

Automatic Monitoring of Web Accessibility: the Validity Issues

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

June 7, 2006

Abstract

This paper highlights some of the difficult issues related with validity of data that occur when dealing with results produced by automatic tools for accessibility testing. In particular the paper shows some experimental results collected by repeatedly applying a testing tool to a number of web sites at two time points and by drawing some conclusions about comparisons between web sites and across time.

1 Introduction

Monitoring web accessibility is an activity that is going to be more and more important in the future, as common perception of web quality factors increase and correspondingly does the role of accessibility.

There are many reasons why one should monitor web accessibility of a single web site or by comparing accessibility levels of different web sites; they include (i) to rank tested web sites with respect to some quality factor related to accessibility; (ii) to show how the accessibility level reached by a web site changes over time; (iii) to show how the accessibility level changes across sections of a web site; (iv) to show which are the most typical accessibility defects within a web site; (v) to provide and show the distribution of the different kinds of problems across different kinds of web sites; and (vi) to estimate what is the effort required to fix some of the defects present in a web site.

In order to reliably perform such monitoring activities, a service should be capable of periodically analyzing a set of web sites and populate a data warehouse that can be mined.

Although human intervention is needed, such a monitoring service can only be built upon automatic tools. The frequency of the analysis, the sheer amount of pages to test, the level of details at which most of the tests have to go, the large number of issues that are typically raised for each page, can only be dealt

with by automatic means. In this way one could update the warehouse quickly, often, at a low cost, and be able to analyze the web sites in a systematic way and be able to produce results that are as standard as possible, and therefore as comparable as possible.

The fact that accessibility of a web site is an external property of the web site¹, and especially one that is depends upon human cognitive processes, makes it very difficult to measure in a reliable way.

As a consequence, there are many open issues related with monitoring methods.

In this paper I want to highlight some of the issues dealing with validity of the results that could be produced by mixed-methods (*i.e.* based on automatic tools and integrated with human judgments) and present some experimental data collected in a small-scale experiment that was run in 2005 and 2006.

2 Related work

There are several observatories for accessibility and web quality (for example [12; 6; 1]). However all of them are based on unreliable or unknown methods for measuring accessibility, and they often mix several properties with accessibility (like quality, usability, searchability, content adequacy, interactivity, navigability). Figure 1 shows an example of the comparisons that can be drawn from data collected through monitoring methods.

A notable exception is the ongoing European project on the Accessibility Observatory [8].

3 The validity issues

There are several methods that can be adopted to test for accessibility. It can be tested based on guidelines (like WCAG 1.0, or Section 508) through a *standards review* method, or other methods can be employed, like user testing [5], usability inspection methods [11; 7; 10], barrier walkthrough [3] or those suggested in [9].

Automatic monitoring of web sites require the ability to automate the following activities, each of which can introduce factors that can invalidate or reduce the quality of the results: (i) spidering web sites; (ii) collecting appropriate features² from web pages and related files (html, css, javascript, flash, pdf, ...); (iii) classification of features as failure modes (*i.e.* as accessibility barriers); (iv) classification of features into defects (diagnosis); (v) assignment of severity scores to

¹External properties of a software system are properties that can only be measured with respect to how the system relates to its environment.

²In the following I will use these definitions: a *feature* is a pattern of DOM elements (like an IMG tag with no ALT attribute, a table with no TH, a spacer image, a data table, etc.) or absence of a pattern; a *guideline violation* is a feature that does not agree with the guideline requirement; a *failure mode* is a hindrance, for a user in given situations, to the achievement of a user goal; a *defect* is a feature 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 *frequency* (extent to which the failure mode is encountered during a task execution).

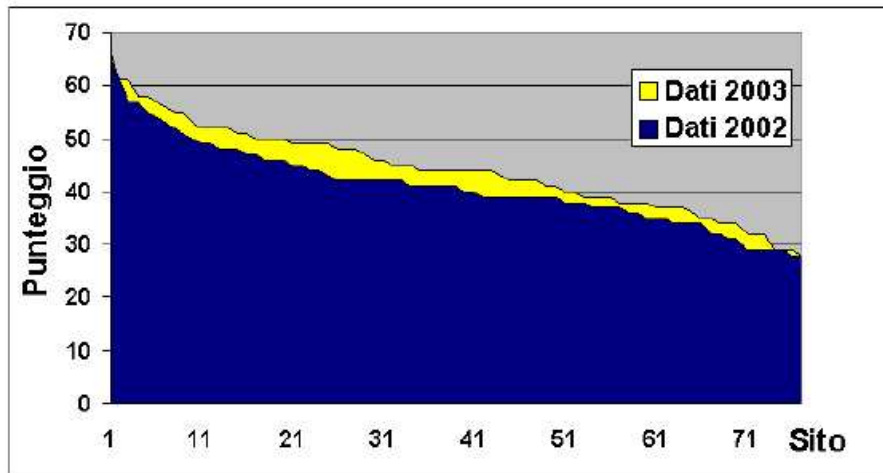


Figure 1: Taken from [1], this diagram shows the distribution of accessibility scores assigned to about 100 web sites when they were evaluated in 2002 and in 2003.

