Avoiding to dispense with accuracy: A method to make different DTDs documents comparable

Matteo Battistelli  
Corso di Laurea in Scienze e Tecnologie informatiche  
Via Sacchi, 3  
Cesena - Italy  
+39 0547 338829  
matteo.battistelli4@unibo.it

Silvia Mirri  
Dipartimento di Scienze dell'Informazione  
Via Mura Anteo Zamboni, 7  
Bologna - Italy  
+39 0547 338813  
silvia.mirri@unibo.it

Ludovico Antonio Muratori  
Dipartimento di Scienze dell'Informazione  
Via Mura Anteo Zamboni, 7  
Bologna - Italy  
+39 0547 338813  
ludovico.muratori3@unibo.it

Paola Salomoni  
Dipartimento di Scienze dell'Informazione  
Via Mura Anteo Zamboni, 7  
Bologna - Italy  
+39 0547 338813  
paola.salomoni@unibo.it

Simone Spagnoli  
Corso di Laurea in Scienze e Tecnologie informatiche  
Via Sacchi, 3  
Cesena - Italy  
+39 0547 338829  
s.spagnoli@unibo.it

ABSTRACT
The Italian regulations about Web content accessibility impose HTML 4.01 strict, XHTML 1.0 strict or superior grammar validity. Such markup constraints are complied by a low percentage of Public Institution sites. In order to include the largest amount of not strict DTD sites on a synthetic, realistic representation of accessibility degree, this paper presents an experimental approach to set up parameters for making validity measures uniform, despite they are taken from evaluating different grammars. The proposed method allows for effectiveness in computing and storing information which otherwise will be quite unfeasible. Finally, the generalization of such an approach is shown to be suited for markup quality evaluations, beyond explicit law requirements.

Categories and Subject Descriptors
H.5.4 [Hypertext/Hypermedia]: User issues;  
I.7.2 [Document Preparation]: Markup languages;  
I.7.5 [Document Capture]: Document analysis;  
K.4.2 [Social Issues]: Assistive technologies for persons with disabilities.

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

Keywords
Accessibility evaluation, Standards, DTD compliance, Monitoring, Metrics.

1. INTRODUCTION
In assessing accessibility of Web contents, international scientific researches have definitively assumed the necessity of continuous and homogeneous measures. Metrics about barriers on Web pages have been proposed and justified both experimentally and theoretically, to compare or order accessibility degrees. Consistency of results, i.e. their capability to be meaningfully compared, are guaranteed by constraints and goals the international guidelines such as [15, 16] and National regulations [6, 9] state. Once barriers are identified according to some rules and methods, they can be counted and processed to provide a synthesis on some ordered dominion. Admitted variations and all-out presence of violations in some requirements limit or prevent, the possibility of realistic assessment procedures. This is the case of the Italian regulations about Web content accessibility [9] and their pursuance by the Italian Public institutions Web sites. On the one hand, the law imposes the markup being HTML 4.01 strict, XHTML 1.0 strict or superior, and thus it provides distinct dominions which are comparable in a difficult way. On the other hand, as the authors of this paper have experienced on the VaMoLà monitor [7], just about 43% of Italian Public institutions homepages are declared DTD strict, while the others refer to a transitional or frameset grammar. Any meaningful synthesis about code validity cannot simply exclude 57% of monitored Web contents. Hence, a normalizing parameter or distance as regards
strictness, in order to homogenize markup, has to be hypothesized to include not strict sites on a synthetic and down-to-earth code validity estimation.

The VaMoLà project is devised as a collaboration between the University of Bologna and the Emilia-Romagna Region, gathering two integrated applications. The first one is a validator; its goal is getting together all the automatic procedures the law imposes and providing some guidelines to the subjective manual evaluation the regulations state as well. The second one is a monitor, whose aim is periodically querying the validation tool and reporting the accessibility of a given set of Web sites. Such a monitor utilizes messages and errors coming from the validator, in order to compute some sort of measures, on the strength of metrics. The goal of the whole project is aiding public institutions to effectively survey their Web contents and services and to point out a global, synthetic portrait of accessibility and its dynamics in time [7].

