# Web Accessibility Testing: When the Method Is the Culprit

Giorgio Brajnik

Dip. di Matematica e Informatica
Università di Udine
Italy
www.dimi.uniud.it/giorgio

**Abstract.** Testing accessibility of a web site is still an art. Lack of appropriate definitions of accessibility and of standard testing methods are some of the reasons why Web accessibility is so difficult to achieve.

The paper describes a heuristic walkthrough method based on barriers; it then discusses how methods like this can be evaluated, and it shows experimental data about validity and usefulness of the method when compared to *standards review*.

## 1 Introduction

Unless one has a clear notion of the property that has to be tested, the testing process is rarely successful. And when the property depends on human cognitive processes, then also the testing method may reduce the effectiveness of testing. This is indeed the case for web accessibility. On the one hand, there are several (more or less practical) definitions of accessibility. Sometimes accessibility is defined in terms of effectiveness; now and then it is defined in terms of usability; but unfortunately there are too often claims that a web site is accessible simply because an automatic testing tool yielded no error.

While accessibility can be tested based on guidelines (like WCAG 1.0, or Section 508) through a *standards review* method, other methods can be employed, like user testing [4] or usability inspection methods [14,7,12] or those suggested in [8]. Knowledge about the benefits and shortcomings of these methods applied to accessibility is still missing. If it were available, then methods appropriate to the specific situation at hand could be used, and results could be easily compared (over time for the same web site and type of users, or between different web sites or populations with differing characteristics).

To be really useful, evaluation methods should constrain the way in which the evaluator identifies problems and how they are graded in terms of importance. Only when these two kinds of decisions can be standardized, then the results produced can be used to rank web sites and to prioritize their bugs. Prioritization of defects is of paramount importance as any web developer required to fix them works always in a *scarce resource* mode.

The purpose of this paper is to propose and evaluate a heuristic walkthrough method to fill what I consider to be a gap in evaluation methods for accessibility

K. Miesenberger et al. (Eds.): ICCHP 2006, LNCS 4061, pp. 156–163, 2006.

assessments. Heuristics are interpretations and extensions of well known accessibility principles; within this method they are linked to user characteristics, user activities, and situation patterns so that appropriate conclusions about user effectiveness, productivity, satisfaction and safety can be drawn, and appropriate severity scores can be consequently derived.

## 2 Accessibility Testing

### 2.1 Problems in Accessibility Testing

As discussed in [1] at least three definitions of accessibility exist: some refer to usability, some to effectiveness[1] and some to other principles like perceivability, understandability and operability. The problem is that depending on the definition different methods have to be used to investigate. Running a user test doesn't make sense if we want to determine conformance to guidelines, and conversely conformance testing cannot be used to determine usability of the web site with respect to disabled users.

More confusion exists regarding the methods to use. For example, the current Italian regulation for web accessibility [13,6] specifies a number of technical requirements similar to WCAG 1.0 and Section 508 points. However, in order to achieve a certification mark evaluators have to perform a *cognitive walkthrough* (an analytical method for early–on usability investigations, that is yet unproven as a method for accessibility testing). In addition, the regulation specifies 12 general usability principles that are generally employed with a different method, namely *heuristic evaluation*. It also requires that evaluators classify identified problems into 5 severity levels, without specifying how severity should be determined. It then suggests empirical methods that have no proved effectiveness (*e.g. subjective assessments*) and finally it requires that evaluators compute a final score for the site on the basis of mean averages of severity levels associated to problems (an ineffective aggregation technique). Although such a regulation sets a certification framework for web accessibility, in my view it is unlikely to succeed because of extreme subjectivity and variability, poor practicality and measure–theoretical shortcomings.

Recently, other initiatives towards accessibility certifications took place [5]: but unless adoption of proper methods is prescribed, the desired connection between quality marks and actual accessibility levels will be missing, consistently with those that argue against any accessibility certification program [10].

