# Quality Models for Automated Evaluation of Web Sites Usability and Accessibility

**Francisco Montero[1,2], Pascual González[1], María Lozano[1], Jean Vanderdonckt[2]**

[1]Grupo de Investigación LoUISE, EPSA – UCLM, 02071 Albacete, Spain
{fmontero, mlozano, pgonzalez}@info-ab.uclm.es

[2]Belgian Lab. of Computer-Human Interaction – UCL, 1348 Louvain-la-Neuve, Belgium
{montero,vanderdonckt}@isys.ucl.ac.be

**ABSTRACT**

When usability evaluation is performed on web sites, many different evaluation methods can be used that are analytical or empirical, depending if they are conducted with or without end users, on a real web site (part or whole) or on a representation of it. A classification on evaluation methods is given on these parameters so as to assess the relevance and appropriateness of each evaluation method. For this purpose, the quality models for web usability evaluation need to be characterized. When a method is applied (semi-) automatically on a web site, the characterization of these quality models become even more preeminent and crucial to really know the relevance and appropriateness of the results provided by the automated method. Towards this end, different quality models are compared based on guideline review to show their various levels of precision, their advantages and shortcomings

**Author Keywords**

Accessibility, automated evaluation, criteria, evaluation method, factor, guidelines, metric, quality model, quality of web sites, usability.

**ACM Classification Keywords**

H.5.2.e Evaluation/methodology and H.5.2.n Style guides.

**INTRODUCTION**

Despite the abundance of usability and accessibility knowledge, web site quality continues to be a pressing human-computer interaction issue. The majority of web sites have usability and/or accessibility problems, which can result in confusing users, and ultimately, loss of revenue. One of biggest signs that a site has quality problems is that users struggle to find the information they are looking for or they simply cannot manipulate the contents as they wish. Some web pages are so cluttered that users can easily miss the link or the feature that they are looking for. Given that an estimated 90% of web sites provide inadequate usability, a projected growth of 196 million new sites within the next five years, and a severe shortage of user interface professionals to ensure usable sites: tools and methodologies are needed to accelerate and improve the web site design process [3]. Additionally, not every organization can afford to spend millions to hire professionals to design their sites.

To achieve software quality in a system, the software's attributes must be clearly defined. Otherwise, assessment of quality is left to the intuition or the responsibility to the persons who are in charge of the process. In this sense, a quality model must be built and evaluation methods should be used during design and implementation stages based on these quality models. These web quality models can take many different forms depending on the emphasis unconsciously or consciously put on some part:

- Software Engineering pays a lot of attention to the software quality in terms of factors such as correctness, robustness, extendibility, and reusability.
- Software Performance tends to privilege performance factors such as rapidity, compactness, and efficiency.
- Assistive technologies consider that the most important quality factor of a web site resides in its accessibility as any web site should be accessed by the widest audience possible (perhaps with different disabilities or specific needs), in different contexts of use, with various computing platforms.
- Mobile Computing is interested with factors affecting the code and the contents of a web site so as to transcode it to a mobile phone, a Personal Digital

Assistant (PDA), a handbag PC, a tablet PC or a laptop.

- The area of Graphic Design includes skills and requirements for ensuring the quality of the visual design of web sites.
- Web Engineering applies systematic, disciplined and quantifiable approaches to development, operation, and maintenance of Web-based applications. Not only developers are busy with the well-formedness of code with respect to established standards, but also the validity of the code is an integral part of the quality.
- Last, but not least, Human-Computer Interaction has tackled how to assess the usability and the accessibility of a web site according to user-centered methods.

