Getting one voice: tuning up experts' assessment in measuring accessibility

Silvia Mirri
Department of Computer Science
University of Bologna
Via Mura Anteo Zamboni 7
40127 Bologna (BO), Italy
silvia.mirri@unibo.it

Ludovico A. Muratori
Polo Scientifico-Didattico Cesena
University of Bologna
Via Sacchi 3
47023 Cesena (FC), Italy
ludovico.muratori3@unibo.it

Paola Salomoni
Department of Computer Science
University of Bologna
Via Mura Anteo Zamboni 7
40127 Bologna (BO), Italy
paola.salomoni@unibo.it

Matteo Battistelli
Polo Scientifico-Didattico Cesena
University of Bologna
Via Sacchi 3
47023 Cesena (FC), Italy
matteo.battistelli4@unibo.it

ABSTRACT
Web accessibility evaluations are typically done by means of automatic tools and by humans' assessments. Metrics about accessibility are devoted to quantify accessibility level or accessibility barriers, providing numerical synthesis from such evaluations. It is worth noting that, while automatic tools usually return binary values (meant as the presence or the absence of an error), human assessment in manual evaluations are subjective and can get values from a continuous range.

In this paper we present a model which takes into account multiple manual evaluations and provides final single values. In particular, an extension of our previous metric BIF, called cBIF, has been designed and implemented to evaluate consistence and effectiveness of such a model. Suitable tools and the collaboration of a group of evaluators is supporting us to provide first results on our metric and is drawing interesting clues for future researches.

Categories and Subject Descriptors
H.1.2 [Models and Principles]: User/Machine Systems – *human factor*. H.5.2 [Information Interfaces and Presentation]: User interfaces - *Evaluation/methodology*. H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia - *User issues*. K.4.2 [Social Issues]: Handicapped persons/special needs.

General Terms
Measurement, Performance, Experimentation, Human Factors, Verification.

W4A2012 – Communication, April 16–17, 2012, Lyon, France.

Keywords
Web Accessibility, Evaluation, Monitoring, Metrics.

1. INTRODUCTION
The Italian mathematician Renato Caccioppoli said that "If you are scared of something, measure it". Without referring to any specific metric or measure, such a quote states that numerical synthesis can reveal specific aspects of a given phenomenon, which are helpful to understand it.

Unfortunately, getting a quantitative estimation can be a complex process: as for measuring accessibility two main issues have to be addressed to compute a quantitative estimation that combines both automatic and manual evaluations, done by humans, and involving also tests conducted by users with disabilities. On the other hand, exploiting suitable metrics can be strategic in analyzing and comparing the accessibility levels and barriers of a large amount of Web sites.

Manual evaluation of any accessibility barrier is a task performed by one or more experts, with the aim of evaluating how much the barrier afflicts the navigation by users with disabilities. This estimation is usually expressed by using a range of possible values (in our work we have chosen the [0, 1] real numbers interval).

The first necessary step to obtain a unique final value is suitably combining the set of numerical assessments coming from different human evaluations. Due to the subjective nature of this kind of evaluation activity, values can be distributed into the given range. Finally, the more experts' assessments contribute to compute a value, the more this value can be considered stable and reliable. Synthesize a plethora of assessments on the same barrier as a single value is the due prologue to "tune up many voices to the same tone".

The second step consists in mixing up the manual evaluation together with the automatic ones, which are performed by an accessibility evaluation system (in our case we have used AChecker [4]). Some barriers should be evaluated both with an

automatic parsing and a manual assessment. Each automatic control detects well-known errors, defined by specific syntactic patterns (in our case they are directly referred to WCAG 2.0 techniques). Then the automatic evaluation system outputs 1 for each detected barrier, 0 otherwise. As an example, let us consider the combination between the **IMG** element and its **ALT** attribute:

1. If the **ALT** attribute is omitted the automatic check outputs 1.
2. If the **ALT** attribute is present the automatic check outputs 0.

In both cases a manual evaluation might state that:

- there is no lack of information once the images are hidden (this can happen in case 1, if the image is a pure decorative one);
- there is a lack of information once the image is hidden.

Value combinations are possible and to mix all the cases in a metric is another issue which has to be addressed. We could assert that the manual assessment is more important than the automatic one. We could assert that they identically quantify the height of a barrier afflicting the navigation done by the user with disabilities. Finally, we could say that the automatic assessment is more important than the manual one. There are many reasons supporting all these points of view and dealing with this aspect is out of the scope of this paper.

Summing up, in this work we have investigated the above issues and we approached a solution to provide a metric about many experts' assessments on Web accessibility, also taking into account its integration with automatic assessment on the same target. The very general principles which led us are feasibility and simplicity.

