

USEFul: A Framework to Mainstream Web Site Usability Through Automated Evaluation

Alexiei Dingli

*Intelligent Computer Systems
University of Malta
Msida, MSD 2080, Malta*

alexiei.dingli@um.edu.mt

Justin Mifsud

*Computing & Information Systems
Goldsmiths University of London
London, SE14 6NW, England, UK*

justinmifsud@gmail.com

Abstract

A paradox has been observed whereby web site usability is proven to be an essential element in a web site, yet at the same time there exist an abundance of web pages with poor usability. This discrepancy is the result of limitations that are currently preventing web developers in the commercial sector from producing usable web sites. In this paper we propose a framework whose objective is to alleviate this problem by automating certain aspects of the usability evaluation process. Mainstreaming comes as a result of automation, therefore enabling a non-expert in the field of usability to conduct the evaluation. This results in reducing the costs associated with such evaluation. Additionally, the framework allows the flexibility of adding, modifying or deleting guidelines without altering the code that references them since the guidelines and the code are two separate components. A comparison of the evaluation results carried out using the framework against published evaluations of web sites carried out by web site usability professionals reveals that the framework is able to automatically identify the majority of usability violations. Due to the consistency with which it evaluates, it identified additional guideline-related violations that were not identified by the human evaluators.

Keywords: Usability, Automated Usability Evaluation, Usability Guidelines, Usability Problems.

1. INTRODUCTION

The International Standards Organization's ISO9241 standard, defines usability as the "effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments" [1].

Whilst there exists a general agreement about the importance of web site usability, especially within the technical communication professional and the academic communities [2, 3, 4], it is given less priority in the commercial sector [3]. In fact, even in its early years, it was noted that on average, the web sites on the World Wide Web were of a poor quality [5, 6, 7, 8].

Problems related to poor usability and accessibility in software and web sites, also prompted some countries to also have their own guidelines and legislation for usability and accessibility of web sites [9, 10]. Additionally, there is less accord about what constitutes usability [2] particularly because some argue that usability is perceived in different ways by different users based on their characteristics such as age, gender, education level, technology skills and culture [11, 12].

Thus, it can be observed that a paradox exists whereby web site usability is proven to be an essential element in a web site, the absence of which confuses users and results in loss of revenue [13, 14] and at the same time it is not commonly applied with the commercial sector. The main question that this study aims to address is "How can web site usability be automated and as

a result mainstreamed?"- meaning how can an automated tool be developed that can make it easier and more possible for more web designers and developers to produce usable web sites.

2. BACKGROUND

2.1 Current limitations that Prevent Web Site Usability from Going Mainstream

Usability Evaluation (UE) is the process of measuring usability and recognizing explicit usability problems [15]. Its main goal is to identify the main issues in the user interface that may lead to human error, terminate the user interaction with the system and cause user frustration [16].

Although there exist a number of widely accepted usability evaluation techniques such as Heuristic Evaluation [17], Cognitive Walkthrough [18], Think aloud testing and Query techniques [15, 19], the development of usable web sites is not common because of the following limitations:

- Usability evaluation requires the engagement of experts to conduct it, and there is a shortage of such experts [20, 21, 13, 22].
- The process of conducting usability evaluation is expensive and some companies do not have the finance to afford it [23, 24].
- Conducting usability evaluation and improvement of web sites is becoming increasingly difficult because of the number of web sites being developed, their size and the regularity at which they are updated [25].
- Time is an issue since the web site life cycle is fast due to market pressure and absence of distribution barriers [5]. So as to meet such demanding deadlines, evaluation many be overlooked, thus resulting in less usable web sites.
- Tobar et al. [24] state that all forms of quality measurement of a web site such as usability evaluation can only be carried out up to a limited depth.
- Studies also show inconsistencies in the reported usability violations when the same web sites were evaluated by different usability experts [26, 20, 27, 28].

2.2 How Automation Helps in Mainstreaming Web Site Usability

In this study, automation is being chosen as the primary method to mainstream web site usability. This is because our proposed framework is based on research carried out by Beirekdar et al. [20], Ivory and Hearst [29] and Brajnik [5] who identify automated usability evaluation as a viable approach that can overcome the limitations of its manual counterpart. Since most of the advantages of automating web site usability that they propose actually overcome the current limitations, outlined in Section 2.1 of this document, then automation has been chosen as the technique to mainstream web site usability. This is because they state that automated web site evaluation:

- **Reduces the costs of usability evaluation:** Through automation, the evaluation can be done more quickly and hence more cheaply.
- **Reduces or eliminates the need for usability experts to carry out the usability evaluation:** The use of such a tool will be of assistance to designers and developers who do not have such expert skills in web site usability.
- **Overcomes inconsistency in the usability problems that are identified:** By removing the human element, automation removes the inconsistencies in the usability problems that are detected as well as any misinterpretations and wrong application of usability guidelines.
- **Enables the prediction of the time and costs of errors across a whole design:** Since automated evaluation tools perform usability evaluation methodically, they are more consistent and cover a wider area in their evaluation and thus, one can better predict the time and cost required to repair usability errors that are identified

- **Increases the coverage of the usability aspects that are evaluated:** Automation overcomes commercial constraints such as those associated with time, cost and resources, which typically limit the depth of evaluation
- **Enables the evaluation between different potential designs:** Commercial constraints limit the evaluation against one design or a group of features. Automated evaluation software provides designers with an environment where alternative designs can be evaluated.
- **Facilitates the evaluation in various stages of the design process:** An interface can be evaluated and any usability issues identified and resolved early, thus saving time and costs that would be incurred should it be addressed at a later stage.
- **Is of immediate value in the web design and development domain:** Brinck and Hofer [25] state that due to the large number of web designers and developers, a tool that enables the evaluation of a web site is something that will appeal to this large community.

2.3 Attempts at automating web site usability evaluation

Chi et al [6] state that there are two types of tools that can perform automated usability evaluation. These categories refer to tools that:

- Make use of **conformance to standards**
- Try to **predict** the usage of a web site

Through the research carried out for this study, four tools have been identified that perform the usability of a web site, all of which fall in the second category. These are Cognitive Walkthrough of the Web - CWW [30] Web Tango [31], WebCriteria Site Profile [5] and Bloodhound [6]. All three solutions base their usability evaluation through usage prediction - something which various researchers such as Groves [32], Winckler et al. [33] and Murray and Costanzo [34] argue against since this method is based on prediction algorithms that can provide misleading data.

3. MAINSTREAMING WEB SITE USABILITY THROUGH USEFUL

3.1 The Components of the USEFUL Framework

The framework that is being proposed in this study has been named USEFUL (USability Evaluation Framework). Unlike the previous attempts at automated usability evaluation, USEFUL falls in the first category proposed by Chi et al [6].

This is because it is based on research conducted by Jeffries et al. [21], Otaiza et al. [19] and Tobar et al. [24], who identify Heuristic Evaluation, that is, the evaluation of the interface with respect to a set of usability principles [35] as the usability evaluation technique that manages to detect the majority of global usability issues, from all the usability evaluation techniques that they evaluated. Similarly, studies by Comber [7], Ivory and Hearst [29] and Tobar et al. [24], show that adherence to guidelines can effectively contribute towards making a web site more usable.

Thus, the USEFUL framework will reference web site usability guidelines and use them to automatically assess the usability of a web site that is being evaluated.

