Towards a sustainable web accessibility

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

Abstract

Cases like Target.com show that there is still uncertainty about the precise meaning of "web accessibility", about how to identify accessibility problems, and about when to claim that a web site is accessible. In this paper I suggest that the only way to change this is by focussing on sustainable accessibility, *i.e.* by shaping and establishing effective accessibility processes that can be sustained mainly by their own return on investment and by minimal additional resources. To achieve a sustainable accessibility, at least two issues need 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 big difference in accessibility. Therefore, evaluators and developers can focus on these problems first, thus optimizing their resources. Secondly, accessibility processes (involved when conceiving, developing, maintaining, revamping web sites) should be effective and efficient. After discussing the role that a well defined accessibility model plays, the paper illustrates recent findings concerning quality factors of each type of accessibility process, including: sensitivity, correctness, specificity and reliability of elicitation processes; reliability and validity of processes for sampling pages/resources of a website and of accessibility measurement processes. Several of these factors were experimentally studied in recent and still ongoing experiments, and the paper discusses the overall conclusions.

1 Introduction

Although awareness of web accessibility has been growing over the last decade, it is still in an unsatisfactory state. After 9 years since the official release of the WAI/WCAG guidelines, 7 years since enactment of the Section 508 requirements, 4 years since the issue of the Italian accessibility law, the number of accessibility barriers included in web sites is still very high. There is evidence that uncertainty exists about the precise meaning of "web accessibility", about how to identify accessibility problems, and about when to claim that a web site is accessible. Accessibility testing tools, though useful and necessary, rather than help very often further obfuscate these concepts.

My claim is that the only way to change this state of things is by focussing on sustainable accessibility, i.e. by shaping and establishing effective accessibility

processes that can be sustained mainly by their own return on investment and by minimal additional resources.

To achieve a sustainable accessibility, at least two issues need 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 big difference in accessibility. Therefore, evaluators and developers can focus on these problems first, and optimize their resources. Secondly, processes upon which accessibility engineering rests (involved when conceiving, developing, maintaining, revamping web sites) should be effective and efficient; therefore an appropriate choice should be made. When these two conditions are met, accessibility processes can be evaluated, compared and chosen on an informed basis, and they 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.

2 Issues in accessibility evaluations

2.1 Confusion on methods

Confusion exists regarding the methods to use to evaluate accessibility. For example, the current Italian regulation for web accessibility (Italian Government, 2005) specifies a number of technical requirements similar to WCAG 1.0 and Section 508 points. However, in order to certify 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) to be used to determine if the website is accessible. Regulations also require evaluators to 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.

There is also confusion about the concept of accessibility and how to decide when a website is accessible. Consider the Target legal case in the U.S., for example[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 analysed: 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 labelled; and

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

there was no accommodation for screen reader or keyboard navigation, such as skip links or HTML headings."

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

This suggests that there is substantial variability, and lack of standardization, in the way pages to be analysed were selected, in the way accessibility was investigated, and in the way a conclusion was drawn. Both side's witnesses were correct, but witnesses of one side were referring mainly to user behaviour and performance indicators, the others to conformance features.

2.2 Complexity of conformance

In fact, the relation among accessibility, conformance and usability is a complex and somewhat fuzzy one. As discussed by Kelly et al. (2007), 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 addition, empirical evidence shows that the link between conformance and accessibility is missing, *i.e.* even conformant websites may fail in being accessible (DRC, 2004; Petrie and Kheir, 2007). Which accessibility problems are identified and how their severity is rated are two aspects of accessibility investigations that lack substantial standardization, leading to low reproducibility of results. Yet these two aspects are the core of many accessibility guidelines.

In the last two years the second version of the WCAG were repeatedly published for review, which raised some criticism. For example, they were criticized for being too complex and abstract, containing too many obscure definitions, being too long and complex as a specification document, and essentially being void of new information if compared to previous WCAG version (Clark, 2006).

