

Evaluating Conformance to WCAG 2.0: Open Challenges

Fernando Alonso, José L. Fuertes, Ángel L. González, and Loïc Martínez

Facultad de Informática, Universidad Politécnica de Madrid, Spain
{falonso, jfuertes, agonzalez, loic}@fi.upm.es

Abstract. Web accessibility for people with disabilities is a highly visible area of work in the field of ICT accessibility, including many policy activities in several countries. The commonly accepted guidelines for web accessibility (WCAG 1.0) were published in 1999 and have been extensively used by designers, evaluators and legislators. A new version of these guidelines (WCAG 2.0) was published in 2008. In this paper we point out the main challenges that WCAG 2.0 raises for web accessibility evaluators: the concept of “accessibility supported technologies”; success criteria testability; technique and failure openness, and the aggregation of partial results. We conclude the paper with some recommendations for the future.

Keywords: Web accessibility, Web accessibility evaluation.

1 Introduction

The web is an essential component of the information society, and as such has attracted special attention from the promoters of accessibility for people with disabilities.

The commonly accepted guidelines for web accessibility are the Web Content Accessibility Guidelines (WCAG) 1.0, published in 1999 by the World Wide Web Consortium (W3C) [1]. These guidelines have been in use for several years by designers, evaluators and legislators, and there is an arguably large consensus among practitioners about how to interpret and evaluate them.

A new version of WCAG has been under development for several years, and was published in December 2008 as WCAG 2.0 [2]. This new version had two main goals. Firstly, it aimed to be technology-independent, so it could be applied to current and future web technologies from either the W3C or other sources. Secondly, it was to be testable, that is, practitioners should agree about how to evaluate the conformance of a web site with WCAG 2.0.

WCAG 2.0 has a different language, a different structure and a different rationale. All of these influence how the evaluation of conformance to WCAG 2.0 is to be performed in the future, either manually or with the support of evaluation tools [3].

In this paper we present some challenges that we think are currently present when evaluating the accessibility of web sites with respect to WCAG 2.0. This discussion is based on our experience in teaching and evaluating web accessibility [4], [5], [6].

The content of this paper is structured as follows. Section 2 will provide an overview of WCAG 2.0. This is followed by four sections describing what we consider to be the key challenges: a) the concept of “accessibility supported technologies”;

b) success criteria testability; c) technique and failure openness, and d) the aggregation of partial results. Finally, the paper concludes with some recommendations for the future.

2 WCAG 2.0 Overview

WCAG 2.0 is a W3C Recommendation [2] that contains three layers of guidance: principles, guidelines and success criteria.

- The *principles* provide the foundation for web accessibility. There are four principles: perceivable, operable, understandable and robust.
- The 12 *guidelines* provide the basic goals that web designers should work toward in order to make content more accessible to users with different disabilities. The guidelines are not testable, but provide the framework and overall objectives to help web designers understand the success criteria and better implement the techniques.
- For each guideline, testable *success criteria* are provided to allow WCAG 2.0 to be used where requirements and conformance testing are necessary such as in design specification, purchasing, regulation and contractual agreements. In order to meet the needs of different groups and different situations, three levels of conformance are defined: A (lowest), AA, and AAA (highest).

There are additional layers of guidance provided by an external document that complements WCAG 2.0 [7]. This document is “informative” and provides three additional layers, referred to as sufficient techniques, advisory techniques and common failures.

- The *Sufficient Techniques* offer guidance and examples for meeting the guidelines using specific technologies. These techniques are considered sufficient to meet the success criteria. Most success criteria list multiple sufficient techniques, and any of the sufficient techniques can be used to meet the success criterion. There may be other techniques not documented by the W3C that could also meet the success criteria. As new sufficient techniques are identified, they can be added to the listing.
- The *Advisory Techniques* can enhance accessibility, but did not qualify as sufficient techniques because they are not sufficient to meet the full requirements of the success criteria, they are not testable, and/or are good and effective techniques in some circumstances but not effective or helpful in others.
- The *Common Failures* are examples of bad practices that cause web pages to fail to meet the success criteria. Failures during evaluation are interpreted differently than for techniques: if a common failure is found in a web page, then that web page fails the respective success criterion.

