Beyond Conformance: The Role of Accessibility Evaluation Methods

Giorgio Brajnik

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

1 Introduction

The topic I want to address is the role that accessibility evaluation methods can play in helping the transition from accessibility viewed as standard conformance, to a user-centered accessibility. As we will see, this change sets additional requirements on how evaluations of websites should be carried out.

This paper first discusses different problems that occur while dealing with accessibility. We will see that different people have radically different views of accessibility and how it should be assessed.

The first requirement is a clear definition of what accessibility is and how it should be assessed. The accessibility model discussed in Section 2.1 has precisely this role.

Several existing evaluation methods are then reviewed and discussed, a simple taxonomy is presented, and differences that occur when evaluating accessibility rather than usability are pinpointed.

1.1 Problems in Managing Web Accessibility

As discussed by Kelly et al. [22], the W3C/WAI model of accessibility aims at *universal accessibility*, it assumes that website conformance to WCAG (Web Content Accessibility Guidelines) is the key precondition to that, and it hypothesizes that accessibility is entailed by a conformant website if two other conditions are met. Namely, that the tools used by the web developer (including CMSs) are conformant to ATAG (Authoring Tools Accessibility Guidelines), and that browser and assistive technology used by the end user are conformant to UAAG (User Agent Accessibility Guidelines). However, since both these two conditions are not under control of the web developer, the conclusion is that the developer cannot guarantee accessibility, whatever efforts s/he may put it.

In fact, empirical evidence shows that the link between conformance and accessibility is missing, *i.e.* even conformant websites may fail in being accessible [13, 29].

Confusion exists regarding the methods to use. For example, the current Italian regulation for web accessibility [19] specifies a number of technical requirements similar to WCAG 1.0 and Section 508 points. However, in order to certify

S. Hartmann et al. (Eds.): WISE 2008, LNCS 5176, pp. 63–80, 2008.

accessibility evaluators have to perform a *cognitive walkthrough*, that is an analytical method generally used for early-on usability investigations, whose effectiveness as a method for accessibility evaluations is yet unproven. In addition, the regulation specifies 12 general usability principles that are generally employed with *heuristic evaluation*. It also requires that evaluators classify identified problems into 5 severity levels, without specifying how severity should be determined. It then suggests using an empirical method that again has no proved effectiveness (*i.e. subjective assessments*) and finally it requires that evaluators compute a final score for the site on the basis of mean averages of severity levels (an ineffective aggregation technique of ordinal variables). 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.

As evidence of further confusion, consider the Target legal case in the U.S.A.[1]. The National Federation for the Blind (NFB) claimed that `target.com` is not accessible since some NFB's witnesses gave up when using the site; on the other hand, Target's witnesses testified that they were able to navigate, shop and that they actually enjoyed it; in addition, an NFB's expert declared in court that `target.com` fails to address accessibility since:

> ... 15 of the site's pages were analyzed: six top-level pages as well as nine pages that had to be navigated in order to complete a purchase. In those fifteen pages, alt-text was missing on 219 active images (links); none of the form controls were properly labeled; and there was no accommodation for screen reader or keyboard navigation, such as skip links or HTML headings.

Finally, the Court concluded that the question of the accessibility of `target.com` was not decided and so it refused to grant a preliminary injunction.

We can see that there is substantial variability, and lack of standardization, in the way pages were selected, in the way accessibility was investigated, and in the way a conclusion was drawn. Witnesses of one side were referring to user performance indicators, the others to conformance features.

Additional evidence exists showing that accessibility evaluation based on a sample of pages (sampling is necessary for all but trivial websites) can be affected by the criteria used to select the sample. There is interdependence between the sampling criteria and the purpose of the accessibility analysis [8], leading to large differences in accuracy. If the evaluation aims at conformance, then the most frequently used sampling criterion (selecting predefined pages: home, contact, site map, etc.) may lead up to a 38% inaccuracy rate, *i.e.* 38% of the checkpoints may be wrongly estimated.

My claim is that to change this state of things we have to focus on how to standardize methods, and through them aim at an accessibility that is sustainable; in other words, we need to shape and establish effective accessibility processes that can be sustained mainly by their own return on investment.

[1] See `www.jimthatcher.com/law-target.htm` for details.

At least two issues have to be addressed. First, accessibility evaluations have to produce sets of accessibility problems that are prioritized by their impact: in other words, evaluations should identify problems whose solution makes a difference in accessibility as viewed by stakeholders. Therefore, evaluators and developers can focus on these problems first, and optimize their resources. Secondly, accessibility processes (taking place when conceiving, developing, maintaining, revamping websites) should be effective and efficient, and these properties should be the result of scientific investigations. When these two conditions are met, then accessibility methods can be compared and chosen on an informed basis, and this will lead to more accessible websites/web applications that in turn will positively affect key performance indicators related to the underlying business the website should support.

The relation between accessibility and usability is also controversial. According to Thatcher et al. [36], accessibility problems affect only disabled people and have no effect on others. Petrie and Kheir [29] mention that they noticed that disabled and non-disabled people often encounter the same problems, but are affected by them differently. Slatin and Lewis [33] performed an experiment aimed at determining whether accessibility features of a website positively affect non-disabled users. While vision-impaired subjects improved their success rate and productivity when using the accessible version of a website, no statistically significant difference was found for non-disabled subjects. The implication of this study is that accessibility does not necessarily lead to higher usability. On the basis of a comparative experiment between vision-impaired and sighted users, Petrie and Kheir [29] reach the conclusion that the problems faced by the two groups were overlapping, but the overlap was small, and the majority of the problems were found only by disabled users. This study did not detect significant differences in the severity of the problems found by the two groups.

