Following the WCAG 2.0 Techniques: Experiences from Designing a WCAG 2.0 Checking Tool

Annika Nietzio[1], Mandana Eibegger[2], Morten Goodwin[2], and Mikael Snaprud[2,3]

[1] Forschungsinstitut Technologie und Behinderung (FTB)
der Evangelischen Stiftung Volmarstein
Grundschötteler Str. 40, 58300 Wetter (Ruhr), Germany
egovmon@ftb-esv.de
http://www.ftb-esv.de

[2] Tingtun AS, PO Box 48, N−4791 Lillesand, Norway
m.eibegger@schoener.at, {morten.goodwin,mikael.snaprud}@tingtun.no
http://www.tingtun.no

[3] University of Agder, PO Box 509, N−4898 Grimstad, Norway

Abstract. This paper presents a conceptual analysis of how the *Web Content Accessibility Guidelines (WCAG) 2.0* and its accompanying documents can be used as a basis for the implementation of an automatic checking tool and the definition of a web accessibility metric. There are two major issues that need to be resolved to derive valid and reliable conclusions from the output of individual tests. First, the relationship of Sufficient Techniques and Common Failures has to be taken into account. Second, the logical combination of the techniques related to a Success Criterion must be represented in the results.

The eGovMon project has a lot of experience in specifying and implementing tools for automatic checking of web accessibility. The project is based on the belief that web accessibility evaluation is not an end in itself. Its purpose is to promote web accessibility and initiate improvements.

1 Web Accessibility Benchmarking with WCAG 2.0

Benchmarking goes beyond the presentation of conformance results such as: “The web content conforms to the guidelines.” or “The web content does not conform to the guidelines.” If the outcome of the evaluation is represented as numerical value the descriptive power is much higher. It becomes possible to distinguish between web sites which “almost conform” to the guidelines (i.e. sites that have only very few accessibility problems) and web sites that don’t meet the guidelines at all. Several metrics have been proposed , such as:

- Failure rate [12]
- Unified Web Evaluation Methodology (UWEM) score [7]
- Web Accessibility Quantitative Metric (WAQM) [5]

K. Miesenberger et al. (Eds.): ICCHP 2012, Part I, LNCS 7382, pp. 417–424, 2012.

- Barrier Impact Factor (BIF) [3]
- Accessibility Score used by WCAG 2.0 Web Assessment Tool (WaaT) [1]

The latter two of the metrics are tailored towards WCAG 2.0 [9] while the others were constructed for WCAG 1.0[8] or can be applied to any kind of accessibility evaluation results. At present, a generally accepted practice for reporting WCAG 2.0 evaluation results does not exist.

On the contrary, the results of tools which claim to check according to WCAG 2.0 often are not comparable. This problem is mainly caused by the varying granularity of tests (Some tools implement several tests per Success Criterion while others only have one test.) and the differences in counting the instances of potential barriers (Some tools count every checked HTML element while others only count each barrier type once.). The tools also differ in how outcomes are grouped into categories like "error", "potential error", or "warning". Some tools only report the absolute numbers for each outcome category, while other tools use some kind of score function. Alonso et al. [2] describe the consequences of this challenge: "This could lead to a situation where different evaluators use different aggregation strategies and thus produce different evaluation results."

In this paper we propose the introduction of aggregation on the level of Success Criteria, as a first step to increase the comparability of the results from different tools. Fortunately, we don't have to re-invent the wheel. As WCAG 2.0 already is constructed to guide and support evaluation of web pages and sites, we can follow the instructions and test procedures in WCAG 2.0 and its supporting documents.

To ensure the validity of our approach, we base the method on WCAG 2.0 and its supporting documents. This contributes to the usefulness of the tool because WCAG 2.0 is widely adopted and much used as a reference for public procurement and quality assurance. Furthermore, it has the additional advantage that the implementation of the tool can keep up with the latest developments in web accessibility by incorporation the regular updates of the WCAG 2.0 Techniques[1] by W3C.

2 Development of the WCAG 2.0 Checker and Score

The score function should be tailored to the structure of the test set (in this case WCAG 2.0). Therefore we start the design of the score function with an analysis of WCAG 2.0, taking also the structural differences between WCAG 1.0 and 2.0 into account.