In addition, there is another important part of WCAG 2.0: the conformance section. This section lists five requirements for conformance to WCAG 2.0: (1) one conformance level is met in full; (2) conformance is for full web pages; (3) all web pages in a process conform to the same level; (4) only accessibility-supported ways of using the technologies are relied upon to satisfy the success criteria; and (5) technologies

that are used in a way that is not accessibility-supported do not interfere with the accessibility of the page.

The conformance section also gives information about how to make conformance claims, which are optional. Finally, it describes what “accessibility supported” means. This “accessibility supported” concept is the key to the first challenge for WCAG 2.0 conformance evaluation.

3 Challenge 1: Accessibility Supported Technologies

The first challenge is to apply the key concept “accessibility supported ways of using technologies”. The fourth WCAG 2.0 conformance requirement states: “Only accessibility supported ways of using technologies are relied upon to satisfy the success criteria. Any information or functionality that is provided in a way that is not accessibility supported is also available in a way that is accessibility supported” [2].

According to W3C, a Web content technology is “accessibility supported” when users’ assistive technologies will work with the Web technologies and when the accessibility features of mainstream technologies will work with the technology [8].

The problem is that this is an open definition and, in particular, the W3C does not specify which or how many assistive technologies a Web technology must support in order for it to be classified as accessibility supported. This is a complex, environment- and language-dependent topic. The W3C states in WCAG 2.0 that there is a need for an external and international dialogue on this point.

This is the first and main challenge: a definition of which technologies are considered to be “accessibility supported” in a given context is needed in order to get consistent evaluation results. This definition should be provided by organizations that are industry independent and should be internationally agreed upon to prevent market fragmentation.

This is especially important for web sites that are obliged to be accessible by law. For instance, the web sites of the public administrations are obliged to conform to WCAG 1.0 (and, in the future, to WCAG 2.0) in many European countries. Each country defining its own set of accessibility-supported technologies would lead to a fragmented market where a web site is considered to be accessible in one country and non-accessible in another.

4 Challenge 2: Testability of Success Criteria

The second challenge deals with the actual degree of testability of the success criteria. According to W3C, the success criteria are written as testable sentences.

W3C provides definitions of testability for techniques [9]: a technique is testable if it is either machine testable or reliably human testable. It is machine testable if there is a known algorithm (regardless of whether that algorithm is known to be implemented in tools) that will determine absolutely reliably whether or not the technique has been implemented. It is reliably human testable if the technique can be tested by human inspection, and it is believed that at least 80% of knowledgeable human evaluators would agree on the finding.

There was some debate about the implications of testability during the development of WCAG 2.0 [10], [11], but now that WCAG 2.0 is complete the real challenge is to find out whether the success criteria and the techniques are actually testable.

Many success criteria have a more precise and objective language in WCAG 2.0 (for example, the language used for success criteria 1.4.3 on color contrast is much more precise –providing exact values to check against– than WCAG 1.0 checkpoint 2.2). But it is not clear that this is true for all the success criteria.

During our course on web accessibility taught in March 2009 within the ATHENS Program, we used WCAG 2.0 for the first time. It was an intensive one-week course for international students, and they were set the exercise of evaluating the same web page according to WCAG 2.0. Given that at the time there were no automated tools providing support for WCAG 2.0, all the evaluations were done manually.