We can see that accessibility and usability are different; my view is that currently a good accessibility model is missing.

2 The Role of Accessibility Models

2.1 An Accessibility Model

A model of accessibility specifies what accessibility is, how it is achieved and managed, and which boundary conditions can influence it. A model not only helps to plan and perform activities like diagnosing the defects of a website, monitoring it and comparing it to other websites, measuring its accessibility level to determine whether it is conformant to certain standards.

More specifically the accessibility model I propose addresses the following questions and comprises the following components.

Properties. Which properties should be central in the notion of accessibility? To respond to this question, if we look at definitions that were proposed for accessibility in the past (see Table 1), it's clear that very different properties are taken into account. Some definitions focus on user performance

Table 1. Existing definitions of accessibility

Source	A website is accessible if ...
W3C/WAI [45]	... its pages transform gracefully despite constraints caused by physical, sensory, and cognitive disabilities, work constraints, and technological barriers, and its content is understandable and navigable.
Slatin and Rush [34], U.S. Dept. of Justice [39]	... disabled people can use it with the same effectiveness as non-disabled people.
Thatcher et al. [35]	... it is effective, efficient and satisfactory for more people in more situations.
ISO [18]	... it is usable by people with the widest range of capabilities
Italian Parliament [20]	... deploys services and information so that they can be exploited with no discrimination also by disabled persons.
Thatcher et al. [36]	... people with disabilities can perceive it, understand it, navigate it and interact with it.
Petrie and Kheir [29]	... it can be used by specific users with specific disabilities to achieve specific goals with effectiveness, efficiency and satisfaction in a specific context of use.
W3C/WAI [46]	... its content must be perceivable to each user; user interface components in its content must be operable by each user; content and controls must be understandable to each user; content must be robust enough to work with current and future technologies.
College of Design, North Carolina University [10]	... it is usable by all people, to the greatest extent possible, without the need for adaptation or specialized design.

indicators that can be experimentally measured (*e.g.* effectiveness, usability[2]), one definition sets appropriate relative levels (*e.g.* same effectiveness), other definitions focus on properties that are more difficult to define and measure (*e.g.* navigability, understandability, exploitation); sometimes even properties unrelated to user-performance properties are considered (*e.g.* robustness, degradation). The last definition refers to Universal Design, which is often considered to be the same as accessibility. Such a definition excludes many contextual elements that are central in the definition of usability, reducing in such a way the power of AEMs, as we will see below.

In this paper I will assume that a website is accessible when

> specific users with specific disabilities can use it to achieve specific goals with the same effectiveness, safety and security as non-disabled people.

[2] *Effectiveness* is the accuracy and completeness levels that can achieved by specified users when aiming at specified goals under specified conditions. *Usability* is the effectiveness, productivity, satisfaction and security with which specified users can achieve specified goals under specified conditions; *productivity* is related to the resources expended (time, effort, skills, infrastructure) in achieving those goals at given levels of effectiveness (ISO 9241). See books like [7, 31] for relevant metrics.

This definition points to measurable user-performance parameters, sets viable, relative thresholds and restricts the claim to certain users and goals.

Context. Which additional factors influence accessibility and how can they be detected, isolated and controlled?

As we move from a viewpoint were accessibility is equated to conformance to some standard, to a view were accessibility becomes user-centered, then context plays an increasingly important role and needs to be considered whenever accessibility is evaluated. Context should provide enough information to enable evaluators to determine possible hurdles for users and their consequences.

Ideally context should address the "who", "what" and "how" questions: (i) the type of user disability, (ii) the experience level in using the browser, the Web, the assistive technology, and possibly the specific website and domain of operation, (iii) the short-term user goals, (iv) the physical environment the user is working in (posture, light and noise conditions), (v) input and output devices and interaction modalities (media used, possible user actions and operations, user agents, assistive technologies and infrastructure).

Methods. Given that we know on what properties to focus, and how to characterize boundary conditions, how are we going to detect and measure these properties accurately and reliably?

This ingredient of the accessibility model comprises techniques, methods and methodologies used to evaluate, assess and manage accessibility. As we will see later on, there are a number of evaluation methods usually put in operation for accessibility; some of them are adapted from usability methods, others are specific to accessibility. Nonetheless, few studies are available to shed light on how well these methods work for accessibility, making the choice of the evaluation method and the choice of the metric to use for measuring accessibility very uncertain.

2.2 The Importance of Context

Context is more crucial for accessibility than it is for usability. Besides being dependent on users' experience, goals and physical environment, accessibility of a website depends also on the platform that's being used. It is the engine of a transformation process that is not under control of the web developer. In fact, accessibility of digital media requires a number of transformations to occur automatically, concerning the *expression* of the *content* of the website [1, 7]. Content is the meaning that a person (*e.g.* visitor, developer, evaluator) associates to perceivable elements of a web page, which constitute its expression. See Table 2 for a brief taxonomy. Examples of transformations include text that could be read aloud; images that could be "transformed" into spoken words (via their textual equivalents); scenes of a video that could be enriched with textual captions describing them; audio content that could be transformed into textual transcripts; changes in font attributes.

