

# Making Sense of Automatic Web Accessibility Evaluation: A Closer Look at the EIAO Data

Annika Nietzio  
Forschungsinstitut Technologie-Behindertenhilfe (FTB)  
der Evangelischen Stiftung Volmarstein, Grundschötteler Str. 40  
58300 Wetter (Ruhr), Germany  
eiao@ftb-net.de

## Abstract

We are anticipating the first large scale results delivered from the European Internet Accessibility Observatory (EIAO). So far only preliminary experiences of handling the data exist. In this paper we report the findings of two evaluations of the results, each addressing a different facet of the data. Furthermore, we show how the analyses can be used to guide the selection of an appropriate aggregation function for EIAO.

## 1 Introduction

In the Information Society, where a lot of information is made available on the web, it is essential to make this content accessible to all people including people with disabilities.<sup>1</sup> To get an overview of the accessibility status of a large number of sites manual evaluation by experts or disabled users can produce the most reliable results but often turns out to be too time-consuming and expensive. An automatic assessment of web accessibility is an alternative even though it can not perform all the necessary tests for a conformance claim. However, it can measure certain features that can be utilised as indicators for accessibility and allows the monitoring of a large number of web sites.

The web crawling and subsequent automatic accessibility assessment of EIAO<sup>2</sup> produce a large amount of data that is stored in a data warehouse. The data has to be preprocessed in order to enable disabled users or policy makers to draw meaningful conclusions. The relevant questions are amongst others:

- How accessible is this web site for a certain disability group?
- How accessible is this web site compared to previous versions or compared to other sites?

A conformance approach that summarises the evaluation result into a conformance category (WCAG A, AA or AAA [2]) is too coarse to answer these questions. Therefore, it is useful to present the results instead as a continuous quality measure that allows comparison and grading.

The remainder of this paper is organised as follows. After a short overview of web accessibility evaluation we describe the aggregation model currently used in EIAO and the different

---

<sup>1</sup>The European Union has defined eInclusion as part of the Lisbon strategy: “Ensure that every citizen should have the appropriate skills needed to live and work in a new Information Society for all.”

<sup>2</sup>The EIAO project is co-funded by the European Commission, under the IST contract 2003-004526-STREP.

aggregation functions that have been implemented and tested (section 2.1). The following sections present the analysis of some EIAO data with respect to two different aspects. The first experiment (section 3) studies the relevance of the results. To answer the question how good the aggregation model simulates the user experiences, we compare the ratings given by disabled users to the automatically generated results from the data warehouse. The second experiment (section 4) assesses the reliability of the aggregation results. We investigate whether the aggregation results provide an indicator for the accessibility situation. In this case expert evaluations serve as comparison data.

The discussion about the EIAO aggregation model is still ongoing. We expect that the results presented here will be valuable input to this discussion. In the conclusion we summarise the findings that can be used to guide the selection of an appropriate aggregation function for EIAO. We also point out some open questions and give prospects for future research directions.