Since then WCAG 2.0 have improved, but in my view they are still showing two problems: one with priorities of success criteria (*i.e.* the new "checkpoints") and the other with claims of testability. The three priority levels of success criteria don't have a clear definition at the moment[2]; however they were defined on the basis of generic expectations of impact on end-users *and* on how easily the requirements entailed by the criterion can be implemented. This is problematic because it is difficult to rate severity of a checkpoint violation without contextualizing the circumstances under which the problem actually occurs. Unless a concrete description of who the user is, what are his/her capabilities, what kind of devices are used and under which conditions, no reliable conclusion can be derived regarding the impact that the problem has on the task the user is

[2]The current WCAG 2.0 document suite, dated 30 April 2008, does not explain what the three levels mean.

carrying out. Therefore setting the priority in advance is likely to lead to incorrect ratings, as was demonstrated by (Petrie and Kheir, 2007). In addition, defining severity as a function of both the ease of implementation and the impact on users, merges two very different triage criteria, both important, but that in many cases should be kept separate to allow more informed decisions to be taken. Ease of implementation should be considered only when deciding which problems to fix first, by those people that are in charge of the maintenance team. By merging the two criteria into a single priority it is impossible to reason about impact on users separately from ease of implementation.

The second problem is that WCAG 2.0 are claimed to testable, but no empirical evidence is given as to what are the expected levels of agreement between evaluators. For example, success criterion 1.3.1[3] could be evaluated differently on a page that uses visual cues (like horizontal spacing between two sections) depending on whether evaluators agree on what is to be considered as "structure" and whether the visual clue should be interpreted as carrying information that is essential.

2.3 Universal design approach

Besides problems with guidelines, another factor that hampers the practicality of accessibility engineering is the implicit assumption that to achieve an accessible web site one has to follow the universal design approach (College of Design, North Carolina University, 1997), that is designing a website with a single user interface that fits all requirements. As claimed in (Brajnik, 2004; Kelly et al., 2005) and partially demonstrated in (Brajnik et al., 2005) with customized on-the-fly text-only pages, the "one size fits all" viewpoint is not always appropriate. If accessibility is meant as "such that disabled people can use a website with the same effectiveness as non-disabled people", as long as a person can use effectively the content of interest, bearing information, navigation tools contained in web pages and is given ways to enter user-supplied information, there's no reason (other than technological limitations) why the user interface shouldn't be personalized, either through explicit customization, implicit profiling or automatic adaptation. A personalized interface could only improve effectiveness, productivity, and satisfaction of users, while simplifying at the same time the conceptual design and the assessment process since each type of interface would be geared towards a more uniform audience.

2.4 Return on investment

The consequences of these problems affecting accessibility engineering are that accessibility is difficult to understand, learn and measure, and its return on investment (ROI) is severely hampered. One factor is that since resources are expended sub-optimally, the ROI is smaller than it could be. In addition, accurate estimations of the ROI are difficult to obtain, due to being a sensible information to be released, and due to the magnitude of the effect of accessibility features. Features that are well implemented tend to have a significant effect on few individuals (relative to the entire user population of a web site), and a minor but diffused effect on most of the other users. For example, many

[3] 1.3.1 Info and Relationships: Information, structure, and relationships conveyed through presentation can be programmatically determined or are available in text. (Level A)

users will take advantage of the improved usability caused by keyboard access, of increased flexibility of the layout, of improved indexing and ranking by search engines. These features will have a positive, but generally minor effect, on productivity and satisfaction of those users. For few users, however, these features may prove to be essential in order to use the website. In terms of accessibility engineering, these are significant effects that can be easily measured in user testing sessions or that could be reliably estimated in analytical methods; they are often not economically relevant though, due to the small fraction of the user base that can take advantage of them. The former effects would be more economically relevant given the large user base, but unfortunately they are difficult to measure reliably and even more difficult to convert into monetary units.

While these difficulties in calculating the ROI cannot be easily overcome, at least the other issues we discussed in this section could be mitigated by better focussing the activities upon which accessibility engineering is based. To this end, a crucial role is played by the accessibility model.

3 The 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 helps to plan and perform activities like diagnosing the defects of a website, comparing it to other websites, and after measuring its accessibility level, monitoring its accessibility to determine whether it is conformant to certain standards.

In (Brajnik, 2008c) I defined the Properties-Context-Processes (PCP) model, shown in Figure 1, which describes three groups of factors that affect the way accessibility is managed. The *Properties* part describes how to characterize accessibility, and in practice this results in a definition of accessibility and identification of the factors that are to be used to assess it. For example, a viable definition can be the one mentioned before[4], which refers to the *effectiveness* factor only. Other definitions are of course possible; see (Brajnik, 2008c) for nine different ones.