Many works about accessibility measurements are available in literature [2, 6, 7]. They propose metrics (derived from automatic controls and manual evaluations) measuring syntactic correctness, its implicit semantics (referring to actual barriers in accessing Web resources), or integrating these two aspects. However, the critical problem of choosing unique values from multiple manual evaluation results is never dealt with. The metrics we present here has been inspired by such previous and related works (in particular by Giorgio Brajnik's Barrier Walkthrough method [2]), but it has been defined and adapted so as to take into account manual evaluations done by different human operators.

The remainder of this paper is organized as follows. Section 2 (*Gathering and Reporting Data – The VaMoLà System*) will describe the systems which have been exploited in order to gather and report data about Web sites accessibility, by means of automatic and semi-automatic evaluation and monitoring activities. Section 3 (*The CBIF metric*) will detail our proposed metric, measuring automatic and manual evaluations and it will present how we model our metric to take into account different manual evaluations done by many experts. Section 4 (*Assessment and experimental results*) will present an experiment we are assessing and briefly discuss some first results. Finally Section 5 (*Conclusions*) will close the paper, by showing some final considerations and future work.

2. GATHERING AND REPORTING DATA – THE VAMOLÀ SYSTEM

The CBIF metric is based on a previous metric, named BIF [1, 5], which has been designed and proved thanks to the huge amount of data coming from a suitable automatic system for accessibility evaluation and monitoring. Such a system, that is called VaMolà (an acronym standing for Accessibility Validator and Monitor in the Italian language) is born from a collaboration between the University of Bologna and The Emilia-Romagna Region [5]. Its design and development have been led by the necessity of building a system capable to:

- evaluate Web contents accessibility according to the constraints of different sets of guidelines and requirements (including the Italian regulation),
- automatically, periodically and parametrically gather data about accessibility from a huge amount of URLs,
- provide a geo-political view of monitored contents.

The first instance of the above list has been satisfied with the implementation of a specific validator, starting from Achecker by IDRC [4]. The VaMoLà validator extends and customizes controls to the ones the Italian Regulation state, letting users set up a variety of parameters, thereby making them able to focus on specific checks or groups of checks. In its latest release, WCAG 2.0 controls have been exhaustively added to the application [8].

In order to accomplish the second and third instances of the list above, a specific application has been designed and implemented. It has been called AMA (Accessibility Monitoring Application) and integrates the VaMolà validator as its accessibility evaluation engine. AMA lets the users define a series of parameters to monitor accessibility of a wide amount of sites: from the depth of analysis (in terms of links and pages), to the checks to be done, up to the time interval among accessibility evaluations. A suitable database can be populated with the URLs to be monitored, their geographical position and their role inside the structure of public administration [1, 5]. AMA automatically and periodically launches the VaMoLà validator for such URLs, gathering results for reporting. Tabular and graphics views of results can be shown on a Web browser. Reports are completely customizable by users as well. Finally, a mashup with GoogleMaps service let users to see results on a geo-political map.

VaMoLà supports the integration between the measurement of accessibility on the syntactic domain and on the semantic one (human-evaluation). In fact, AMA provides a long series of warnings about the necessity of human controls on specific elements of a Web page, thereby letting experts to focus on them.

3. THE CBIF METRIC

The goal of our metric is measuring how far a Web page is from its accessibility version. In other words, it is a quantitative synthesis about how much accessibility barriers affect user's browsing by means of assistive technologies. Hence, the lower is the resulting value and the better is the accessibility level of the evaluated Web page. To associate errors to barriers in the most effective way, we analyzed WCAG 2.0 and their related Techniques [8]. For each error, success criteria and techniques have been used to identify disabilities/assistive technologies it affects.

A first version of our metric (named Barriers Impact Factor, BIF) is computed on the basis of a barrier-error association table [1]. This table reports, for each error detected in evaluating WCAG 2.0, the list of assistive technologies/disabilities affected by such an error. Barriers have been grouped into 7 sets, which impact in the following assistive technologies and disabilities: screen reader/blindness; screen magnifier/low vision; color blindness; input device independence/movement impairments; deafness; cognitive disabilities; photosensitive epilepsy. Details about BIF metric are

available in [1, 5]. For the sake of simplicity, manually checked controls are not taken into account in this version of the metric.