## 2 Overview of Web Accessibility Evaluation

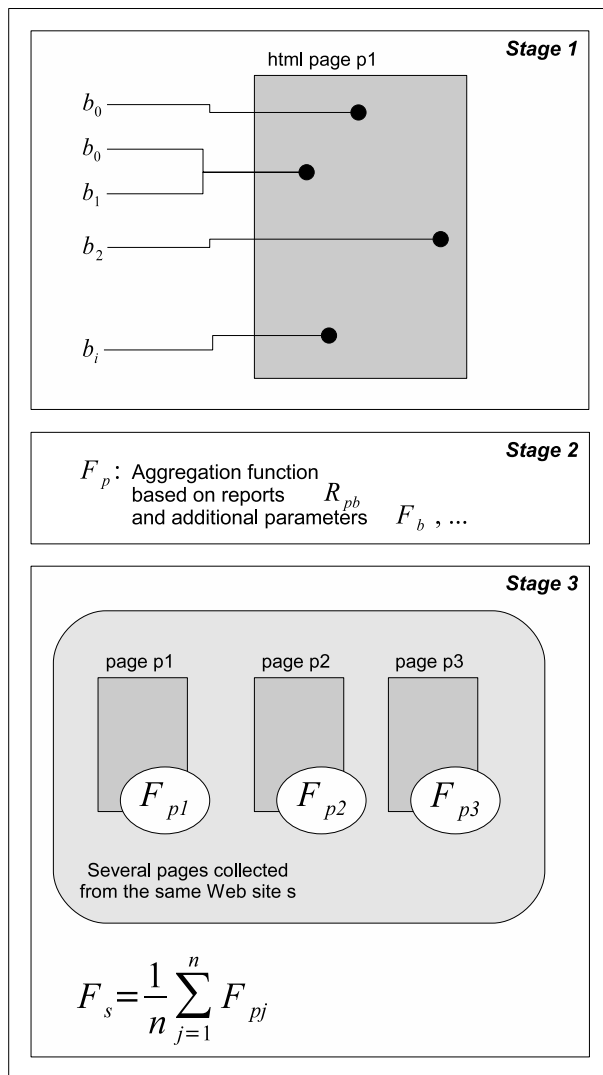


Figure 1: The web accessibility evaluation process

The adjoining figure gives an overview of the web accessibility evaluation process underlying EIAO. In the first stage a number of accessibility tests are performed. In EIAO these tests correspond to the subset of UWEM 0.5 tests [3] that can be performed automatically. However, it is also possible to include the results of manual (expert) testing.

The second stage performs aggregation of the individual test results into one comprehensive figure for a web page. The calculation can be based on different models. The model currently used in EIAO is presented in more detail in section 2.1. The first experiment addresses the analysis of the the result data from the second stage.

To present the results to the public the third stage of web accessibility evaluation needs to provide means of interpreting the results. This includes statistical analyses of the findings (e.g. average value and sample variance).

### 2.1 The EIAO Aggregation Model

We will use the following notation to refer to the quantities that are involved in the calculations.

$b$  is a **barrier type** (Each of the tests in stage one assesses a single barrier type.)

$u$  is a **disability group** (e. g. blind, hard of hearing, physically disabled)

$p$  is the web **page** that was inspected. The total number of reports for barrier type  $b$  within page  $p$  is denoted by  $N_{pb}$ .  $B_{pb}$  denotes the number of fail reports for barrier type  $b$ .<sup>3</sup>

A sum over all barrier types  $b$  yields  $N_p = \sum_b N_{pb}$  the total number of reports for  $p$  and  $B_p = \sum_b B_{pb}$  the total number of fail reports for  $p$ .

The **severity** of barrier type  $b$  for disability group  $u$  is given by  $F_{ub} \in [0; 1]$ .<sup>4</sup> If no special disability is considered the severity is denoted by  $F_b$ .

A web **site**  $s$  is represented by a set of pages  $s = \{p_1, p_2, \dots, p_n\}$ . Note that this set usually does not contain all the pages from the site but only a subset sampled by the crawler.

## 2.2 Aggregation Functions

An aggregation function computes a single value that summarises all the test results for a web page. This value can be used as an indicator for accessibility. In this study the following functions are included.

**UWEM 0.5.** The first release of EIAO is based on UWEM 0.5 which defines the following aggregation function:

$$F_p^{\text{UWEM05}} = 1 - \prod_{\text{all } b} (1 - F_b)^{B_{pb}}$$

The number of failed tests  $B_{pb}$  has a strong influence on the end result. This can be problematic because one fail result out of one contributes much less to the end result than ten fails out of ten. Even if both cases constitute a complete failure for the inspected barrier type.

**UWEM 0.9.** The next version of UWEM defines an indicator to solve this problem.

$$R_{pb} = \begin{cases} 1, & \text{if } B_{pb} > 0 \\ 0, & \text{if } B_{pb} = 0 \end{cases}$$

This leads to an aggregation function that is independent of the number of fail results for the same barrier type.

$$F_p^{\text{UWEM09}} = 1 - \prod_{\text{all } b} (1 - F_b)^{R_{pb}}$$

In this function the number of tests has no influence at all on the end result. The size and complexity of the web page are not taken into account. One fail result out of two contributes the same amount to the end results as one out of ten.