The second component of the PCP model is *Context*, or a description of additional factors that affect accessibility though not contributing to it; *Context* describes also how these factors can be detected, isolated and controlled. For example, *Context* could be defined in terms of the type of user disability, the experience level the user has (in using the browser, the Web, the assistive technology, and possibly the specific website and domain of operation). Context could also include the short-term user goals, the physical environment the user is working in (posture, light and noise conditions), input and output devices and interaction modalities (media used, possible user actions and operations, user agents, assistive technologies and infrastructure). It is important to isolate and circumscribe context because it supplies important information about how the user behaves. Unless this information is represented explicitly, it will stay in the background of evaluators and developers, leading to inconsistent, poorly focused and unreliable analyses. Just like context plays an important role in the definition and management of usability, there is no reason why it

[4] "An accessible website is such that disabled people can use it with the same effectiveness as non-disabled people".

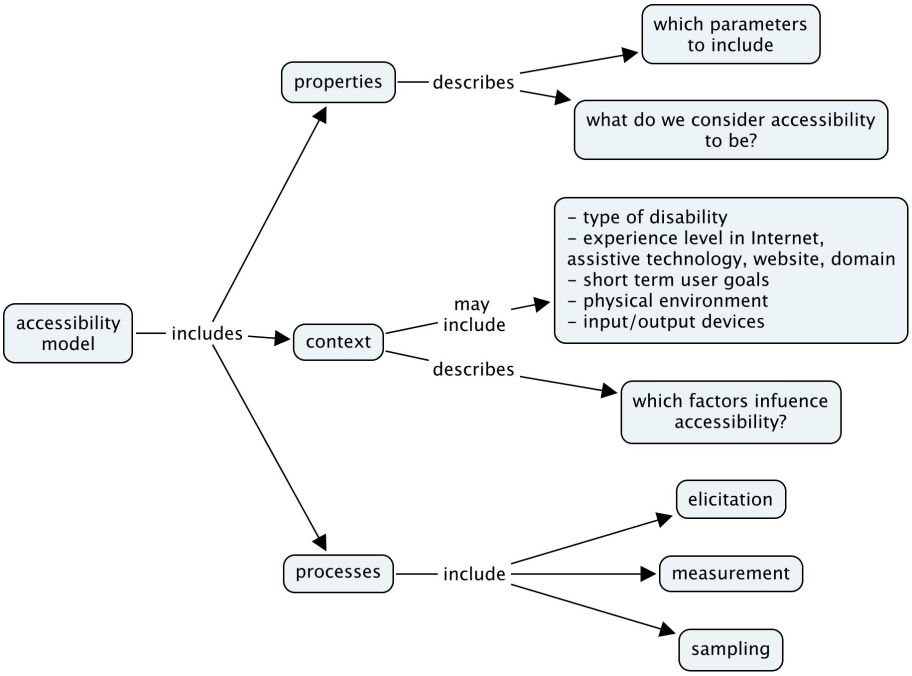

Figure 1: The PCP model of accessibility

shouldn't do that for accessibility as well. In fact, context is even more important for accessibility than for usability as there are additional factors that web developers cannot control but that affect users behaviors, including at least the assistive technology they use, their experience level in using it, the accessibility architecture/platform they use.

The third part of the model (*Processes*) describes which key processes are involved in assessing accessibility and how they are carried out. For example, it could prescribe which rules to use to sample pages, which methods (and possibly which guidelines) to use to evaluate them and how to measure accessibility (Figure 2).

These processes are important because they always take place, even when they are not explicitly defined and established. Except for trivial cases, assessments of accessibility require selection of pages to be analyzed. The most often used criterion is the "ad-hoc" selection which, as we will see below, is not the optimal one. Secondly, while conformance review is the most often used method for eliciting accessibility defects, it is just one of several methods that can be used within the elicitation process. Finally, when an auditor states a conformance level, then a implicit measurement process is carried out (in this case, one that transforms a set of checkpoint violations into ordinal values like "A", "AA" or "AAA"); other types of measurements can take place which can be more appropriate.

Methods that can be used within each type of process have their own pros and cons, and before choosing them, they should be evaluated against the requirements set by the accessibility model.