defects and/or guideline violations; (vi) aggregation of these defects and their severity scores so that appropriate and valid conclusions can be drawn with respect to global accessibility assessments. Once a warehouse of data is available, then also the activities of trend analysis, drill-down inspections, statistical inferences, data visualizations need to be supported.

There are several open issues that need to be solved before reliable, valid, and useful monitoring services can be automated:

- detection of useful features might be complex to implement (for example, detection of data tables, or of skip-links solutions);
- detection might not be correct;
- classification of features into failure modes might not be valid (i.e. some feature is erroneously related to some failure mode, for example an image with an empty ALT attribute, or viceversa some failure mode is not detected, like a badly chosen ALT text);
- classification of features into defects (i.e. diagnosis) is not valid, for example because not all the causes are identified;
- invalid assignments of severity are made (for example when priority levels are defined a priori, with no relationship to specific goals, users and situations);

- appropriate methods for aggregating ratings, for identifying data mining dimensions, for proper data visualization techniques are still missing.

4 An experimental exploration

In order to investigate on some of the outlined open issues, a small scale experiment was performed using a mixed analysis method, where results produced by an automatic tool were also judged by humans. The goal was to determine how valid, useful and efficient could such a method be.

The adopted method is based on an automatic tool (LIFT Machine v. 1.7.1). Selected web sites were crawled (on May 18–19, 2005) by downloading about 50 pages for each of them, starting from the home page and following a breadth–first strategy. The same criteria were used for all the web sites (*e.g.* same timeout, same URL filters, ...).

The evaluation of these pages was based on the same kind of tests, which are a subset of the tests that can be used to assess conformance with respect to WCAG 1.0 level AA. No specific customization was performed on the adopted tests (*i.e.* the default built–in preferences were adopted for all the web site).

Some results are shown in table 1 (see [2] for additional details). The numbers represent the percentages of instances of specific features (like number of link images, or number of navigation bars) that resulted in a violation of a checkpoint (absolute numbers ranged from 26 to 1592). Results are grouped into comprehension, operability or flexibility barriers in order to provide an easy way to estimate the kind of effect that they can have on visitors.

While such a data can be used as a basis for comparing the accessibility status of different web sites, or for trend analysis of single web sites, there are three major shortcomings.

First of all, any diagnostic tool implements a number of tests whose ability to capture violations of accessibility guidelines is less than complete. This limit is fundamental, as several guidelines refer to external properties of the web site, that require human judgment to be assessed. Therefore the validity issue is "how much are those numbers related to actual guideline violations?"

Second, except for the most trivial tests, any tool is bound to produce results that include false positives. This limit is related to the complexity and variability of features in pages (think, for example, at the many ways in which "skip links" can be implemented and the many places in which those links can be employed). Therefore the second validity issue is "how much of those numbers refer to true guideline violations?"

Third, even assuming that the acquired data are valid with respect chosen guidelines, not all violations have the same impact. For example, the vast majority of the images in web pages have only a decorative role. If their ALT text is missing, the consequence in the worst case is that a screen reader user will hear the URL of the SRC attribute being read aloud. But if images contain information, then the consequence wrt screen reader users worsen dramatically. This needs not to be the case for other user categories, though. For example,

Region	Compr.		Operability				Flexibility		
	img	frm	skipl	areas	imglnk	lbl	evnt	pop	unt
1. liguria	100	0	0	0	0	98	100	0	100
2. piemonte	100	0	1	0	1	0	5	7	97
3. lombardia	87	0	0	0	90	78	1	0	91
4. basilicata	64	0	39	31	67	69	95	0	97
5. calabria	54	5	22	40	49	23	2	5	69
6. campania	97	0	0	0	100	96	96	0	98
7. emilia	70	0	0	1	38	3	18	0	95
8. fvg	46	0	7	55	53	12	81	0	80
9. lazio	95	0	0	1	98	1	1	1	96
10. marche	100	94	0	20	100	98	18	0	100
11. molise	97	0	0	0	3	0	0	73	96
12. puglia	96	85	0	3	98	85	85	0	91
13. sardegna	100	0	0	0	52	1	0	0	100
14. sicilia	67	0	6	30	46	6	68	0	82
15. taa	60	0	0	37	47	13	2	0	92
16. toscana	79	0	42	1	37	1	44	0	87
17. umbria	98	0	0	93	91	46	62	0	94
18. vda	100	0	9	96	98	98	27	5	100
19. veneto	98	0	0	0	38	5	3	0	100
20. abruzzo	90	1	32	66	68	32	95	0	90
means it	85	9	8	24	59	38	40	4	93
means de	92	8	18	40	51	66	83	11	91
means at	82	19	3	38	66	67	79	10	92
overall	87	11	10	32	57	54	63	8	92

