The Troubled Path of Accessibility Engineering: an Overview of Traps to Avoid and Hurdles to Overcome

Giorgio Brajnik
Dep. of Mathematics and Informatics
University of Udine, Italy
brajnik@uniud.it

Abstract

Web accessibility entails far more than just paying attention to a handful of HTML tags and attributes: it deals with human behavior. As consequence we have a number of pitfalls that hamper accessibility evaluation of web pages. In this paper I review some of the research my coauthors and I did in the last four years that provides some experimental evidence. In fact, the three fundamental processes of (1) selecting the pages to be investigated, (2) finding their problems, and (3) measuring the corresponding accessibility levels are ridden with potential traps which affect reliability and even validity of evaluations. Knowing which traps are there and figuring out how to overcome them should rank high in the priority list of researchers and practitioners in accessibility. This is what is needed in order to move towards an *engineering* of accessibility.

Introduction

Web accessibility engineering improves the accessibility level of web sites and applications through a systematic, disciplined and quantifiable approach to their development and analysis. While at first sight accessibility seems to be a relatively objective property, grounded on mark-up languages and seemingly syntactic properties, like the Alt attribute of the IMG tag, or the LABEL/@FOR construct for form controls, or the TH/@SCOPE and TD/@HEADER mark-up for data tables, it is a difficult goal for developers to reach. Not least because it has to do with human behavior.

One of the reasons may be that so far, as a research and practice community, we have not been so systematic and disciplined and rigorous as the definition given above assumes. We have however to improve such a status and become better accessibility engineers. In fact, on one side the push of new technologies and of new user interfaces techniques and paradigms, such as multimodal ones, lead to more and more potential and diverse accessibility barriers that need to be dealt with. We need more effective design and analysis processes to cope with such a moving frontier. On the other side, in the current evaluation practice there are several traps which undermine our ability to identify and address the true problems that possibly residing in a web application. Again, more effective analysis processes are needed badly.

In this short paper I would like to recap some of the results found in experiments that others and I did in the last four years concerning methodological aspects of web accessibility evaluations. We discovered that errors and subjectivity can creep in at different steps of an

evaluation: during page selection you can introduce up to 20% of the errors if you follow the wrong page selection criterion; during elicitation of accessibility problems, even if you are an expert, something more than 20% of your ratings might be wrong, you might produce more than 29% of false positives and miss more than 29% of the true problems; finally, if you want to measure the level of accessibility, and adopt one of the published automatic accessibility metrics, you are likely to use one that does a very poor job in discriminating high vs. low accessibility pages.

The stakes are high: whether you are a web developer that needs to know what the accessibility status of your application is, or you are working in quality assurance and are comparing accessibility of the current version of an application to previous ones, or you are an accessibility consultant that has to determine if an application is conformant to WCAG 2.0, or you are a chief information officer worried about the conformance status of your sites, or you are a legal consultant for somebody who is in the process of writing a law concerning accessibility of certain types of sites, the consequences of unreliable or invalid conclusions regarding accessibility might be very costly.

My conclusions are easy to summarize: First, web accessibility is not an objective property, and in order to be dealt with in engineering terms it has to be contextualized. Second, I would like this research community to spend more efforts in studying and improving the methods that we (scholars and practitioners) adopt when designing or analyzing applications, so that subjectivity and errors can be estimated and kept under control. After all, what De Marco said 30 years ago for software engineering – "you cannot control what you cannot measure" – nowadays applies equally well to accessibility engineering (De Marco, 1982).

Background

Let's start from the words "web accessibility". In (Brajnik, 2008) I listed as many as nine different definitions, ranging from the one used in the WCAG 2.0 (a web site is accessible if it is perceivable, understandable, operable by users despite their impairments and if it is robust with respect to user agents and assistive technologies) to the one proposed by Petrie and Kheir, which reduces accessibility to usability (… if 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, Petrie & Kheir, 2007). The fact that there are so many definitions means that the notion of accessibility by itself is likely to stir disagreement. In addition, the fact that most of these definitions are not operational (i.e. cannot be easily mapped to more objective criteria) make the whole issue ill-defined. For example, how would you decide if a site is robust? Or what does "people with disability" mean?

Comment n. 1: accessibility is relative.