We also investigate how other implementations of WCAG 2.0 tools address the challenges mentioned above and try to identify potential shortcomings and possible solutions.

[1] The development of the eGovMon WCAG 2.0 tool described in this paper is based on the latest version of the *Techniques for WCAG 2.0*, which was published on 3 January 2012.

2.1 Combining Sufficient Techniques and Common Failures

The WCAG 1.0 tests (as defined in UWEM 1.2 and by other tools) are independent: failure of a test also means failure of a WCAG 1.0 Checkpoint. The structure of WCAG 2.0 is different. The techniques with their detailed test procedure provide a natural starting point for the implementation of an evaluation tool. But in the presentation of results the dependencies of the techniques must be taken into account. On the one hand there are Common Failures which directly cause the web content to fail a Success Criterion. On the other hand, conclusions can be drawn from the presence or absence of Sufficient Techniques.

The majority of existing tools implements a strategy that is either based only on Common Failures or only on Sufficient Techniques. While this avoids the problem of reporting the same issue twice, it also misses a number of issues and thus leads to incomplete results.

Some Success Criteria don't define redundant techniques. For instance *3.1.1: Language of Page*, does not have related Common Failures, instead the absence of an implementation of the corresponding Sufficient Techniques is interpreted as failure of the Success Criterion. This case is missed if the evaluation is based solely on Common Failures.

If only the Sufficient Techniques were used, the tool might ignore some specific problems that are only described as Common Failures but can not be derived directly from the Sufficient Techniques. For instance *F3: Failure of Success Criterion 1.1.1 due to using CSS to include images that convey important information* requires the checking of images included via CSS. The Sufficient Techniques for short text alternatives in Success Criterion 1.1.1 target only HTML img elements. Thus a tool that only looks at Sufficient Techniques will miss accessibility barriers caused by CSS background images.

Incorrect results can also occur if the tests for accessibility barriers are derived from negated Sufficient Techniques. This is also described in the *Techniques for WCAG 2.0* document [11]:

> "However, failure of a test procedure for a sufficient technique does not necessarily mean that the success criterion has not been satisfied in some other way, only that this technique has not been successfully implemented and can not be used to claim conformance."

For instance Technique *G134: Validating Web pages* for Success Criterion *4.1.1: Parsing* includes the validation of CSS in its test procedure. However, this does not mean that CSS validation errors cause the web content to fail the Success Criterion. This happens only if none of the other Sufficient Techniques is implemented on the web page either.

An evaluation result can only capture the full image of WCAG 2.0 if both Sufficient Techniques and Common Failures are taken into account. The section *Understanding conformance* of the Understanding WCAG 2.0 document [10] describes the logic required to combine the techniques.

To overcome the problems discussed above, we suggest the following approach to derive an implementation of tests for a Success Criterion (SC):

1. Applicability: Does the web page contain any HTML elements to which the SC is applicable? — No: SC passes. Yes: continue.
2. Does a Common Failure occur? — Yes: SC fails. No: continue.
3. Are the Sufficient Techniques successfully implemented for every element? — Yes: SC passes or SC passes with warning ("human input required"). No: continue.
4. Does the checker provide tests for all techniques related to the SC? — Yes: SC fails with warning ("no Sufficient Technique used"). No: warning ("not checked").

2.2 Combining Results for Techniques

The WCAG 2.0 tool developed by eGovMon specifies simple checks for situations that can easily be captured by a single question. These checks correspond to the WCAG 2.0 Techniques or sometimes even to the individual steps in the test procedures of the technique.

The results of these checks are then combined as described in the Understanding WCAG 2.0 document [10]. The following example illustrates our approach:

Success Criterion 3.3.2 Labels or Instructions. The following individual tests are applied to check that the purpose of a form control can be identified:

H44: Is there is a **label element** that identifies the purpose of the control?
H65: Is there is a **title attribute** that identifies the purpose of the control?
G167: Is there is an **adjacent button** that identifies the purpose of the control?

The report for Success Criterion 3.3.2 is the result of a logical combination:

```
IF ((H44:cause=no_label OR H44:cause=label_empty)
    AND (H65:cause=title_missing OR H65:cause=title_empty)
    AND G167:cause=empty_button_as_label OR G167:not_applicable)
THEN return SC3.3.2 failed
```

The web content fails Success Criterion 3.3.2 only if neither of the related techniques was successfully implemented.