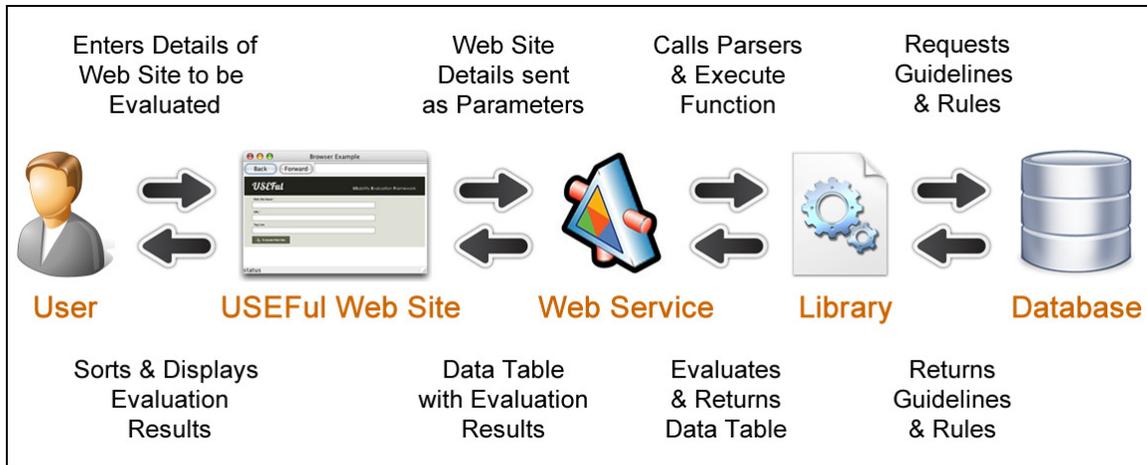


FIGURE 1: A visual representation of the USEful framework (Source: Authors)

The different components of the USEful framework as illustrated in Figure 1 are the:

- **User** - the person who is conducting the usability evaluation of a web site
- **Web site** - the means by which the user can interact with the USEful framework. To specify which web site needs to be evaluated, the user needs to key in the web site's name, its tag line and URL.
- **Web service** - The web service communicates with the library by calling the execute function from the library and passes it the parameters that it needs to evaluate the web site. Once the evaluation is complete, the web service passes the results of the evaluation back to the web site.
- **Library** - The library contains the program that carries out the evaluation. In order to carry out the evaluation, it retrieves the data from the database.
- **Database** - The database is an SQL relational database that contains 4 tables
 - **Usability Category table** - stores the usability categories available (Section 3.1)
 - **Implementation Level table** - stores the implementation levels available (Section 3.2)
 - **Guidelines Definitions table** - stores the guidelines that will be used in the framework, expressed in natural language. This table also stores the Priority Rating (Section 3.2) of each guideline and references the usability and implementation level tables.
 - **Rule Type 1 table** - In this table, the guidelines from the guidelines definitions table that have green or amber implementation level (Section 3.2) are expressed in a form that the library can interpret to carry out the evaluation. For each record, the fields contain the HTML tag along with its additional data such as its attribute and size that the library needs to search for so as to find the pattern that identifies a specific guideline. The rule type 1 table also allows the comparison or searching of two HTML tags or tags within tags. An important column in this table is the "ruleSuccess" column as the fields in it store a value that the library interprets as the conditions under which the guideline is considered to have been violated or not. This is important since it distinguishes between guidelines that must be adhered to and thus must be present in the web site and those that must never be found or can only be found once as otherwise they would cause a usability violation.

At this point, it is worth mentioning that the current build of the USEful framework contains just 1 rule type (Rule type 1 table), that enables the library to identify the guidelines that relate to HTML tags or CSS selectors. It is envisaged that future builds will

enable the addition of new rule types such as ones that enable the library to evaluate usability guidelines related to images and other resources used by the web site. Also, should new structuring tools such as a JavaScript parser be incorporated into the framework, then these would require new rule types to be incorporated.

3.2 The Set of Guidelines That Will be Used by the USEful Framework

Over the years, a number of usability guidelines have been published such as those by Smith and Mosier [36], Norman [37], Nielsen [38], Comber [7], Sano [39], Borges et al. [27], Spool et al. [40], Fleming [41], Rosenfeld and Morville [42], Shneiderman [43], Nielsen [44], Dix et. al. [15] and Nielsen and Loranger [45].

However, the problem with Usability guidelines is that there is no set of guidelines that has been established as a standard [15]. Thus, a set of 240 guidelines has been compiled for this study from the results of usability studies carried out by researchers and experts in the fields of cognitive psychology, technical communication, computer science, human factors and usability.

Since most of the proposed guidelines have been retrieved from the U.S Department of Health and Human Services' (HHS) Research-Based Web Design & Usability Guidelines [46], the same categorization has been used, that is, each guideline was placed in 1 of the 15 categories shown in Table 1

Usability Category	Number of Guidelines
Optimizing the user experience	29
Hardware and software	4
The homepage	12
Page layout	9
Navigation	27
Scrolling and paging	3
Headlines, titles and labels	18
Links	21
Text appearance	18
Lists	13
Screen based controls (widgets)	27
Graphics, images and multimedia	17
Writing web content	18
Content organization	8
Search	16
Total Guidelines	240

TABLE 1: The number of guidelines used by the USEful framework in each category (Source: Authors)

3.3 Assigning the Priority Rating and Level of Implementation to each Guideline

Each guideline used in the USEful framework has been assigned an **Implementation Level** which denotes the ability (or otherwise) to translate that guideline into a form which can then be referenced by the program. This gives an indication as to what automation level each guideline has. The parameters that will be used for this classification are as shown in Table 2:

Implementation Level Category	Interpretation
Green	<ul style="list-style-type: none"> • Guideline can be fully implemented in the database within the USEFUL framework. • The framework is able to automatically determine whether this guideline applies to the web site being evaluated. • The results returned by the framework when referring to this guideline are conclusive since these types of guidelines are typically measurable, with clearly defined parameters.
Amber	<ul style="list-style-type: none"> • Guideline is harder to fully implement in the USEFUL framework. • Certain patterns that automatically identify if this guideline may apply to the web site being evaluated have been implemented in the database. • This guideline can be converted into a "green" guideline by incorporating within the USEFUL framework additional Artificial Intelligence algorithms. • The results outputted by the framework when referring to this guideline consist of data that can assist the human evaluator in checking whether it applies to the web site being evaluated
Red	<ul style="list-style-type: none"> • This guideline is typically abstract and requires user intervention or very advanced algorithms from the field of Artificial Intelligence or additional technology to make it possible for it to be implemented in the framework. • Through the use of advanced algorithms or technology, it can be converted into "amber" or "green" guideline • In its current build, the framework lists this guideline so that the human evaluator can manually check if it applies to the web site being evaluated

TABLE 2: How the guidelines should be interpreted in terms of their Implementation Level (Source: Authors)

Since the resources to tackle usability violations are typically scarce, the evaluator carrying out manual usability evaluation prioritizes them so that the violations that cause the highest problems are addressed first [47, 45].

One of the most used prioritization techniques is the severity scale, whereby each guideline is given a severity rating [48]. For this project a severity scale called **Priority Rating (PR)** is proposed, whereby each guideline is assigned a PR from 1 to 5 where a guideline with PR 5 is very important in terms of its contribution towards making a web site usable, whilst a guideline with PR 1 provides minor contribution. This prioritizes the list of usability violations identified by the program.