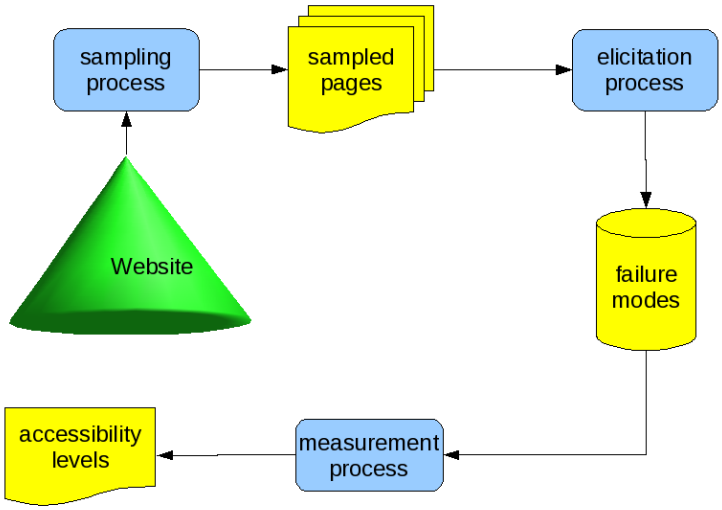

Figure 2: The key processes involved in accessibility assessments

4 Accessibility assessment processes

4.1 Sampling processes

Sampling processes are necessary when dealing with large or very dynamic websites and the problem elicitation process requires human judgments (which normally should be).

Several methods have been suggested so far. *Ad hoc sampling* methods are suggested by W3C/WAI and by UWEM (W3C/WAI; Velleman et al., 2007). In general these methods can be based on predefined criteria that consider the type of pages (*e.g.* home page, site map, contact page, representative pages with respect to content, pages featuring purchase forms, etc.).

The *uniform random sampling* method, which guarantees that each page of the web site has the same probability to be included in the sample, is conceptually simple, but for large and dynamic web sites it is not practical since rarely one has an exhaustive list of all the pages ready to choose from and often the same URL may lead to content that changes over time. However, this method can be easily approximated by having a tool download a large set of the pages and then implement a random sampling with no replacement from such a set.

(Henzinger et al., 2000; Ulltveit-Moe et al., 2006) discuss two methods based on *random walks* that emulate the behavior of a generic visitor of a website, that can either keep following random links from the visited page, or randomly jump to another page.

Sampling methods based on *distributions of violations* have the advantage of selecting pages on the basis of information relevant to accessibility evaluations. King et al. (2005) describe a sampling method based on clustering pages according to similarity of the distribution of checkpoint violations (called *error profile*) and to URLs sharing a prefix. From each cluster they randomly sampled pages. In this way, the resulting pages reflect the distribution of checkpoint

violations as found by an automatic tool.

In terms of quality factors, sampling processes can be characterized by their *effectiveness*, defined as how representative the sample is of the entire population (of pages or accessibility issues). Effectiveness can be measured by accuracy, namely the difference in values of a specific metric obtained from a sample of pages and from the entire website. Another quality factor is the *appropriateness* of the sample, defined as its ability to support statistical generalizations of results (which typically are dependent on sample size and variance of the variable of interest). Sampling processes can be also characterized by their *reliability*, *i.e.* the extent to which their results can be easily reproduced.

In (Brajnik et al., 2007) we discussed experimental results showing that accuracy of different sampling methods depends on sample size and the metric being used (13 methods were tested belonging to the categories mentioned above, implemented with automatic tools). As expected, the larger the sample the better is the accuracy: when testing for conformance (*i.e.* measuring the number of checkpoints that fail) a sample of 50 pages yields accuracy equal to 95%[5]. However for other metrics the sample size is not so important: with a sample of just one page we can get an accuracy equal to 96% for WAQM and to 95% for UWEM. There is a significant interaction between the metric used to measure accessibility and accuracy. If the metric is conformance, then accuracy can be as low as 62% with the "ad-hoc" method suggested by the W3C; with other metrics (*e.g.* WAQM), accuracy increases to more than 98%.

The accessibility model is related to the quality of the sampling process. How accessibility is defined, and especially how it is measured, affects the effectiveness of the sampling process: we have seen how accuracy and appropriateness of the sampling methods depend on the metric being used. Furthermore, if we know in advance what kind of tasks are more important or frequent for users of the investigated website (part of the *Context*), then the effectiveness of the sampling process can improve.

