ICSE 99 - 2nd Workshop on Software Engineering over the Internet

Web-site Quality Evaluation Method: a Case Study on Museums

*Luis **Olsina Santos***
GIDIS, Department of Computer Science,
Engineering School at UNLPam;
also, at UNLP – Argentina
E-mail olsinal@ing.unlpam.edu.ar
URL http://gidis.ing.unlpam.edu.ar

ABSTRACT
In this work, we present a methodology for the quantitative evaluation and comparison of Web site quality called Web-site Quality Evaluation Method (QEM). The core models and procedures for artifact evaluation are supported by the Logic Scoring of Preference (LSP) model and continuous preference logic as mathematical background. We discuss the process steps that the evaluators should follow by applying the Web-site QEM, namely: (a) Selecting a site or a set of competitive sites specific to a domain, (b) Specifying goals and the user view, (c) Specifying in a standard-compliant way, Web-site quality characteristics and attributes, (d) Defining the evaluation criterion for each attribute, and applying attribute measurement, (f) Aggregating elementary attributes to yield the global quality preference, and (g) Analyzing, assessing, and comparing partial and global outcomes. In order to illustrate the methodology we focus on a case study on typical museum sites where more than ninety components were involved regarding the general visitor view. The process results may be useful to understand, control, and improve the Web artifacts quality in small, medium and large-scale projects.

KEYWORDS: Web-site QEM, Quantitative Evaluation, Quality, LSP, Museum, Case Study.

1. INTRODUCTION
Web-based Information Systems (WIS) are growing at a rapid pace, both in terms of the increasing acceptability of Web sites, and in terms of the complexity of such artifacts. However, a much defined product process model that leverage the effective development, and evaluation process model that promote the Web-site quality assessment and improvement are not being accompanied by that sites growth [16, 20]. Therefore, a systematic and disciplined utilization of engineering methods, models, and techniques for the understanding, assessment, and improvement of this kind of software should be considered a mandatory requirement. One of the primary goal for Web-site quantitative evaluation is to understand the extent which a given collection of quality characteristics fulfills a selected set of needs regarding a specific user view.

On the one hand, Web site domains like electronic commerce, museums, academic sites, etc., are becoming increasingly complex systems. Hence, an integral quantitative evaluation process regarding all relevant quality characteristics is also a complex issue. The evaluation complexity is caused by the large amount of intervening characteristics and attributes (about a hundred in the case study presented here), and by the complex logic relationships among attributes and characteristics. Besides, some relevant attributes to evaluate can not objectively be measured so that only can be included after a subjective measurement made by expert evaluators.

On the other hand, evaluation methods and techniques fall in two main categories: qualitative and quantitative. Even if software evaluation has more than three decades as a discipline and well-known methods and techniques were applied to evaluate hardware and software systems [1, 2, among others], the systematic and quantitative quality evaluation of Hypermedia applications and, particularly of Web sites, is rather a recent and frequently neglected issue. In fact, Garzotto et al. [5] have introduced some evaluation criteria like richness, ease, consistency, etc., to evaluate in a qualitative way hypermedia systems. However, this approach is only well

suited when the evaluation problem is rather simple and intuitive. In cases with many elementary attributes, it is difficult to evaluate accordingly and it is hard to identify minor differences between similar competitive systems. In addition, in the last three years Web-site style guides and design principles have emerged to assist developers in the development process [7, 15, 19]; also, list of guidelines that author should follow in order to make sites more accessible [21]. These guidelines and techniques have shed light on essential characteristics and attributes and might improve the Web-site designing and authoring process but, obviously do not constitute an evaluation method by themselves. Finally, quantitative surveys [15] and domain-specific evaluations are right now emerging. In this direction, Lohse and Spiller [10] identified and measured 32 attributes that influence store traffic and sales. However, we need a broad, integrated, engineering-based evaluation method and process model for the assessment and comparison of complex Web-site quality requirements.

The main aim of this work is to show our methodology, utilized for the quantitative evaluation and comparison of sites in the operational phase. The core evaluation models and procedures are grounded in the LSP model and continuous preference logic as mathematical background [1]. We discuss the general process steps that evaluators should follow by applying Web-site QEM. So that, to illustrate it, we include a detailed case study on museum sites; e.g., Louvre [11], Prado [13], Metropolitan [12], and National Gallery of Art [14] sites. These museums are internationally well known and placed in three different countries and in four big cities. Therefore, our goal is to evaluate the level of accomplishment of required characteristic as usability, functionality, reliability, and efficiency, and compare partial and global preferences to analyze and draw conclusions about the state-of-the-art of Web-site quality. These quality characteristics and attributes were outlined considering IEEE and ISO/IEC standards for software quality metrics and guidelines [8, 9]. In order to effectively select quality characteristics we should consider different kind of users. In this case study, we consider the visitor standpoint.

At the end of the evaluation and comparison process, we obtain for each selected Web-site system a global quality indicator using the scale from 0 to 100%. Such cardinal rating will fall in three categories or preference levels, namely: *unsatisfactory* (from 0 to 40%), *marginal* (from 40 to 60%), and *satisfactory* (from 60 to 100%). The global preference can be approximately interpreted as the degree of satisfied user requirements.