2.3 Aggregation Beyond Success Criterion

Some techniques are used by several Success Criteria. Therefore an interpretation of the results below the level of Success Criteria is not meaningful. This is a major difference from UWEM 1.2 (and other WCAG 1.0 tools), which has an erratic number of tests per WCAG 1.0 Checkpoint. Each test contributes equally to the score result causing Checkpoints with many tests to be over-represented in the result. Using Success Criteria as intermediary aggregation level addresses this shortcoming and also has several other advantages. The influence of Success Criteria with many techniques is balanced. In an automated tool it becomes easy to highlight which Success Criteria need human judgement or were not tested. Disadvantages of the approach are that the number of instances of a specific feature (such as form control) does not influence the score if each Success Criterion gets the same weight. If a tool does not implement all techniques related to a Success Criterion, no valid conclusion can be drawn. However, this limitation does not occur if the tool results can be complemented by expert evaluations.

2.4 eGovMon Score Function

The theoretical considerations given above lead to the following definition of the score function. Let p denote a web page, c a Success Criterion, and C the set of all Success Criteria that are covered by the evaluation, i.e. all Success Criteria for which at least one test has been carried out.

The instances of application of a test for Success Criterion c on page p are denoted by:

$f_c(p)$ = number of instances where tests for c failed
$n_c(p)$ = number of all instances where tests for c were applied

The intermediary result for c on page p is defined as:

$$S_c(p) = 1 - \frac{f_c(p)}{n_c(p)} \tag{1}$$

Page Score. The page score S is calculated from the intermediary results per Success Criterion.

$$S(p) = \frac{1}{|C|} \sum_{c \in C} S_c(p) \tag{2}$$

Site Score. A web site is a set of web pages $s = \{p_1, p_2, \ldots\}$. The simplest way to calculate a score for the web site is to take the average of the page scores.

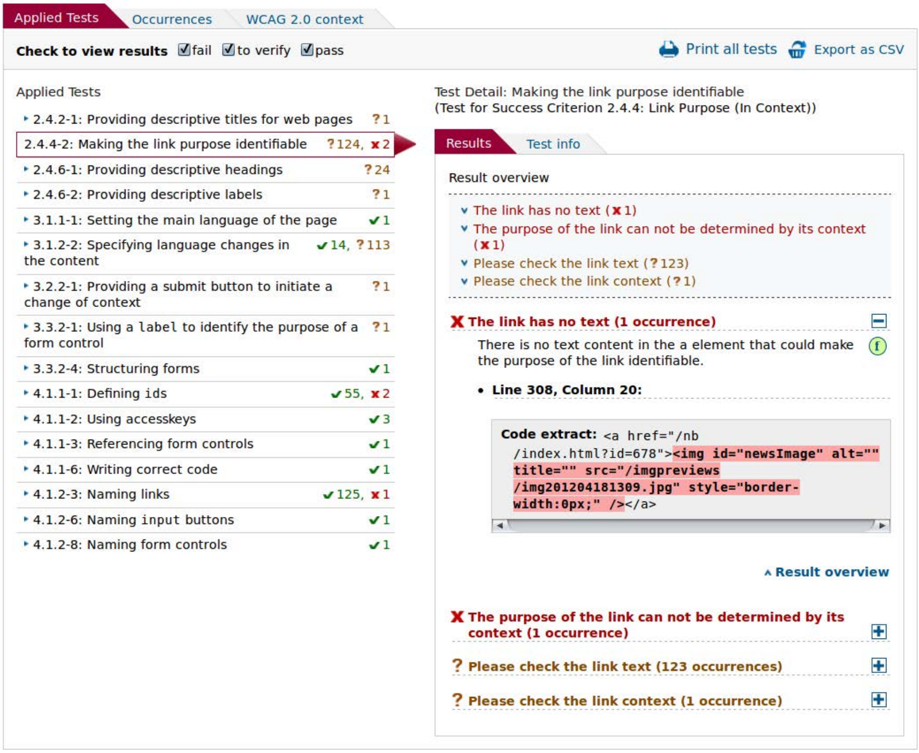
$$S_{mean}(s) = \frac{1}{|s|} \sum_{p \in s} S(p) \tag{3}$$