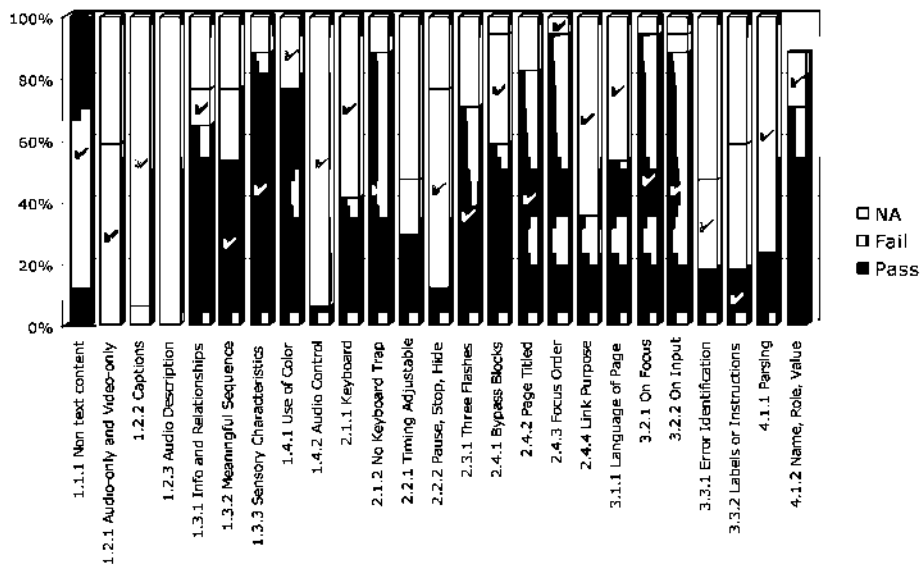


Fig. 1. Results of evaluating Level-A success criteria in an intensive web accessibility course

We examined our students' and our own results, and we compared the results for each success criterion. Figure 1 shows, for each success criterion, the number of times that each possible value (pass, fail, not applicable) appeared in the results provided by our students. It also shows what was considered to be the right result, that is, the result generated by the instructors, indicated by a check mark (✓), provided that at least one student entered the correct result. If a bar does not reach a value of 100%, it means that some students did not evaluate that success criterion. Of the 25 level-A success criteria (the ones that all the students were able to evaluate), only nine were reliably human testable, that is, 79% or more of the population agreed on the result (which is almost the 80% used in the W3C's definition). Agreement among evaluators was much less for the other 16 success criteria. The outcomes for some criteria, such as 1.3.2 (meaningful sequence of elements), 3.1.1 (language of page), 3.3.1 (error identification) and 3.3.2 (labels or instructions), were especially divergent [12].

In a repetition of the experiment on a recent non-intensive course (October 2009 to January 2010), better results were obtained. In this course, the students also had to evaluate the same web page (different to the one used in the previous experiment) according to WCAG 2.0. Figure 2 shows the results provided by the students. We examined our students' and our own results, and we compared the results for each success criterion. Of the 25 level-A success criteria, 17 were reliably human testable. The outcomes for some level-A criteria, such as 1.3.3 (sensory characteristics), 2.3.1 (three flashes), 2.4.3 (focus order) and 3.3.2 (labels or instructions), were especially divergent. In addition, of the 36 level-AA and level-AAA success criteria, 17 were reliably human testable, showing that students found the success criteria of these two levels were quite a lot harder to test reliably compared with level-A success criteria.

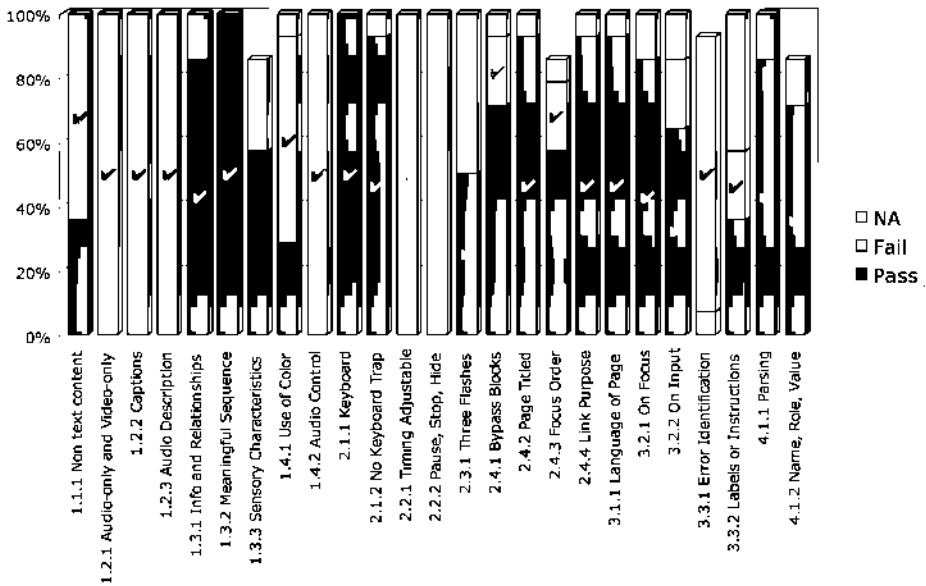


Fig. 2. Results of evaluating Level-A success criteria in a non-intensive web accessibility course

In addition, our experiments and results agree with Brajnik's experiment [13]. Some success criteria show up as being particularly weak in the studies: 1.3.3 (sensory characteristics), 2.1.1 (timing adjustable), 2.2.2 (pause, stop, hide), 2.4.1 (bypass blocks), 2.4.4 (link purpose in context) and 3.3.2 (labels or instructions). It would be interesting to see whether this trend keeps up in further experiments run to complement Brajnik's and our studies.