Since there is no universal method for measuring the usability and the accessibility of a web site in all circumstances that guarantees the representativeness of these notions, many attempts exist to approximate usability and accessibility by different models, involving different types of functions. As quality, usability, and accessibility are concepts requiring interpretation, we assign the following goals to this paper:

- We must identify the underlying quality models used for (semi-) automated evaluation of web sites for usability and accessibility. For example, the ISO 9126 [16] standard decomposes quality into six factors (*functionality, reliability, efficiency, portability, usability, maintainability*) that are in turn decomposed into sub-factors, for instance usability contains *learnability*, *operability* and *understandability*. The standard does not prescribe how these sub-factors can be effectively and efficiently measured.
- We need to identify what are the potential shortcomings of usability and accessibility evaluation performed on these models. For example, web designers, developers and evaluators need to understand general characteristics of web quality evaluation tools as they need to know that there is more to ensuring the quality of a website than merely checking it with a tool.
- We need to identify the strengths and weaknesses of tools for automated evaluation. For example, many accessibility tools are based on accessibility guidelines, e.g. WCAG [24] or Section 508 [22], but vary slightly in their interpretation of guidelines. So these results still require human judgment.

To address the three above goals, this paper first surveys some automated or semi-automated tools for usability and accessibility evaluation. This section ends up with a table comparing the salient characteristics of these tools so as to identify their strengths and weaknesses. Then, a framework is introduced to identify and express their underlying quality models that are further analyzed and discussed in the fourth section.

## RELATED WORK

Several surveys of evaluation methods for user interfaces exist: Hom [15], Zhang, [25] or Ivory [17], provide a detailed discussion of inspection, inquiry or testing methods and automatic tools. Ivory introduced a taxonomy for classifying evaluation methods [17]. She groups methods along four dimensions: method class, method type, automation type and effort level. Guidelines and automatic testing tools for websites play an important role in determining a set of attributes and measurement methods that are both viable and reliable. However, guidelines and tools need to be validated. A general classification of guidelines that can be used to position a tool for Web evaluation by guideline review has also been introduced [18] in which guidelines are considered during operationalization and evaluation.

Automatic tools for websites analysis can be used to inspect the source code, to inspect the live web pages, to inspect the web server logs of usage of a website, to test the performance of the web server and back ends, to test the positioning of a site on search engines. Automatic tools are able to detect only features related to internal attributes; there's no way for them, in a totally automatic way, to determine external attributes. Automatic evaluation tools are potentially beneficial in the web site design process [11].

### Related work in software engineering

Pioneering work on quality factors was performed by McCall [19] and by Boehm [7]. These early models are usually criticized because they lack rationales used to decide which factors should be included in the definition of the quality and which criteria should associated with specific factors. In addition, the metrics used in the bottom layer are not specified, which makes the application of these models difficult in practice. McCall's model includes eleven quality factors to describe three aspects of the product quality: operation, revision, transition. Boehm model uses seven quality factors to evaluate the maintainability and utility.

The ISO 9126-1 [16] model improves the McCall model by defining six factors but does not elaborate on criteria and metrics layers. All these models suppose the quality factors are higher level and not easily directly measured.

### Related Work in Web Engineering

Quality of web sites and their associated factors like usability, accessibility or performance are very interesting, and there are many ongoing research and development projects in this scope. WebTango [17], WebQEM [20] or Kwaresmi [5] are only a few relevant examples in evaluating quality on the web. So, before introducing our proposal in this issue, we would like to comment briefly some of these meaningful previous proposals. So, let us focus now on the usability factor of quality. A summary table of these proposals and tools is shown in Table 1.

**WebTango** of Ivory [17] presents a synthesis of usability and performance evaluation techniques, which together build an empirical foundation for automated interface evaluation. WebTango is able to capture up to 157 different metrics of each web site, depending on its type. However, only six of them are effectively used to assess the usability of a web site in the underlying model: the actual value of each of this metric is then compared with a reference value of this metric depending on the site type. The reference values has been previously empirically validated by several user testing experiments on different types.