3.4 Incorporating the Guidelines into the SQL Database

Usability guidelines are occasionally abstract and difficult to interpret and apply [26, 20, 27, 28, 49]. This has been addressed in the USEFUL framework through the use of the guidelines definitions and rule type 1 tables (Section 3.1). The process through which a guideline is entered into these 2 tables is illustrated below through the use of one of guideline#81 which is one of the guidelines used in the USEFUL framework:

3.4.1 Guideline Expressed in Natural language

Guideline#81: URLs should not be complex and should ideally be less than 50 characters. This is beneficial for both usability and SEO [45]

3.4.2 Guideline as Entered in the Guidelines Definitions Table

When the guideline is entered in the guidelines definition table, its primary key is 81. The guideline and its explanation have been split into the fields under the "Guideline" and "Reason" columns respectively. The value in the field "ruleType" is 1 since this guideline needs to be evaluated using the rule type 1 rule. The guideline is a green guideline, hence the value 1 under the "ruleCat" column and its Priority Rating is 5, hence the reason why there is a "5" under the

ruleSeverity column. Since this guideline belongs to the "navigation" usability category, the value "5" has been entered in the field under the ruleGroup column.

pk	Guideline	Reason	ruleType	ruleCat	rule Severity	rule Group
81	URLs should not be complex	URLs should ideally be less than 50 characters. Such URLs are beneficial for both usability and SEO	1	1	5	5

TABLE 3: How guideline#81 is represented in the guidelines definition table (Source: Authors)

3.4.1 Guideline as Entered in the Rules Type 1 Table

To check whether the guideline is found in the parsed HTML, the execute function needs to search for the following pattern:

```
<a href="any text as long as it is less than 70 characters"> .. </a>
```

Therefore, the guideline is converted to a form that the execute function can understand and this is stored as a record in the rule type 1 table (Section 3.1) as shown below:

pk	ruleFk	tagA	attributeA	valueA	sizeA	tagB
12	81	a	href	NULL	70	NULL

attributeB	valueB	sizeB	rule Command	compare Operator	rule Success	must Succeed
NULL	NULL	NULL	NULL	<	True	0

TABLE 4: How guideline#81 is represented in the rules type 1 table (Source: Authors)

Thus, the value "81" is a foreign key that references the primary key "81" in the Guidelines Definition table. The execute function will look for is the "a" tags which have an "href" attribute as stated by the contents in the fields under the "tagA" and "attributeA" columns respectively. The exact content in between the inverted commas of the "href" attribute is irrelevant, hence the reason for the *NULL* value in the field under the "valueA" column.

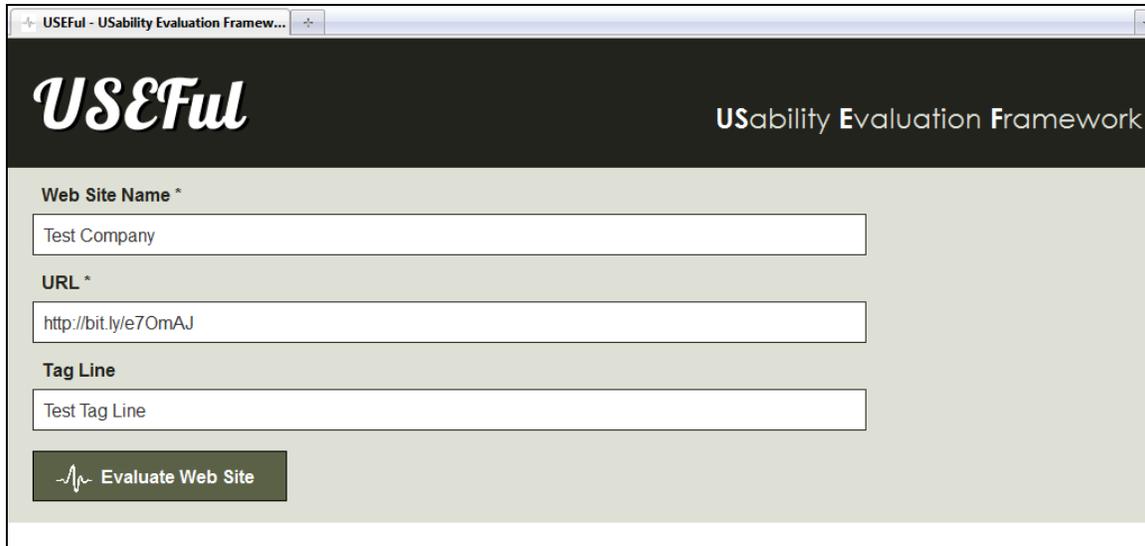
However, for the guideline not to be violated, this content needs to be less than 70 characters long, as stated by the contents in the fields under the "compareOperator" and "sizeA" columns respectively. Since the guideline is not dependent on any other HTML tags, the fields under the four tag B columns are all set to *NULL*. If the guideline matching the pattern in this record is found, then it is a good thing, hence the reason why the value under the "ruleSuccess" column is True. Also, for the guideline to succeed, all the content of all "href" attributes within all the "a" tags found must be less than 70 characters. This is set by the "0" value in the field under the "mustSucceed" column.

4. HOW THE PROPOSED FRAMEWORK EVALUATES WEBSITE USABILITY

This section will describe the process that takes place from when the user accesses the USEFUL web site and keys in the parameters pertaining to the web site they would like to evaluate to when the results of the usability evaluation are reported on the web site. This description will thus discuss on a high level the interactions that take place between the various components within the framework represented in Figure 1.

4.1 Step 1 - The user passes the data to the web site

This step refers to when the user states what web site the USEful framework will need to evaluate. They do this by filling in the text fields pertaining to the web site's name, URL and (optionally) the tag line. The website's GUI can be seen in the screenshot below (Figure 2)



The screenshot shows a web browser window titled "USEful - USability Evaluation Framew...". The page has a dark header with the "USEful" logo on the left and "USability Evaluation Framework" on the right. Below the header, there are three text input fields: "Web Site Name *" containing "Test Company", "URL *" containing "http://bit.ly/e7OmAJ", and "Tag Line" containing "Test Tag Line". At the bottom of the form is a dark button with a pulse icon and the text "Evaluate Web Site".

FIGURE 2: Screenshot of the USEful web site (Source: Authors)

4.2 Step 2 - The Web Site Passes the Data to the Web Service

When the user fills in the data in Step 1 and presses the "Evaluate Web Site" button, the web site passes the data as parameters to the web service.

4.3 Step 3 - The Web Service Configures the Library

When the web service receives the parameters from the web site, it communicates with the library and creates a new instance by setting the configuration values in the library according to these parameters. It is important to note that in reality, the library is actually contained within the web service. The only reason why the library was illustrated as a component outside the web service in Figure 1 is to create a distinction between the two components for explanation purposes. Therefore, the phrase "communicates" is being used to illustrate the flow of data between the library and the web service.

Thus, the web service sets the path of the web site, the company name and the tag line. The web service then uses the library's functionalities to load and parse the web site to create parsed HTML and CSS documents. These parsed documents are stored in the web service in static variables.