Table 1: For each Italian Region, the percentage of features that lead to a checkpoint violation is given. Means (*it*, *at*, *de* stand for Italy, Austria and Germany) are also given as a reference value. *Img* represents tests on images that are not decorative nor links/buttons and that do not have appropriate ALT; *frm* represents frames without appropriate TITLE; *skipl* represents navigation bars without hidden links for skipping around them; *areas* represents hot-spots with no proper ALT; *imglnk* represents images used as links or buttons without ALT; *lbl* represents forms with no explicit labeling; *evnt* represents events handlers that cannot be activated by keyboard; *pop* represents the use of *JavaScript* for opening new windows; *unt* represents CSS dimensions specified with absolute units rather than relative ones.

a motor disabled person is not affected by the lack of such alternative text. Context of usage of the web site (defined by the category of the user, his/her goal, the operating situation) has to be considered in order to be able to draw such conclusions on the impact of a failure mode. At the moment it is not clear how to consider such a context within the results produced by tools (it is not even clear how to use such a context in evaluations performed by human evaluators!). And therefore inability to draw such conclusions prevents us to properly rank web sites according not only to the number of guideline violations, but also according to the impact that these violations might have on certain users, under certain circumstances.

To address the first issue (incompleteness) I suggested in 2004 [4] a comparative method between pairs or tuples of testing tools. When used only with so-called automatic tests, LIFT's effectiveness is characterized by less than 7% of false positives (warnings being raised incorrectly) and less than 17% false negatives (issues that are missed by the tool³).

To address the second (incorrectness) and partly the third issue I propose a method based on a stratified sampling of the results produced by tools and a subsequent human judgment of those results in order to determine which is a true problem and which is not. In January 2006 I ran the same tool, with the same configuration, on the same web sites mentioned in figure 1, producing similar data. Then, using a proportional stratified sampling methods on the resulting set of issues (the strata were given by the different tests being deployed), a random sample of issues was selected. These issues were then manually judged and classified as true or false according to whether they are an accessibility barrier or not [3].

As expected⁴ the the proportion of true positive problems over positive test results ranged widely. For example on web site A for test T1 (images w/o alt) the (lower bound of a 95% confidence interval of the) proportion was 72%, whereas for the same test on web site B it dropped to 7%. A similar variation can be seen also when looking at different tests for the same web site: in one case the proportion was 100% whereas for a different test it dropped to 33%. Table 2 shows more details along with statistical significance figures. These figures (*i.e.* the lower bound of the confidence interval around the proportions) are important since the proportion is determined on the basis of a sample of the issues.

The same experiment yielded some information about trend analysis (how the status of a web site changes over time). I compared the number of violations detected by test T_i over web site W_j obtained in May 2005 and in January 2006 and some of the results are shown in table 3.

In many cases, although the two numbers were different, their difference was statistically insignificant (at $\alpha = 5\%$). Statistical significance here is important if we consider that for each web site only a subset of the pages were tested (albeit not through a random sample). If we want to draw some conclusions

³With respect to the checkpoint the test refers to, and with respect to what other tools were able to identify as true problems for the same checkpoint; see [4] for more details.

⁴Since no site-oriented customization was performed on the tool.

Site	Compr.		Operability				Flexibility		
	img	frm	skipl	areas	imglnk	lbl	evnt	pop	unt
fvg	1 (.72)	1 (.75)	?	1 (.63)	1 (.25)	1 (.82)	1 (.37)	1 (.92)	1 (.75)
ktn	.54 (.36)	?	1 (.46)	?	1 (.37)	1 (.37)	1 (.46)	1 (.14)	1 (.95)
lo	1 (.37)	?	0 (0)	?	?	1 (.84)	?	?	1 (.87)
mw	.33 (.07)	?	0 (0)	?	?	1 (.14)	1 (.63)	0 (0)	0 (0)

Table 2: For selected web sites, the numbers show the proportion of identified issues that were judged as true defects and in parenthesis the lower bound of the 95% confidence interval on the proportion. Tests were the same as those used in table 1; fvg=www.regione.fvg.it, ktn=www.ktn.gv.at, lo=www.land-oberoesterreich.gv.at, mw=www.magwien.gv.at.