The structure of this paper is as follows: In section 2, we present an overview of the process steps that decision-makers should follow by applying the Web-site QEM. In Section 3, we make some general considerations about the case study. In Section 4, we represent quality characteristics and attributes regarding the general visitor standpoint and we discuss some elementary criteria and rating levels. We show the process of aggregating elementary criteria to yield the global quality preference, in Section 5. Next, we analyze some partial and global outcomes; and, finally, we consider concluding remarks and future work.

2. BASIC PROCESS STEPS OF THE WEB-SITE QEM

Figure 1 shows a high-level view of major steps required for quality evaluation and comparison. In addition, it depicts the Quality Requirement Definition, Elementary and Global Evaluation, and Analysis, Conclusions and Documentation phases. Next, we describe the major process steps that evaluators should follow by applying the Web-site QEM, namely:

- ✓ *Selecting a site or a set of competitive sites to evaluate or compare*
- ✓ *Specifying goals and the user viewpoint*
- ✓ *Defining the Web-site quality characteristics and attributes requirement tree*
- ✓ *Defining criterion function for each attribute, and applying attribute measurement*
- ✓ *Aggregating elementary preferences to yield the global Web-site quality preference*
- ✓ *Analyzing, assessing, and comparing partial and global outcomes*

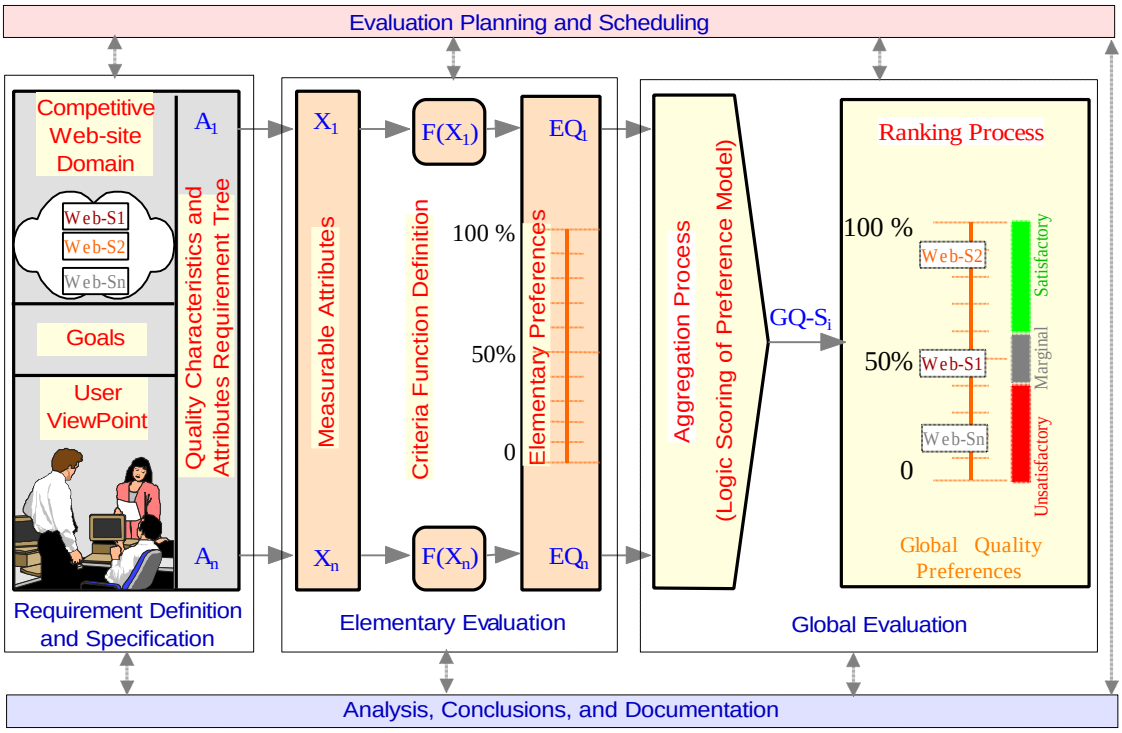

Figure 1 *A high-level view of the Web-site quantitative evaluation and comparison process*

Step one. *Selecting a site or a set of competitive sites to evaluate or compare:* in this task, decision-makers should know what is the evaluation domain and select the systems to assess. For instance, regarding WIS or sub-systems we should emphasize more usability than security characteristic or both, depending on the specific situation. In e-commerce domain, security is an essential characteristic, but in academic domain could be less important. In addition, if the purpose of the assessment is the comparison of competitive sites, it should be selected an appropriate sample in order to be successful throughout the evaluation process.

Step two. *Specifying goals and the user viewpoint:* in this step, evaluators should define the goals and scope of the evaluation process. They could assess a new running Web project, or an operational project, and could evaluate the quality of a set of attributes or sub-characteristics, a sub-system, an overall system, or compare characteristics or global preferences of competitive systems. The results might be useful to understand, assess, control, forecast, or improve the quality of artifacts. On the other hand, the formulation of a list of goals and, consequently, the relative importance of quality characteristics and attributes vary depending on different users. Therefore, we define three generals views of quality: visitor´s view, developer´s view, and manager´s view. In turn, we could decompose the visitor category in two audiences regarding museums: *general visitors* and *expert visitors* [4, 6]. The former could represent casual or intentional audience maybe having minimum domain knowledge or, perhaps, a general interest on museums. The latter represents, a specialist or researcher in museum domains.