My first conclusion is that accessibility needs to be contextualized, as it is done for usability. Within a specific context (for example, restricted to a certain group of people, a certain group of goals/tasks, and a certain group of computing platforms and operating situations) we could define accessibility as the extent to which effectiveness is achieved, when measured in terms of success rates and errors. After obtaining similar results for a corresponding control group with no disability, we could say that a site is accessible if the two groups achieve the same level of effectiveness. As a consequence it would make no

sense to say "site www.example.com is accessible" without declaring the context (people, tasks, situation) of such a claim.

At the moment there are a few accessibility models (frameworks that establish how accessibility depends on which factors). A well known model is the one proposed by W3C/WAI, which is based on three fundamental tiers and two additional hypotheses: to support accessibility, web content must be accessible according to the WCAG; the user agent used by the user has to be accessible according to the UAAG (User Agent Accessibility Guidelines); the web authoring tool used by the developer has to produce code that is accessible (Authoring Tool Accessibility Guidelines); assistive technologies have to be compatible with WCAG and UAAG, and finally the operating system has to provide the appropriate support (in terms of the accessibility architecture) to allow proper interaction between user agents and assistive technologies.

Another model was suggested by Kelly et al. (2007), based on a model of the accessibility stakeholders and their motivations combined with an extensible list of accessibility criteria and methods.

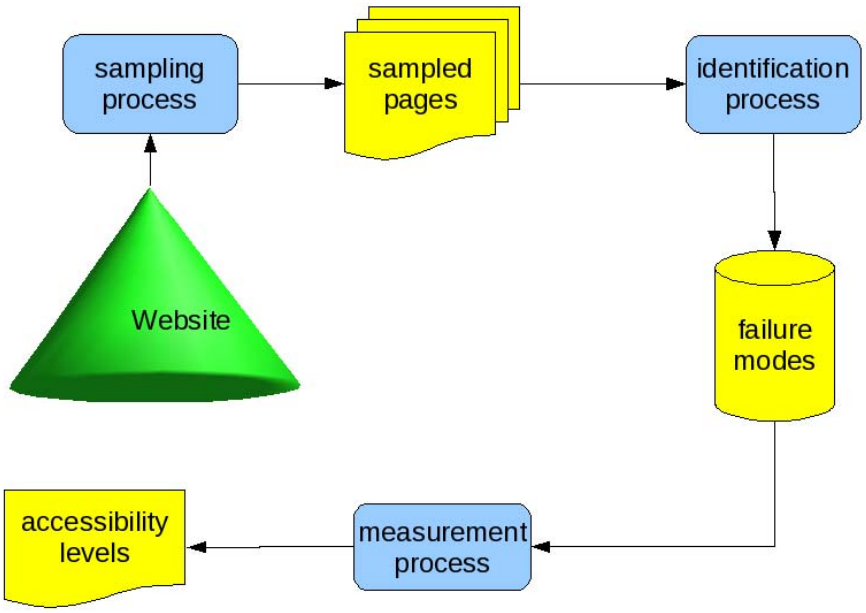

Figure 1: The three processes of selection (sampling) of pages, problem elicitation (identification) and accessibility measurement.

Yet another model is the simple one I suggested in (Brajnik, 2008) based on the notions of *properties*, *context*, and *processes*. By properties I mean an operational definition of "web accessibility". By context I mean a characterization of the context of use, in terms at least of user and user platform profiles). By processes I mean a definition of how the following three processes are to be carried out: selection of the pages to be analyzed, elicitation of accessibility problems, measurement of accessibility levels (see Figure 1). Notice that this is nothing new compared to a well written accessibility policy which, among other things should state details about properties, context and processes.

Comment n. 2: define and stick to concrete and operational procedures.

Let's now consider each of these three processes, and in particular the pitfalls that they hide.

Selection of pages

Because many interesting web sites have a huge number of pages, and in many cases the content of those pages is likely to change very frequently, and sometimes it is the users who add/change content, when an evaluation is carried out one has to select the pages/content to investigate. There are different ways to do so; one is *ad-hoc*, which suggests to select pages such as the home page, the site map, the contact page, and a representative one for each of the subsites. Other methods discussed in the literature are probabilistic in nature, variations of *random walks* over the links of the site. This kind of methods are used, for example, in some of the automatic accessibility observatories that have to repeatedly scan (part of) a site. Another category of selection methods are based on the *error profile* of pages (the error profile is a vector of the outcome of all accessibility criteria applied to the page) and the idea of clustering profiles that are similar so that *different* pages are considered; these methods are used when using automatic tools to monitor the accessibility status of a website.