Site	Compr.		Operability				Flexibility		
	img	frm	skipl	areas	imglnk	lbl	evnt	pop	unt
fvg05	.54 (302)	.64 (348)	.47 (257)	.07 (221)	.56 (397)	.13 (236)	.82 (808)	0 (213)	.81 (278)
fvg06	.58 (297)	.62 (318)	.47 (235)	.11 (216)	.65 (447)	.16 (230)	.84 (842)	0 (201)	.79 (246)
p	.42	.67	.98	.22	.01	.36	.24	1	.79
ktn05	.86 (111)	.10 (59)	.47 (55)	.04 (56)	.79 (147)	.16 (61)	.79 (205)	0 (55)	.51 (83)
ktn06	1 (540)	0 (51)	.99 (70)	0 (51)	.99 (67)	.98 (51)	.71 (120)	.06 (51)	1 (1473)
p	0	.05	0	.52	0	.27	0	.13	.22
lo05	.08 (53)	0 (51)	.98 (217)	0 (51)	.02 (51)	.96 (89)	.38 (80)	0 (51)	.98 (100)
lo06	.18 (56)	0 (51)	.99 (223)	0 (51)	0 (51)	.96 (83)	0 (51)	0 (51)	.99 (101)
p	.19	1	.97	1	1	1	0	1	.99
mw05	.69 (64)	.72 (58)	0 (51)	0 (51)	0 (51)	.38 (80)	.02 (51)	0 (51)	.98 (58)
mw06	.63 (62)	.99 (139)	0 (51)	0 (51)	0 (51)	.12 (57)	.49 (96)	.02 (51)	.98 (120)
p	.61	1	0	1	1	0	0	1	1

Table 3: For selected web sites, at two time points (2005 and 2006), the numbers show the proportion of features that were labeled as a failure, and in parenthesis the total number of features. For example, in 2005 on fvg there were 297 features tested by the *img* test, out of which 58% were labeled as guideline violation. *p* gives the p-value that the two proportions are the same (*e.g.* p=.42 for fvg05 and fvg06 means that with probability 42% the two proportions differ only by chance). Numbers in boldface are the only statistically significant ones.

about the accessibility trends of the entire web site, then only statistically significant differences should be considered. On the other hand, to compare only the specific sample of web pages that were downloaded and analyzed, then the p-values are not relevant.

5 Conclusions

This explorative study demonstrates that it is not easy to compare data produced by testing tools in a valid way. In many cases, conclusions drawn on shallow analysis of data are too naive and wrong. Yet, a viable model of which measures to compute and how to use them to characterize the validity of the results produced by tools is missing.

References

- [1] Bernardini A. Rapporto annuale sull'accessibilità dei siti web italiani. <http://www.webxtutti.it/documenti/Report%20Annuale%20accessibilita'%202003.ppt>, 2003.
- [2] G. Brajnik. Engineering accessibility through corporate policies. In *Congresso Annuale AICA 2005, Comunità Virtuale dalla Ricerca all'Impresa, dalla Formazione al Cittadino*, Udine, Italy, Sept. 2005.
- [3] G. Brajnik. Web accessibility testing: When the method is the culprit. In Prof. Miesenberger, editor, *ICCHP 2006, 10th International Conference on Computers Helping People with Special Needs*, Lecture Notes in Computer Science, Linz, Austria, July 2006. Springer Verlag. to appear.
- [4] Giorgio Brajnik. Comparing accessibility evaluation tools: a method for tool effectiveness. *Universal Access in the Information Society*, 3(3-4):252–263, Oct 2004. www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s10209-004-0105-y.
- [5] DRC. Formal investigation report: web accessibility. Disability Rights Commission, www.drc-gb.org/publicationsandreports/report.asp, April 2004. Visited Jan. 2006.
- [6] Formez. Guida utile metodologia arpa-l — valutare la qualità dei siti web della p.a. <http://db.formez.it/guideutili.nsf/Arpa>, 2002.
- [7] 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.
- [8] EIAO Working Group. European internet accessibility observatory. <http://www.eiao.net/>. Visited June 2006.

- [9] S.L. Henry and M. Grossnickle. *Accessibility in the User-Centered Design Process*. Georgia Tech Research Corporation, Atlanta, Georgia, USA, 2004. On-line book, www.UIAccess.com/AccessUCD.
- [10] Jakob Nielsen. *Usability engineering*. Academic Press, Boston, MA, 1993.
- [11] J. Preece, Y. Rogers, and H. Sharp. *Interaction design*. John Wiley and Sons, 2002.
- [12] UIC. Unione Italiana dei Ciechi, Osservatorio Siti Internet. www.uiciechi.it/osi/Lavori/index.asp, 2005.