The general approach of WebTango [17] involves:

1. Identifying an exhaustive set of quantitative interface measures, building a quality model. Under nine characteristics, 157 highly-accurate, quantitative page-level and site-level measures are proposed. The measures assess many aspects of Web interfaces (Eq. 1). Each criteria has associated several quantitative metrics, these are related with text, link, and graphic elements or formatting, page formatting or performance and site architecture measures. Ivory proposes a **statistical quality model** (Eq. 1);
2. computing measures for a large sample of rated interfaces;
3. deriving statistical models from the measures and ratings;
4. using the models to predict ratings for new interfaces, and
5. validating model predictions.

$$\mho(p) \approx f_{\text{WebTango}}(\textit{content, structure and navigation, visual design, functionality, interactivity, overall experience}) \qquad \text{(Eq.}\mathbf{1}\text{)}$$

Olsina [20] follows common practice in describing software quality in terms of quality characteristics as defined in the ISO/IEC 9126-1 [16] standard. **WebQEM** [20] starts from the ISO model and customizes it at the sub-factor level. WebQEM sees attributes as measurable properties of an entity and propose using a **linear quality model** (Eq. 2) following linear additive and non-linear multimedia scoring criteria to specify them.

$$\mho(p) \approx f_{WebQEM}(\text{ISO 9126-1 customizing it at the subcharacteristics level}) = \sum_{i=1}^{attributes} a_i x_i + \left(\sum_{j}^{attributes} \left(b_j y_j^r\right)\right)^{1/r} \qquad \text{(Eq. }\mathbf{2}\text{)}$$

where $a_i$ and $b_j$ are weight and $x_i$ and $y_j$ are elemental indicators

**Kwaresmi** [5] is a framework that defines a systematic and consistent way for structuring guidelines in order to enable their automatic evaluation; a Guideline Definition Language (GDL) able to express guideline information in a sufficiently rich manner to enable an evaluation engine to perform automated evaluation of any GDL-compliant guideline and a tool to support the proposal.

$$\mho(p) \approx f_{kwaresmi}(\text{Web\_page}, \text{UES}_{i,j}) =$$

$$\textbf{EXEC}\ (\text{EC}_{i,j}\{\text{INST\_UES}_{i,j}\}) =$$

$$\{\text{“}\textit{Respected}\text{”} \mid \text{“}\textit{Violated}\text{”} \mid \text{“}\textit{Partially Respected}\text{”}\} \qquad \text{(Eq. }\mathbf{3}\text{)}$$

where $UES_{i,j}$ be the set of evaluation sets associated to the guideline i in the source j and that will be used for the evaluation of the evaluated page. $EC_{i,j}$ be the set of evaluation conditions associated to $UES_{i,j}$. $INST\_UES_{i,j}$ be the set of captured instances of $UES_{i,j}$ in the evaluated page

In practice, the $f$(Web_page, $UES_{i,j}$) (Eq. 3) executes each $EC_{i,j}$ condition, and then it combines the results to have the overall result for the guideline i. We say that a web page satisfies a guideline $G_{i,j}$, if the execution of all $EC_{i,j}$ on all the $INST\_UES_{i,j}$ is true. Using the above evaluation parameters allows us to define a kind of quality model to balance the evaluation result. Contrary to the binary model used by most existing evaluation tools, Beirekdar uses a weight concept to express the evaluation result. Kwaresmi uses a **linear quality model**.

In the accessibility field, **Bobby** [6], **Valet** [2] and **EvalIris** [1, 13] are representative examples of accessible evaluation tools. All these tools are based on accessibility guidelines, so for example Bobby helps authors determine if their sites are accessible. It does this through automatic checks as well as manual checks. It also analyzes web pages for compatibility with various browsers (Eq. 4). Accessibility tools use a **binary model** to evaluate the accessibility of web pages (Eq. 4)

$$\textit{Accessibility errors} = \sum_{i=1}^{guidelines} a_i x_i \qquad \text{(Eq. }\mathbf{4}\text{)}$$

where $a_i$ is 0 when guideline is violated and 1 when guideline is not violated and $x_i$ is a guideline.