These changes in expression involve *inter-media* transformations (*e.g.* text to spoken words), *intra-media* transformations (*e.g.* by changing the geometric

Table 2. Taxonomy of content and expression

Content elements	
Interest information	Concepts, questions, answers that can satisfy users goals
Bearing information	Location information (*e.g.* breadcrumbs, headings), direction information (*e.g.* link labels)
Access information	Supports user actions (*e.g.* navigation bars, sequential paging, filters)
Functional information	Provided by users and necessary to achieve the goal (*e.g.* address data provided when completing an on-line order)
Expression elements	
Expression media	Text, image, sound, video
Expression style	Font, size, colors, texture, orientation, ...
Compositional structure	Spatial, temporal, spatio-temporal, or hyper-medial

properties of space when using a screen magnifier to enlarge the screen, or when changing the text size), *temporal* transformations through new synchronization of events (*e.g.* by using audio signals to notify a user of a screen reader that a certain feedback message has appeared in a location that differs from the current focus of interaction) or slow-down of an animation/simulation; finally *de-contextualization* of information occurs (*e.g.* when the user of a screen reader extracts and lists all the links in a page, so that each link is rendered out of its original context).

Some of these transformations affect interaction modality. For example, new operations are made available, like the ability to extract and scan links in a page, or page headings, or the ability to jump directly to the content of the page, or the ability to move back and forth through items of a list. In a sense, this perspective on transformations occurring for the sake of accessibility is close to the notions of *plasticity* and *retargeting* of the user interface[3]. Notice however that these transformations occur on the fly and solely on users' platform.

As a consequence of the definition of accessibility I gave before, a website is accessible only if the transformation of web pages from one expression to another is such that *invariance of content* is preserved, in specified contexts. In other words, regardless of the expression and interaction modality used by the visitor and within given contexts, the same content is rendered, reaching the same level of effectiveness. Finally, many diverse transformations have to be enabled for each "target" interface required by the assistive technology that is considered in the model.

Invariance of content holds if a number of enabling conditions are met. First, the platform should support all required transformations and the technologies

[3] Plasticity is the ability of the system to produce a user interface that is adapted to the device being used and possible context of use. Retargeting means to statically analyze a web page, to automatically derive an abstract user interface (*e.g.* by inferring the existence of an abstract object called "RadioButton"), to transform such an abstract interface into the abstract interface for another platform (*e.g.* on a mobile device the "RadioButton" object is mapped into a "Listbox"), and finally to generate an appropriate and running user interface on the selected platform [9, 38].

used by the website (*e.g.* HTML, CSS, JavaScript, SMIL, Flash, PDF). Second, website developers need to provide the required redundant expression in the different media that might be needed (*e.g.* textual descriptions of video scenes). They also have to provide specifications to support transformation of expression (*e.g.* synchronization constraints so that captions are rendered at the right time). Thus, from the perspective of authors, accessibility requires them to clearly identify all the content units and make sure that (i) interest and bearing information can be transformed into all possible media that might be available in users platforms, (ii) that the transformations are complete (*e.g.*, all bearing information is transformed) and (iii) that operability is guaranteed, *i.e.* all functional information and access information can be operated in the transformed interaction modality.

Context can affect all three of these conditions, which don't usually occur when dealing with usability; this is why it plays a more important role in accessibility and why it poses more challenges to evaluators and developers. Therefore, in order to be accurate and produce relatively standardized results, evaluation methods need to consider context.

3 Accessibility Evaluation Methods

With *accessibility evaluation method* (AEM) I mean a procedure aimed at finding accessibility problems, such as guideline violations, failure modes, defects[4], or user performance indexes. More specifically an AEM:

1. prescribes which steps, which decisions, which criteria should be used under which conditions, so that accessibility problems can be detected;
2. may prescribe how to classify and rate problems (in terms of severity, priority, or else);
3. may prescribe how to aggregate data about problems, as well as how to describe and report them;
4. may prescribe how to select web pages for evaluation.

3.1 A Taxonomy of Accessibility Evaluation Methods

Several methods can be used to find accessibility problems; they are reviewed in Section 3.2. Before discussing each of them in detail, however, I provide a taxonomy highlighting criteria that can be useful to contrast them; see Figure 1. Some of the criteria illustrated below were discussed also by Hartson et al. [15].

[4] *Failure mode* to the way in which the interaction fails; the *defect* is the reason for the failure, its cause; *effects* are the negative consequences of the failure mode. In this context, an *error* is a wrong design/implementation decision taken by developers. For example, a failure mode may be the inability of a screen reader user to swiftly navigate around elements of a web page; a corresponding defect may be the absence of *skip links* links and of page headings; effects include a reduction of user productivity, satisfaction and a dramatic reduction of effectiveness if the user — each time a new page is reloaded — has to repeatedly press the TAB key to get to the desired content of the page.