Since the VaMoLà monitor is expected to gather data on 8-10000 Web sites and other 20 requirements have to be verified, effectiveness in evaluating DTD conformance is necessary. Forcing validation on strict DTDs, in order to provide any distance from strictness and homogenize results, implies the loss of information about actual DTD conformance or necessarily increase costs of computation and storage. Here, we propose an experimental approach, starting from a large sample of Web sites and the appliance of suitably (once) computed parameters on the strength of it. This way, analysis of really declared grammar is preserved without increasing costs of computation and storage in monitoring large amount of pages. Fallouts of gathered information on the sample are also analyzed in this paper, whose reminder is organized as follows: Section 2 summarizes notable works about accessibility metrics; Section 3 details scenario and hypothesis of our approach about DTDs variety; Section 4 works about accessibility metrics; Section 5 finally concludes the paper, describing settings and results of experimental assessment about hypothesis of our approach about DTDs variety; Section 3 details scenario and hypothesis of our approach about DTDs variety; Section 4 works about accessibility metrics; Section 5 finally concludes the paper, describing settings and results of experimental assessment about hypothesis of our approach about DTDs variety.

The accessibility barrier target is taken into account by Unified Web Evaluation Methodology (UWEM) [12]. On its first release, it assumes to group barriers on the strength of WCAG checkpoints and synthesizes their presence with the probability that a pages sample contains barriers for a certain kind of disability. Such a formula assumes the statistical independence of barriers (i.e., that they are not correlated). Some limitations, here are own to the discretion in evaluating the probability that a barrier affects a disability and to the lack of constraints about pages complexity. Bühler et al. [3] proposed a parameter to describe the pages sample complexity (for a certain barrier). It is defined as the sum of ratios respectively between actual and potential error on a certain barrier and between actual errors on such a barrier and the amount of all the errors (of all the barriers). Hence, the following formula is proposed to evaluate the probability that a pages sample \( p \) presents barriers for disability of type \( u \):

\[
A_u(p, u) = 1 - \prod_{b \in u} (1 - S_{ub})^{B_{pb}}
\]

Where \( C_{job} \) is the previously hinted parameter, \( S_{ub} \) is the probability that \( b \) is a barrier for \( u \). Notwithstanding such improvements, Freire et al. [4] show that the UWEM and \( A_3 \) are strictly correlated and they produce similar results.

On a further release of UWEM [14], given the test \( t \), the amounts of appliances and their failures on a page \( p \) are computed to provide the formula:

\[
f(p) = \frac{B_p}{N_p}
\]

Where \( N_p \) is the amount of all the tests (chosen from a set), and \( B_p \) is the sum of all the failures for all the tests.

This formula produces values between 0 (tests never fail) and 1 (all the tests fail). Once again, the above hinted limitations about page complexity and disability target impact are present on UWEM, even though it is really very simple to be computed. The Web Accessibility Quantitative Metric (WAQM) metrics has been proposed by Vigo et al. [11] to overcome previous measures limitations. It takes into account each accessibility attribute (according to the WCAG 2.0 vocabulary), each checkpoint, its priority and how many times it has been tested, every warning (or potential error to be manually verified) and finally, the ratio between errors and tests (as a normalized value). In order to enlarge the range of possible measures in case of a low ratio, a suitable multiplier has been proposed to obtain a parametric final formula (see [11] for details). WAQM provides values from 0 to 100, where 100 is an index of complete accessibility. Some limitations come from the amount of variables to be computed.
and from the lack of a methodology to mediate manual and automatic evaluations. Brajnik and Lomuscio [2] have defined the Semi-Automatic Method for measuring Barriers of Accessibility (SAMBA), to integrate manual and automatic evaluations on the strength of barriers harshness and of tools errors rate. Through a three phases process a “disability vector” and a “confidence intervals severity matrix” are created, whose values report the probability that each barrier affects a particular disability. Finally, a barrier density factor is computed to state the probability that each line of markup may contain a barrier (it is the ratio between the amount of potential barriers and the number of markup lines). Accessibility factors (called Raw Accessibility Index and Weighed Accessibility Index) are calculated from the previous values. The main advantage of SAMBA is related to its user-centered approach to accessibility and to its precision in considering human and automatic evaluations.