In an experiment my students and I did (Brajnik, Mulas, & Pitton, 2007), we found that across 13 different sampling methods (1 ad-hoc, 3 random-walks and 9 variations of the error profile ones), the error in accessibility when using the worst-behaving method is close to 20%. More specifically, we downloaded 1000 pages from each of the 32 web sites we considered; an accessibility testing tool was then applied to find out the violations of WCAG 1.0 checkpoints in each of these pages and the number of checkpoints that failed on each page was computed (which is our measure of accessibility). Then, for different sample sizes (ranging from 1 to 100), we selected several samples using all 13 sampling methods. Finally, we compared the average accessibility level of the sample against that of the entire site.

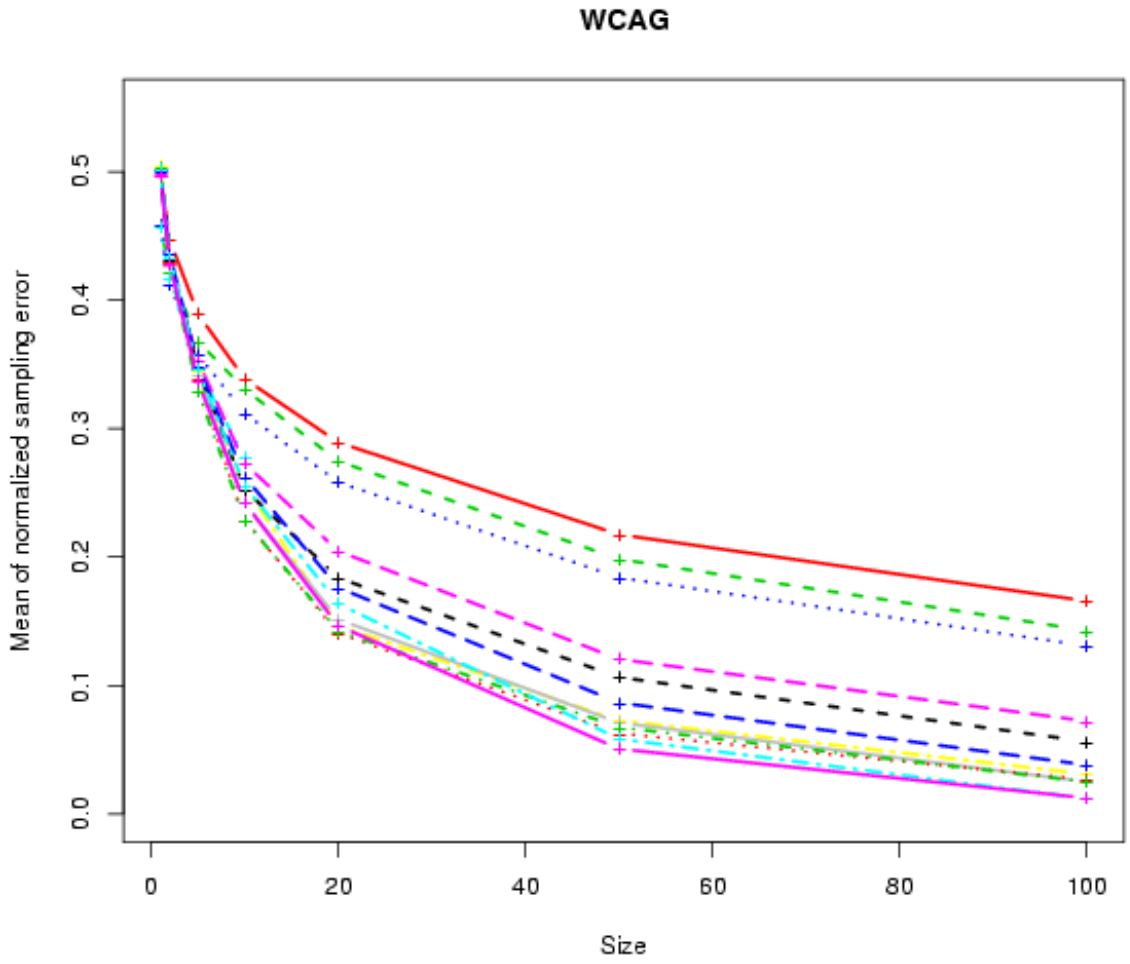

Figure 2: Influence of the sample size on the WCAG 1.0 conformance when using different sampling methods.