**EIAO.** Further development within the EIAO project [1] resulted in the definition of a special complexity parameter for the aggregation function.

$$C_{pb} = \frac{B_{pb}}{N_{pb}} + \frac{B_{pb}}{B_p}$$

This function takes into account the ratio of potential and actual barriers, and in addition the ratio of all failures to the number of failures for one barrier type. This additional contribution ensures that barriers are considered according to their overall proportion of occurrences within the web page.

$$F_p^{\text{EIAO}} = 1 - \prod_{\text{all } b} (1 - F_b)^{C_{pb}}$$

---

<sup>3</sup>In the current EIAO design the tests from stage one have only two possible outcomes: pass and fail.

<sup>4</sup>The severity is sometimes denoted by the term **barrier probability** (i. e. the probability that a barrier of type  $b$  is a barrier for disability group  $u$ ).

All three functions described above are based on the same statistical assumptions. Each barrier type  $b$  has a *fixed severity*. The overall barrier probability of a page  $F_p$  is the probability that the user encounters *any barrier* within the page  $p$ . Additionally, we assume that the barrier types are independent.

**Barrier Ratio.** For further comparison we also consider the simple barrier ratio as proposed for example by Sullivan and Matson in [4].

$$\text{Ratio}_p = \frac{B_p}{N_p}$$

By including this function in the comparison we can assess whether the complexity of the probabilistic model is justified or whether a simpler aggregation model would also be sufficient.

### 3 Relevance of aggregation results

To ensure that the results of the automatic accessibility assessment are relevant for disabled users the EIAO project has undertaken a user testing study. Twenty people with different disabilities and ten control group participants were given a set of tasks on web pages with known accessibility problems. The study consisted of ten tasks.

The participants rated the accessibility of the web pages on a seven value Lickert scale where 7 corresponds to “very satisfied” and 1 to “not at all satisfied”. For comparison we compute the average rating  $L$  and apply a linear transformation to map the rating to a barrier probability value.

$$F_p^{\text{USER}} = 1 - \frac{L - 1}{6}$$

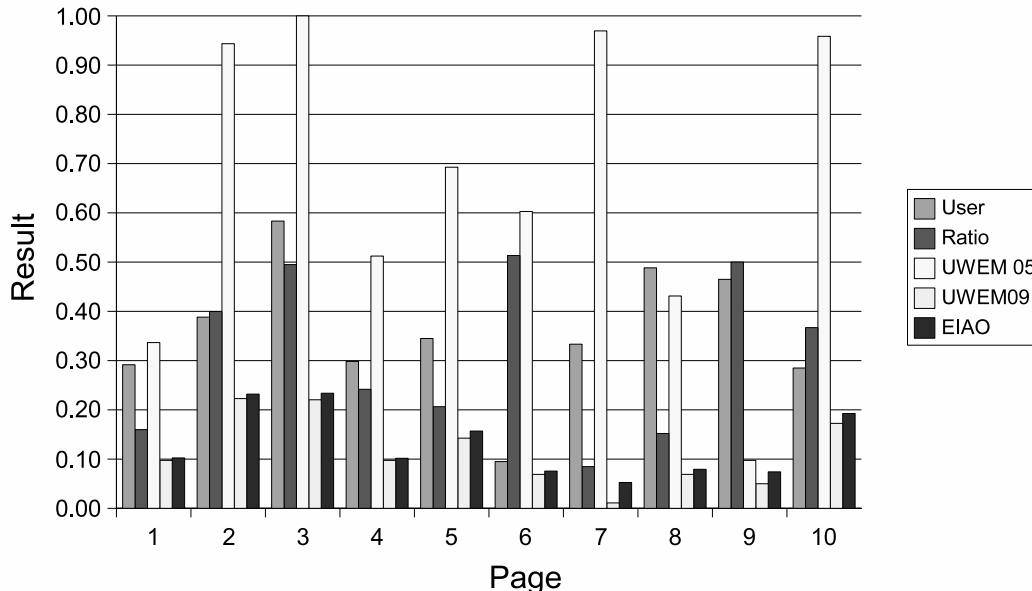


Figure 2: Comparison of aggregation results to user ratings (all users)

Figure 2 shows the results for all aggregation functions defined above. The single test results are provided by an automatic assessment performed by EIAO. The user results represent the ratings of all participants.