This approach gives equal weight to all pages, irrespective of their size. To take the size of pages into account, we define:

$N(p) = \sum_{c \in C} n_c(p)$ = number of all instances within page p
$N(s) = \sum_{p \in s} N(p)$ = number of all instances within site s

Then the page scores can be weighted by the number of instances.

$$S_{weighted}(s) = \sum_{p \in s} \frac{N(p)}{N(s)} S(p) \tag{4}$$



Check to view results ☑fail ☑to verify ☑pass

Applied Tests

2.4.2-1: Providing descriptive titles for web pages	? 1
2.4.4-2: Making the link purpose identifiable	? 124, ✗ 2
2.4.6-1: Providing descriptive headings	? 24
2.4.6-2: Providing descriptive labels	? 1
3.1.1-1: Setting the main language of the page	✓ 1
3.1.2-2: Specifying language changes in the content	✓ 14, ? 113
3.2.2-1: Providing a submit button to initiate a change of context	? 1
3.3.2-1: Using a `label` to identify the purpose of a form control	? 1
3.3.2-4: Structuring forms	✓ 1
4.1.1-1: Defining `ids`	✓ 55, ✗ 2
4.1.1-2: Using accesskeys	✓ 3
4.1.1-3: Referencing form controls	✓ 1
4.1.1-6: Writing correct code	✓ 1
4.1.2-3: Naming links	✓ 125, ✗ 1
4.1.2-6: Naming input buttons	✓ 1
4.1.2-8: Naming form controls	✓ 1

Test Detail: Making the link purpose identifiable
(Test for Success Criterion 2.4.4: Link Purpose (In Context))

Result overview

- The link has no text (✗ 1)
- The purpose of the link can not be determined by its context (✗ 1)
- Please check the link text (? 123)
- Please check the link context (? 1)

✗ The link has no text (1 occurrence)

There is no text content in the a element that could make the purpose of the link identifiable.

- **Line 308, Column 20:**

Code extract: `<a href="/nb/index.html?id=678"><img id="newsImage" alt="" title="" src="/imgpreviews/img201204181309.jpg" style="border-width:0px;" /></a>`

Result overview

✗ The purpose of the link can not be determined by its context (1 occurrence)

? Please check the link text (123 occurrences)

? Please check the link context (1 occurrence)

Fig. 1. Screenshot of the detailed results in the eGovMon WCAG 2.0 checker

3 Results and Impact

In the course of the eGovMon project, we were able to gain many insights into how different target groups such as people working in public administration, politicians, accessibility experts, and software developers, use the results from web accessibility evaluation and benchmarking. There are few occasions were the presentation of a single number per web site was used for actual improvements (rather than mere blaming and shaming). Site owners who are really interested in improving their web site need more information.

EGovMon has developed an online interface[2] that presents a detailed report from the accessibility evaluation and explains why the identified issues might cause an accessibility problem and how to fix the issues. It also includes cross-references to the related WCAG 2.0 documents. Figure 1 shows an example of a report related to Common Failure *F89: Failure of Success Criteria 2.4.4,*

[2] The eGovMon WCAG 2.0 checker is available at `http://accessibility.egovmon.no/en/pagecheck2.0/`. The version of April 2012 provides tests for nine Success Criteria.

2.4.9 and 4.1.2 due to using null alt on an image where the image is the only content in a link. The interface was developed in close collaboration with the eGovMon users – a group of Norwegian municipalities – to ensure its usability and usefulness.

In addition to the detailed results, eGovMon also offers aggregated benchmarking results. So that national authorities in charge of monitoring web accessibility (such as the Agency for Public Management and eGovernment (Difi) in Norway) can make better use of the results.

4 Conclusion and Future Work

The main contribution of this paper is the suggestion to insert a new aggregation layer between individual tests and aggregated accessibility score for a page. This supports the understanding of the different concepts necessary for accessibility for instance "text alternative", "purpose of input elements", and increases the transparency of the evaluation because users of the results can trace how the results are computed. It also supports the validity of the results by establishing a strong link to WCAG 2.0 and its supporting documents.

Furthermore, we have presented some important aspects that have to be considered when building a tool that is consistent with WCAG 2.0. This includes the implicit and explicit definitions of when web content passes or fails a Success Criterion. The implementation of the eGovMon checking tool establishes a proof of concept for the approach.