As can be seen in Figure 2, the relative error decreases as the sample size increases, as we all would expect. However, even for size=100, when the wrong method is chosen, the error rate is close to 20% (meaning that 20% of checkpoints are not correctly estimated using the sample). Among the worst-behaving methods were the random-walk and ad-hoc ones.

Comment n. 3: page selection affects the outcome of the audit significantly.

Even though in this study we made very strong assumptions (that accessibility errors are exactly those identified by the tool; that 1000 pages are considered the same as the entire web site; that a difference in the number of failed checkpoints bears upon accessibility), the overall conclusion is that the method used to select pages is likely to affect the outcome (even to a large extent) of an accessibility audit, especially in terms of conformance.

Elicitation of accessibility problems

There are several methods that can be used to elicit accessibility problems: user testing with people with impairments, subjective assessments (asking a panel of people to report back what works and what doesn't), screening techniques (artificially reducing one's sensory or motor capabilities, like when forcing a developer to use a screen reader to navigate in the web site that s/he develops), barrier walkthrough (using descriptions of known barriers mapped to disability types when inspecting pages) or conformance reviews (with respect to guidelines such as WCAG 2.0 or Section 508).

In my view, an elicitation method is a conceptual tool that should help you in predicting which accessibility problems will show up on a site when it will be used by real people for real purposes. Ideally, the best tools are those that help you find almost only true problems (few false positives), almost every true problems (few false negatives), with high reliability.

In a series of experiments, we (Yesilada, Brajnik, & Harper, 2009; Brajnik, 2009; Brajnik, Yesilada, & Harper, 2010) explored these issues for the barrier walkthrough and conformance review methods, figuring out also how much the evaluators' experience accounts for. In two new, soon to be published, journal papers we provide a much deeper analysis of the outcomes, including estimates of the optimal number of judges to employ.

We found out that reliability of conformance review based on WCAG 2.0 is low, even when evaluators are experts in the fields. About half of the success criteria got an agreement among experts that did not exceed 75%; some success criterion never, over the pages we considered, reached a value higher than 75%. See Figure 3.

In terms of validity, the picture is not much better: on average, experts produce 29% of false positives (correctness is 71%) *and simultaneously* 29% of false negatives (sensitivity is 71%, i.e. they missed 29% of the true problems). See Figure 4.

Expertise accounts for a slight increase in reliability (about 6%) and a 20% increase in validity.

Similar results were found for the barrier walkthrough method.

As with any experiment, also in our case the assumptions we made limit how far we can generalize the results. First of all, we recruited 23 experts among our friends and colleagues (most were attendees or authors of ASSETS) and we asked them to evaluate one page each, and fill-in a spreadsheet of success criteria: this is clearly an artificial setting, since they knew no client was waiting for their answers and that their answers would not be used to decide what to do on those pages. We also assumed that the *true problems* were the success criteria whose most frequent rating was *fail* (over all the ratings given by experts on a given page).

Comment n. 4: conformance to WCAG 2.0 is subjective.

Comment n. 5: substantial conformance errors cannot be avoided, not even with experts.

Thus, even considering the underlying assumptions, I believe we should definitely forget to treat accessibility, but also conformance, as a binary absolute property. First, because it is subjective; second, because we need to cope with error margins.