4.2 Elicitation processes

Elicitation processes are based on accessibility evaluation methods, which are procedures aimed at finding accessibility problems, such as guideline violations, failure modes[6], defects, or user performance indexes. A good elicitation method is a dependable tool that yields accurate predictions of all the accessibility problems that may occur in a website. This is why methods can and should be compared in terms of such criteria as *correctness* (the percentage of problems reported when applying a method that are true problems), *sensitivity* (the percentage of the true problems being reported), *reliability* (the extent to which independent evaluations produce the same results), *usefulness* (the usability of the produced results with respect to stakeholders including people involved in QA, in managing accessibility and in fixing defects), *efficiency* (the amount of resources expended to carry out an evaluation that leads to specified levels of

[5]Meaning that 5% of the checkpoints might not be correctly represented in the selected sample.

[6]A *failure mode* is a hindrance, for a user in given situations, to the achievement of a user goal; a *defect* is a feature of the website that causes a failure mode; *severity* of a failure mode is a function of the *impact* of the failure mode (extent to which the goal cannot be achieved) and the *persistence* (extent to which the failure mode is encountered repeatedly during a task execution).

effectiveness and usefulness) and the method's *usability* (how easily it can be understood, learned and remembered by evaluators); for more details on these criteria, please refer to (Sears, 1997; Gray and Salzman, 1998; Hertzum and Jacobsen, 2001; Lang, 2003; Hartson et al., 2003; Brajnik, 2008c).

Elicitation methods most often used include *conformance reviews* (when an evaluator uses a set of guidelines to determine which pages violate some of the guidelines), *screening techniques* (when the evaluator uses a website in a way that some sensory, motor or cognitive capabilities are artificially reduced, for example after unplugging the mouse), *subjective assessments* (where a panel of persons are asked to use a website and then report back what worked and what did not), *barrier walkthrough* (where an evaluator uses a list of predefined barriers, a user profile and a scenario, to determine which barriers actually exist and to rate their severity), and *user testing* (where a panel of persons are asked to use a website to accomplish certain tasks while being observed and asked to think aloud). (Brajnik, 2008c) provides more details on pros and cons of different methods.

The PCP model plays an important role for elicitation processes. To start with, the *Properties* component is closely related to the elicitation method. In fact, if accessibility is defined in terms of differences in effectiveness, then conformance reviews and screening techniques are suboptimal methods while user testing is more likely to yield optimal results when accessibility is defined as "usability with respect to people with disabilities". The *Context* also plays an important role. While it is virtually absent when performing conformance reviews, context gets richer and richer when methods change from barrier walkthrough to user testing, where it includes user profile, user goals, assistive architecture and technologies. And the more context is brought into the elicitation process, and better results will likely ensue, with higher correctness and reliability (Brajnik, 2008a).

4.3 Measurement processes

Accessibility metrics are rules and procedures to analyze accessibility problems in order to yield a value representing the accessibility level of pages being analyzed. They are needed for a number of goals, including Vigo et al. (2007): ranking web pages within search engines results according to their accessibility level; monitoring adoption of accessibility regulations and standards; monitoring penetration of accessibility in given areas or site genres; quality management and monitoring of accessibility levels of a single web site over time.

Several accessibility metrics have been discussed in the literature. Sullivan and Matson (2000) defined the *failure rate* on the basis of a subset of WCAG 1.0 checkpoints as the number of violations of any checkpoint divided by the maximum number of violations of any of those checkpoints that can take place (*i.e.* by the number of *possible* violations). Zeng (2004); Velleman et al. (2007); Arrue et al. (2005) defined respectively the metrics called WAB, UWEM, WAQM, all based on failure rate. Also the evaluation form used by the Accessibility Internet Rally[7] judges implements an accessibility metric, where a website is ranked according to penalty points rated by a panel of judges.

In many cases the measurement process is based on automated testing tools,

[7]See `www.knowbility.org`.

capable of systematic application of an array of tests covering some or all of the requirements of a standard. The advantages behind such a solution is that tools are systematic scanners of web pages, efficient processors and reliable evaluators (in the sense that they produce repeatable results). However, tools are plagued with the problems of generating issues that are not accessibility problems (false positives), of missing certain true accessibility problems (false negatives), and of being incapable of estimating the severity of a requirement violation. Since the tools do conformance testing, they often compute the failure rate and support therefore automation of the metrics. However, the failure rate has a drawback when used for measuring accessibility. For example, two pages that include 10 and 20 images respectively, one with 2 properly defined "alt-text", the other with 8, have FR = 0.8 and FR = 0.6 respectively. Even though the second page has a larger number of violations, hence a larger number of potential obstacles to users, it has a smaller FR. Therefore, in addition to wrong estimates due to false positives and negatives, the values produced by accessibility metrics based on automatic tools should not be used to measure accessibility.