| Tool | WebQEM | WebTango | Kwaresmi | EvalIris | Valet | Weblint |
|---|---|---|---|---|---|---|
| **Author** | Olsina [20] | Ivory [17] | Beirekdar [5] | Abascal [1, 13] | WebThing Ltd. [2] | Bowers [8] |
| **Application** | Quality | Usability | Accessibility | Accessibility | Accessibility | HTML validator |
| **Model** | Linear and non-linear | Statistical | Linear | Binary | Binary | Binary |
| **Scope of approach**[3] | Product | Product | Product | Product | Product | Product |
| **Defines factor**[4] | Yes | No | No | No | No | No |
| **Defines criteria**[4] | Yes | No | No | No | No | No |
| **Defines sub-criteria**[4] | Yes | Yes | Yes | No | No | No |
| **Suggests metrics**[4] | No | Yes | No | No | No | No |
| **Aspect measured** | Metrics | Metrics | Errors | Errors | Errors | Errors |
| **Point of view** | Dev.[1] / User | Dev. /User | Dev. | Dev. | Dev. | Dev. |
| **Interpreting data** | Quantitative | Qualitative | Qualitative | Quantitative | Quantitative | Quantitative |
| **Evaluation technique** | Semi-autom. inspection | Automated inspection | Automatted inspection | Automated inspection | Automated inspection | Automated inspection |
| **Implementation** | Hard coded[2] | Hard coded | GDL | XML | Hard coded | Hard coded |
| **Based on** | ISO 9126 | Webby score | Guidelines | Accessibility guidelines | Accessibility guiielines | HTML specification |

**Table 1. Some meaningful usability and accessibility evaluation methods and tools**

[1] Dev ≈developer's point of view

[2] All *hard coded* means is that it is constant and very hard to change.

[3] Each method or tool can be product or process-oriented. All proposes are product-oriented.

[4] These rows (*defines factor, criteria, sub-criteria and metrics*) refer to the possibility of using these elements as building blocks for quality model construction

## A FRAMEWORK FOR QUALITY MODELS IN WEB AUTOMATED EVALUATION

Consider the examples presented in the previous section, we can see that the quality of a web site is a property difficult to define and capture in an operational way. We observe two meaningful things: there is not a universal quality model and there is a little relationship between quality errors and how to prevent them. To improve the understanding of these shortcomings, a reference framework for quality models in web automated evaluation is introduced (Fig. 1) that decomposes the quality into several factors and links them using experience in form of guidelines.

A ***quality model*** specifies which properties are important for ensuring the quality of a web site [9, 10], it is a description of which criteria are important for the analysis, which one is more important than others, and which measurement methods have to be used to assess the criteria. Quality may depend on task-related factors, performance-related factors and development-related factors. A ***factor*** is a statement of a general evaluation dimension which is expressed symptomatically by intrinsic qualities and deficiencies and which could be measured and/or estimated.

For instance, the ISO 9126 [16] definition of quality for software products is *the totality of features and characteristics of a software product that bear on its ability to satisfy stated or implied needs*. In this definition, quality is decomposed into six factors which are each in turn further decomposed into *criteria*. ***Criteria*** are recognized and accepted dimensions that are empirically proved to influence the quality. For instance, usability is decomposed in ISO 9126 into three criteria: *learnability*, *operability*, and *understandability*.

Alternatively in the ISO 9241 standard, usability is decomposed into *effectiveness*, *efficiency*, and *user satisfaction*. We are focus on usability and in [21], usability is decomposed into 8 ergonomic criteria: *compatibility*, *consistency*, *workload*, *adaptability*, *dialog control*, *representativeness*, *guidance*, and *error management*. *Ergonomic criteria* are primarily considered as design criteria because they can serve at design time, but they can also serve as evaluation criteria at evaluation time. For instance, it is interesting to see the impact of ergonomic criteria (e.g., *consistency*) on factors (e.g., *usability*). When appropriate, criteria can be recursively decomposed into a taxonomy of sub-criteria. For instance, consistency can be refined into consistency of location, of presentation formats, of dialogue (Fig. 1).