In order to better quantify accessibility barriers within a Web page, thereby providing a more realistic synthesis, we have decided to take into account the whole amount of controls (including manual assessments). First of all, we have analyzed the validation checks, comparing them with WCAG 2.0 success criteria and then we have identified relationships among them. Whenever a validation check fails, it means that a certain accessibility error occurs or that a manual control is necessary (to certify the effective presence of an error or not). Success criteria suggest checks on the basis of Techniques and Failures [8]. Some of them are devoted to identify different aspects and shapes of the same accessibility barrier, showing some intersections in checks, which have to be manually controlled. In order to avoid overlapping controls on the same accessibility error, we have grouped all the checks into disjoined sets, on the basis of each barrier. Whenever at least one check of a specific group fails, then the related accessibility barrier is actually found in the analyzed Web page. Each barrier is related to one (and only one) success criterion and then to one level of conformity: A, AA or AAA. We have assumed that, differently from syntactic shortcomings which take binary values, manual evaluations take values on the [0, 1] real numbers interval. In particular, 1 means that an accessibility error occurs, 0 means the absence of that accessibility error. The cBIF value for each barrier is computed as follows:

$$cBIF(i) = \sum_i \frac{(m(i) * E_m(i) + a(i) * E_a(i))}{m(i) + a(i)} * \frac{weight(i)}{\#check(i)}$$

where:

- i represents an accessibility barrier, according to detected errors;
- $cBIF(i)$ is the Barrier Impact Factor referred to i, which takes into account both manual and automatic checks;
- $E_a(i)$ represents the number of detected errors which causes the i barrier and which are automatically controlled;
- $E_m(i)$ represents the number of detected errors which causes the i barrier and which are manually controlled by an accessibility expert;
- $m(i)$ is a parametric weight assigned to the manual evaluation related to the i barrier;
- $a(i)$ is a parametric weight assigned to the automatic evaluation related to the i barrier.
- $weight(i)$ represents the weight which has been assigned to the i barrier (related to the corresponding level of WCAG 2.0 conformity, A, AA or AAA);
- $\#check(i)$ represents the number of checks (related to the i barrier) that the system has actually performed (to normalize the number of errors in terms of evaluated checks).

In our proposed metric, each i barrier is evaluated by:

$$\frac{(m(i) * E_m(i) + a(i) * E_a(i))}{m(i) + a(i)}$$

where parameters $m(i)$ and $a(i)$ aim to weight respectively the role of manual and automatic evaluations. Such parameters can be differently assessed for each i barrier and all the cases can be classified as follows:

1. $m(i)=a(i)$: in this case the formula is a mere average between $E_m(i)$ and $E_a(i)$;
2. $m(i)>a(i)$: in this case the failure in manual assessment is considered more significant than the automatic one;
3. $m(i)<a(i)$: in this case the failure in automatic assessment is considered more significant than the manual one.

The following tables represent a sort of path-matrix or directed graph, to relate manual and automatic checks. Cells contain expected values of integrated evaluations whenever syntactic or semantics checks fail or not. Narrows point out a possible order in weighting final values: in particular, as for Figure 1, we can say that failure on manual checks is considered more significant than the automatic one (case 2, in the above list). On the contrary, Figure 2 shows the case that failure on automatic checks is considered more significant than the manual one (case 3, in the above list). Hence, accessibility level is assumed to increase or decrease by moving from one cell to another of the same row/column. Cell values are maximum and minimum values for each weight in our metric.

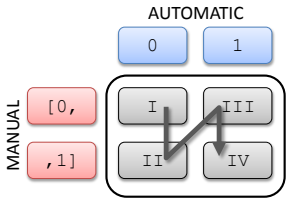

Figure 1 – Path-matrix, case 2

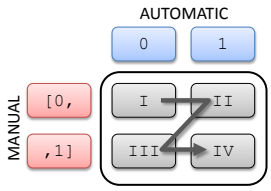

Figure 2 - Path-matrix, case 3

One of the aspects we had to face in providing the "semantic part" (i.e. any final value for manually checked barriers) for cBIF, is addressed to the way of arranging evaluations done by many experts. On the one hand, the presence of more than one manual evaluation about the height of any given obstacle in accessing Web content is expectable. On the other hand such a multiplicity implies any rule or model to consistently synthesize all the given rates. The more human operators provide evaluation about an accessibility barrier and the more the value they express (in terms of accessibility level) can be meant as reliable. This behavior is very similar to the online rating systems ones. It has become a popular feature in the Web 2.0 applications and it typically involves a set of reviewers assigning rating scores (based on various evaluation criteria) to a set of objects [3]. In particular, reviewers can develop trust and distrust on rated objects depending on a few rating and trust related factors [3]. It is worth noting that new reviewers rating can be influenced by already expressed evaluations from other reviewers. Moreover this new additional weight of manual evaluations cannot be a mere average values, because variance must be considered in order to reinforce or not the whole

computed accessibility level. All these aspects have to be taken into account in defining our metric.

As for cBIF, we adopted a very simple statistical model and some very general assumptions about gathered rates. The following list summarizes the issues of such a model:

- Each human evaluation about accessibility of a given Web content is requested to be quantified as a value over the real numbers interval [0, 1].
- Every human evaluation is done without previously knowing the other ones referring to the same content and/or the same barrier.
- The μ average and σ^2 variance are computed for the set of evaluations about such assessment (i.e. for the rates that a given sample of experts has assigned). They can be exploited in measuring the highness of the referring barrier. Such values are used as $E_m(i)$ on the cBIF formula, which we detailed above.