Step three. *Defining the Web-site quality characteristics and attributes requirement tree:* in this process step, evaluators should define and specify the quality characteristics and attributes, grouping them into a requirement tree. In order to follows well-known standards, we use the same conceptual characteristics (also known as factors in [8]) like usability, functionality, reliability, efficiency, portability, and maintainability. From these characteristics, we derive sub-characteristics, and from these, we can specify attributes with a minimal overlap. For each quantifiable attribute A_i, we can associate a variable X_i, which can take a real value: the measured value. That hierarchical decomposition from characteristics in sub-characteristics and

measurable attributes could be also considered in the IEEE software quality metric framework.

Step four. *Defining criterion function for each attribute, and applying attribute measurement:* in this process, the evaluators should define the basis for elementary evaluation criteria and perform the measurement sub-process. Elementary evaluation criteria say how to measure quantifiable attributes. The result is an elementary preference, which can be interpreted as the degree or percentage of satisfied requirement. For each variable X_i , i = 1, ...,n it is necessary to establish an acceptable range of values and define a function, called the elementary criterion. This function is a mapping of the measured value in the empirical domain [3] into the new numerical domain. Next, the final outcome is mapped in a preference called the elementary quality preference, EQ_i. We can assume the elementary quality preference EQ_i as the percentage of requirement satisfied by the value of X_i . In this sense, EQ_i = 0% denotes a totally unsatisfactory situation, while EQ_i = 100% represents a fully satisfactory situation [1]. Ultimately, for each quantifiable attribute, the measurement activity should be carried out.

Step five. *Aggregating elementary preferences to yield the global quality preference:* in this task, the evaluators should prepare and enact the evaluation process to obtain an indicator of global preference for each competitive system, or for a single evaluated system. For n attributes, the mapping produces n elementary quality preferences. Applying a stepwise aggregation mechanism, the elementary quality preference can be grouped accordingly, resulting the global quality preference. The global quality preference represents the global degree of satisfaction of all involved requirements. In the museum case study, we use a logical scoring model to get the global preference. Specifically, we use the LSP model grounded in the continuous preference logic. The strength of LSP resides in the power to model simultaneity, replaceability, and other relationships using aggregation operators based on weighted power means. At the end of the evaluation and comparison process we obtain for each competitive system a global indicator using the scale from 0 to 100%. Such cardinal rating level will fall in one out of three categories of preferences: unsatisfactory, marginal, and satisfactory (as shown by colored bars in figure 1).

Step six. *Analyzing, assessing, and comparing partial and global quality preferences:* in this final step, the evaluators analyze, assess, and compare the elementary, partial and total quantitative results regarding the established goals and user view. Feedback cycles could be necessary.

In the following sections, we will focus on specific aspect of the museum study.

3. SOME CONSIDERATIONS ABOUT THE CASE STUDY

Primary, we want to draw some general issues and assumptions to the present case study. One of the main goal for museums assessment is to understand the extent which a selected set of quality attributes fulfill a given set of stated requirements. Particularly, in this work we focus on the operational phase of Websites.

Speaking in a wide sense, software artifacts are generally produced to satisfy specific user's needs, and Web-site artifacts are not the exception. In designing Web-site artifacts, there are many challenges that frequently are minimized. For instance, when users enter the first time at a given home page they often can figure to find a piece of information quickly. There are two mechanisms to help them in doing that: browsing and searching. Thus, to get a time-effective mental model of the overall site (that is, structure and content), there are attributes like a site map, an index, or a table of contents, that help in getting a quick global site understandability. These attributes facilitate browsing. However, a global searching function provided in the main page could effectively help retrieving the desired piece of information and avoid browsing. Moreover, both functions could be complemented at any moment. There are many attributes and characteristics that contribute to site quality such as usability, functionality, reliability among others, that a designer should take into account when designing for intended audiences.

So, to analyze and draw conclusions about the state-of-the-art of essential quality characteristics and attributes we chose museum's domain. Typical sites already established like Louvre, Prado, Metropolitan, and National Gallery of Art museums were chosen. Even if there are major differences between physical museums and their Internet counterpart, these museums are internationally well known and were posted on an average two year ago. Figure 2 shows a snapshot of home pages.

Figure 2 *From upper-corner left to right, Prado and Gallery of Art museums' home pages; and from bottom-corner left to right, Metropolitan and Louvre museums' home pages.*

Finally, we should deal with this important consideration: web sites are artifacts that can evolve dynamically and users always access the last on-line version. By the time of data collection (from September 15 to October 15, 1998), we did not perceive changes in these Web sites that could affect the evaluation process. However, since late November (after data collection), Louvre museum has radically changed its look & feel feature and consequently attributes of interface and navigational control object characteristics have changed too. In addition, has recently incorporated virtual tours. This attribute do not was included in the requirement at evaluation time and is beginning to appear in some other museums' sites around the world. On the other hand, Metropolitan museum has incorporated the global search function, which was not available at the time of evaluation. Fortunately, some raw data were downloaded.

4. THE ELEMENTARY MEASUREMENT

4.1 Outlining the Quality Requirement Tree. In this step, we define and categorize a wide set of museum quality attributes grouping them into a requirement tree. The primary goal is to group characteristics and attributes by performing the third step of the Web-site QEM. As previously said, to follows well-known standards we use the same high-level characteristics like usability, functionality, reliability, efficiency, portability, and maintainability. These characteristics give evaluators a conceptual and general description of software quality and provide a baseline for further decomposition. From these characteristics, we could derive sub-characteristics, and from these, we could specify measurable attributes and variables.