3. COMPARING SITES ACCESSIBILITY WITH DIFFERENT DTD

The VaMolà project, which has been detailed on [7], consists of two integrated services (through a web-centered interface) to respectively evaluate and monitor the accessibility degree of Public Institutions Web sites according to the Italian Law [6]. The validator (based on AChecker [1]) lets the users assess actual and potential barriers referring to the 22 requirements the regulations impose. The monitor allows to automatically gather and store data from the validator, depending on the geographical location of Web sites institutions, their role and all the changes the evaluated contents undergo in time.

Since the validator provides –page by page– an analytical list of errors and warnings, the monitor has to present different views of aggregated data related to groups of sites. In particular, the monitor shows different combinations derived from analytical sources (see [7] for details):

- a whole evaluation of law conformance, obtained aggregating data related to a single requirement.
- Site assessment, computed on the basis of values associated to single pages.
- Evaluation of specific groups of sites (on the basis of the geographical location or the type of institution) computed starting from single site values.

Relationship between evaluation results and pages complexity (e.g. the number of elements, or the amount of text or media contents), as well as the appliance of weights to each of them, which are provided by experts, can refine measures. Also the most part of the formulas from previous works about accessibility metrics can be applied to the gathered data, by customizing some parameters to the law provisions.

In this scenario, we have started working on the first requirement of Italian Law about Web accessibility which admits three different strict grammars: HTML 4.01 strict, XHTML 1.0 strict and XHTML 1.1. This requirement represents one of the main difference between the Italian Web accessibility law and all the other international acts and guidelines which allows pages compliant to a larger set of grammars. Furthermore, from an analysis on more than 1000 Homepages, VaMolà has revealed that only 43% of them are declared to be strict. These figures have implied some reflections about the opportunity of not simply excluding a large amount of sites with unlawful grammar from a meaningful synthesis about validity degree. Hence, in order to provide a more complete evaluation, we have considered three different properties (resulting from three different evaluations):

A. validity, to the declared DTD; this value is computed by dividing the number of errors in the declared DTD by the number of DOM elements inside a page $p$;

B. strictness, i.e. compliance to the correspondent strict DTD. Pages declared as transitional or frameset in a specific markup language (HTML 4.01 or XHTML 1.0) are evaluated as declared with the correspondent strict DTD;

C. markup quality, i.e. compliance to XHTML 1.1. Pages declared with different DTDs are evaluated as declared with the XHTML 1.1 DTD.

Measures from such properties are obtained by forcing the validation of DTDs which can be considered as restrictions in comparison with that one a page declares. Managing such an amount of data on a wide set of pages (as the monitor periodically does) is too heavy, both in terms of computation time and storage dimension.

Thus, a couple of parameters we called Errors Sprin ging up Rates (ESUR, indicated in the following as $X'_i$ and $X''_i$) can be computed to relate Web contents with different grammars starting from a single validity evaluation (A). Properties B and C are evaluated by using the following formula for a page $p$:

$$\tilde{n}_j = x'_i n_j + x''_i (n_j + 1)$$

where:

- $i$ indicates the declared DTD and $j$ indicates the destination DTD, i.e. the correspondent strict DTD in case B and the XHTML 1.1 DTD in case C.
- $n_j$ denotes the number of errors in the declared DTD divided by the number of DOM elements inside that page $p$. Analogously $\tilde{n}_j$ is the expected value calculated on the $j$ DTD divided by the number of DOM elements inside that page $p$.
- $X'_i$ and $X''_i$ are the ESUR parameters which depend from the property B or C and are computed as follows.