The aggregation functions – with the exception of the barrier ratio – support also the usage of a special severity parameter set targeted at a specific disability group. Figure 3 shows the results computed with the “blind” severity parameter set compared to the ratings from the blind participants.

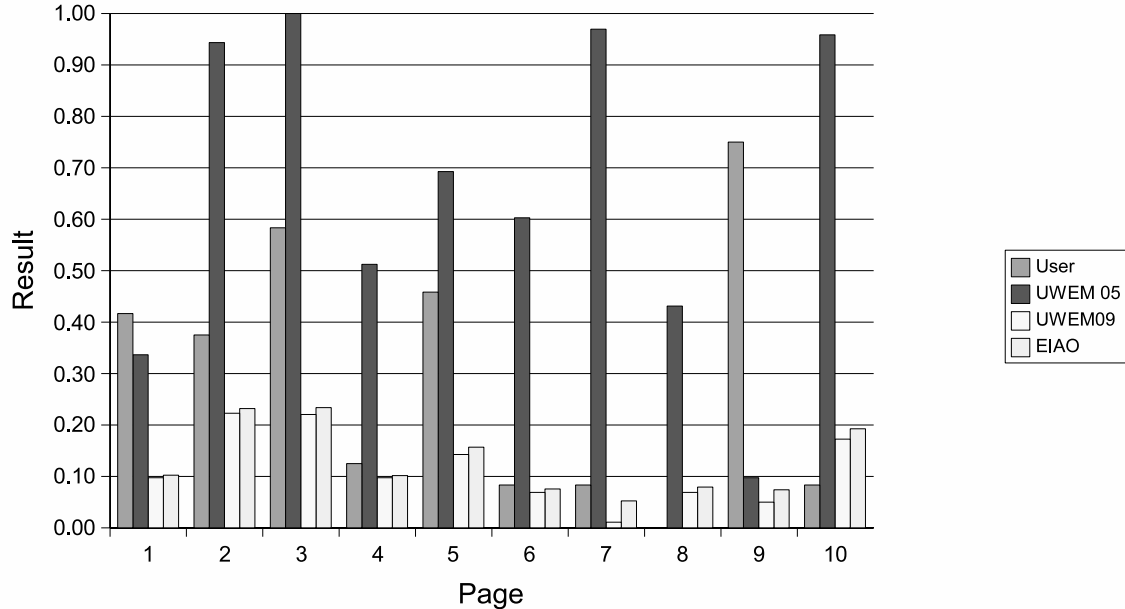


Figure 3: Comparison of aggregation results to user ratings (blind users)

The analysis is based on the data pairs (user rating & aggregation result) available for each web page. As measure for the similarity we use the average distance between the two values.<sup>5</sup> In table 1 the mean and variance of the distance values are summarised. Further statistical analysis (e.g. hypothesis testing) has not been undertaken because of the limited amount of data currently available.

		Ratio	UWEM05	UWEM09	EIAO	
					$(F_b = 0.05)$	$(F_b = 0.1)$
<b>all</b>	mean distance	0.1547	0.3820	0.2421	0.2272	0.1439
	variance	0.0185	0.0495	0.0175	0.0164	0.0126
<b>blind</b>	mean distance	n/a	0.5051	0.2122	0.2035	0.1778
	variance		0.0655	0.0461	0.0439	0.0288

Table 1: Average distance of user rating and automatic aggregation results

The evaluation shows that the values from the EIAO aggregation function have the smallest average distance. It is interesting to see that the simple barrier ratio measure also yields good results.

## 4 Reliability of aggregation results

For the credibility of the EIAO observatory it is crucial that the collected data is reliable. The presentation of incorrect or invalid results might lead the wrong conclusions.

<sup>5</sup>We use the distance  $|x_1 - x_2|$  instead of the simple difference  $x_1 - x_2$  to avoid that positive and negative deviations cancel out in the calculation of the average.