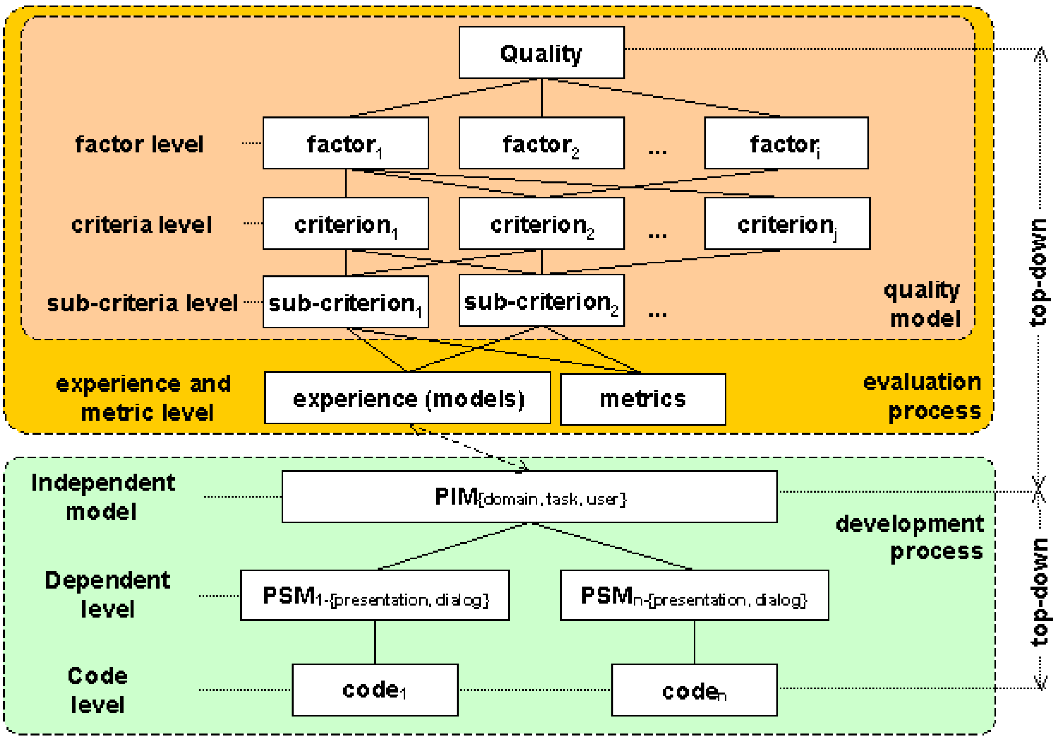


**Figure 1. Reference framework for automated evaluation of web sites**

Our reference framework is aimed at providing explicit guidance in measuring and applying the criteria. For this purpose, the concept of experience is introduced (Fig. 1) in the definition of our quality model. ***Experience*** are advices on web design based on good practices which have web design professionals. This definition highlights that following experience is a necessary but sufficient condition to reach the goal. In this framework experience is modeled using different notations and this experience is platform-independent and a same experience can participate in several criteria, thus introducing a multi-criteria approach. For instance, to promote the criteria of consistency, experience in form of models can be manipulated that govern the consistency within a web page or across web pages, for every type of contents such as objects, menus, icons, controls, their locations, their presentation, and their dialogs.

To concretely assess the verification of these models, one or several metrics are required. In our framework, a ***metric*** is a quantifiable dimensions that explicitly participates into the verification of a guideline. For instance, to check the guideline "Location of controls should be consistent", the different controls need to be identified and their location, computed. Then, the guideline may be assessed in some automated way. An *object* is any widget composing the user interface of a web page. This includes banners, images, illustrations, videos, controls (e.g., edit boxes, radio buttons, push buttons), panes, etc. Pages when assembled form the web site submitted to evaluation or a presentation of it, typically a significant subset of it.

Consequently, this framework combines two top-down approaches, in one of them quality is progressively refined into models and metrics, and in the other approach those models are used to specify a software product where quality criteria of interest for the evaluation are explicitly considered.