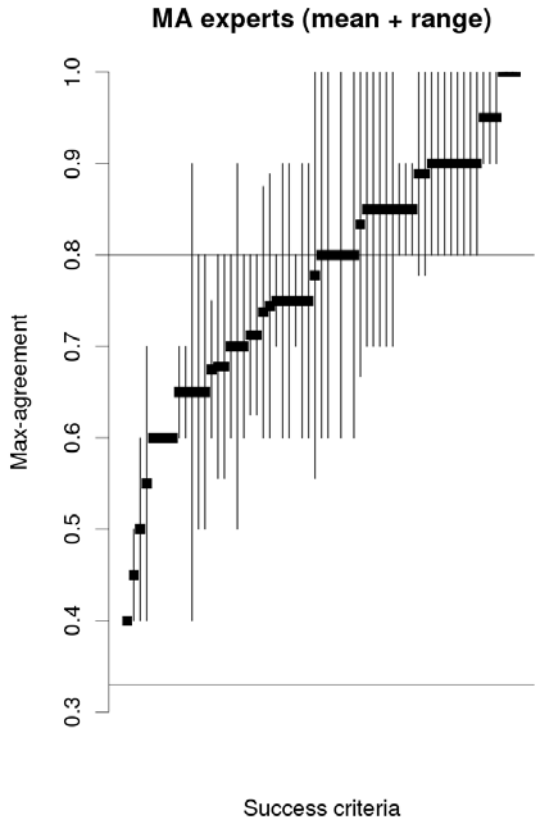

Figure 3: Agreement (relative number of experts that agreed) on the most frequent outcome of each success criterion of WCAG 2.0 over four given pages.

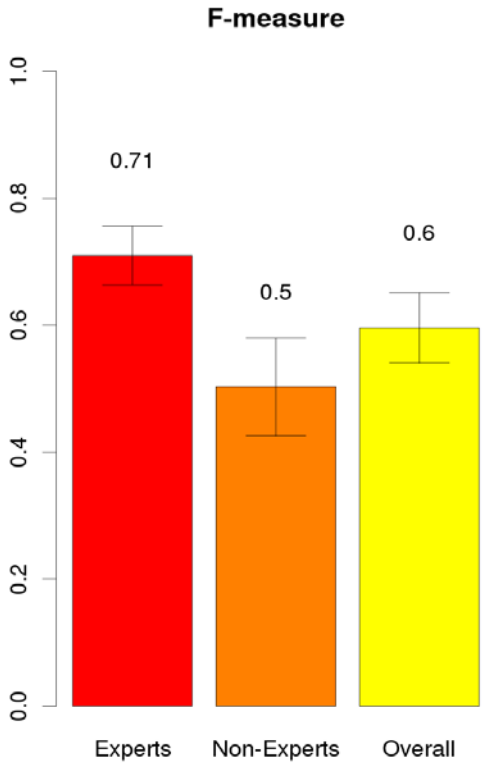

Figure 4: Overall measure of validity (f-measure is the harmonic mean of correctness and of sensitivity).

Measuring accessibility

In (Vigo & Brajnik, 2011) we explored the issue of automatic accessibility metrics. This issue is important not only when one is implementing an accessibility observatory, or interpreting its data, but also when one is comparing two versions of the same site (as in regression testing), or when one is implementing a personalized user interface which illuminates users on potential accessibility problems, or when a search engine has to rank results also in terms of accessibility. Even when determining conformance levels, one has to count how many success criteria of a certain group fail, and then summarize such counts into values taken from {not-conformant, A, AA, AAA}.

Comment n. 6: measurements of accessibility occur very often.

We implemented 6 metrics defined in the literature (failure rate, UWEM, WAQM, WAB, PM, A3); they are all based on the results that an accessibility testing tool can provide. We used EvalAccess and applied it to 1500+ pages (15 sites, 100+ pages each), some of which were labeled as "high-" or "low-accessibility" on the basis of a manual inspection we made.

Figure 5 shows the kind of disagreement between metrics that we found (more detailed charts are available in the paper). In addition, we saw that the disagreement is higher for low accessibility sites, and that most metrics do a very poor job in discriminating high-accessibility sites from low-accessibility ones.