In addition, the relative importance of characteristics varies depending on the different users and application domains. According to this, we define three views of quality: visitor, developer, and manager views [9]. Figure 3, outline the major characteristics and measurable attributes regarding the visitor standpoint. Specifically, from the point of view of general visitors, artifacts characteristics such as maintainability and portability will not be necessary to evaluate. General visitors are mainly interested in the ease of use and communicativeness of the Web site, in its browsing and search mechanisms, in its coherent navigation mechanisms and dependent-domain expected functionality, and also, in the site reliability and efficiency.

1. Usability
1.1 Global Site Understandability
1.1.1 Global Organization Scheme
1.1.1.1 *Site Map*
1.1.1.2 *Global Index (Subject, Alphabetic)*
1.1.1.3 *Table of Content*
1.1.2 Quality of Labeling System
1.1.2.1 *Textual Labeling*
1.1.2.2 *Iconic Labeling*
1.1.3 *Guided Tours*
1.1.3.1 *Conventional Tour*
1.1.3.2 *Virtual Tour* (*)
1.1.4 *Floor and Room Image Map*
1.2 Feedback and Help Features
1.2.1 Quality of Help Features
1.2.1.1 *Web-site Explanatory Help*
1.2.1.2 *Search Help*
1.2.2 Web-site Last Update Indicator
1.2.2.1 *Global*
1.2.2.2 *Scoped (per sub-site or page)*
1.2.3 Addresses Directory
1.2.3.1 *E-mail Directory*
1.2.3.2 *Phone-Fax Directory*
1.2.3.3 *Post mail Directory*
1.2.4 *FAQ Feature*
1.2.5 *Survey/Questionnaire Feature*
1.3 Interface and Aesthetic Features
1.3.1 *Cohesiveness to Group Main Control Objects*
1.3.2 Presentation Permanence and Stability of Main Controls
1.3.2.1 *Direct Controls Permanence*
1.3.2.2 *Indirect Controls Permanence*
1.3.2.3 *Stability*
1.3.3 *Aesthetic Preference*
1.3.4 *Style Uniformity*
1.4 Miscellaneous Features
1.4.1 *Foreign Language Support*
1.4.2 *Download Feature*
2. Functionality
2.1 Searching Issues
2.1.1 Web-site Search Mechanisms
2.1.1.1 *Scoped Search (Collection sub-site)*
2.1.1.2 *Global Search*
2.2 Navigation (and Browsing) Issues
2.2.1 Local Navigability
2.2.1.1 *Level of Local Interconnection (for a Collection sub-site)*
2.2.1.2 *Orientation*
2.2.1.2.1 *Indicator of Path*
2.2.1.2.2 *Label of Current Position*
2.2.2 Global Navigability
2.2.2.1 *Coupling among Sub-sites*
2.2.3 Navigational Control Objects
2.2.3.1 Presentation Permanence and Stability of Contextual Controls
2.2.3.1.1 *Contextual Controls Permanence*
2.2.3.1.2 *Contextual Controls Stability*
2.2.3.2 Level of Scrolling
2.2.3.2.1 *Vertical Scrolling*
2.2.3.2.2 *Horizontal Scrolling*
2.2.4 Navigational Prediction
2.2.4.1 *Link Title (link with explanatory help)*
2.2.4.2 *Quality of Link Phrase*
2.3 Domain Specific and Miscellaneous Functions
2.3.1 *Content Relevancy (this attribute could be decomposed)*
2.3.2 *Link Relevancy*
2.3.3 Electronic Commerce
2.3.3.1 Purchase Features
2.3.3.1.1 *Shopping Basket Facility*
2.3.3.1.2 *Quality of Product Catalog*
2.3.3.2 *Secure Transaction*
2.3.4 Image Features
2.3.4.1 *Image Size Indicator*
2.3.4.2 *Zooming*
3. Site Reliability
3.1 Nondeficiency
3.1.1 Link Errors
3.1.1.1 *Broken Links*
3.1.1.2 *Invalid Links*
3.1.1.3 *Unimplemented Links*
3.1.2 Miscellaneous Errors or Drawbacks
3.1.2.1 *Number of deficiencies or absent features due to different browsers*
3.1.2.2 *Number of Web-site deficiencies or malfunctions (e.g. non-trapped search errors) or unexpected results independent of browsers*
3.1.2.3 *Number of Dead-end Web Nodes*
3.1.2.4 *Number of Destination Nodes (unexpectedly) under Construction*
4. Efficiency
4.1 Information Accessibility
4.1.1 *Support for Web-site text-only version*
4.1.2 Readability by deactivating Browser Image Feature
4.1.2.1 *Image Title*
4.1.2.2 *Global Readability*
4.2 Performance behavior
4.2.1 *Page Size* (*)

Figure 3 *Requirement tree regarding the museum domain and the general visitor view. (The attributes marked with the * sign, do not were considered in this study. Particularly, the page size attribute was not included because a lot of museum images are zoomed, which is a good feature for visitors though affects performance. However, it was included in an academic study)*

Following we discuss some characteristics and attributes and the decomposition mechanism. The *Usability* characteristic is decomposed in sub-factors such as *Global Site Understandability*, *Feedback and Help*, *Interface and Aesthetic Features*, and *Miscellaneous Features*. The *Functionality* characteristic is decomposed in *Searching*, *Navigation*, and *Domain Specific* issues. The same decomposition mechanism is applied to *Reliability* and *Efficiency* factors. With regard to *Site Understandability* sub-characteristic, in turn we have decomposed it in *Global Organization Scheme*, *Labeling*, *Guided Tours* sub-characteristics, and *Image Map* attribute. These features are mainly available in the home page (and could stay during sub-site navigation), and contribute to a global and quick Web-site understanding of both the structure and the content. However, for instance, the *Global Organization Scheme* factor is still too general to be directly measurable; many attributes can still be grouped in this sub-category. So, we found attributes like *Table of Content*, *Site Map*, and *Global Index*, that could contribute to the *Global Organization Scheme* factor. These attributes are finally quantifiable.