Summing up, for any given analyzed element of code, once it is associated to any i_{th} barrier, a unique value emerges, corresponding to the experts' rates. And this value represents a suitable approximation of common sense (where with the adjective "common" we mean the experts opinion). Moreover, the more experts' assessments contribute to compute a value, the more this value can be considered stable and reliable.

4. ASSESSMENT AND EXPERIMENTAL RESULTS

As a first assessment of our proposed metric, we have evaluated the accessibility of a set of Web pages, by means of the AMA monitoring system, involving a group of experts (composed by 5 people). We have evaluated 10 Web sites of Italian Public Administration according to WCAG 2.0 success criterion 1.1.1, by using the automatic validator of the AMA system. Then, we asked to the experts group to rate accessibility barriers for the same pages (mainly related to adequateness of image textual alternatives), according to a range over the real numbers interval [0, 1]. During the whole experiment we have assigned 2 to the m parameter and 1 to the a parameter (see the cBIF parameters as discussed in the previous Section).

Experts group has faced different kind of situations and errors in the textual alternatives of images, from too long alt text for pure decorative images to the absence of alt, title or any other textual clue for images used as links.

Resulting variance shows that the experts have assigned different values for some images evaluation, thereby expressing that they disagree. In this experiment, manual evaluations have always been conducted by the whole group of experts. In future works we will compare results coming from different number of experts' manual assessment, so as to better discuss about evaluation reliability. It is worth mentioning that such results are related to the mere evaluations of 1.1.1. success criterion. Results evaluations must be completed, by extending assessments to the whole set of accessibility barriers and related automatic checks.

It is worth mentioning that this is an ongoing work. Our experts are still conducting manual evaluations and we are still appraising how to adjust our proposed metric on the basis of preliminary results obtained by applying it and of experts' work, so as to better represent results of evaluations done by many experts.

5. CONCLUSION AND FUTURE WORKS

This paper presents an accessibility metric, which has been designed with the aim of evaluating barriers as a whole, combining results provided by using automatic tools and manual evaluations done by experts. We have identified different issues in combining values gathered from these two different sources and suggested solutions to obtain a feasible barrier evaluation. The defined metric has been preliminary tested by measuring the barriers in several local public administration sites. Five experts are supporting the evaluation by manually assessing barriers related to WCAG 2.0 1.1.1 (as specified by the WCAG 2.0 techniques). We used the automatic monitoring system AMA both to verify the page content and to collect data from manual evaluations.

There are two main open issues that we want to address as future works: (i) propose and discuss weights for the whole WCAG 2.0 set of barriers; (ii) investigate how the number of experts involved in the evaluation, together with their rating variance, could influence the reliability of the computed values.

6. ACKNOWLEDGMENTS

The authors would like to thank Catia Prandi and Mauro Donadio.

REFERENCES

[1] Battistelli, M., Mirri, S., Muratori, L.A., Salomoni, P., Spagnoli, S. Making the Tree Fall Sound: Reporting Web Accessibility with the VaMoLà Monitor, in *Proceedings of the 5th International Conference on Methodologies, Technologies and Tools enabling e-Government*, Camerino (Italy), 30th June - 1st July 2011.

[2] Brajnik, G. and Lomuscio, R. SAMBA: a Semi-Automatic Method for Measuring Barriers of Accessibility. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility 2007*, pp. 43-50.

[3] Chua, F.C.T., Lim, E. Trust network inference for online rating data using generative models. In *KDD'10 Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington D.C. (USA), 2010.

[4] Gay, G.R., Li, C. AChecker: Open, Interactive, Customizable, Web Accessibility Checking. In *Proceedings 7th ACM International Cross-Disciplinary Conference on Web Accessibility (W4A 2010)* Raleigh (North Carolina, USA), April 2010, ACM Press, New York, 2010.

[5] Mirri, S., Muratori, L.A. and Salomoni P. Monitoring accessibility: large scale evaluations at a geo political level. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS'11)*, Dundee (Scotland, UK), October 2011.

[6] Parmanto, B. and Zeng, X. Metric for Web Accessibility Evaluation. *Journal of the American Society for Information Science and Technology*, 56(13):1394–504, 2005.

[7] Vigo, M., Arrue, M., Brajnik, G., Lomuscio R. and Abascal, J. Quantitative Metrics for Measuring Web Accessibility. In *Proceedings of the W4A2007* (Banff, Alberta, Canada, May 7-8, 2007) ACM Press, New York, NY, 2007, 99-107.

[8] World Wide Web Consortium. Web Content Accessibility Guidelines (WCAG) 2.0. Available at: http://www.w3.org/TR/WCAG20/, 2008.