Our plans for future development include the extension of the suggested framework to the evaluation of complete web sites. This involves addressing the following open questions:

- How can the score calculation accommodate results of tests that are applied on site level?
- How to distinguish Success Criteria that are not applicable from those for which no tests are available ("undocumented techniques")?
- How does the score function deal with conforming alternate versions?

In parallel, the checker user interface will be enhanced. It is our particular interest to add support for user input so that the report about the accessibility of a web page can combine results produced by the checking tool and findings entered by a human expert. The tool will actively prompt the users to enter their judgement for the results that are "to be verified".

Finally, to define a truly unified WCAG 2.0 score and thus achieve actual inter-tool reliability – as demanded by Vigo and Brajnik [6] a dedicated collaboration between tool developers and researches could be envisaged. We feel that credibility of web accessibility tools and metrics is crucial. There are many tools which all present different results for the same web pages. If the users are confused or alienated, the whole purpose of web accessibility evaluations is jeopardised. The users might stop caring about accessibility or they could be tempted to select the tool which gives the best results, and that, of course, is not the right way to improve accessibility.

Acknowledgements. The eGovMon project is co-funded by the Research Council of Norway under the VERDIKT program. Project no.: VERDIKT 183392/S10.

The ideas described in this paper were first presented [4] in the *Online Symposium on Website Accessibility Metrics* organised by the W3C/WAI RDWG in December 2011. The feedback received during the Symposium provided us with useful input to the consolidation and elaboration of our approach.

References

1. ACCESSIBLE Project: Web accessibility assessment Tool (WaaT) (2011), `http://www.accessible-project.eu/` (retrieved January 30, 2012)
2. Alonso, F., Fuertes, J.L., González, Á.L., Martínez, L.: Evaluating Conformance to WCAG 2.0: Open Challenges. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) ICCHP 2010. LNCS, vol. 6179, pp. 417–424. Springer, Heidelberg (2010)
3. Battistelli, M., Mirri, S., Muratori, L.A., Salomoni, P.: Measuring accessibility barriers on large scale sets of pages. In: Proceedings of W3C Online Symposium on Website Accessibility Metrics (2011)
4. Nietzio, A., Eibegger, M., Goodwin, M., Snaprud, M.: Towards a score function for WCAG 2.0 benchmarking. In: Proceedings of W3C Online Symposium on Website Accessibility Metrics (2011)
5. Vigo, M., Abascal, J., Aizpurua, A., Arrue, M.: Attaining Metric Validity and Reliability with the Web Accessibility Quantitative Metric. In: Proceedings of W3C Online Symposium on Website Accessibility Metrics (2011)
6. Vigo, M., Brajnik, G.: Automatic web accessibility metrics: where we are and where we can go. Interacting with Computers 23(2), 137–155 (2011)
7. Web Accessibility Benchmarking Cluster: Unified Web Evaluation Methodology (UWEM 1.2) (2007), `http://www.wabcluster.org/uwem1_2/` (retrieved January 30, 2012)
8. World Wide Web Consortium: Web Content Accessibility Guidelines 1.0. W3C Recommendation (May 5, 1999), `http://www.w3.org/TR/WCAG10/` (retrieved April 17, 2012)
9. World Wide Web Consortium: Web Content Accessibility Guidelines 2.0. W3C Recommendation (December 11, 2008), `http://www.w3.org/TR/WCAG20/` (retrieved January 30, 2012)
10. World Wide Web Consortium: A guide to understanding and implementing Web Content Accessibility Guidelines 2.0. W3C Working Group Note (January 3, 2012), `http://www.w3.org/TR/2012/NOTE-UNDERSTANDING-WCAG20-20120103/` (retrieved January 31, 2012)
11. World Wide Web Consortium: Techniques and Failures for Web Content Accessibility Guidelines 2.0. W3C Working Group Note (January 3, 2012), `http://www.w3.org/TR/2012/NOTE-WCAG20-TECHS-20120103/` (retrieved January 31, 2012)
12. Zeng, X.: Evaluation and Enhancement of Web Content Accessibility for Persons with Disabilities. Ph.D. thesis, University of Pittsburgh (2004)