Focusing on attributes we easily could see that no necessarily all of them should exist at the same time; it could be necessary a *Table of Content*, or an *Index* attribute. Moreover, an index type like index by subject, chronological, or alphabetical, could be replaceable according the domain. For instance, subject-oriented indexes could be better in some circumstances that chronological-oriented indexes. One important thing to point out is that LSP model allows to deal with simultaneity and replaceability relationships taking into account weights and levels of and/or polarization.

On the other hand, and regarding the *Web-site Search Mechanism*, it could be better for a visitor counting with a scoped search and global search (the simultaneity relationship), under some circumstances. In fact, by searching a museum collection regarding author and school could be necessary a customized *Scoped Search* as long as a *Global Search* could also be necessary. Sometimes, specific areas of a site are highly coherent and distinct from the rest of the site that makes sense to give a scoped or restricted search to users [15]. However, a basic and advanced global search mechanism is generally enough.

Finally, regarding *Reliability* factor we discuss the *Nondeficiency* sub-factor. That is, the degree to which artifacts do not contain undetected errors. In this category, and considering *Link Errors*, we found attributes like *Broken, Invalid, and Unimplemented Links.* The *Broken Links* attribute counts dangling links out of the total site links leading to absent destination nodes. Similarly, the *Invalid Links attributes* counts the founded links that drive into wrong or unrelated nodes; and the *Unimplemented Links* attribute counts links that unexpectedly drive to the same origin node. The higher the detected number of links errors, the lower the site *Reliability*. Consequently, the quality is debased.

4.2 Establishing Elementary Criteria As aforementioned, for each attribute A_i we can associate a variable X_i which can take a real value by means of the elementary criterion function. The final result represents a mapping of the function value into the elementary quality preference, EQ_i. The value of EQ_i is a real value that 'fortunately' belong to the unit interval. As stated by Dujmovic et al. in [1]:

"*the elementary preference is interpreted as a continuos logic variable. The value 0 denotes that X_i does not satisfy the requirements, and the value 1 denotes a perfect satisfaction of requirements. The values between 0 and 1 denote a partial satisfaction of requirements. Consequently, all preferences are frequently interpreted as a percentage of satisfied requirements, and defined in the range [0, 100%]*".

In turn, that preference can be categorized in three rating levels in the Web-site QEM, namely: satisfactory (from 60 to 100%), marginal (from 40 to 60%), and unsatisfactory (from 0 to 40%). For instance, a marginal score for an attribute could indicate that a correction action to improve the attribute quality should be taken into account by the manager or developer.

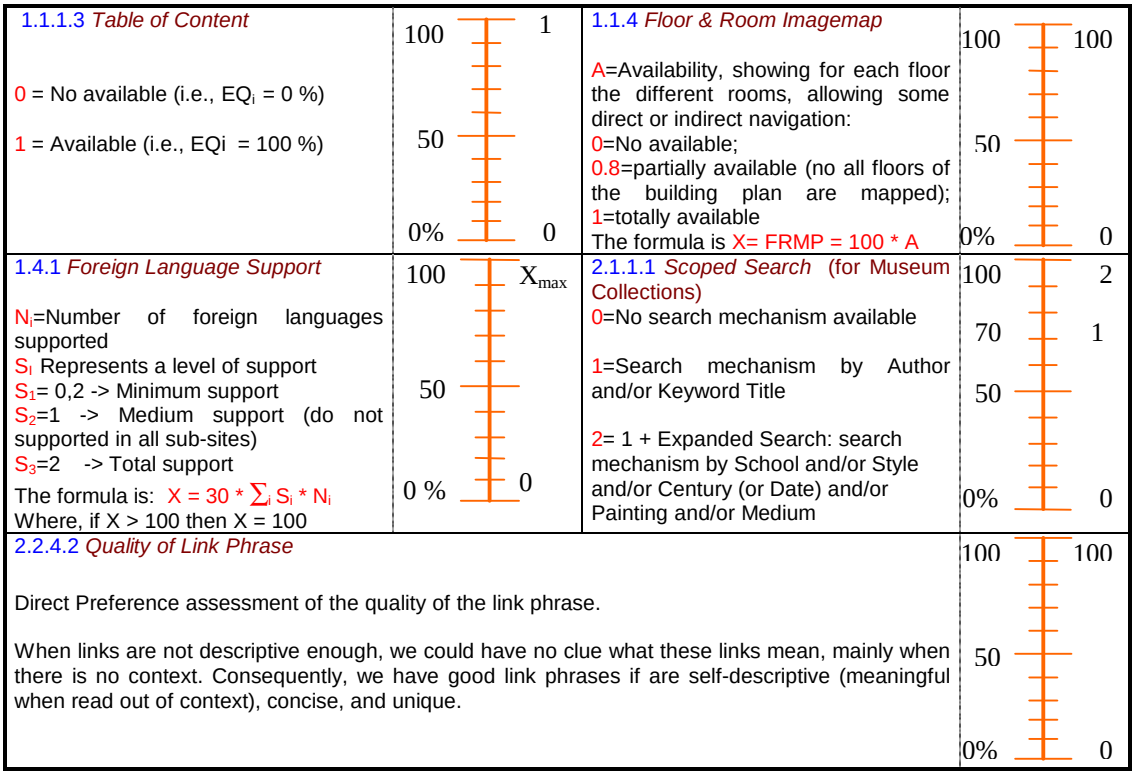

Figure 4 *Set of five elementary criteria defined as preference scales taken from the tree*