This framework is inspired in different experiments and tendencies. First of all, usability criteria need to be integrated into web site development training. In addition, usability criteria need to be adapted such that designers, who do not have a background in human factors, can apply them. In this sense, there are different experiments [12] where is showed that a student group identified 51% more usability problems in the condition with the usability criteria than in the conditions without guidelines or with the ISO/DIS 9241-10 standard. In these experiments, participants were familiar with ergonomic and usability aspects (they were in a cognitive ergonomics university program), but in these experiments was determined that people, who do not have a background in ergonomics or human factors, can similarly apply usability criteria. So in our framework, in the criteria level ergonomic criteria are considered. These ergonomic criteria are related with quality factors from quality and usability international standards (i.e. ISO 9126, ISO 9241).

Second, in this moment Model-Driven Architecture (MDA) is the tendency in software development. The MDA defines an approach to modeling that separates the specification of system functionality from the specification of its implementation on a specific technology platform. In short it defines a guidelines for structuring specifications expressed as ***models***. The MDA promotes an approach where the same model specifying system functionality can be realized on multiple platforms through auxiliary mapping standards, or through point mappings to specific platforms. It also supports the concept of explicitly relating the models of different applications, enabling integration and interoperability and supporting system evolution as platform technologies come and go.

Third, in many occasions quality is additional non-explicit functionality (copy, paste, undo, redo, feedback, error management, help, assistance, information, etc.). This functionality can be modeled and this models can be associated with usability criteria.

With our framework quality model and software development process are linked using models. Those models are platform-independent because users need quality and usability in whatever platform.

So, in this workshop several questions can be introduced, for instance, Can the quality be modeled?, Is the quality platform-independent?. In this moment, quality is adequately considered in the software development process?. Is the quality a quantitative or a qualitative feature?, etc.

## CONCLUSION

Quality web applications need to be usable, functional, reliable, maintainable, scalable and secure. A wide range of evaluation techniques have been proposed and a subset of these techniques is currently in common use. In this sense, we need two ingredients: development process and quality evaluation. There is thus a strong need for Web Engineering. We need to get a better understanding of the development process itself, we need to gain a much clearer understanding of the development process and how it relates to the resultant qualities of web sites. Developing Web based systems is significantly different from traditional software development and poses many additional challenges, some of them were introduced in this paper.

Some techniques and tools pretend to assess the usability or the accessibility of we sites, but they actually do not: they replace usability by metrics that are not necessarily related to usability. Some others replace usability by a mix of usability-oriented metrics and a set of non-usability-oriented metrics. Sometimes, these metrics are not directly related to usability, but they may influence it. For example, the time to download a web page considered per se is a performance factor than a usability factor. But when one knows that usability guidelines recommend that the system response time stays within the limits of 2 sec for a simple task and 5 sec for a more complex task, then the downloading time becomes interesting. The problem is that when this situation occurs, there is little or no guidance on how to turn the computed metrics into interpretable data.

In this paper we introduced a framework where experience and web site development are linked using qualitative good practices. This experience is modeled and is organized using ergonomic criteria These models can be used in software analysis and design stages. Using this experience novice programmers can find usability problems and they can do changes in the specification of theirs software products to achieve quality improvements in theirs products.

## ACKNOWLEDGMENTS

This work was supported by the Spanish CICYT projects TIN2004-08000-C03-01 and PBC-03-003. We gratefully acknowledge the support of the SIMILAR network of excellence (www.similar.cc) and the COST European Action n°294 (MAUSE).