Even simpler methods like *conformance testing* (called also *standards review*) are not problem–free. As discussed in [11], WCAG 1.0 guidelines suffer from their theoretical nature, dependency on other guidelines, ambiguity, complexity, their closed nature and by some logical flaws. Most of these comments apply as well to the current WCAG 2.0 draft, to Section 508 and to the Italian official

[1] In the following I assume that accessibility is defined as: "*. . . web sites are accessible when individuals with disabilities can access and use them as effectively as people who don't have disabilities*" [15, p. 3].

technical requirements. Guidelines are intentionally defined in terms that are independent from the technology used in implementing and in visiting web sites. As a consequence, guidelines are often too abstract to be directly applicable to a web site, creating a gap that has to be filled by the evaluator. In addition, they don't help the evaluator in distinguishing important problems from trivial ones. For example, few of the images in a web site that lack an appropriate alternative text are a true barrier: most of the images are used for emotional purposes, which in textual alternatives would almost always be lost anyway. But an important function asked to an evaluator is to tell what the consequences of such defects are: this, however, can be done only if appropriate use scenarios are considered. The *standards review* method does not tell how to choose scenarios and how to rate the defect, except for static priorities that cannot reflect these scenarios.

### 2.2 Requirements on Methods

Evaluation methods should be valid, reliable, useful and efficient. *Validity* is "the extent to which the problems detected during an evaluation are also those that show up during real–world use of the system" whereas *reliability* is "the extent to which independent evaluations produce the same result" [7,9]. *Usefulness* is the effectiveness and usability of the results produced (with respect to users that have to assess, or to fix, or otherwise to manage accessibility of a web site). Finally, *efficiency* is given by the resources being utilized during an evaluation (in terms of time, persons, skill level, facilities, money) related to some level of validity, reliability and/or usefulness.

Notice that especially validity and reliability are negatively affected by a number of factors: different evaluators may have different levels of knowledge in web accessibility, in assistive technology, in the relevant standards, in the behavior of browsers. Hence it is important that the method constrains as much as possible the subjectivity of the evaluators in identifying failure modes[2], in diagnosing them and in judging them.

## 3 Barriers Walkthrough

An *accessibility barrier* is any condition that makes it difficult for people to achieve a goal when using the web site through specified assistive technology (see figure 1 for an example). A barrier is a failure mode of the web site, described in terms of (i) the user category involved, (ii) the type of assistive technology being used, (iii) the goal that is being hindered, (iv) the features of the pages that raise the barrier, and (v) further effects of the barrier.

Barriers to be considered are derived by interpretation of relevant guidelines and principles [4,16]. A complete list can be found in [2].

[2] I use the term *failure mode* to mean the way in which the interaction fails; the *defect* is the reason for the failure, its cause; the *effects* are consequences of the failure mode.

To apply the *barriers walkthrough* method (BW), a number of different scenarios need to be identified. A scenario is defined by user characteristics, settings, goals, and possibly tasks of users belonging to given categories. At least categories involving blind users of screen readers, low–vision users of screen magnifiers, motor–disabled users of a normal keyboard and/or mouse, deaf users, and cognitively disabled users (with reading and learning disabilities and/or attention deficit disorders) should be considered.

| | |
|---|---|
| **barrier** | users cannot perceive nor understand the information conveyed by an information rich image (*e.g.* a diagram, a histogram) |
| **defect** | an image that does not have accompanying text (as an ALT attribute, content of the OBJECT tag, as running text close to the picture or as a linked separate page) |
| **users affected** | blind users of screen readers, users of small devices |
| **consequences** | users try to look around for more explanations, they spend substantial time and effort; effectiveness, productivity, satisfaction are severely affected |

**Fig. 1.** Example of barrier

User goals and tasks can be defined only with reference to the site being tested. For a web application, one should consider some of the possible goals and tasks usually documented in *use cases* and cross these goals with user categories to obtain the relevant scenarios. For information web sites, a sample of possible information needs can be considered and crossed with user categories. In this way, most of the times, each user goal/task will be associated to different sets of pages to test, and these will be crossed to user categories.

Evaluators then analyze these pages by investigating the presence of barriers that are relevant to the particular user category involved in the scenario.

Cross–checking a barrier to a set of pages in the context of a scenario enables evaluators to understand the impact that the barrier has with respect to the user goal and how often that barrier shows up when those users try to achieve the goal.