The metrics SAMBA and MAMBO, defined in (Brajnik and Lomuscio, 2007; Brajnik, 2008b), attempt to address this kind of limitations. SAMBA is a method for measuring accessibility that uses the output of testing tools and couples it with opinions of experienced human evaluators, so that correct estimates of tool errors can be assessed, and that estimations of severities of barriers are used. MAMBO is applied in a similar way, but in cases when no automatic tools can or should be used.

Quality factors related with measurement processes include *validity*, *i.e.* the extent to which the adopted metric actually measures the property one is interested into; another factor is *reliability*, *i.e.* the extent to which independent applications of the metric produce the same result.

All components of the accessibility model are related with the metrics. We have already seen that the quality of sampling methods does depend on the metric being used. Since application of any metric is based on the results produced by elicitation methods, quality of metrics necessarily depends on the quality of the results produced by former processes. In principle, most metrics can be used with any elicitation method, but this coupling might affect validity of the results. For example, using MAMBO with results obtained with user testing is likely to yield a value that reflects a definition of accessibility seen as usability for the disabled ones. If the property we were interested into was accessibility defined in terms of equal effectiveness between disabled and non disabled persons, then the results produced in that way will have low validity.

Context also plays an important role; unless it is carefully defined and isolated, it will negatively impact both validity and reliability of the metric. Validity can be damaged if ratings of severity are based on static and generic priority levels; reliability can be reduced if too much context is left unspecified so that each evaluator is making his/her own hypotheses about user behavior and severity of barriers.

5 Conclusions

We have seen that there are several issues that limit the success of accessibility initiatives. There is confusion in the methods to be used to assess it,

conformance is a complex property to ensure, universal design may interfere negatively, and computing the return on investment is difficult or may even prove to be impossible.

The definition of an accessibility model can however help in improving such a situation. The PCP model specifies that proper definitions of accessibility, of context to be used when evaluating, and of processes for sampling pages, eliciting problems and for measuring. Such a model helps in focussing on important characteristics of methods, so that among all the available methods the most appropriate ones can be chosen.

I claimed that sustainable accessibility could be a reality if problems are prioritized by the impact they have on users and if selection of the methods to deploy is based on an informed comparison. The accessibility model I suggest should enable these two conditions, firstly by moving beyond the view where accessibility is seen as conformance, and secondly by emphasizing a number of quality factors that can be used to compare methods and processes. However, at this moment, few studies have concentrated on accessibility management processes, and this is an important research avenue to follow since it appears there are confusing factors and interactions that rule out a straightforward selection procedure.

References

M. Arrue, M. Vigo, and J. Abascal. Quantitative metrics for web accessibility evaluation. In *Proc. of the ICWE 2005 Workshop on Web Metrics and Measurement*, 2005.

G. Brajnik. A comparative test of web accessibility evaluation methods. In S. Harper, editor, *10th Int. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*, Halifax, Canada, Oct. 2008a. ACM Press.

G. Brajnik. Measuring web accessibility by estimating severity of barriers. In S. Hartmann et al., editor, *WISE 2008: 9th Int. Conference on Web Information Systems Engineering – 2nd International Workshop on Web Usability and Accessibility IWWUA08*, LNCS 5176, pages 112–121, Auckland, New Zealand, Sept. 2008b. Springer-Verlag.

G. Brajnik. Beyond conformance: the role of accessibility evaluation methods. In S. Hartmann et al., editor, *WISE 2008: 9th Int. Conference on Web Information Systems Engineering – 2nd International Workshop on Web Usability and Accessibility IWWUA08*, LNCS 5176, pages 63–80, Auckland, New Zealand, Sept. 2008c. Springer-Verlag. Keynote speech.

G. Brajnik. Achieving universal web access through specialized user interfaces. In C. Stary and C. Stephanidis, editors, *Lecture Notes in Computer Science Proc. of the 8th ERCIM UI4ALL Workshop*, Vienna, June 2004. Springer Verlag. `www.dimi.uniud.it/giorgio/publications.html#specui`.

G. Brajnik and R. Lomuscio. SAMBA: a semi-automatic method for measuring barriers of accessibility. In S. Trewin and E. Pontelli, editors, *9th Int. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*, Tempe, AZ, Oct. 2007. ACM Press.