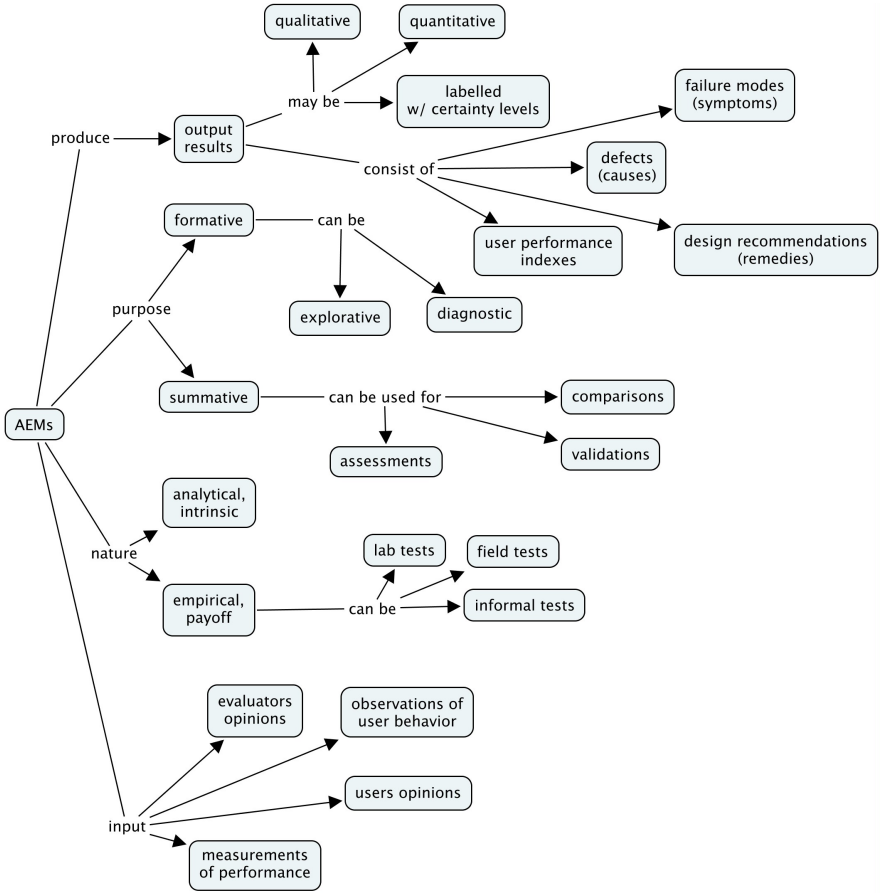

Fig. 1. The taxonomy of accessibility evaluation methods

Methods can be *analytic*, *empirical*, or both: the former are based on inspections of web pages usually carried out by experienced evaluators, without putting pages in a real work context. Empirical methods, sometimes used to perform so called *payoff* evaluations, require that an interaction takes place between users and the studied website. Empirical methods may be *laboratory based*, when potential disturbances to the user interaction are reduced to the minimum, or *informal ones*, when the strict "aseptic" conditions are not needed.

Methods differ also according to the *information used* to derive accessibility problems: some methods are based on observations of the behavior of users; others on opinions expressed by users or by evaluators.

In terms of results produced, AEMs can yield descriptions of *failure modes*, or may produce also corresponding *defects* and even design recommendations, *i.e. solutions*. Some methods produce synthetic measures of users' performance indicators, called payoff functions (*e.g.* effectiveness measured as success rate).

Results can be qualitative or quantitative; they may also be qualified with a confidence degree (like the probability of being a wrong result) and can support generalization to a wider population of Web users and/or to a wider set of conditions.

Regarding their purpose, methods can be used to perform *formative* evaluations aimed at identifying lists of problems. These methods can be used to explore failure modes that are the accessibility obstacles to a smooth interaction (*explorative* evaluations); they can also identify defects and solutions so that problems can be fixed (*diagnostic* evaluations). Formative evaluations are (or should be) used during the development, supporting iterative design. *Summative* evaluations, on the other hand, are carried out to assess the accessibility level of an interface; differently than formative evaluations, summative AEMs may produce only aggregated results regarding user performance measures (*e.g.* global effectiveness, productivity, user satisfaction figures). They can be deployed to estimate the accessibility of an interface (*assessment*), to *validate* it or to *compare* one interface against different versions or different systems. If summative evaluations are used for comparing different websites that may be used within different contexts, which is what happens when conformance is assessed, then some sort of standards of conditions should be defined. Only in these cases the results of the evaluation can be compared safely.

3.2 Review of Existing Methods

Although close to usability, accessibility has its own evaluation methods, and few are in common. I will briefly review the most typical ones, highlighting their benefits and disadvantages. Table 3 summarizes this discussion.

Ideally, a good method is a dependable tool that yields accurate predictions of all the accessibility problems that may occur in a website. This is why methods are compared in terms of such criteria as *correctness* (the percentage of reported problems that are true problems), *sensitivity* (the percentage of the true problems being reported), *reliability* (the extent to which independent evaluations produce the same results), *efficiency* (the amount of resources expended to carry out an evaluation that leads to specified levels of effectiveness and usefulness), *usefulness* (the effectiveness and usability of the produced results) and the method's *usability* (how easily it can be understood, learned and remembered by evaluators); for more details the reader is referred to [14, 15, 17, 23, 32].

Conformance Reviews. Called also *expert, standards, or guidelines review* or *manual inspection* [16, 42], this is by far the most widely used AEM [12]. It is based on checking whether a page satisfies a checklist of criteria. It is an analytic method, based on evaluators' opinions, producing failure modes (in the form of violated checkpoints) possibly with defects and solutions. Conformance reviews are used in both formative and summative evaluations: the former when defects are diagnosed in order to be fixed, the latter when a formal conformance statement is needed (for assessment, validation or comparison).

The method often entails the following steps [19, 42]: (i) determining an appropriate sample of web pages (including different sorts of tables, forms, images

and scripts), (ii) running markup validators on selected pages, (iii) cross-checking selected pages against all applicable checkpoints, (iv) examining selected pages with a range of graphical, textual and voice browsers, and finally (v) summarizing the results.

Benefits of this method include the ability to identify a large range of diverse problems for a diverse audience (albeit this depends on the quality of the underlying checkpoints); it is relatively cost-effective, especially when coupled with automatic testing tools, and, by being diagnostic in nature, it can be used to identify the defects underlying the checkpoint violations, hence assisting those who have to fix them.

Conformance reviews are dependent on the chosen checklist, that range from standards issued by international bodies (like the Web Content Accessibility Guidelines, WCAG, published by the W3C), to national or state-level guidelines, like [19, 40], to individual organizations guidelines (like those issued by IBM, SUN or SAP, for example). Guidelines of course affect the quality of this method. As discussed in [21], WCAG 1.0 guidelines suffer from their theoretical nature, dependency on other guidelines, ambiguity, complexity, their closed nature and by some logical flaws.