Figure 4, shows five elementary criteria for museum attributes. There are two major categories to classify elementary criteria, that is, absolute and relative criteria. Moreover, regarding the absolute elementary criteria, these are further decomposed in continuous and discrete variables. For instance, the evaluation criterion for a *Table of Content* attribute is an absolute and discrete binary criteria: we only ask if it is available or not available (the corresponding preference scale is shown in fig. 4). The preference scale for the *Floor and Room Imagemap* attribute is a generalization of the binary criteria. The X variable takes three different discrete values and each possible value (0, 80, 100) is mapped in the respective preference. Similar discrete criterion was followed to *Image Title* attribute at the time of elementary measurement. (However, this criterion now is a continuous one due to the automation of data collection -we are just using the SiteSweeper tool to automate some attributes).

On the other hand, the evaluation criterion for *Collection Search* attribute is a multi-level discrete absolute criterion defined as a subset, where 0 implies no search mechanism available; 1 implies a basic search mechanism (accomplishing 70% of the requirement); and 2 implies both the basic search and advanced (expanded) search mechanism (accomplishing 100% of the requirement). Instead, the evaluation criterion for a *Foreign Language Support* attribute is according the formula shown in the same figure. We consider here, the number of foreign languages supported by the international museum sites (the N_i variable) and the level of support, e.g. total, partial, or minimum (the S_i variable). The resulting value of this discrete multi-variable absolute criterion could be between 0 (completely unsatisfactory) and X_{max} (completely satisfactory). If the measured value of X is above $X_{max,}$ the corresponding elementary preference X will be equal to X_{max} (we can do some interpretation regarding the cost-benefit ratio).

Finally, the *Quality of Link Phrase* attribute is measure by a direct preference assessment. Whenever the above criteria (and other criteria do not showed here) are not applicable because is hard to define variables and their corresponding preference scale, we then use a direct-subjective assessment. Generally, we need an expert in the field to assess such attribute. For

example, in the case of *Quality of Link Phrase* the evaluators should consider if Web-page links are enough descriptive, hence intended users could have clue what these links mean, and predict navigation mainly when there is no context. Consequently, there are good link phrases if they are self-descriptive (meaningful when read out of context), concise, and unique. Specifically, the W3C in the WAI Accessibility Guidelines [21], states: "*When links are not descriptive enough, do not make sense when read out of context, or are not unique, the auditory user must stop to read the text surrounding each link to identify it*".

4.3 Performing Attribute Measurement. Once all elementary criteria were agreed, and data collected, then we can compute the variable value and the elementary quality preference for each attribute of each system. Table 1 shows some results of elementary preferences after computing the corresponding criteria function to the *Usability* characteristic. This activity should be performed for each characteristic (such as *Functionality*, etc.) for each museum.

Table 1 *Partial results of elementary quality preferences (for Usability characteristic) after computing the corresponding criteria function for each museums' attribute*

1. Usability Ai	Louvre Museum EQ_i	Prado EQ_i	Metropolitan EQ_i	Gallery of Art EQ_i
1.1.1.1	0	0	0	0
1.1.1.2	0	0	0	100
1.1.1.3	0	0	0	100
1.1.2.1	80	80	80	100
1.1.2.2	0	80	50	0
1.1.3	50 F=0.5	100 F=1	50 F=0.5	100 F=1
1.1.4	100 A=1	80 A=0.8	80 A=0.8	0 A=0
1.2.1.1	60	60	60	100
1.2.1.2	0	100	0	100
1.2.2.1	100	100	0	0
1.2.2.2	0	0	0	0
1.4.1	90 N=3; S=1	60 N=1; S=2	0 N=0	24 N=4; S=0.2
1.4.2	0	0	0	0

Finally, we should make some considerations with regard to data collection. Data collection activity can be done manually, semi-automatically, and automatically. Most of the attributes values were collected manually because there is no way to do it otherwise. For instance, it is the case to check if there exists a *Table of Content*, a *Site Map*, or a *Guided Tour* attributes. Likewise, to compute the level of *Foreign Language Support*, or to check the availability of a *Secure Transaction Facility*, or a *Scoped Search*, or a *Support for Web-site text-only version*. In many cases, the data is ease to collect and verify. Moreover, for all attributes measurable by a direct preference criterion the unique way to draw an outcome is by means of an expert human judgment. In these cases, the assessment could be harder. On the other hand, automatic data collection is also in many cases the more reliable and almost unique mechanism to collect data for a given attribute. This is the case to measure the *Broken Links*, or to measure the *Image Title* attribute. Nevertheless, the detection of the *Number of Nodes under Construction* could be semi-automated, among others. By the time of data collection for museums only a couple of metrics were automated.