To estimate $X'_i$ and $X''_i$, we used a sample of more than 1000 Italian public institution Homepages. For each page, we considered all the three conformance evaluations listed above. The resulting data have been used to compute $X'_i$ and $X''_i$ as follows:

$$X'_i = \frac{1}{2^{\#P_e}} \sum_{k=1}^{\#P_e} n_{k_i}; \quad X''_i = \frac{1}{\#P_v} \sum_{k=1}^{\#P_v} n_{k_j}$$

Where $\#P_e$ and $\#P_v$ are the amounts of pages (declared with the i DTD) with and without errors. $S$ has been experimentally assessed
to minimize the variance between estimated and computed values. In particular:

- for strictness, S is 2 whenever the declared DTD is HTML and 1 for XHTML;
- for markup quality, S is 3 whenever the declared DTD is HTML and 2 for XHTML.

Let us notice that actual declared grammar validation is preserved when a page is monitored, since our parameters are computed once on the strength of statistical investigations. Computation costs (for revalidating twice in case of transitional grammars) and storage space increase are prevented, since ESUR is simply computed on effective gathered data. The cost or distance from law recommendation to plan updates in case of not strict grammar can also be assessed analyzing results we obtained from our sample and used on the VaMoLà monitor through the ESUR appliance. In the next section we will show the scenario where we gather data and all their characteristics.

4. EXPERIMENTAL ASSESSMENTS

Pages sample which has been used to obtain ESUR has been evaluated on August 2009. It consists of about one-thousand Homepages geographically distributed over the Italian territory as follows: 521 from the north, 252 from the centre, 195 from the south and 142 from the islands. Such URLs are homepages of Italian Public Administration Web sites. Due to the kind of subjects which have to the Stanca Act requirements [6], this amount of monitored homepages is an adequate starting point for our evaluations. We are still collecting data through experimental assessments and we are enlarging this pages sample, in order to obtain more complete evaluation of the Italian Public Administration Web homepages and to better cover all the kinds of Italian Public Administrations.

The complete list of monitored homepages is available at: http://polo-rer.polocesena.unibo.it/matteo/lista$url/.

Figure 1 depicts a screenshot of the VaMoLà Monitor synthesizing accessibility situation of such a sample (in Italian language). The map on the right shows the percentage of valid sites (independently from their declared DTD), from lighter to darker blue. Table 2. reports declared and valid DTDs (in percentage). Validity percentage (case A of the previous Section) is referred to the amount of pages with their declared DTD. ESUR values for strictness (case B on the previous Section) are summarized on Table 3. Variance of expected $\tilde{H}_j$ and re-validated $n_j$ is less than 0.1. The couple ($x'_j$,$x''_j$) for markup quality measure, estimated for each not strict declared grammar is reported on table 4. Variance is less than 0.07. The symbol $\sim$ points out that values are computed from a small subsample.

Table 2. Sites per declared and valid DTD

<table>
<thead>
<tr>
<th>DTD</th>
<th>Declared</th>
<th>Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML 3.2</td>
<td>0.18%</td>
<td>0%</td>
</tr>
<tr>
<td>HTML 4.0 Frameset</td>
<td>0.09%</td>
<td>0%</td>
</tr>
<tr>
<td>HTML 4.0 Transitional</td>
<td>4.47%</td>
<td>4%</td>
</tr>
<tr>
<td>HTML 4.0 Strict</td>
<td>0.09%</td>
<td>0%</td>
</tr>
<tr>
<td>HTML 4.0 Frameset</td>
<td>0.89%</td>
<td>0%</td>
</tr>
<tr>
<td>HTML 4.01 Transitional</td>
<td>34.7%</td>
<td>6.79%</td>
</tr>
<tr>
<td>HTML 4.01 Strict</td>
<td>7.42%</td>
<td>36.58%</td>
</tr>
<tr>
<td>XHTML 1.0 Frameset</td>
<td>0.09%</td>
<td>0%</td>
</tr>
<tr>
<td>XHTML 1.0 Transitional</td>
<td>16.91%</td>
<td>19.25%</td>
</tr>
<tr>
<td>XHTML 1.0 Strict</td>
<td>5.37%</td>
<td>56.67%</td>
</tr>
<tr>
<td>Other DTD</td>
<td>0.18%</td>
<td>0%</td>
</tr>
</tbody>
</table>