The study [13] found fundamental limits of conformance review with respect to WCAG 1.0: "as many as 45% of the problems experienced by the user group were not a violation of any checkpoint, and would not have been detected without user testing". The study identified gaps in the guidelines such as reducing deep structures in websites, improving search mechanisms, preserving links to home pages, reducing the number of existing links; the study suggested also a reordering of checkpoint priorities.

The project by Theofanos and Redish [37], performed as a field study with 16 users over a period of months, highlights several guidelines that address usability for screen reader users and that go beyond conformance: starting links with significant words, rewording questions with the main topic first, writing "home page" rather than "homepage", avoiding page refresh, synchronizing alt-text with text in the page, are some of the 32 suggested (additional) guidelines. This holds true also for the guidelines proposed by Nielsen Norman Group [27, 28]. In agreement with [24], the study found that an additional shortcoming of conformance review is the large number of possible guidelines and principles to choose from.

Rating the severity of problems through analytical methods appears also to be a source of methodological weakness. Petrie and Kheir [29] showed that while participants and experimenters agreed substantially on assigning severities to problems found via empirical methods, the agreement on these severities with checkpoint priorities in WCAG 1.0 was extremely poor. The same happened with respect to usability guidelines. This result suggests that it's extremely inaccurate to use fixed predefined priorities/severities. For example, few of the images in a website that lack an appropriate alternative text are a true barrier: most of the images are used for emotional purposes, which in textual alternatives would be lost anyway. But an important function of an evaluator is to find out what the

Table 3. Summary of pros and cons of AEMs

pros	cons
conformance review (CR)	
– low cost – diagnostic – suitable for formative and summative eval. – identifies a large spectrum of problems	– requires skilled evaluators – does not support the evaluator in assigning severities – not practical with lots of pages – conformance does not mean accessibility – unable to catch important usability problems – guidelines may be complex to read, too abstract, too many – with inexperienced evaluators, it is less effective than other methods
subjective assessment (SA)	
– low cost, low difficulty – can be done remotely – good correctness	– not systematic (problems and/or pages) – highly dependent on users' experience – users may not be aware of certain problems – poor description of problems – low thoroughness – requires users with different disabilities
screening techniques (ST)	
– low cost – suitable for formative eval.	– time consuming for web developers – singles out certain disabilities – yields developers opinions – highly dependent on developers experience – cannot be used for summative eval.
barrier walkthrough (BW)	
– low cost, low difficulty – supports learning – higher correctness than CR – yields severity ratings	– lower sensitivity than CR – dependent on evaluators experience – less reliable than CR
user testing (UT)	
– highlights important problems – leads to correct severity ratings	– higher cost than analytical methods – logistics is complicated – mixes accessibility with usability problems – should not be done remotely

consequences of such defects are: this, however, can be done only if appropriate use scenarios are considered. Conformance review does not prescribe how to choose scenarios nor how to rate the defect, except for static priorities that cannot reflect specific usage scenarios.

Automated Tests. Though closely related to conformance reviews and very popular, methods that are based exclusively on automated testing tools, like those listed in [44], should not be considered *evaluation* methods. The reason is that these tools have to rely on heuristics to determine violations of several checkpoints. The quality of these heuristics is not satisfying: in a previous study [4] we found false positives to be up to 33% and false negatives to 35%; Thatcher et al. [36] found that of 40 different benchmark tests the best and worst of six tools passed respectively 23 and 38 tests, a failure rate between 5% and 42%.

Therefore, using *only* automated tools is not by itself a viable solution to the problem of evaluating accessibility. The W3C/WAI puts it nicely[5]: "Web accessibility evaluation tools can not determine the accessibility of Web sites, they can only assist in doing so." On the other hand, because they are systematic and fast, tools yield other important advantages, namely effectiveness, productivity, and wide coverage of web pages. They are therefore an important option in the portfolio of a careful evaluator.

An interesting approach is to consider context by tailoring the automatic evaluation to peculiarities of certain types of users in the form of personal accessibility profiles; see for example [41].

Screening Techniques. These are informal empirical techniques based on using an interface in a way that some sensory, motor or cognitive capabilities are artificially reduced [16]. For example, an evaluator would use a website through a screen reader with the monitor turned off; or after unplugging the mouse; or by using the mouse with the left hand (for a right-handed person). After carefully selecting the screening conditions so that they match the characteristics of the target population, the evaluator explores the website and tries to accomplish selected goals. Hindrances that occur during such a process are accessibility problems that these empirical, informal, explorative techniques can detect.

Their benefits include the relatively low performing costs (the evaluator has to install and learn how to use a number of assistive technologies, but this is a one-time cost). However it is a method that is not systematic, and we should expect it to show low effectiveness since it depends heavily on the experience level of the evaluator in using the assistive technology, which rarely would match the experience of users.

Mankoff et al. [24] report that web developers using the screen monitor together with the screen reader were able to reach good levels of sensitivity which are comparable to conformance review.

Subjective Assessments. Rubin [31] calls them *self-reporting methodologies*. When applying this method, the evaluator involves a panel of users (sharing

[5] `www.w3.org/WAI/eval/selectingtools.html`

characteristics with the reference audience), instructs them to explore and use a given website, which they do by individually or jointly with other users. Then the users are interviewed, directly or through a questionnaire — that can be submitted during the usage —, providing feedback on what worked for them and what did not. The evaluator extracts the list of accessibility problems from this body of self-reported user opinions. Depending on users' experience in accessibility, the method can be not only empirical, but also analytical and diagnostic; it is based on users' opinions, and can yield failure modes, defects and possible solutions. Therefore it can be adopted for explorative or diagnostic formative evaluations.