In the end, evaluators produce a list of problems, associating each problem to a barrier showing up in a given scenario, to a severity level, and possibly to performance attributes that are affected (*e.g.* effectiveness, productivity, satisfaction, safety).

## 4 Evaluating Testing Methods

Evaluations of a testing method should determine its validity, reliability, usefulness and/or efficiency. In this paper I report an experimental evaluation of validity and to some extent of usefulness of the BW method compared to conformance test (CT).

Validity and usefulness are determined by computing certain metrics on reports[3] written by $3^{rd}$ year university students doing their term projects in a class on web design. A comparative study was set up between reports produced using the BW method and reports of conformance tests (CT) based on WCAG 1.0 AA checkpoints. Nineteen different reports produced by 8 different student teams were analyzed; 8 reports were based on conformance test and 11 on barriers walkthrough. These reports were about 6 different public web sites; each web site was evaluated by at least 2 teams using both methods.

The metrics are based on whether a problem is a true or false one, and whether there are false negatives (*i.e.* problems missed by the report). In order to determine the true problems of a web site, the instructor[4] took all the reports regarding a web site and marked out the issues that are wrongly raised (*e.g.* students not having properly understood a checkpoint or barrier). Repetitions of the same problems were not counted, leading to a list of unique (true or false) problems. For reports produced by the CT method, that don't have the severity score, a new score was added by the instructor: when the issue is mentioned in the executive summary of the report it gets a severity score of 3, otherwise 1. A further severity score was added for each issue, representing the severity of the issue that the instructor deemed appropriate. Such a score is 0 for issues that are not a problem, and 1 to 3 otherwise.

Each issue therefore consists of the tuple (team, site, type of problem, method, team–severity, instructor–severity). Data were then aggregated so that at most a single tuple results for any combination of site, team, method and problem type. Severities were aggregated by computing the maximum (*i.e.* by deriving the highest value for the aggregated issues, which corresponds the worst case).

The set of true problems for a web site is then the union of all the problem types that were labeled as "true" by the instructor.

Comparison metrics include:

- *precision* (P), the percentage of reported problems that are true problems; in the following, $P$ is precision, whereas $P_3$ is the percentage of reported problems that have an instructor–assigned severity of 3;
- *sensitivity* (S), the percentage of the true problems being reported (notice that P and S are totally independent; for example, $P = 1$ also in the extreme case where only 1 true problem was reported, but the web site contains many other unreported problems); $S_3$ is the sensitivity with respect to the problems having instructor–assigned severity of 3;
- *fallout* (F), the percentage of false problems that are reported;
- *E–measure* (E): a combination of precision and sensitivity in the range $[0, 1)$: $E = PS/(\alpha P + (1 - \alpha)S)$; I set $\alpha = 0.5$ and then $E = 2PS/(P + S)$; $E$ is a monotonic and symmetric function on both arguments; $E_3$ is the combination of $P_3$ and $S_3$;

[3] Available on–line at `www.dimi.uniud.it/giorgio/dida/psw/galleria/galleria.html`. The experimental data used in this paper is available online at [2]. The BW method description that students followed is [3].

[4] And author of this paper.

- *mean severity* ($\overline{sev}$): the mean of the severity assigned by students to true problems;
- *n. of problems with severity=3*: the number of true problems associated to the highest severity level.

Precision, sensitivity, fallout and e–measure are related to method validity; mean severity, the number of high severity true problems, $P_3, S_3$, and $E_3$ are related to the usefulness of the method (*i.e.* its ability to focus the evaluator resources onto most important problems).

### 4.1 Results

Collectively the reports raised 303 problems (respectively 166 for CT and 137 for BW), with 260 true problems (86%).