The detailed results that we have obtained led us to identify three sources of unreliability: comprehension, knowledge and effort [12].

Of course these are not conclusive experiments, but they point out that the testability of the WCAG 2.0 success criteria is not to be taken for granted for beginners and that support material and tools will be needed to help evaluators to provide consistent results in the future. Our experiments have been carried out by small groups of beginners

(17 and 14, respectively) and on non-diverse web pages, but we plan to repeat the experiments with more beginners and with a group of experts in web accessibility.

5 Challenge 3: Openness of Techniques and Failures

The third challenge is related to the openness of the techniques. The techniques belong to a non-normative document [7] that is intended to be a living document that will change as new techniques are defined, either inside or outside the W3C.

One reason for this incompleteness is that the W3C only documents techniques for non-proprietary technologies; the W3C hopes that vendors of other technologies will provide similar techniques to describe how to conform to WCAG 2.0 using those technologies. Another reason is that new web technologies will require new techniques to make web sites conform to WCAG 2.0.

In fact, the W3C acknowledges this fact and has established a process for updating the techniques document. The W3C encourages submission of new techniques so they can be considered for inclusion in the document. This should make the set of techniques maintained by the W3C as comprehensive as possible.

From the viewpoint of an evaluator this is a new challenge. If for a given web page element and a given success criterion, none of the documented techniques apply and none of the common failures apply, then the evaluator cannot decide whether or not the success criterion has been met. There could be a technique for that particular element in that particular case that makes the content accessible. However, if the W3C has not yet documented the technique, it will be difficult to provide a reliable result.

Of course this is not an issue with the most basic web technologies (such as standard HTML 4.01 elements), because, given all the experience gained since WCAG 1.0 was written, it is well known how they can be made accessible.

6 Challenge 4: Aggregation of Partial Results

Once evaluators know the result for each technique and failure for each element and each situation, how do they aggregate this result to get the final conformance result for the success criteria?

In WCAG 1.0 the situation was simpler because the checkpoints were restrictive: if one element fails to comply with one checkpoint, then this checkpoint fails for that page.

In WCAG 2.0 there are potentially different ways of making some specific element accessible. There are complex cases, where one of many different situations applies for each element, each situation includes several techniques, and several common failures are documented. In these cases, the evaluator first has to identify the situation that applies for each element, and then evaluate the sufficient techniques and common failures for that situation.

The complexity in WCAG 2.0 stems from the permissiveness of the techniques: if none of the documented techniques pass for any given element, the web page could still be accessible due to the application of a non-documented technique. And in a few cases, it is difficult to aggregate the values of the techniques because it is not clear if an 'or' or an 'and' operator should be used.

The challenge here is that the W3C has not documented how to combine the results of techniques, failures and situations to produce an aggregated result for each success criterion. This could lead to a situation where different evaluators use different aggregation strategies and thus produce different evaluation results. We are in the process of updating the Hera-FFX tool [6] to cover WCAG 2.0, and we have found that this is a real challenge that should be solved.

7 Some Recommendations for the Future

This paper has presented some challenges that, in our opinion, have to be faced when evaluating the conformance of web pages to WCAG 2.0. These challenges could explain why there are not many evaluation tools for WCAG 2.0 at present. In fact we are acquainted with just two tools: AChecker [14] and TAW [15]. As concluding remarks, we list the following recommendations for the future:

- When web accessibility is mandatory, public authorities should define the accessibility-supported ways of using technologies that apply in each context. Furthermore, this definition should be internationally harmonized to prevent market fragmentation.
- More experimentation is needed to assess the testability of techniques, failures and success criteria. This could help to create a common understanding and support material that could enable evaluators to produce consistent results. It could even lead to proposals of changes in the wording of some of the success criteria. We plan to work on further experiments, including comparisons between expert and non-expert evaluation, to gather more data and get more insight into success criteria testability.
- The W3C should provide recommendations about how to deal with the openness of techniques and about how to aggregate partial evaluation results to improve the consistency of evaluations produced by different tools and people.