Its benefits include the low cost, the fact that it does not require experienced evaluators, and the ability to carry it out remotely in space and time (*i.e.* asynchronously). In addition, participants may be allowed to explore areas of the website that most suit them, with corresponding increase of their motivation in using the website.

However there are important drawbacks: the method is very unsystematic, not only regarding the pages that are being tested, but also the criteria used to evaluate them. In addition, different users with different experience levels and different attitudes will report very different things about the same page. Subjects have to remember what happened during an interaction, they often rationalize their behavior, and may be distracted without being aware of it. Mankoff et al. [24] discovered that this method ranks well in terms of correctness, but poorly in terms of sensitivity when compared to conformance review and to the screening technique mentioned above.

Barrier Walkthrough. The barrier walkthrough method is an analytical technique based on heuristic walkthrough [32] that I proposed in [3, 5]. An evaluator has to consider a number of predefined possible barriers which are interpretations and extensions of well known accessibility principles; they are assessed in a context which potentially includes the elements described in Section 2.1 so that appropriate conclusions about user effectiveness, productivity, satisfaction, and safety can be drawn, and severity scores can be derived. For BW, context comprises user categories (like blind persons), website usage scenarios (like using a given screen reader), and user goals (corresponding to *use cases*, like submitting an IRS form). An *accessibility barrier* is any condition that makes it difficult for people to achieve a goal when using the website in the specified context. A barrier is a failure mode of the website, 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 on payoff functions.

The BW method prescribes that severity is graded on a 1–2–3 scale (minor, major, critical), and is a function of *impact* (the degree to which the user goal cannot be achieved within the considered context) and *persistence* (the number of times the barrier shows up while a user is trying to achieve that goal). Therefore the same type of barrier may be rated with different severities in different contexts; for example, a missing *skip-links* link may turn out to be a nuisance for

a blind user reading a page that has little preliminary stuff, while the same defect may show a higher severity within a page that does a server refresh whenever the user interacts with a sequence of select boxes. Compared to suggestions on how to rate problems given by Nielsen [26], I believe that "cosmetic" problems should not be considered when evaluating accessibility.

Potential barriers to be considered are derived by interpretation of relevant guidelines and principles [13, 37, 43]. A complete list can be found in [5].

We should expect two major benefits of BW compared to conformance review: by listing possible barriers grouped by disability type, evaluators should be more constrained in determining whether the barrier actually occurs. Secondly, by forcing evaluators to consider usage scenarios, an appropriate context is available to rate severity of the problems found.

In fact, a preliminary experimental evaluation of the BW method [3] showed that this method is more effective than conformance reviews in finding more severe problems and in reducing false positives; however, it is less effective in finding all the possible accessibility problems. Some of these results agree with findings reported by Sears [32], who compared heuristic walkthrough with other inspection-based usability evaluation methods, heuristic evaluation and cognitive walkthrough.

Other studies showed how BW can be used as a basis for measuring the accessibility level of a website rather than measuring the conformance level. In particular [6] illustrates how the output of an accessibility testing tool can be sampled so that an assessment similar to BW is performed by one or more judges. On the basis of these sampled barriers, estimates of tool errors and of the accessibility of the website can be computed. These computations can be performed also on conformance review reports, again on the basis of a judging step based on BW [2].

User Testing. Even though empirical methods like laboratory and field testing can in principle be used for evaluating accessibility, the method more often chosen is the lightweight *informal user testing* through the think-aloud protocol [11, 16, 25, 26, 31]. Once a panel of users is selected (representing the target audience in terms of disability, user roles with respect to the website, experience levels in the Internet, in assistive technologies, and in the specific domain and website), they are required to perform given tasks while being observed and being asked to think aloud. In the end, from notes, audio and video recording taken during the test run, evaluators generate the list of problems and assign severity levels.

To ensure effectiveness, the protocol used by evaluators to identify problems should be carefully defined to reduce what Hertzum and Jacobsen [17] call the "evaluator effect", which influences the kind of problems that are detected, at which level of abstractions, and how they are rated for severity. Furthermore, users should be asked to use applications and assistive technologies they are familiar with, and they should be screened according to their level of experience in using these tools.

One benefit of user testing is important [23]: its capability to accurately identify usability problems that are usually experienced by real users, and that have

potentially catastrophic consequences. Conversely, this method is not suitable to identify low-severity problems.

More important drawbacks include its higher costs compared to analytic methods and its inability to highlight defects in addition to failure modes. Furthermore, problems may be missed if predefined scenarios are not well chosen or if user disabilities, experience levels or roles are not representative of the target audience. In addition, given users' requirement in terms of appropriate assistive technology and room facilities, setting up a user testing session with disabled participants may be challenging; similarly for recruiting a group that represents the target audience. Results of performing user testing are likely to be a set of usability problems that are general for all the users of the website, rather than being specific for disabled persons (*e.g.* a misleading link label). In other words, the method is likely to identify a number of true problems, but irrelevant with respect to accessibility (as defined in Section 2.1).

Finally, as reported by Petrie et al. [30], remote user testing for accessibility eases the logistic difficulties, but raises additional issues concerning the validity of results. In two studies comparing local *v.s.* remote user testing, they found out that asynchronous remote user testing, where users work at home on given tasks and websites, and take notes of problems, is a method that can be used with some care for summative evaluations, but is unlikely to be useful for formative evaluations. The reason is that the level of details used to describe problems is much higher when the evaluator observes, and perhaps, challenges the user. Secondarily, care must be payed so that reliable data is gathered concerning the success levels achieved by users. The problem is that often users are not aware of missing part of the goal.