5. AGGREGATION MECHANISM

5.1 Logic Aggregation of Elementary Preferences to yield Global Preferences. In this process step, the evaluators should define and prepare the evaluation process to obtain a quality indicator for each competitive system. Applying a stepwise aggregation mechanism, the elementary quality preferences can be accordingly structured to allow the computing of partial preferences. In turn, repeating the aggregation process at the end can be obtained the global preference. The global quality preference represents the global degree of satisfaction of all involved requirements. In the museum study, we use a logical scoring model called LSP model. A broad treatment of LSP relationships and Continuos Logic Preference (CLP) operators could be found in [1, 2], as well as the mathematical background.

The strength of LSP resides in the power to model different logical relationships to reflect the stakeholders' needs, namely:

- ✓ simultaneity, when is perceived that two or more input preferences must be present simultaneously
- ✓ replaceability, when is perceived that two or more attributes can be replaced (there exist alternatives, i.e., a low quality of an input preference can always be compensated by a high quality of some other input).
- ✓ neutrality, when is perceived that two or more input preferences can be grouped independently (neither conjunctive nor disjunctive relationship)
- ✓ symmetric relationships, when is perceived that two or more input preferences affect evaluation in the same logical way (tough maybe with different weights)
- ✓ asymmetric relationships, when mandatory attributes are combined with desirable or optional ones; and when sufficient attributes are combined with desirable or optional ones.

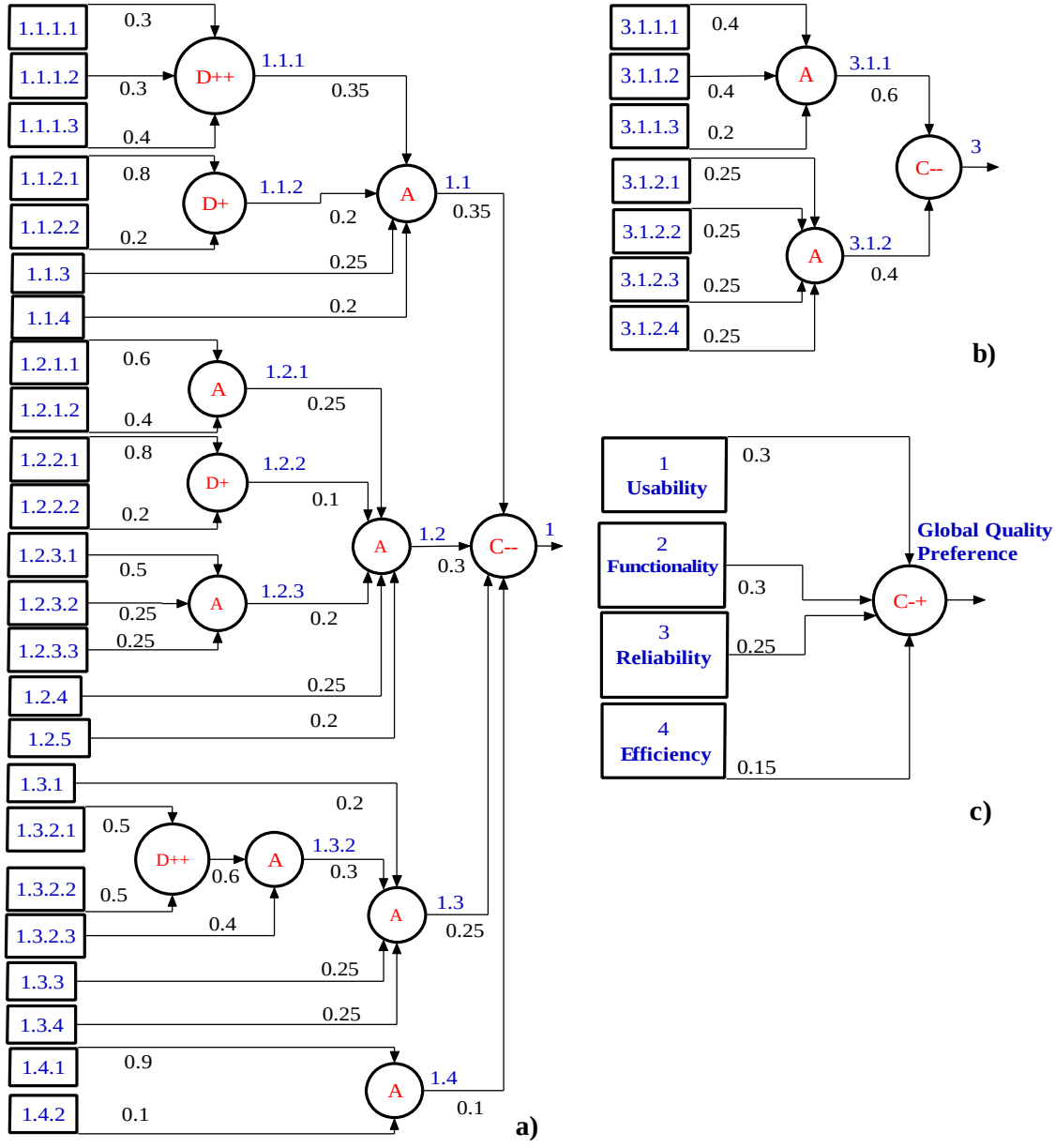

Figure 5. *Structure of partial and global logic aggregation of preferences using the LSP model. Partial logic aggregation for: a) Usability, and b) Reliability characteristics. The c) part, shows the global aggregation of preferences for museum study.*