Figure 2 shows the validity and usefulness metrics split by web site.

| **calabria** | **BW** | **CT** |
|---|---|---|
| P | 0.85 | 0.80 |
| $P_3$ | 0.08 | 0 |
| **S** | **0.48** | **0.70** |
| $S_3$ | 1 | 0 |
| F | 0.33 | 0.66 |
| **E** | **0.61** | **0.74** |
| $E_3$ | 0.14 | 0 |
| $\overline{sev}$ | 1.91 | 1 |
| n sev=3 | 3 | 0 |

| **campania** | **BW** | **CT** |
|---|---|---|
| P | 1 | 0.76 |
| $P_3$ | 0.24 | 0.12 |
| S | 0.75 | 0.46 |
| $S_3$ | 0.83 | 0.33 |
| F | 0 | 1 |
| E | 0.86 | 0.58 |
| $E_3$ | 0.37 | 0.17 |
| $\overline{sev}$ | 1.69 | 1.62 |
| n sev=3 | 6 | 4 |

| **fvg** | **BW** | **CT** |
|---|---|---|
| P | 1 | 0.69 |
| $P_3$ | 0.20 | 0 |
| **S** | **0.29** | **0.85** |
| $S_3$ | 1 | 0 |
| F | 0 | 1 |
| **E** | **0.45** | **0.76** |
| $E_3$ | 0.33 | — |
| $\overline{sev}$ | 2.2 | 1 |
| n sev=3 | 4 | 0 |

| **molise** | **BW** | **CT** |
|---|---|---|
| P | 1 | 0.73 |
| $P_3$ | 0.36 | 0 |
| S | 0.74 | 0.42 |
| $S_3$ | 1 | 0 |
| F | 0 | 1 |
| E | 0.85 | 0.53 |
| $E_3$ | 0.53 | 0 |
| $\overline{sev}$ | 2.43 | 2.25 |
| n sev=3 | 8 | 5 |

| **puglia** | **BW** | **CT** |
|---|---|---|
| P | 1 | 0.81 |
| $P_3$ | 0.60 | 0.41 |
| **S** | **0.57** | **0.63** |
| $S_3$ | 0.63 | 0.58 |
| F | 0 | 1 |
| E | 0.73 | 0.71 |
| $E_3$ | 0.62 | 0.48 |
| $\overline{sev}$ | 2.34 | 1 |
| n sev=3 | 12 | 0 |

| **toscana** | **BW** | **CT** |
|---|---|---|
| P | 0.95 | 0.84 |
| $P_3$ | 0.38 | 0.03 |
| **S** | **0.56** | **0.72** |
| $S_3$ | 0.89 | 0.11 |
| F | 0.11 | 0.89 |
| **E** | **0.70** | **0.78** |
| $E_3$ | 0.53 | 0.05 |
| $\overline{sev}$ | 2.10 | 1.17 |
| n sev=3 | 11 | 3 |

**Fig. 2.** Results for the validity and usefulness metrics split by web site; $\overline{sev}$ is the mean severity. Values highlighted in boldface are the only ones where CT scores better than BW. Notice that for each web site the sum of fallout is always 1 since there are only two methods being considered that can produce false positives.

In most of the cases BW is more valid and useful than CT. In fact:

**Precision.** In 4 cases out of 6, precision for BW is absolute (100%); it is never less than that of CT, never smaller than 85% and it is between 5 to 45%

higher than that obtained through CT. This means that while CT produces between 31 and 16% of false positives, BW yields 15% at the most.
When restricted to severity 3 problems, precision of BW drops, but it never goes below that of CT.

**Sensitivity.** In 4 cases BW yields a sensitivity lower (i.e. worse) than CT; the minimum for BW is 29% while for CT it is 42%. This means that BW yields a smaller proportion of true problems, which is explainable because some guidelines do not correspond to any barrier in the scenarios that were considered by students. On the other hand, by reducing the analysis to true problems with severity 3 then in all 6 cases BW scores higher than CT, and in 3 cases it reaches 100%. This means that BW is more effective than CT in identifying more severe problems.

**Fallout.** In all cases BW is better than CT (a smaller fallout means a smaller proportion of errors being produced). The highest fallout for BW is 33%, and in 4 cases fallout is 0.

**E–measure.** In 3 cases BW is better than CT, and viceversa. This metric ranges from 45 to 86%, and BW tends to stay in the lower part of the range. This is because e–measure is a balanced combination of precision and sensitivity, and the lower sensitivity of BW reflects also here.
But when we consider $E_3$ then BW scores always better than CT.

**Severity values.** In all cases the mean severity and the number of problems with highest severity are higher with BW. Therefore BW is more suitable to identify important problems.