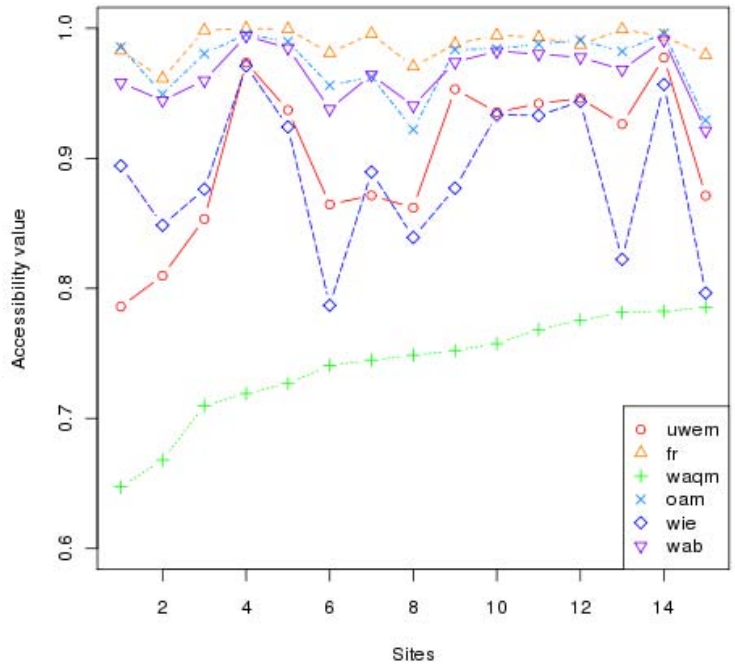

Figure 5: Correlation of different metrics on each site. The scale ranges from 0.6 to 1.0, 1.0 being the highest value of non-accessibility.

The assumptions underlying this experiment are also strong: we worked with automatic metrics only, used a single tool to elicit potential problems; used a partial gold-standard. Despite this, my conclusion is that while automatic metrics are appealing because easy to use and to interpret, and very efficient, we don't know what they measure. For example, *failure-rate* is defined as the number of failures of a criterion/test (like a form control missing its label tag in a page) over the total number of potential failures (the number of all form controls in that page). Notice that the same failure-rate (e.g. 0.30) for two pages may mean very different numbers of accessibility barriers (if page X has 3/10 and page Y has 6/20, then page Y has 12 barriers of that kind, whereas page X has "only" 6 of them). Failure-rate is thus a measure of how well developers addressed accessibility features, it's not a measure of the impact that the accessibility barriers have.

Comment n. 7: automatic metrics measure something different than accessibility.

Because these metrics are based on data produced by automatic means only, no estimation of errors can be made. Another approach would be to "use" human judges to

measure accessibility. This is what judges do when involved by Knowbility Inc. (www.knowbility.org) in the Accessibility Internet Rallies: they use a spreadsheet with penalty points for sites that fail to implement certain accessibility features. In this way, each judge computes the total number of penalty points, and then together with other judges they smooth out potential disagreements, determining a ranked list of sites.

In the past, we suggested to adopt a hybrid approach, named SAMBA (Brajnik & Lomuscio, 2007): the data produced by a testing tool are sampled and given to one or more judges, which are asked to rate the severity of the sampled problems (and implicitly to say if a problem is a false positive). After a relatively simple statistical computation, the overall index of accessibility for the site can be derived, together with an estimation of the margin error. For example, Figure 6 shows the mean value of accessibility for the 15 web sites we considered in that experiment, along with an estimation of the amount of uncertainty. The uncertainty can be used to determine if site X is more accessible than site Y. For example, the first 5 sites on the left overlap in terms of the confidence intervals: it is safe to assume that we don't know which one is more accessible. However, the leftmost site is definitely less accessible than the four ones on the right (calgary, pretoria, bolton, london), since the intervals do not overlap.

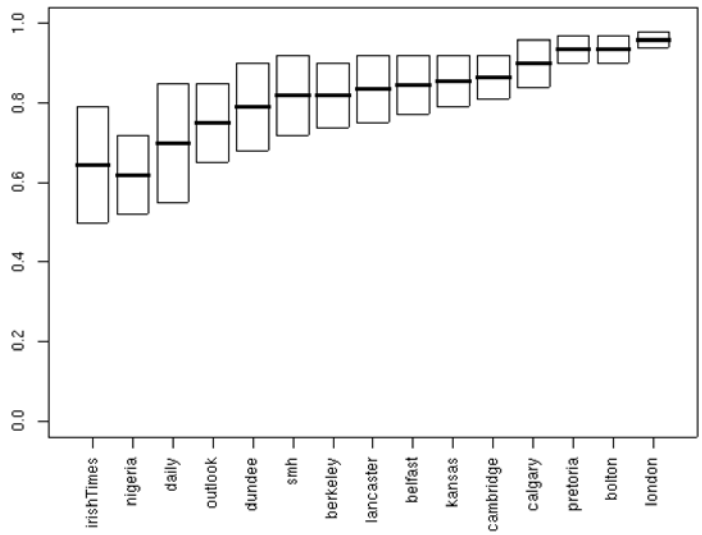

Figure 6: Accessibility levels for 15 websites along with their 95% confidence intervals.