## REFERENCES

1. Abascal J., Arrue M., Fajardo I., Garay N., Tomás J.: Use of Guidelines to Automatically Verify Web Accessibility. Universal Access in the Information Society (2004)
2. Valet. http://valet.webthing.com/access/url.html
3. Barry, C., Lang, M.: A survey of multimedia and web development techniques and methodology usage. IEEE 8, 2 (2001)
4. Basili, V., Caldiera, G., Dieter H.: The Goal Question Metric Approach. In: Caldiera, G., Rombach, D.H. (eds.): Encyclopedia of Software Engineering. John Wiley: New York, (1994)
5. Beirekdar, A., Vanderdonckt, J., Noirhomme-Fraiture, M.: A Framework and a Language for Usability Automatic Evaluation of Web Sites by Static Analysis of HTML Source Code. In: Kolski, Ch. Vanderdonckt, J. (eds.): Proc. of 4th Int. Conf. on Computer-Aided Design of User Interfaces CADUI (2002)
6. Bobby. http://webxact.watchfire.com/
7. Boehm, B.W., Brown, J.R., Kaspar, H., Lipow, M., Macleod, G.J., Merritt, M.J.: Characteristics of Software Quality. North-Holland: Amsterdam (1978)
8. Bowers, N. Weblint: Quality Assurance for the World-Wide Web. Journal reference: *Computer Networks and ISDN Systems*, Volume 28, issues 7–11, p. 1283. (1996)
9. Brajnik, G.: Towards Valid Quality Models for Websites. In: Proc. of 7th Conf. on Human Factors and the Web HFWeb'01 (2001)
10. Brajnik, G.: Quality Models based on Automatic Webtesting. In: Proc. of the CHI'2002 Workshop "Automatically Evaluating Usability of Web Sites" (2002)
11. Brajnik, G.: Comparing Accessibility Evaluation Tools: Results from a Case Study. In: Ardissono, L., Goy, A. (eds.), Proc. of Symposium on Human-Computer Interaction HCITALY'2003 (2003)
12. Chevalier, A., Ivory, M. Can novice designers apply usability criteria and recommendations to make web sites easier to use ? , in Proc. of Int. Conf. on Human-Computer Interaction HCI International. (2003)
13. EvalIris. http:// www.sc.ehu.es/acwbbpke/evaliris.html
14. Farenc, Ch., Palanque, Ph., Vanderdonckt, J., User Interface Evaluation: is it Ever Usable?, in Proc. of 6th Int. Conf. on Human-Computer Interaction HCI International. (1995).
15. Hom, J.: The Usability Methods Toolbox. http://jthom.best.vwh.net/usability/ usable.htm. (1996)
16. ISO/IEC 9126-1. Software Engineering–Product Quality–Part 1: Quality Model. International Organization for Standardization: Geneva (2001)
17. Ivory, M.: Automated Web Site Evaluation. Researchers' and practitioners' perspectives. Human-Computer Interaction Series, Vol. 4, Kluwer Academic Pub.: Dordrecht (2003)
18. Mariage, C., Vanderdonckt, J., Pribeanu, C.: State of the Art of Web Usability Guidelines. In: Proctor, R.W., Vu, K.-Ph.L. (eds.): The Handbook of Human Factors in Web Design. (2004)
19. McCall, J.A., Richards, P.G., Walters, G.F. Factors in Software Quality, Vols. I, II, and III. NTIS. (1977)
20. Olsina, L., Rossi, G.. Measuring Web Application Quality with WebQEM. IEEE Multimedia (2002)
21. Scapin, D.L., Leulier, C., Vanderdonckt, J., Mariage, C., Bastien, Ch., Farenc, Ch., Palanque, Ph., Bastide, R.: A Framework for Organizing Web Usability Guidelines. In Kortum, Ph., Kudzinger, E. (eds.): Proc. of Conf. on Human Factors and the Web HFWeb'2000 (2000).
22. United States Rehabilitation Act Section 508 standard. http://www.section508.gov
23. Vanderdonckt, J., Beirekdar, A., Noirhomme-Fraiture, M., Automated. Evaluation of Web Usability and Accessibility by Guideline Review, Proc. of 4th Int. Conf. on Web Engineering ICWE'04. (2004)
24. World Wide Web Consortium. Evaluating Web Sites for Accessibility. http://www.w3c.com
25. Zhang, Z., Basili, V., and Shneiderman, B.: An empirical study of perspective-based usability inspection. Human Factors and Ergonomics Society Annual Meeting, Chicago. (1998)