4.4 Step 4 - The Web Service Fetches the Guidelines

The web service communicates with the library which in turn communicates with the SQL database to fetch the data stored in the guidelines and rule type tables. The returned data is stored inside the web service in a **data table** as shown in Table 5 below:

1 data row \Rightarrow

Guideline Definition	Rule Type Properties	Results Fields

TABLE 5: The Structure of the Data Table stored in the Web Service (Source: Authors)

The components of the data table shown in Table 5 are the following:

- **Guideline definition:** This contains a copy of the data found in the guidelines definitions table (Section 3.1)
- **Rule type properties:** This field contains a copy of the rules type 1 table (Section 3.1)
- **Results fields:** The results fields are additional fields initially set as empty when the data table is created in the web service. These fields will eventually contain the results that the web service will communicate back to the web site after the evaluation is completed. These are:
 - Tags: will contain the number of times the HTML tag or CSS selector found in the field under the rule type properties column in the data row is found.
 - Success: The number of tags or selectors found whose properties match the properties of the guideline being referenced
 - Fail: The number of tags or selectors whose attributes match the properties of the guideline being referenced but their value or size properties do not match
 - Null: The number of tags or selectors whose attributes, sizes or properties do not match with the property of the guideline
 - Success%: This value is the result of the equation $\text{Success}/(\text{Tags}-\text{Null}) \times 100$
 - Passed: This field will eventually contain a True/False value that will indicate whether the guideline has been violated or not

4.5 Step 5 - The Web Service Uses the Library to Evaluate the Web Site

At the end of Step 4, the web service contains a copy of the parsed HTML and CSS documents which are stored in static variables. It also has a data table as shown in Table 5.

The web service then takes the first data row and calls the execute function from the library. The execute function takes 1 data row (1 row of the data table as indicated in Table 5) as a parameter. When the execute function in the library receives the data row from the web service, it sees what rule it has to work on in order to evaluate whether the usability guideline defined in that data row is being violated.

We will use guideline#81 as an example to illustrate the steps the library performs to assess whether the web site being evaluated adheres to a guideline or if it violates it. It is being assumed that guideline#81 has been entered in both the guidelines definition table and the rules table as indicated in Tables 3 and 4 respectively.

When the web service passes calls the execute function with the data row pertaining to guideline#81, each time the execute function is run, and it finds an "a" tag it follows the logic tree shown in Figure 3 below. As it can be seen in Figure 3, the execute function can take different paths, depending on what the fields contain and the type of pattern matching that it needs to perform.

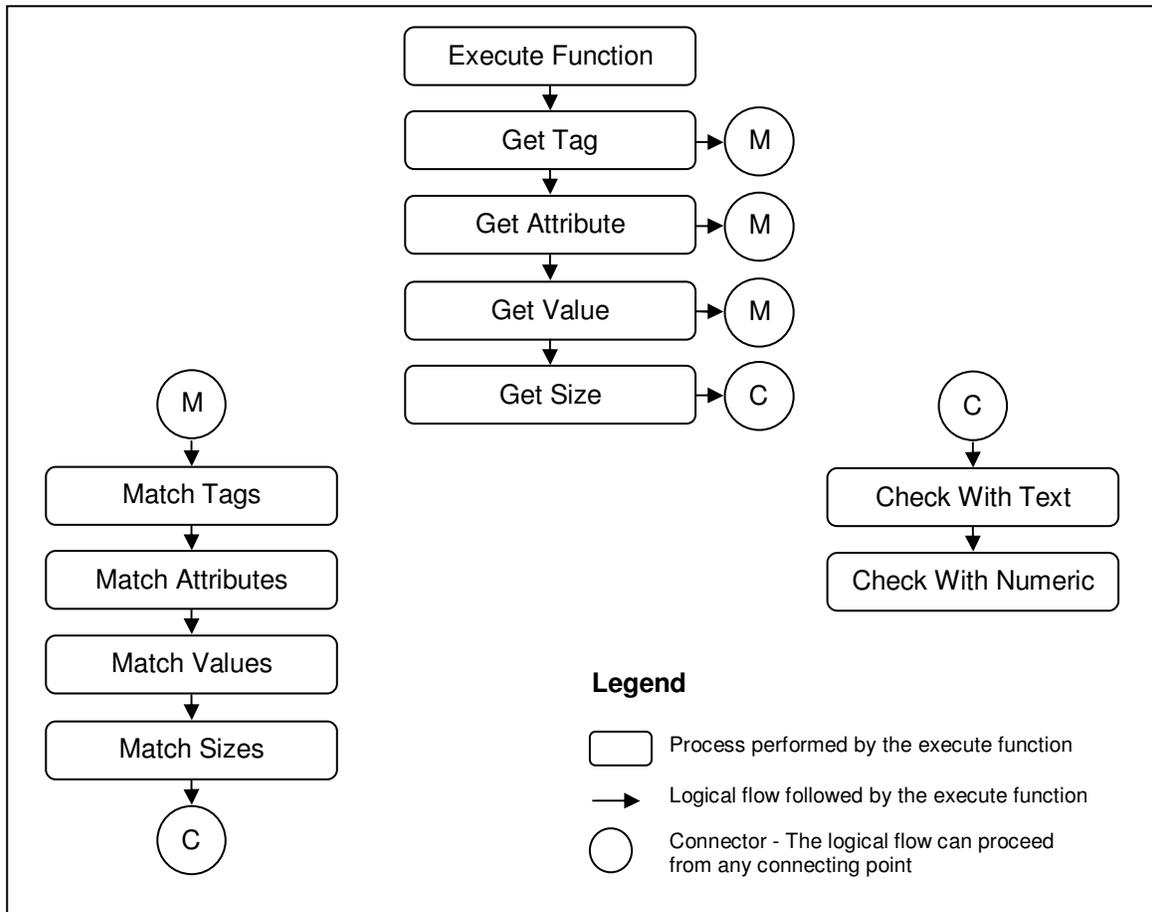


FIGURE 3: The Logic Tree for Rule Type 1 (Source: Authors)

At the end of this process, the execute function returns the evaluation result which can have 1 of 3 values: True, False or Null. In the case of guideline#81, these are interpreted as follows:

- True: The "a" tag has an "href" attribute whose content is less than 70 characters
- False: The "a" tag has an "href" attribute whose content is 70 characters or more
- Null: The "a" tag does not have an "href" attribute

Suppose that the execute function finds 3 instances of the "a" tag in the parsed HTML document and these are as follows:

- True: 2
- False: 1
- Null: 0

For each evaluation result, the execute rule compares it with the value of the field under the "ruleSuccess" column in Table 4. During the same process, it uses the result of the comparison to increment the counters of the values that will be written in the results fields of the data table in the web service. The method of comparison is modeled on the XNOR truth table and can be seen in Table 6 below:

Input A (Evaluation Result)	Input B (ruleSuccess)	Output (A XNOR B)
False: Guideline not found	False: Bad guideline	True: Increment success counter
False: Guideline not found	True: Good guideline	False: Increment fail counter
True: Guideline found	False: Bas guideline	False: Increment fail counter
True: Guideline found	True: Good guideline	True: Increment success counter
Null: Guideline not applicable	True: Good guideline	Null: Increment null counter
Null: Guideline not applicable	False: Bad guideline	Null: Increment null counter

TABLE 6: How guideline#81 is represented in the guidelines definition table (Source: Authors)

Thus, assuming that in the case of the guideline#81 example, the sequence in which the evaluation results are issued by the execute function are 2 True, 1 False and 1 null, then the execute rule would make the following comparisons:

Input A (Evaluation Result)	Input B (ruleSuccess)	Output (A XNOR B)
True	True	True: Increment success counter
True	True	True: Increment success counter
False	True	False: Increment fail counter

TABLE 7: The comparisons made by the execute function for the guideline#81 example (Source: Authors)

In this way, the values for the results fields of the data row for guideline#81 in the web service (Table 5) would be as shown in Table 8 below:

Results Field	Value
Tags	3
Success	2
Fail	1
Null	0
Success%	66.7
Passed	<i>NULL</i>

TABLE 8: The values in fields of the data row for guideline#81 (Source: Authors)

As can be seen in the table above, at this stage, it has not been determined whether the guideline has passed or not. In fact, to set this value, the execute function looks up the value of the field under "mustSucceed" in Table 4 to see under which conditions it can be stated that guideline#81 has not been violated.

