels of accessibility to assess the validity of the different metrics. Even though the set of pages was small, and the metrics were computed from the outcomes of an automated tool (i.e., not all accessibility problems were caught), we were able to identify which metrics were consistent with the expected levels of accessibility of the pages, and which were not. This information can be also relevant in assisting which metrics to employ.

Acknowledgements This work was supported by FCT through the LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020.

Funding The research leading to these results received funding from FCT through the LASIGE Research Unit under Grant Agreements ref. UIDB/00408/2020 and ref. UIDP/00408/2020.

Data availability statement The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. W3C: Introduction to Web Accessibility. (2005). Accessed in 28 of December of 2021. https://www.w3.org/WAI/fundamentals/accessibility-intro/
2. Henry, S.L.: Web Content Accessibility Guidelines (WCAG) Overview. (2021). Accessed in 28 of December of 2021. https://www.w3.org/WAI/standards-guidelines/wcag/
3. Vigo, M., Brajnik, G., Connor, J.O.: Research report on web accessibility metrics. In: Vigo, M., Brajnik, G., eds., J.O.C. (eds.) W3C WAI Symposium on Website Accessibility Metrics, First public working draft edn. W3C WAI Research and Development Working Group (RDWG) Notes. W3C Web Accessibility Initiative (WAI), ??? (2012). http://www.w3.org/TR/accessibility-metrics-report
4. Vigo, M., Brajnik, G.: Automatic web accessibility metrics: Where we are and where we can go. Interacting with Computers 23(2), 137–155 (2011) https://arxiv.org/abs/https://academic.oup.com/iwc/article-pdf/23/2/137/2238626/iwc23-0137.pdf. https://doi.org/10.1016/j.intcom.2011.01.001
5. Lopes, R., Gomes, D., Carriço, L.: Web not for all: A large scale study of web accessibility. In: Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A). W4A '10. Association for Computing Machinery, New York, NY, USA (2010). https://doi.org/10.1145/1805986.1806001
6. Kimmons, R.: Open to all? nationwide evaluation of high-priority web accessibility considerations among higher education websites. J. Comput. Higher Educ. (2017). https://doi.org/10.1007/s12528-017-9151-3
7. Acosta-Vargas, G., Acosta-Vargas, P., Jadán-Guerrero, J., Salvador-Ullauri, L., Gonzalez, M.: Improvement of accessibility in medical and healthcare websites. In: Nunes, I.L. (ed.) Advances in Human Factors and System Interactions, pp. 266–273. Springer, Cham (2021)
8. Snaprud, M., Sawicka, A.: Large scale web accessibility evaluation - a european perspective. In: Stephanidis, C. (ed.) Universal Access in Human-Computer Interaction. Applications and Services, pp. 150–159. Springer, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-73283-9_18
9. Costa, D., Fernandes, N., Neves, S., Duarte, C., Hijón-Neira, R., Carriço, L.: Web accessibility in africa: A study of three african domains. In: Kotzé, P., Marsden, G., Lindgaard, G., Wesson, J., Winckler, M. (eds.) Human-Computer Interaction - INTERACT 2013, pp. 331–338. Springer, Berlin, Heidelberg (2013)
10. Sirithumgul, P., Suchato, A., Punyabukkana, P.: Quantitative evaluation for web accessibility with respect to disabled groups. In: Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibililty (W4A). W4A '09, pp. 136–141. Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1535654.1535687
11. Freire, A.P., Bittar, T.J., Fortes, R.P.M.: An approach based on metrics for monitoring web accessibility in brazilian municipalities web sites. In: Proceedings of the 2008 ACM Symposium on Applied Computing. SAC '08, pp. 2421–2425. Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1363686.1364259
12. Song, S., Bu, J., Shen, C., Artmeier, A., Yu, Z., Zhou, Q.: Reliability aware web accessibility experience metric. In: Proceedings of the 15th International Web for All Conference. W4A '18. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3192714.3192836
13. Freire, A.P., Fortes, R.P.M., Turine, M.A.S., Paiva, D.M.B.: An evaluation of web accessibility metrics based on their attributes. In: Proceedings of the 26th Annual ACM International Conference on Design of Communication. SIGDOC '08, pp. 73–80. Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1456536.1456551
14. Parmanto, B., Zeng, X.: Metric for web accessibility evaluation. J. Am. Soc. Inform. Sci. Technol. **56**(13), 1394–1404 (2005)
15. Abuaddous, H.Y., Jali, M.Z., Basir, N.: Quantitative metric for ranking web accessibility barriers based on their severity. J. Inform. Commun. Technol. **16**(1), 81–102 (2017)
16. Sullivan, T., Matson, R.: Barriers to use: Usability and content accessibility on the web's most popular sites. In: Proceedings on the 2000 Conference on Universal Usability. CUU '00, pp. 139–144. Association for Computing Machinery, New York, NY, USA (2000). https://doi.org/10.1145/355460.355549
17. Vigo, M., Brajnik, G., Arrue, M., Abascal, J.: Tool independence for the web accessibility quantitative metric. Disabil. Rehabil. Assist. Technol. **4**(4), 248–263 (2009). https://doi.org/10.1080/17483100902903291
18. Vigo, M., Arrue, M., Brajnik, G., Lomuscio, R., Abascal, J.: Quantitative metrics for measuring web accessibility. In: Proceedings of the 2007 International Cross-Disciplinary Conference on Web Accessibility (W4A), pp. 99–107. Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1243441.1243465
19. Velleman, E., Strobbe, C., Koch, J., Velasco, C.A., Snaprud, M.: A unified web evaluation methodology using wcag. In: Stephanidis, C. (ed.) Universal Access in Human-Computer Interaction. Applications and Services, pp. 177–184. Springer, Berlin, Heidelberg (2007)
20. Freire, A.P., Power, C., Petrie, H., Tanaka, E.H., Rocha, H.V., Fortes, R.P.: Web accessibility metrics: Effects of different computational approaches. In: International Conference on Universal Access in Human-Computer Interaction, pp. 664–673 (2009). https://doi.org/10.1007/978-3-642-02713-0_70. Springer
21. Martínez, A.B., Juan, A.A., Álvarez, D., del Carmen Suárez, M.: Wab*: A quantitative metric based on wab. In: Gaedke, M., Grossniklaus, M., Díaz, O. (eds.) Web Engineering, pp. 485–488. Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02818-2_44
22. Bühler, C., Heck, H., Perlick, O., Nietzio, A., Ulltveit-Moe, N.: Interpreting results from large scale automatic evaluation of web accessibility. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) Computers Helping People with Special Needs, pp. 184–191. Springer, Berlin, Heidelberg (2006). https://doi.org/10.1007/11788713_28
23. Hackett, S., Parmanto, B., Zeng, X.: Accessibility of internet websites through time. In: Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility. Assets '04, pp. 32–39. Association for Computing Machinery, New York, NY, USA (2003). https://doi.org/10.1145/1028630.1028638
24. Brajnik, G., Vigo, M.: Automatic web accessibility metrics: Where we were and where we went. In: Web Accessibility, pp. 505–521 (2019). https://doi.org/10.1007/978-1-4471-7440-0_27
25. Bailey, J., Burd, E.: Tree-map visualisation for web accessibility. In: 29th Annual International Computer Software and Applications Conference (COMPSAC'05), vol. 1, pp. 275–2802 (2005). https://doi.org/10.1109/COMPSAC.2005.161
26. Bailey, J., Burd, E.: Towards more mature web maintenance practices for accessibility. In: 2007 9th IEEE International Workshop on Web Site Evolution, pp. 81–87 (2007). https://doi.org/10.1109/WSE.2007.4380248
27. Brajnik, G., Lomuscio, R.: Samba: A semi-automatic method for measuring barriers of accessibility. In: Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility. Assets '07, pp. 43–50. Association for Computing

Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1296843.1296853

28. Brajnik, G.: Web accessibility testing: When the method is the culprit. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) Computers Helping People with Special Needs, pp. 156–163. Springer, Berlin, Heidelberg (2006)
29. Song, S., Wang, C., Li, L., Yu, Z., Lin, X., Bu, J.: Waem: A web accessibility evaluation metric based on partial user experience order. In: Proceedings of the 14th Web for All Conference on The Future of Accessible Work. W4A '17. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3058555.3058576
30. Battistelli, M., Mirri, S., Muratori, L.A., Salomoni, P.: Measuring Accessibility Barriers on Large Scale Sets of Pages. (2011). Accessed in 28 of December of 2021. https://www.w3.org/WAI/RD/2011/metrics/paper2/
31. Vigo, M., Abascal, J., Aizpurua, A., Arrue, M.: Attaining Metric Validity and Reliability with the Web Accessibility Quantitative Metric. (2011). Accessed in 28 of December of 2021. https://www.w3.org/WAI/RD/2011/metrics/paper6/
32. Fukuda, K., Saito, S., Takagi, H., Asakawa, C.: Proposing new metrics to evaluate web usability for the blind. In: CHI '05 Extended Abstracts on Human Factors in Computing Systems, pp. 1387–1390. Association for Computing Machinery, New York, NY, USA (2005). https://doi.org/10.1145/1056808.1056923
33. Lopes, R., Carriço, L.: The impact of accessibility assessment in macro scale universal usability studies of the web. In: Proceedings of the 2008 International Cross-Disciplinary Conference on Web Accessibility (W4A), pp. 5–14. Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1368044.1368048
34. Benavidez, C.: Libro Blanco de eXaminator, (2012)
35. Mirri, S., Muratori, L.A., Salomoni, P.: Monitoring accessibility: Large scale evaluations at a geo political level. In: The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility. ASSETS '11, pp. 163–170. Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2049536.2049566
36. Lazar, J., Goldstein, D., Taylor, A.: Ensuring Digital Accessibility Through Process and Policy. Morgan kaufmann, ??? (2015)
37. Fernandes, N., Costa, D., Neves, S., Duarte, C., Carriço, L.: Evaluating the accessibility of rich internet applications. In: Proceedings of the International Cross-Disciplinary Conference on Web Accessibility. W4A '12. Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2207016.2207019
38. Statstutor: Spearman's Correlation. (2021). Accessed in 28 of December of 2021. https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf
39. Wikipedia: Hierarchical Clustering. (2021). Accessed in 28 of December of 2021. https://en.wikipedia.org/wiki/Hierarchical_clustering
40. Hackett, S., Parmanto, B.: Homepage not enough when evaluating web site accessibility. Internet Research (2009)
41. Abascal, J., Arrue, M., Valencia, X.: Tools for web accessibility evaluation. In: Yesilada, Y., Harper, S. (eds.) Web Accessibility: A Foundation for Research, pp. 479–503. Springer, London (2019). https://doi.org/10.1007/978-1-4471-7440-0_26

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.