G. Brajnik, D. Cancila, D. Nicoli, and M. Pignatelli. Do text transcoders improve usability? In S. Harper, Y. Yesilada, and C. Goble, editors, *Int. Cross–Disciplinary Workshop on Web Accessibility, W4A*, pages 9–17, Chiba, Japan, April 2005. ACM.

G. Brajnik, A. Mulas, and C. Pitton. Effects of sampling methods on web accessibility evaluations. In S. Trewin and E. Pontelli, editors, *9th Int. ACM SIGACCESS Conference on Computers and Accessibility, ASSETS*, Tempe, AZ, Oct. 2007. ACM Press.

J. Clark. To Hell with WCAG 2. A LIST apart, `alistapart.com/articles/tohellwithwcag2`, May 2006.

College of Design, North Carolina University. Principles of Universal Design. The Center for Universal Design, `http://www.design.ncsu.edu/cud/about_ud/udprinciplesshtmlformat.html`, Feb 1997. Visited May 2008.

DRC. Formal investigation report: web accessibility. Disability Rights Commission, `www.drc-gb.org/publicationsandreports/report.asp`, April 2004. Visited Jan. 2006.

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.

H. R. Hartson, T. S. Andre, and R. C. Williges. Criteria for evaluating usability evaluation methods. *Int. Journal of Human-Computer Interaction*, 15(1): 145–181, 2003.

M. R. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork. On near-uniform URL sampling. In *Proceedings of the 9th international World Wide Web conference on Computer Networks: the International Journal of Computer and Telecommunications Networking*, pages 295–308, Amsterdam, The Netherlands, 2000. North-Holland Publishing Co. doi: dx.doi.org/10.1016/S1389-1286(00)00055-4.

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.

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.

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, *W4A '05: Proc. of the 2005 international cross-disciplinary conference on Web accessibility*, pages 46–54, Chiba, Japan, April 2005. ACM.

B. Kelly, D. Sloan, S. Brown, J. Seale, H. Petrie, P. Lauke, and S. Ball. Accessibility 2.0: people, policies and processes. In *W4A '07: Proc. of the 2007 international cross-disciplinary conference on Web accessibility (W4A)*, pages 138–147, New York, NY, USA, 2007. ACM. ISBN 1-59593-590-X.

M. King, J.W. Thatcher, P.M. Bronstad, and R. Easton. Managing usability for people with disabilities in a large web presence. *IBM Systems Journal*, 44 (3):519–535, 2005.

T. Lang. Comparing website accessibility evaluation methods and learnings from usability evaluation methods. `http://www.peakusability.com.au/about-us/pdf/website_accessibility.pdf`, Visited May 2008, 2003.

H. Petrie and O. Kheir. The relationship between accessibility and usability of websites. In *Proc. CHI 2007*, pages 397–406, San Jose, CA, USA, 2007. ACM.

A. Sears. Heuristic walkthroughs: finding the problems without the noise. *Int. Journal of Human-Computer Interaction*, 9(3):213–234, 1997.

T. Sullivan and R. Matson. Barriers to use: usability and content accessibility on the web's most popular sites. In *Proc. of ACM Conference on Universal Usability*, pages 139–144, 2000.

N. Ulltveit-Moe, M. Snaprud, A. Nietzio, M. Goodwin Olsen, and C. Thomsen. Early Results from Automatic Accessibility Benchmarking of Public European Web Sites from the EIAO. `eiao.net/publications`, 2006. Visited May 2007.

E. Velleman, C. Strobbe, J. Koch, C.A. Velasco, M. Snaprud, and A. Nietzio. D-WAB4: Unified web evaluation methodology (UWEM)1.2. Technical report, WAB Cluster, 2007.

M. Vigo, M. Arrue, G. Brajnik, R. Lomuscio, and J. Abascal. Quantitative metrics for measuring web accessibility. In *W4A '07: Proceedings of the 2007 International Cross-Disciplinary Workshop on Web Accessibility (W4A)*, pages 99–107, New York, NY, USA, 2007. ACM Press. ISBN 1-59593-590-X. doi: doi.acm.org/10.1145/1243441.1243465. `www.dimi.uniud.it/giorgio/papers/w4a07.pdf`.

W3C/WAI. Conformance evaluation of web sites for accessibility: Determine the scope of the evaluation. `www.w3.org/WAI/eval/conformance.html#scope`. Visited May 2007.

X. Zeng. *Evaluation of Enhancement of Web Content Accessibility for Persons with Disabilities*. PhD thesis, University of Pittsburgh, 2004.