Since the value in this case is "0", the execute rule interprets it that for the guideline not to be violated, this guideline must not fail, meaning that the fail counter in Table 8 should be "0". Since this is not the case, the execute rule sets the value of the "Passed" field in Table 8 to "FALSE". This effectively means that the guideline has been violated.

This value is then stored in the passed field of the data row for guideline#81 in the web service.

4.6 Step 6 - The Web Service Sends the Data Table to the Web Site

Once the data table is complete, the web service sends it to the web site which creates 3 data views, one for each implementation level. It also sorts the violations in each implementation in descending order of priority rating and displays them as shown in Figure 4.

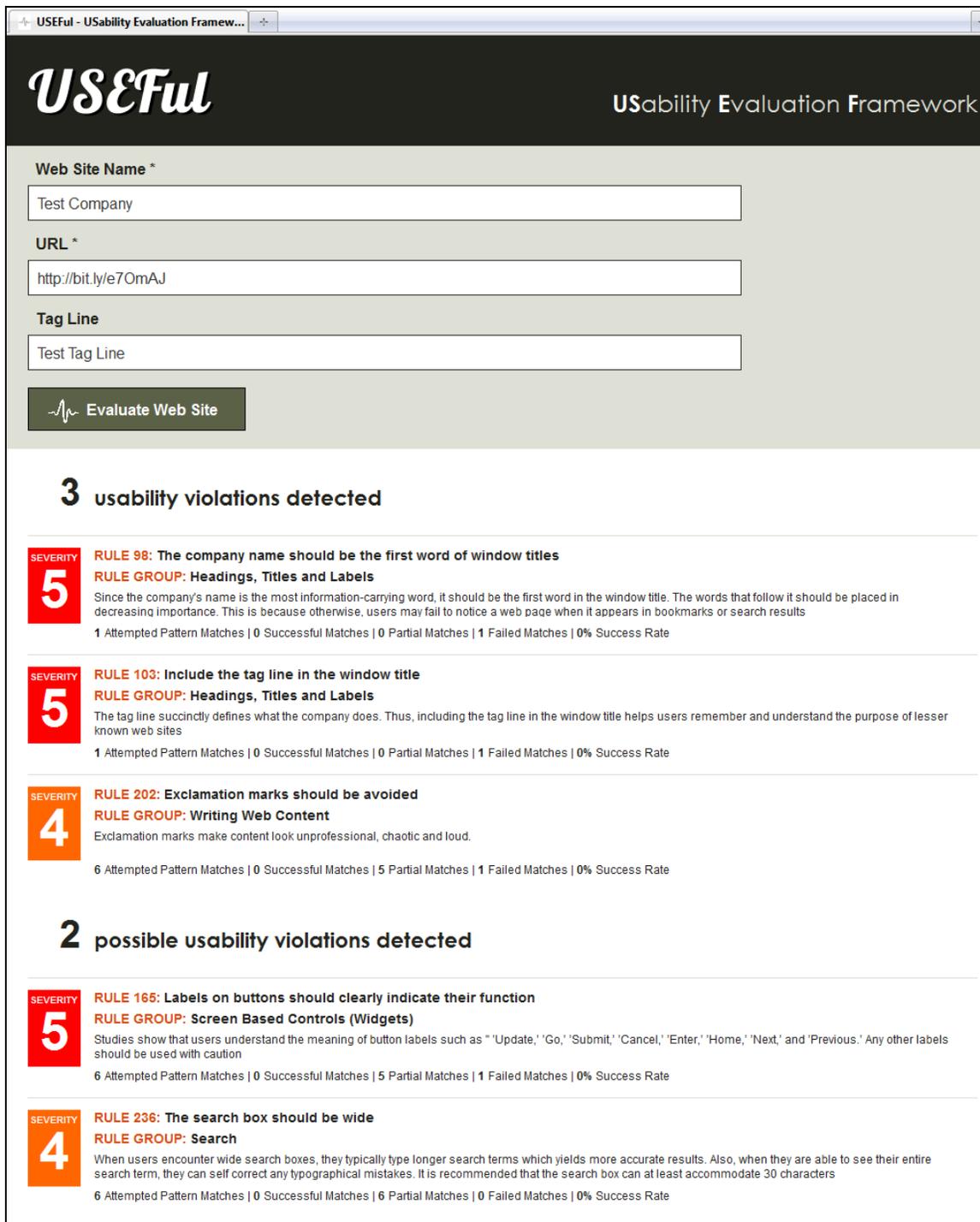


FIGURE 4: Screenshot showing evaluation results carried out by the USEful framework (Source: Authors)

5. EXPERIMENTS AND RESULTS

Since web site usability professionals are scarce (a current limitation mentioned in Section 2.1), it was decided that the effectiveness of the USEful framework will be assessed by comparing the results of the evaluations carried out on web sites against published evaluations of the same web sites carried out by web site usability professionals.

The most reliable source for this evaluation was identified to be the book "Homepage Usability - 50 Websites Deconstructed" by Nielsen and Tahir [50]. The reasoning behind this chosen method of experimentation is based on the following points:

- Dr. Jakob Nielsen is considered to be a web usability guru [51, 52] and has been hailed as "one of the world's foremost experts in web usability" [53].
- The book itself illustrates in a very clear manner the usability violations that have been identified by Nielsen and has received numerous positive reviews [54, 55].
- Nielsen evaluates the web sites featured in this book by referencing web site usability guidelines. This usability evaluation technique is the same technique used by the USEful framework. This eliminates any possibilities that any difference in the list of identified violations is as a result of different techniques being employed.
- The set of guidelines used by Nielsen for this evaluation is a subset of the HHS Research-Based Web Design & Usability Guidelines [46]. In fact, this book is listed as one of the cited sources. As stated in Section 3.2, the majority of the guidelines implemented in the USEful framework are from the HHS guidelines.
- Although the HHS Research-Based Web Design & Usability Guidelines have been retrieved in February 2011 and thus it can be assumed that they are still relevant today, any flaws in these guidelines does not affect the performance of the USEful framework since the guidelines are not hard coded into the library itself. Moreover, using the same set of guidelines as those used by the human evaluator for these tests eliminates the possibility that any discrepancies in the results were due to different sets of guidelines being used.