Table 3. ESUR values for strictness measurement

<table>
<thead>
<tr>
<th></th>
<th>HTML 4.01 Strict</th>
<th>XHTML 1.0 Strict</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML 4.01 Frameset</td>
<td>0.937 ;~0.1</td>
<td></td>
</tr>
<tr>
<td>HTML 4.01 Transitional</td>
<td>0.653 ;0.103</td>
<td></td>
</tr>
<tr>
<td>XHTML 1.0 Frameset</td>
<td>~1.3 ; ~0.074</td>
<td></td>
</tr>
<tr>
<td>XHTML 1.0 Transitional</td>
<td>1.317 ; 0.130</td>
<td></td>
</tr>
</tbody>
</table>

Table 5 points out validity, strictness and markup quality for geographically aggregated data. It is necessary noticing that the current sample is not completely homogeneous (as it concerns to geography). Plenty of sites picked up on some zones, indeed, amplifies not valid (or faraway from strictness and quality) sites.

Let us notice that differences between estimated and real values are acceptably small, either as it concerns to the strictness property or to the Markup quality of sites.
Table 4. ESUR values for markup quality measurement

<table>
<thead>
<tr>
<th>Markup Quality</th>
<th>XHTML 1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(X'_1; x'_1)</td>
</tr>
<tr>
<td>HTML 4.01 Frameset</td>
<td>1.445 ; ~1.03</td>
</tr>
<tr>
<td>HTML 4.01 Transitional</td>
<td>2.482 ; 1.038</td>
</tr>
<tr>
<td>HTML 4.01 Strict</td>
<td>4.062 ; 1.406</td>
</tr>
<tr>
<td>XHTML 1.0 Frameset</td>
<td>~0.662 ; ~0.04</td>
</tr>
<tr>
<td>XHTML 1.0 Transitional</td>
<td>0.693 ; 0.043</td>
</tr>
<tr>
<td>XHTML 1.0 Strict</td>
<td>0.376 ; 0.011</td>
</tr>
</tbody>
</table>

Table 5. Validity, strictness and Markup quality per geographical zones

<table>
<thead>
<tr>
<th>Geographical Zone</th>
<th>Validity (real; estimated)</th>
<th>Strictness (real; estimated)</th>
<th>Markup quality (real; estimated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>0.253</td>
<td>0.329 ; 0.272</td>
<td>1.038 ; 1.099</td>
</tr>
<tr>
<td>Centre</td>
<td>0.319</td>
<td>0.406 ; 0.401</td>
<td>0.951 ; 0.988</td>
</tr>
<tr>
<td>South</td>
<td>0.250</td>
<td>0.331 ; 0.301</td>
<td>1.046 ; 1.112</td>
</tr>
<tr>
<td>Islands</td>
<td>0.377</td>
<td>0.506 ; 0.480</td>
<td>1.148 ; 1.346</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS AND FUTURE WORK

Properties of strictness and markup quality have been introduced in order to support a more complete evaluation of Web pages compliance to the standards. The former can be conceived as a distance between a not strict page and a strict one; the latter generalizes the previous distance to measure quality.

To effectively estimate strictness and markup quality of a Web page with a single markup validation, we have hypothesized a formula, whose a couple of parameters has been experimentally introduced.

We are working to set up a wider experiment by including about 10,000 homepages. Analogous process of evaluation is going to be applied to the other Italian Law requirements, in order to provide an accessibility evaluation beyond code quality.

ACKNOWLEDGMENTS

The authors would like to thank Giovanni Grazia and Jacopo Deyla (from the Emilia-Romagna Region), Marina Vriz and Ennio Paiella (from ASPHI) and Gregory R. Gay (from ATRC, Faculty of Information, University of Toronto).

REFERENCES