The idea of this experiment is to compare the observatory data to the accessibility report given by an expert evaluator. We have at our disposal the expert evaluations of four “prime minister” web sites (from the Netherlands, Norway, France and Germany), which were conducted by the WAB Cluster.

Unfortunately there are some problems with the data.

- The expert reports are quite coarse. Results are reported per WCAG 1.0 checkpoint (only priority 1) on web page level. Therefore the expert results are not very discriminative.
- The expert evaluations were performed by different people who seem to have slightly different interpretations of the methodology.
- The automatic evaluation performed by EIAO is based on UWEM 0.5 tests whereas the experts used UWEM 0.9.
- The automatic and manual evaluation samples differ in size and selection of pages.

We are aware that these facts are limiting the explanatory power of the comparison. Nonetheless, some preliminary conclusions can be drawn.

	Expert	UWEM05	UWEM09	EIAO	
				( $F_b = 0.05$ )	( $F_b = 0.1$ )
<b>minaz.nl</b>	0.1426	0.4072	0.0984	0.1089	0.2079
<b>dep.no</b>	0.1426	0.5907	0.0796	0.0847	0.1656
<b>premier-ministre.gov.fr</b>	0.2649	0.2402	0.0508	0.0553	0.1103
<b>bundestkanzlerin.de</b>	0.1426	0.1468	0.0492	0.0590	0.1164

Table 2: Aggregation results for Primeminister web sites

From the values summarised in table 2 it can be observed that all four aggregation functions introduce similar rankings. The ‘fr’ and ‘de’ sites always occupy first two ranks. The experts on the contrary give the lowest rank to the ‘fr’ site. A possible explanation is that the ‘fr’ site was evaluated by a different expert. This might indicate that the UWEM methodology has some problems with ambiguities that need clarification.

The UWEM 0.5 results are much higher than the other aggregation results due to the different handling of complexity.

The expert evaluation includes more tests than the automatic evaluation. It should be considered to take the number of barrier types into account in the aggregation function. Otherwise it will remain unclear how the comparison should be interpreted.

The second experiment revealed several open issues which have to be addressed by future experiments. The analyses conducted so far should be used as input for the design of the experimental setup.

## 5 Conclusion and Open Issues

The experiments show that the aggregation methods can capture the barrier probability of a web page to some extent. Especially the EIAO aggregation function yields promising results. Further experiments are needed to confirm this conclusion and refine the parameters.

**Reliability checking.** More extensive expert evaluations are needed to assess the reliability and usefulness of the automatic aggregation results.

**Parameter tuning.** The  $F_{ub}$  parameters represent the severity of barrier type  $b$  for disabled user  $u$ . The comparison of different parameter sets can give a hint for the tuning of these parameters. We are aware that the parameter space is huge. A possible first step in parameter tuning could consist of setting all parameters to the same value and try to optimise this value.

**Feedback from user testing.** The first experiment gives an indication how to select an aggregation function. This choice should be verified in the next iteration of EIAO user testing.

**UWEM 0.9 implementation.** The comparison in the second experiment is problematic because different version of the methodology are used. Once an implementation of UWEM 0.9 (1.0) by EIAO is completed the experiment should be repeated.

## Acknowledgements

The authors would like to thank Jenny Craven and Peter Brophy who designed and conducted the user testing experiments.

## References

- [1] Christian Bühler, Helmut Heck, Olaf Perlick, Annika Nietzio, and Nils Ulltveit-Moe. Interpreting results from large scale automatic evaluation of web accessibility. In *Proceedings of ICCHP 2006*, 2006.
- [2] W3 Consortium. Web content accessibility guidelines 1.0. Available at <http://www.w3.org/TR/WCAG10/>, 1999.
- [3] Web Accessibility Benchmarking Cluster. D-WAB2 Unified Web Evaluation Methodology (UWEM 0.5). Available from <http://www.wabcluster.org/uwem05/>, 2005.
- [4] Terry Sullivan and Rebecca Matson. Barriers to use: Usability and content accessibility on the web's most popular sites. In *Proceedings of ACM Conference on Universal Usability - CUU*, 2000.