## 5 Conclusions

Based on the sampled data, the barriers walkthrough method appears to be more valid and more useful than conformance tests.

An additional benefit of the method is its role in educating evaluators: after running evaluations with the BW method students become more knowledgeable of accessibility and assistive technology than when doing a dry checklist evaluation using the conformance test.

Additional investigations are needed to determine the reliability of the BW method and to generalize these results to a wider population of evaluators and web sites than those considered in the sample. Consider also that there are disturbance factors: not all the teams analyzed the same set of pages, web sites may have changed across two different evaluations, the cleverness and knowledge of the teams differ, and there might be judging bias. Only a larger study, involving a larger sample of reports, could lead to significant results. Nevertheless even such a small scale experiment provides encouraging results for a more wide spread adoption of the barriers walkthrough method and hopefully contribute to improved accessibility testing practices.

*Acknowledgments.* I would like to thank Martin Hitz for his help in improving an earlier version of this paper.

## References


1. G. Brajnik. Accessibility assessments through heuristic walkthroughs. In *HCI-taly 2005 — Simposio su Human–Computer Interaction*, Rome, Italy, Sept. 2005. www.dimi.uniud.it/giorgio/publications.html#hcihw05.
2. Giorgio Brajnik. Web accessibility testing with barriers walkthrough. www.dimi.uniud.it/giorgio/projects/bw, March 2006a. Visited Mar. 2006.
3. Giorgio Brajnik. Simulazione euristica per la verifica dell'accessibilità. www.dimi.uniud.it/wq/metodo-barriere.html, Feb 2006b. Visited Mar. 2006.
4. DRC. Formal investigation report: web accessibility. Disability Rights Commission, www.drc-gb.org/publicationsandreports/report.asp, April 2004. Visited Jan. 2006.
5. European Committee for Standardization. Web accessibility certification workshop. www.cenorm.be/cenorm/businessdomains/businessdomains/isss/about_isss/draft_cwas.asp, March 2006.
6. Italian Government. Requisiti tecnici e i diversi livelli per l'accessibilità agli strumenti informatici. www.pubbliaccesso.it/normative/DM080705.htm, July 2005. G. U. n. 183 8/8/2005.
7. W.D. Gray and M.C. Salzman. Damaged merchandise: a review of experiments that compare usability evaluation methods. *Human–Computer Interaction*, 13(3): 203–261, 1998.
8. S.L. Henry and M. Grossnickle. *Accessibility in the User–Centered Design Process*. Georgia Tech Research Corporation, Atlanta, Georgia, USA, 2004. On–line book, www.UIAccess.com/AccessUCD.
9. M. Hertzum and N.E. Jacobsen. The evaluator effect: a chilling fact about usability evaluation methods. *Int. Journal of Human–Computer Interaction*, 1(4):421–443, 2001.
10. European Information and Communications Technology Industry Association. White paper on eAccessibility. www.eicta.org/press.asp?level2=41&level1=6&level0=1&docid=564, Oct 2005.
11. B. Kelly, D. Sloan, L. Phipps, H. Petrie, and F. Hamilton. Forcing standardization or accomodating diversity? A framework for applying the WCAG in the real world. In S. Harper, Y. Yesilada, and C. Goble, editors, *Int. Cross–Disciplinary Workshop on Web Accessibility, W4A*, pages 46–54, Chiba, Japan, April 2005. ACM.
12. Jakob Nielsen. *Usability engineering*. Academic Press, Boston, MA, 1993.
13. Parlamento Italiano. Disposizioni per favorire l'accesso dei soggetti disabili agli strumenti informatici. www.parlamento.it/parlam/leggi/040041.htm, Jan. 2004. Legge del 9 gennaio 2004, n. 4.
14. J. Preece, Y. Rogers, and H. Sharp. *Interaction design*. John Wiley and Sons, 2002.
15. John Slatin and Sharron Rush. *Maximum Accessibility: Making Your Web Site More Usable for Everyone*. Addison-Wesley, 2003.
16. W3C/WAI. How people with disabilities use the web. World Wide Web Consortium — Web Accessibility Initiative, w3.org/WAI/EO/Drafts/PWD-Use-Web/20040302.html, March 2004.