As a final remark, note that all usability evaluation methods can be used to assess accessibility, provided it is understood that in such cases accessibility really means "usability with respect to disabled users and the particular operating conditions determined by the platform used". When this is true then heuristic evaluations and walkthroughs, cognitive and pluralistic walkthroughs, user tests of different sorts can all be used. For analytic methods, the list of principles, guidelines, tasks and basic questions is exactly the same as when dealing with people with no disabilities (*e.g.* the guidelines proposed in [26]).

4 Conclusions

We have seen an accessibility model that clearly defines what accessibility is, how to assess it, and how to represent context. To distinguish accessibility from usability, accessibility should aim at non discriminating users in terms of what they can achieve; accessibility should focus on websites capable of providing equal levels of effectiveness, safety and security in specified contexts. Context should include descriptions of "who" is going to use the website (type of disability, experience level in the Internet, in the specific assistive technology, in using the specific website and its domain), for doing "what" (user goals), and "how" (physical environment and interaction modalities).

Context is necessary when moving from conformance to accessibility, and it has to be considered also in evaluation methods. The methods I reviewed treat context differently. It is virtually absent or very general when performing conformance reviews; it is usually implicitly defined in subjective assessments and screening techniques; it is explicitly characterized in barrier walkthroughs and in user testing. This, in my view, reflects how applicable a method is for evaluating accessibility and affects its correctness, sensitivity and reliability.

More work is needed to provide additional evidence of advantages and disadvantages of methods; but I believe that the adoption of the model and more focus on context would help the web accessibility community to resolve the kind of problems affecting accessibility.

References

[1] Brajnik, G.: Modeling content and expression of learning objects in multimodal learning management systems. In: HCI International 2007, FUITEL: Future Interfaces in Technology Enhanced Learning, Beijing, China (July 2007)

[2] Brajnik, G.: Measuring web accessibility by estimating severity of barriers. In: 2nd International Workshop on Web Usability and Accessibility IWWUA 2008, Auckland, New Zealand (September 2008)

[3] Brajnik, G.: Web Accessibility Testing: When the Method is the Culprit. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A.I. (eds.) ICCHP 2006. LNCS, vol. 4061, pp. 156–163. Springer, Heidelberg (2006)

[4] Brajnik, G.: Comparing accessibility evaluation tools: a method for tool effectiveness. Int. Journal on Universal Access in the Information Society 3(3-4), 252–263 (2004)

[5] Brajnik, G.: Web accessibility testing with barriers walkthrough (March 2006) (Visited May 2008), `http://www.dimi.uniud.it/giorgio/projects/bw`

[6] Brajnik, G., Lomuscio, R.: SAMBA: a semi-automatic method for measuring barriers of accessibility. In: Trewin, S., Pontelli, E. (eds.) 9th Int. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS, Tempe, AZ. ACM Press, New York (2007)

[7] Brajnik, G., Toppano, E.: Creare siti web multimediali: fondamenti di analisi e progettazione, Italy. Addison-Wesley/Pearson Education (2007)

[8] Brajnik, G., Mulas, A., Pitton, C.: Effects of sampling methods on web accessibility evaluations. In: Trewin, S., Pontelli, E. (eds.) 9th Int. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS, Tempe, AZ. ACM Press, New York (2007)

[9] Buillon, L., Vanderdonckt, J.: Retargeting web pages on other computing platforms with vaquita. In: van Deursen, Burd, A. (eds.) Proc. of IEEE Working Conf. on Reverse Engineering WCRE 2002, Richmond, October 2002, pp. 339–348. IEEE Computer Society Press, Los Alamitos (2002)

[10] College of Design, North Carolina University. Principles of Universal Design. The Center for Universal Design (Febuary 1997) (Visited May 2008), `http://www.design.ncsu.edu/cud/about_ud/udprincipleshtmlformat.html`

[11] Coyne, K.P., Nielsen, J.: How to conduct usability evaluations for accessibility: methodology guidelines for testing websites and intranets with users who use assistive technology. Nielsen Norman Group (October 2001), `http://www.nngroup.com/reports/accessibility/testing`