A strong assumption underlying SAMBA is that the uncertainty that can be estimated is related to false positives only; false negatives do not enter into the calculations.

Comment n. 8: manual metrics are the best choice but are expensive. Semi-automatic metrics produce "semi-correct" answers.

Conclusions

My purpose with this quick survey is to convey the message that web accessibility is much more than just checking whether the IMG tag has an ALT attribute, and that this leads to a number of issues. I claimed that:

1. accessibility needs to be contextualized, just like usability;
2. in order to practice a sound accessibility engineering, you need to stick to studied and standardized evaluation procedures;
3. page selection is likely to affect the outcome of an audit;
4. conformance to WCAG 2.0 and barrier walkthrough are subjective;
5. even experts make a relatively large number of errors when determining conformance;
6. even if you are not aware of it, you often measure accessibility;
7. automatic means to measure accessibility are cheap, but not useful;
8. manual metrics are the best choice but they are expensive and face problems 3-4-5 listed above; semi-automatic ones suffer from blindness with respect to false negatives.

I believe that only when these issues are considered and appropriately handled, then we can say that web accessibility is approached with an engineering attitude.

The task ahead of us is clear: we need first to determine how the quality of the accessibility processes (selection of pages, elicitation of problems, measurement, or other ones) can be evaluated; then study evaluation methods, and finally determine how they can be embedded in the development phases and steps. At that point, practitioners could select the best methods for the case at hand, and be relatively sure that results matching a given quality level will ensue, including an accessible web site.

References

Brajnik, G. (2008). Beyond Conformance: the role of Accessibility Evaluation Methods. In S. Hartmann, X. Zhou, & M. Kirchberg (Eds.), WISE 2008: 9th Int. Conference on Web Information Systems Engineering – 2nd International Workshop on Web Usability and Accessibility IWWUA08, LNCS 5176 (p. 63–80). Auckland, New Zealand: Springer-Verlag.

Brajnik, G. (2009). Validity and Reliability of Web Accessibility Guidelines. Proc. of 11th Int. ACM SIGACCESS Conference on Computers and Accessibility – ASSETS 2009 (p. 131–138). Pittsburgh, PA.

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

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

Brajnik, G., Yesilada, Y., & Harper, S. (2010). Testability and Validity of WCAG 2.0: The Expertise Effect. Proc. of the 12th Int. ACM SIGACCESS Conf. on Computers and Accessibility, ASSETS 2010 (p. 43–50). Orlando, Florida, USA: ACM. doi.acm.org/10.1145/1878803.1878813

De Marco, T. (1982). Controlling software projects. Yourdon Press.

Kelly, B., Sloan, D., Brown, S., Seale, J., Petrie, H., Lauke, P., & Ball, S. (2007). Accessibility 2.0: people, policies and processes. W4A '07: Proc. of the 2007 International Cross-Disciplinary Conference on Web Accessibility (W4A) (p. 138–147). New York, NY, USA: ACM.

Petrie, H., & Kheir, O. (2007). The relationship between accessibility and usability of websites. Proc. CHI 2007 (p. 397–406). San Jose, CA, USA.

Vigo, M., & Brajnik, G. (2011). Automatic web accessibility metrics: where we are and where we can go. Interacting with Computers, 23(2), March, (p. 137-155).

Yesilada, Y., Brajnik, G., & Harper, S. (2009). How Much Does Expertise Matter? A Barrier Walkthrough Study with Experts and Non-Experts. Proc. of 11th Int. ACM SIGACCESS Conference on Computers and Accessibility – ASSETS 2009 (p. 203–210). Pittsburgh, PA.

About the Author:

Giorgio Brajnik is a computer scientist with an interest in development and evaluation principles for user interfaces of computer systems; he is currently working on accessibility evaluation methods, on evaluation of user experience, and on interaction design tools and methods. He is assistant professor at the Dept. of Mathematics and Informatics of the University of Udine, in Italy, where he teaches "User Centered Web Development" and "User Experience".