Concerning the aggregation of partial evaluation results, we are currently developing an update of Hera-FFX [6] for WCAG 2.0. This tool considers six evaluation values for a given technique or failure: pass, fail, partial (i.e. “near pass”), not applicable, not verified (the evaluator has yet to examine the technique/failure) and don’t know (the evaluator is unable to decide a value, typically due to disability).

The evaluation strategy for one element in one success criterion is as follows: first we look at the failures, aggregated using a “restrictive” approach in (1). This means that if one failure “fails” then the element fails for the success criterion. Only if the result of the aggregation of failures is “pass” or “not applicable” (i.e. the evaluator is sure that there is no failure) does the strategy look at the techniques.

[Restrictive] fail >> partial >> not verified >> don’t know >> pass >> not applicable (1)

The techniques for one situation are expressed as and/or trees that are evaluated combining the restrictive and permissive approach. We use the “permissive” approach in (2) for an “or” branch (any child is enough), and then they are combined using the “restrictive” approach in an “and” branch (all of the children are needed). In the “permissive” approach, one pass is enough for the corresponding group of techniques.

[*Permissive*] pass >> not verified >> don't know >> partial >> failure >> not applicable (2)

Although WCAG 2.0 success criteria are hard to evaluate objectively [12], [13], everyday practice is showing WCAG 2.0 to be significantly better in flexibility terms than WCAG 1.0.

References

1. Chisholm, W., Vanderheiden, G., Jacobs, I. (eds.): Web Content Accessibility Guidelines 1.0. W3C Recommendation (1999), <http://www.w3.org/TR/WCAG10/>
2. Caldwell, B., Cooper, M., Reid, L.G., Vanderheiden, G. (eds.): Web Content Accessibility Guidelines 2.0. W3C Recommendation (2008), <http://www.w3.org/TR/WCAG20/>
3. W3C. Web Accessibility Evaluation Tools: Overview (2006), <http://www.w3.org/WAI/ER/tools/>
4. Benavidez, C., Fuertes, J.L., Gutiérrez, E., Martínez, L.: Semi-automatic Evaluation of Web Accessibility with HERA 2.0. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2006. LNCS, vol. 4061, pp. 199–206. Springer, Heidelberg (2006)
5. Benavidez, C., Fuertes, J.L., Gutiérrez, E., Martínez, L.: Teaching Web Accessibility with “Contramano” and Hera. In: Miesenberger, K., Klaus, J., Zagler, W.L., Karshmer, A.I. (eds.) ICCHP 2006. LNCS, vol. 4061, pp. 341–348. Springer, Heidelberg (2006)
6. Fuertes, J.L., González, R., Gutiérrez, E., Martínez, L.: Hera-FFX: a Firefox add-on for semi-automatic web accessibility evaluation. In: 6th International Cross-Disciplinary Conference on Web Accessibility, pp. 26–35. ACM Press, New York (2009)
7. Caldwell, B., Cooper, M., Reid, L.G., Vanderheiden, G. (eds.): Techniques for WCAG 2.0 (2008), <http://www.w3.org/TR/WCAG20-TECHS/>
8. W3C: Understanding Conformance (2008), <http://www.w3.org/TR/UNDERSTANDING-WCAG20/conformance.html>
9. Cooper, M. (ed.): Requirements for WCAG 2.0 Checklists and Techniques (2003), <http://www.w3.org/TR/wcag2-tech-req/>
10. Sampson-Wild, G.: Testability Costs Too Much. A List Apart (2007), <http://www.alistapart.com/articles/testability/>
11. Smith, J.: Testability in WCAG 2.0. WebAIM Blog (2007), <http://webaim.org/blog/wcag-2-testability/>
12. Alonso, F., Fuertes, J.L., González, Á.L., Martínez, L.: On the testability of WCAG 2.0 for beginners. In: 7th International Cross-Disciplinary Conference on Web Accessibility. ACM Press, New York (2010)
13. Brajnik, G.: Validity and Reliability of Web Accessibility Guidelines. In: 11th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 131–138. ACM Press, New York (2009)
14. AChecker: Web Accessibility Checker (2009), <http://www.achecker.ca/>
15. TAW tool (2010), <http://www.tawdis.net/ingles.html?lang=en>