[12] Dey, A.: Accessibility evaluation practices - survey results (2004) (Visited May 2008), http://deyalexander.com/publications/accessibility-evaluation-practices.html
[13] DRC. Formal investigation report: web accessibility. Disability Rights Commission (April 2004) (Visited January 2006), www.drc-gb.org/publicationsandreports/report.asp
[14] Gray, W.D., Salzman, M.C.: Damaged merchandise: a review of experiments that compare usability evaluation methods. Human–Computer Interaction 13(3), 203–261 (1998)
[15] Hartson, H.R., Andre, T.S., Williges, R.C.: Criteria for evaluating usability evaluation methods. Journal of Human-Computer Interaction 15(1), 145–181 (2003)
[16] Henry, S.L., Grossnickle, M.: Just Ask: Accessibility in the User-Centered Design Process. Georgia Tech Research Corporation, Atlanta, Georgia, USA, On-line book (2004), http://www.UIAccess.com/AccessUCD
[17] Hertzum, M., Jacobsen, N.E.: The evaluator effect: a chilling fact about usability evaluation methods. Int. Journal of Human-Computer Interaction 1(4), 421–443 (2001)
[18] ISO. Ergonomics of human–system interaction — guidance on accessibility for human–computer interfaces. ISO/TS 16071. Technical report, International Standards Organization (2003), www.iso.ch
[19] Italian Government. Requisiti tecnici e i diversi livelli per l'accessibilità agli strumenti informatici (July 2005) (G. U. n. 183 8/8/2005), www.pubbliaccesso.it/normative/DM080705.htm
[20] Italian Parliament. Disposizioni per favorire l'accesso dei soggetti disabili agli strumenti informatici (January 2004) (Law n. 4, January 9 2004), http://www.parlamento.it/parlam/leggi/040041.htm
[21] Kelly, B., Sloan, D., Phipps, L., Petrie, H., Hamilton, F.: Forcing standardization or accomodating diversity? A framework for applying the WCAG in the real world. In: Harper, S., Yesilada, Y., Goble, C. (eds.) W4A 2005: Proc. of the 2005 international cross-disciplinary conference on Web accessibility, Chiba, Japan, April 2005, pp. 46–54. ACM, New York (2005)
[22] Kelly, B., Sloan, D., Brown, S., Seale, J., Petrie, H., Lauke, P., Ball, S.: Accessibility 2.0: people, policies and processes. In: W4A 2007: Proc. of the 2007 international cross-disciplinary conference on Web accessibility (W4A), pp. 138–147. ACM, New York (2007)
[23] Lang, T.: Comparing website accessibility evaluation methods and learnings from usability evaluation methods (Visited May 2008) (2003), http://www.peakusability.com.au/about-us/pdf/website_accessibility.pdf
[24] Mankoff, J., Fait, H., Tran, T.: Is your web page accessible?: a comparative study of methods for assessing web page accessibility for the blind. In: CHI 2005: Proc. of the SIGCHI conference on Human factors in computing systems, pp. 41–50. ACM, New York (2005)
[25] Nielsen, J.: Usability Engineering. Morgan Kaufmann, San Francisco (1994)
[26] Nielsen, J.: Heuristic evaluation (2002) (Visited January 2008), http://www.useit.com/papers/heuristic
[27] Nielsen Norman Group. Beyond ALT Text: Making the Web Easy to Use for Users with Disabilities (October 2001), http://www.nngroup.com/reports/accessibility/
[28] Nielsen Norman Group. Web usability for senior citizens (April 2002), http://www.nngroup.com/reports

[29] Petrie, H., Kheir, O.: The relationship between accessibility and usability of websites. In: Proc. CHI 2007, pp. 397–406. ACM, CA (2007)
[30] Petrie, H., Hamilton, F., King, N., Pavan, P.: Remote usability evaluations with disabled people. In: CHI 2006: Proc. of the SIGCHI conference on Human factors in computing systems, pp. 1133–1141. ACM, New York (2006)
[31] Rubin, J.: Handbook of usability testing. Wiley, Technical Communication Library, Chichester (1994)
[32] Sears, A.: Heuristic walkthroughs: finding the problems without the noise. Int. Journal of Human-Computer Interaction 9(3), 213–234 (1997)
[33] Slatin, J., Lewis, K.: Challenges of accessible web design: Standards, guidelines, and user testing. In: Technology and Persons with Disabilities Conference, Los Angeles, USA. CSUN, California State University Northridge (2002)
[34] Slatin, J., Rush, S.: Maximum Accessibility: Making Your Web Site More Usable for Everyone. Addison-Wesley, Reading (2003)
[35] Thatcher, J., Waddell, C., Henry, S., Swierenga, S., Urban, M., Burks, M., Regan, B., Bohman, P.: Constructing Accessible Web Sites. Glasshouse (2002)
[36] Thatcher, J., Burks, M., Heilmann, C., Henry, S., Kirkpatrick, A., Lauke, P., Lawson, B., Regan, B., Rutter, R., Urban, M., Waddell, C.: Web Accessibility: Web Standards and Regulatory Compliance (2006); Friends of ED
[37] Theofanos, M.F., Redish, J.: Bridging the gap: between accessibility and usability. Interactions 10(6), 36–51 (2003)
[38] Thevenin, D., Coutaz, J.: Plasticity of user interfaces: framework and research agenda. In: Sasse, A., Johnson, C. (eds.) Proceedings of Interact 1999, Edinburgh, UK, pp. 110–117 (1999)
[39] U.S. Dept. of Justice. Section 508 of the Rehabilitation Act (2001), www.access-board.gov/sec508/guide/1194.22.htm
[40] U.S. Government. SEC. 508. ELECTRONIC AND INFORMATION TECHNOLOGY, 1998 Amendment to Section 508 of the Rehabilitation Act, (1998) (Visited May 2008), www.section508.gov/index.cfm?FuseAction=Content&ID=14
[41] Vigo, M., Kobsa, A., Arrue, M., Abascal, J.: User-tailored web accessibility evaluations. In: HyperText 2007, Manchester, UK, September 2007, pp. 95–104. ACM, New York (2007)
[42] W3C/WAI. Conformance evaluation of web sites for accessibility (2008)(Visited May 2008), www.w3.org/WAI/eval/conformance.html
[43] W3C/WAI. How people with disabilities use the web. World Wide Web Consortium — Web Accessibility Initiative (March 2004) (Visited May 2008), http://w3.org/WAI/EO/Drafts/PWD-Use-Web/20040302.html
[44] W3C/WAI. Web accessibility evaluation tools: Overview. World Wide Web Consortium — Web Accessibility Initiative (2006) (Visited May 2008), http://www.w3.org/WAI/ER/tools/Overview.html
[45] W3C/WAI. Web content accessibility guidelines 1.0. World Wide Web Consortium — Web Accessibility Initiative (May 1999), http://www.w3.org/TR/WCAG10
[46] W3C/WAI. Web content accessibility guidelines 2.0 — w3c candidate recommendation 30 april 2008. World Wide Web Consortium — Web Accessibility Initiative (April 2008), http://www.w3.org/TR/2008/CR-WCAG20-20080430