On inspection of the results reported by evaluation carried by Nielsen and Tahir it was noticed that they also mention some positive usability characteristics. The guidelines that have been observed which have led to these positive traits have also been incorporated in this experiment. Since the USEful framework only reports usability violations, the absence of these guidelines in the list of detected violations was thus interpreted as a positive result. In this regard, since what Nielsen and Tahir reported were both positive as well as negative comments, the term "**usability aspects**" will be used instead of usability violations so as to avoid the negative connotation associated with the word "violation".

Due to their expertise in web site usability, Nielsen and Tahir also list a number of positive as well as negative usability aspects which are specific to the web site being evaluated and their linkage to any of the guidelines could not be established. These **site-specific** aspects were incorporated in this study with the red guidelines since they could not be automatically evaluated by the USEful framework in its present form.

Ten web sites from the book were selected on the basis that their evaluation contained less site-specific recommendations and more green and amber guideline related usability aspects. This means that from the authors' evaluation, it was easier to identify which guidelines from the HHS Research-Based Web Design & Usability Guidelines were being violated. Special attention was taken to select web sites that violated different guidelines so as to increase the set of guidelines that will be incorporated into the database for evaluation.

So as to ensure that the versions and contents of the web sites evaluated are identical to the ones evaluated by Nielsen and Tahir, the Internet Archive's Way Back Machine [56] was used. Using this tool, the exact web sites were loaded by utilizing the dates present on the screenshots in the book. Based on these criteria, the 10 web sites chosen for this experiment are those found in Table 9.

Web Site	Tag Line	URL
About	The Human Internet	http://replay.waybackmachine.org/20010611062521/http://www.about.com/
Accenture	Now It Gets Interesting	http://replay.waybackmachine.org/20010711142024/http://www.accenture.com/
Asia Cuisine	Asia's Leading Food and Beverage Portal	http://replay.waybackmachine.org/20010703030929/http://www.asiacuisine.com.sg/
Barnes & Noble	(No Tag Line)	http://replay.waybackmachine.org/200102031220/http://bn.com/
BBC Online	Welcome to the UK's Favourite Website	http://replay.waybackmachine.org/20010806173705/http://www.bbc.co.uk/
Boeing	Forever New Frontiers	http://replay.waybackmachine.org/20010522194801/http://www.boeing.com/
DIRECTV	America's Leader in Digital Home Entertainment	http://replay.waybackmachine.org/20010629015718/http://www.directv.com/
FedEx	(No Tag Line)	http://replay.waybackmachine.org/20010525031739/http://www.fedex.com/us/
Red Herring	The Business of Innovation	http://replay.waybackmachine.org/20010515222459/http://redherring.com/
The Art Institute of Chicago	(No Tag Line)	http://web.archive.org/web/20010630180812/www.artic.edu/aic/index.html

TABLE 9: The web sites that were used for this experiment (Source: Authors)

So as to be able to identify the usability violations identified by Nielsen and Tahir in these web sites 62 guidelines (36 Green, 27 Amber and 0 Red / Site-Specific guidelines) from the set of guidelines mentioned in Section 3.2 were used for this experiment. The results of the evaluations can be seen side by side in Table 10:

Web Site	Usability aspects identified by Nielsen and Tahir				Nielsen and Tahir's usability aspects identified by USEful	
	Green	Amber	Red	Total	Results 1: As a percentage of implementable usability guidelines (Green & Amber)	Results 2: As a percentage of total usability guidelines (Green, Amber, Red & Site Specific)
About	7	7	11	25	100.00%	56.00%
Accenture	6	8	13	27	85.71%	44.44%
Asia Cuisine	7	3	11	21	100.00%	47.62%
Barnes & Noble	11	4	13	28	93.33%	50.00%
BBC Online	17	6	15	38	100.00%	60.53%
Boeing	9	3	12	24	100.00%	50.00%
DirectTV	14	1	12	27	86.67%	48.15%
FedEx	10	8	11	29	100.00%	62.07%
Red Herring	9	3	17	29	91.67%	37.93%
The Art Institute of Chicago	7	5	10	22	100.00%	54.55%
Total	97	48	125	270	Average: 95.86%	Average: 51.48%

TABLE 10: Usability evaluations carried out by Nielsen & Tahir against USEful (Source: Authors)

As can be seen in the figures presented in the column Results 1 of Table 10, the USEful framework was able to correctly identify the guideline-related usability aspects, 95.86% of the time when compared to Nielsen and Tahir's manual evaluation. When the code was inspected to identify why there was a 4.14% discrepancy it was found that the main reason why the framework

failed was due to bad coding in the web sites being tested. In fact, the primary cause was the use of images to represent text instead of using actual text. So as to minimize the impact of this limitation, most of the guidelines' interpretation in the table also referenced the alt attribute of images. When the alt attribute was not present, then the USEFUL framework was not able to parse the text represented by those images since it does not currently have Optical Character Recognition (OCR) facilities. Additionally, this discrepancy can also be attributed to the lack of proper usage of certain HTML tags such as the use of the <p> tag instead of the <h1>..<h6> tags for headings.

Since only 53.71% of the usability aspects were directly related to green and amber guidelines, it can be seen that overall, the number of violations reported by the USEFUL framework on average relates to just 51.48% of the total usability aspects identified by Nielsen and Tahir. This finding shows why various researchers [6, 57, 29, 58] suggest that any tool that automatically evaluates a web site cannot replace a human being. In this case, because of their experience and expertise in web usability, Nielsen and Tahir were able to identify almost as many usability guidelines that have been classified as red guidelines or site-specific recommendations as those that were classified as green or amber. In this regard, it is clear that the USEFUL framework cannot implemented without the inclusion of a human evaluator.

An interesting observation in these experiments is that since the USEFUL framework checks for the presence of each guideline in the database, it performs a consistent evaluation and thus it was able to identify more usability violations in each of the tested web sites. At this point it is important to note that the term being used is usability violations since the USEFUL program can only report usability violations. The HTML and CSS code was then inspected manually so as to confirm that these additional usability violations were correct.

When the additional usability violations identified by the USEFUL framework are compared to the total aspects identified by Nielsen and Tahir (manually), it can be noted that the framework was able to identify on average 128.15% usability violations (Table 11: column Results 3) i.e. 28.15% more violations than what Nielsen and Tahir actually identified. The total number of additional violations identified are shown in Table 10.

Web site	Additional usability violations detected by USEFUL			RESULTS 3: Total usability aspects detected by USEFUL as a percentage of total usability aspects identified by Nielsen & Tahir
	Green	Amber	Total	
About	11	9	20	136.00%
Accenture	13	9	22	125.93%
Asia Cuisine	16	15	31	195.24%
Barnes & Noble	13	12	25	139.29%
BBC Online	7	10	17	105.26%
Boeing	9	8	17	120.83%
DirectTV	9	7	16	107.41%
FedEx	9	9	18	124.14%
Red Herring	14	12	26	127.59%
The Art Institute of Chicago	6	9	15	122.73%
Total	107	100	207	Average: 128.15%

TABLE 11: Additional usability violations identified by the USEFUL framework (Source: Authors)

This means that despite the fact that only 53.71% of the usability attributes identified by Nielsen and Tahir can be converted into green and amber guidelines, the USEFUL framework was still able to detect 28.15% more usability violations using this limited set of 62 guidelines.

Another interesting find in these experiments suggests that through the USEful framework, a usability expert is still likely to identify more usability violations. Table 12, which summarizes the results of all the tests carried out with the 10 web sites, illustrates this point.

Item #		Usability aspects detected by	
		Nielsen & Tahir	USEful
1	Number of web sites evaluated	10	10
2	Green and Amber guideline-related usability aspects detected	145	139
3	Red and Site-Specific usability aspects detected	125	0
4	Additional usability violations detected	0	207
Total		270	346

TABLE 12: Summary of the results from the experiments

From the results shown in Table 11, Nielsen and Tahir were able to comment on an average of 27 usability aspects per web site ($270 \div 10$), whilst USEful was able to identify 34.6 usability aspects (mainly violations) per web site ($346 \div 10$) which translates to 28.15% more usability violations being detected.

If a usability expert were to make use of the USEful framework, then they would be able to detect the green, amber, red and site-specific usability aspects whilst the framework would still report the additional usability violations. This would mean that using the USEful framework to evaluate the 10 web sites above, the expert would have commented on a total of 477 usability aspects ($145 + 125 + 207$), that is, an average of 47.7 usability aspects per web site ($477 \div 10$). This can also be interpreted as an increase of 76.67% in the number of usability aspects that the expert evaluator can make per web site.

6. LIMITATIONS

As recommended by various researchers [6, 57, 29, 58], the purpose of any tool such as the one being proposed is to provide assistance to human evaluators. Ivory and Chevalier [57] advise that such tools should always be used with caution and one should never completely rely on their results alone. This is because with current technology, it is difficult to develop a tool that can behave like a human and exhibit human attributes such as common sense [16, 58]. This is partially addressed in the USEful framework through the assignment of the Implementation Level to denote the possible level of automation for each guideline. Still, it was observed that the USEful framework was not able to handle Nielsen's site-specific recommendations. Such recommendations can be made by a human evaluator through the application of logic, experience and techniques such as grouping guidelines.

Another limitation is the difficulty encountered with incorporating certain guidelines into the rule types tables in the framework's database, particularly because of their abstract nature. At present, the only way to incorporate such guidelines into the USEful framework is to introduce certain assumptions as recommended by Dix et al. [15] and Vanderdonck and Beirekdar [59].

Additionally, the proposed guidelines are aimed at evaluating the usability of web sites that have business goals, meaning that these web sites serve to promote and/or sell products and/or services either online via the web site itself or through offline channels [45, 46]. Thus, web sites that do not fall in this category require a different set of guidelines. The framework facilitates this

process since it reduces it to incorporating the new guidelines into the database tables without modifying the code.

7. CONCLUSION

From the results obtained in this evaluation, it can be concluded that the USEFUL framework is very effective at identifying usability aspects that violate usability guidelines. However, bad coding practices can adversely affect the results obtained. The USEFUL framework references each guideline that has been implemented in its SQL database to see if the web site has violated it. In the experiments that have been carried out, this factor has enabled it to identify more violations than the usability experts. However, the framework cannot detect what have been classified as red or site-specific violations because it lacks the logic, experience and expertise that an expert in web usability has. In this regard it has been concluded that the framework cannot replace a human evaluator but should be used to assist an evaluator. In fact, the results indicate that using the framework, a usability expert is likely to be able to detect more usability violations

Additional research needs to be carried out to make the framework more flexible so as to be able to implement more abstract guidelines that are currently being classified as having a red implementation level. Currently, the framework can parse HTML and inline and internal CSS code. Therefore, further enhancements that are planned include the ability to parse external CSS stylesheets as well as Javascript parsing since these can considerably affect the way the user sees the web site when rendered through a web browser. So as to overcome the problems with analyzing the content of images in web sites, image processing and Optical Character Recognition algorithms can help in addressing this issue. By implementing these enhancements as well as other algorithms that may be deemed as beneficial, the framework can truly contribute towards mainstreaming web site usability.

8. REFERENCES

1. International Organisation for Standardisation (ISO) DIS 9242-11, (1998) Available at: <http://www.iso.org>
2. P. Paolini. "Hypermedia, the web and usability issues" In IEEE International Conference on Multimedia Computing and Systems. pp. 9-11 IEEE, 1999
3. M. Swaak, M. De Jong, P. De Vries. "Effects of Information usefulness, visual attractiveness, and usability on web visitors' trust and behavioural intentions." In IEEE International Professional Communication Conference. P. 1-5. Waikiki, HI, USA, 2009
4. P. Vora. "Designing for the web: a survey". Interactions. pp. 13-30 (1998)
5. G. Brajnik. "Automatic web usability evaluation: what needs to be done?". In Proceedings of the 6th Conference on Human Factors and the Web, Austin Texas, USA, 2000
6. EH. Chi, A. Rosien, G. Supattanasiri, A. Williams, C. Royer, C. Chow, et al. "The Bloodhound project: automating discovery of web usability issues using the InfoScout™ simulator" In Proceedings of the ACM CHI 03 Conf.. pp. 505-512. Ft.Lauderdale, FL, USA, 2003
7. T. Comber. "Building usable web pages: an HCI perspective". In Proceedings of the Australian Conference on the Web AusWeb'95. pp. 119-124. Ballina, Australia, 1995
8. J. Nielsen. "User interface directions for the web" Communications of the ACM. 1999, pp.65-

9. Cabinet Office. (2007) Available at <http://archive.cabinetoffice.gov.uk/e-government/resources/quality-framework.asp> [Accessed 9 October 2010]
10. European Agency for Safety and Health at Work in European Union Council Directives. (1990) Available at: <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1990L0270:20070627:EN:HTML> [Accessed 9 October 2010]
11. SA. Becker, FE. Mottay. "A global perspective on website usability" IEEE Software, Jan/Feb, 18(1): pp.54-61, 2001
12. P. Fraternali, M. Tisi. "Identifying cultural markers for web application design targeted to a multi-cultural audience" In the 8th International Conference on Web Engineering ICWE'08. pp. 231-239. Yorktown, NJ, USA, 2008
13. F. Montero, P. Gonzáles, M. Lozano, J. Vanderdonckt. "Quality models for automated evaluation of web sites usability and accessibility". In International COST294 Workshop on User Interface Quality Model. Rome, Italy, 2005
14. R. Ruiz-Rodríguez. "An auxiliary tool for usability and design guidelines validation of web sites" In Proceedings of the 15th International Conference on Computing - CIC'06. pp.304-308. Mexico City, Mexico, 2006
15. A.Dix, J. Finlay, GD. Abowd, R. Beale. "Human Computer Interaction" 3rd Ed. Pearson Education Ltd. Essex, (2004)
16. K. Norman, E. Panizzi, "Levels of automation and user participation in usability testing" Interacting with Computers, p. 246-264, 2006
17. J. Nielsen, R. Molich. "Heuristic evaluation of user interfaces" In Proceedings of the SIGCHI conference on Human Factors in Computing Systems: Empowering People. pp. 249-256. Seattle, Washington, USA, 1990
18. P. Polson, C. Lewis, J. Rieman, C. Wharton. "Cognitive walkthroughs: a method for theory-based evaluation of user interfaces" International Journal of Man-Machine Studies, pp. 741-773, 1992
19. R. Otaiza, C. Rusu, S. Roncagliolo. "Evaluating the usability of transactional web sites" In Third International Conference on Advances in Computer-Human Interactions. pp. 32-37. Saint Maarten, Netherlands, Antilles, 2010
20. A. Beirekdar, J. Vanderdonckt, M. Noirhomme-Fraiture. "KWARESMI - knowledge-based web automated evaluation tool with reconfigurable guidelines optimization" In Proceedings of the 9th International Workshop on Design, Specification and Verification of Interactive Systems DSV-IS. pp. 362-376. 2002

21. R. Jeffries, J. Miller, C. Wharton, KM. Uyeda. "*User interface evaluation in the real world: a comparison of four techniques*" In Proceedings of the ACM Computer Human Interaction CHI'91 Conference. pp.119-124. New Orleans, LA, USA, 1991
22. J. Vanderdonckt, A. Beirekdar, M. Noirhomme-Fraiture. "*Automated evaluation of web usability and accessibility by guideline review*". Lecture Notes in Computer Science, pp. 17-30, 2004
23. A. Beirekdar, J. Vanderdonckt, M. Noirhomme-Fraiture. "*A framework and a language for usability automatic evaluation of web sites by static analysis of HTML source code*" In 4th International Conference on Computer-Aided Design of User Interfaces. pp. 337-349. 2002
24. LM. Tobar, PM. Latorre Andrés, E. Lafuente Lapena. "WebA: a tool for the assistance in design and evaluation of websites". Journal of Universal Computer Science, pp. 1496-1512, 2008
25. T. Brinck, E. Hofer. "*Automatically evaluating the usability of web sites*" In Conference on Human Factors in Computing Systems CHI'02. pp. 906-907. Minneapolis, Minnesota, USA, 2002
26. A. Beirekdar, M. Keita, M. Noirhomme, F. Randolet, J. Vanderdonckt, C. Mariage. "*Flexible reporting for automated usability and accessibility evaluation of web sites*" INTERACT, pp. 281-294, 2005
27. JA. Borges, I. Morales, NJ. Rodríguez. "Guidelines for designing usable world wide web pages" In Proceedings of the Conference on Human Factors in Computing Systems: common ground. pp. 277-278. Vancouver, British Columbia, Canada, 1996
28. M. Burmester, J. Machate. "*Creative design of interactive products and use of usability guidelines - a contradiction?*" In J. Jacko, C. Stephanidis, D. Harris. "Human-computer interaction: theory and practice", Lawrence Erlbaum Associates Inc., pp. 43-46 (2003)
29. MY. Ivory, MA. Hearst. "*The state of the art in automating usability evaluation of user interface*" ACM Computing Surveys (CSUR), pp. 470-516, 2001
30. MH. Blackmon, PG. Polson, M. Kitajima, C. Lewis. "*Cognitive walkthrough for the web*" In Proceedings of the SIGCHI Conference on Human factors in Computing Systems: Changing our World, Changing Ourselves. pp. 463-470. Minneapolis, Minnesota, USA, 2002
31. MY. Ivory. "*Web TANGO: towards automated comparison of information-centric web site designs*" In Proceedings of the ACM CHI 00 Conference on Human Factors in Computing Systems, Student Posters. pp. 329-330. 2000
32. K. Groves. "The limitations of server log files for usability analysis" (2007) Available at: <http://www.boxesandarrows.com/view/the-limitations-of> [Accessed 22 January 2011]
33. M. Winckler, C. Freitas, J. Lima. "Usability remote evaluation for the WWW" In Proceedings of the CHI'00 Extended Abstracts on Human Factors in Computing Systems. pp. 131-132

34. G. Murray, T. Costanzo. "*Usability and the web: an overview*" Information Technology Services, #61 National Library of Canada, Network Notes (ISSN 1201-4338)
35. J. Nielsen. "*Enhancing the explanatory power of usability heuristics*" In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: celebrating interdependence. pp. 152-158. Boston, Massachusetts, USA, 1994
36. SL. Smith, JN. Mosier. "*Guidelines for designing user interface software*" Mitre Corporation, Report No.: MTR-9240 (1986)
37. DA. Norman. "*The design of everyday things*", Doubleday (a division of Bantam Doubleday Dell Publishing). (1988)
38. J. Nielsen. "*The usability engineering life cycle*" Computer (IEEE): 12-22, 1992
39. D. Sano. "*Designing large-scale web sites: a visual design methodology*" Wiley Computer Publishing, John Wiley & Sons Inc. (1996)
40. JM. Spool, T. Scanlon, C. Snyder, T. DeAngelo. "*Web site usability: a designer's guide*", Morgan Kaufmann Publishers Inc. (1998)
41. J. Fleming. "*Web navigation: designing the user experience*", O'Reilly & Associates, (1998)
42. L. Rosenfeld, P. Morville. "*Information architecture for the world wide web*", O'Reilly & Associates, (1998)
43. B. Shneiderman. "*Designing the user interface: strategies for effective human-computer interaction*", 3rd Ed. Addison-Wesley, (1998)
44. J. Nielsen. "*Designing web usability: the practice of simplicity*", New Riders Publishing, (1999)
45. J. Nielsen, H. Loranger. "*Prioritizing web usability*", New Riders Press, (2006)
46. U.S. Department of Health and Human Sciences in Research-Based Web Design & Usability Guidelines (2006) Available at: http://www.usability.gov/guidelines/guidelines_book.pdf [Accessed 31 October 2010]
47. M. Hertzum. "*Problem prioritization in usability evaluation: from severity assessments toward impact on design*" International Journal Human Computer Interaction, 21(2): pp. 125-146, 2006
48. G. Sim, JC. Read. "*The damage index: an aggregation tool for usability problem prioritisation*" In Proceedings of HCI 2010. Dundee, Scotland, 2010
49. J. Ratner, EM. Grosse, C. Forsythe. "*Characterization and assessment of HTML style guides*" In Proceedings of the Conference on Human Factors in Computing Systems: common ground. pp. 115-116. Vancouver, British Columbia, Canada, 1996

50. J. Nielsen, M. Tahir. "*Homepage Usability - 50 Websites Deconstructed*", New Riders, (2002)
51. M. Richtel. (1998) from The New York Times: Available at: <http://www.nytimes.com/library/tech/98/07/cyber/articles/13usability.html> [Accessed 27 January 2011]
52. P. Marks. (2001) from CNN: Available at: <http://edition.cnn.com/2001/WORLD/asiapcf/east/02/23/web.usability/index.html> [Accessed on 27 January 2011]
53. J. Hamilton. (2000) from Businessweek: Available at: http://www.businessweek.com/2000/00_47/b3708076.htm [Accessed on 27 January 2011]
54. M. Miller. (2002) from PC Mag: Available at <http://www.pcmag.com/article2/0,2817,1165443,00.asp> [Accessed on 27 January 2011]
55. L. Thomason. (2002) from NetMechanic: Available at http://www.netmechanic.com/news/vol5/review_no21.htm [Accessed on 27 January 2011]
56. Internet Archive. Available at <http://www.archive.org/> [Accessed on 10 January 2011]
57. MY. Ivory, A. Chevalier. "*A study of automated web site evaluation tools*" Technical Report (Report No.: UW-CSE-02-10-01), 2002
58. T. Tiedtke, C. Martin, N. Gerth. "*AWUSA - a tool for automated website usability analysis*" In Pre-Proceedings of the 9th International Workshop on Design, Specification, and Verification of Interactive System DSV-IS'2002. pp. 251-266. 2002
59. J. Vanderdonckt, A. Beirekdar. "*Automated web evaluation by guideline review*" Journal of Web Engineering, pp. 102-117, 2005