Figure 5, depicts the aggregation structure for *usability*, and *reliability* characteristics; the c) part shows the high-level characteristics aggregation to yield the global quality preference. The stepwise aggregation process follows the hierarchical structure of the requirement tree from bottom to top. The major CLP operators are the arithmetic means (A) that models the neutrality relationship; the pure conjunction (C), and quasi-conjunction operators that model the simultaneity one; and the pure disjunction (D), and quasi-disjunction operators that model the replaceability one. With regard to levels of simultaneity, we may utilize the week (C-), medium (CA), and strong (C+) quasi-conjunction functions. In this sense, operators of quasi-conjunction are *flexible and* logic connectives. Also, we can tune these operators to intermediate values. For instance, C-- is positioned between A and C- operators; and C-+ is between CA and C- operators, and so on. The above operators (except A) mean that, given a low quality of an input preference can never be well compensated by a high quality of some other input to output a high quality preference. For example, in the figure 5a), at the end of the aggregation process we have the sub-characteristic coded 1.1 (called *Global Site Understandability* in the requirement tree, with a relative importance or weight of 0.35), *and* 1.2 sub-characteristic (*Feedback and Help Features*, 0.3 weighted), *and* 1.3 sub-characteristic (*Interface and Aesthetic Features*, 0.25 weighted), *and* 1.4 sub-characteristic (*Miscellaneous Features*, 0.1 weighted). All these sub-characteristic preferences are input to the C-- logical function, which produce the partial global preference coded as 1, (called *Usability*). The C-- operator do not model mandatory requirements, that is, a zero in one input do not yield a zero at the output though punishes the outcome. (In table 2, the reader could corroborate this situation for the usability factor of the Met museum. A series of input values 44.13 , 54 , 74.93 , 0 results in a preference of 45.66).

Furthermore, the figure 5c), shows the end of the aggregation characteristics coded as 1, 2, 3, *and* 4, respectively. These serve as input to the C-+ operator, which model mandatory requirements. So, a zero in one input will produce a global quality preference of zero. Thus, the higher the level of conjunctive polarization toward the C operator, the higher the strength of punishment to lower input preferences.

Similarly to the aforementioned conjunctive operators, we can also utilize the quasi-disjunction operators in a range of strong (D+), medium (DA), and week (D-) *or* polarization, and also their intermediate values. For instance, D-- is positioned between A and D- operators; and D-+ is between DA and D- operators; and D+- is between D+ and DA operators; and finally, D++ is between D+ and D operators. D operator represents the pure disjunction. For example, in the figure 5a), at the beginning of the aggregation process for the Global Organization Scheme output (coded as 1.1.1), we can have as input a *Table of Content*, *or* a *Global Index*, *or Site Map* strongly replaceable each other (intervene the D++ operator). So, the availability of one of these attributes is sufficient to yield a high preference –the weights are almost similar.

5.2 Computing the Partial and Global Quality Preferences. Once all aggregation criteria were structured and agreed, the decision-makers should enact the evaluation computer program to obtain partial and global quality preferences for each competitive system. Table 2, shows the detailed outcomes of partial and global quality preferences for each selected museum. The final raw of table 2 indicates that Louvre museum has reached 51.74 % of the quality preference; the Prado museum, 68.40 %; the Metropolitan museum, 50.95 %; and the National Gallery of Art museum, 79.26 %, regarding the general visitor viewpoint. Obviously, we can arrange these values in a ranking as illustrated in the next section.

6. ANALYZING AND COMPARING SITE QUALITY OF SELECTED MUSEUMS

In this Web-site QEM process step, the evaluators analyze, assess, and compare the partial and global outcomes regarding stated goals and users' view. At this moment, results dumped in tables (e.g. tables 1, 2, and 3), and final results shown in graphic diagrams (as the illustrated in figure 6), and schemas depicting models of complex aggregation criteria functions (as in figure 5), are useful tools and sources of information to analyze and draw conclusions about the quality of artifact features.

Table 2 *Detailed results of partial and global quality preferences after computing the corresponding aggregated criteria function for each site museum*

Characteristics and Sub-characteristics	Louvre	Prado	Met.	G.of A.
1. Usability	**59.73**	**57.81**	**45.66**	**70.39**
1.1 Global Site Understandability	48.13	57	44.13	79.03
1.1.1 Global Organization Scheme	0	0	0	98.54
1.1.2 Quality of Labeling System	78.15	80	78.17	97.68
1.2 Feedback and Help Features	58.77	48.77	54	65
1.2.1 Quality of Help Features	36	76	36	100
1.2.2 Web-site Last Update Indicator	97.68	97.68	0	0
1.2.3 Addresses Directory	100	100	100	100
1.3 Interface and Aesthetic Features	70.41	72.53	74.93	90.91
1.3.2 Presentation Permanence and Stability of M. Controls	98.02	78.42	86.42	98.02
1.4 Miscellaneous Features	81	54	0	21.6
2. Functionality	**27.94**	**72.67**	**49.19**	**80.41**
2.1 Searching Issues	0	89.53	0	94.78
2.1.1 Web-site Search Mechanisms	0	89.53	0	94.78
2.2 Navigation (and Browsing) Issues	47.79	62.98	78.88	71.04
2.2.1 Local Navigability	47.97	75	75	75
2.2.1.2 Orientation	15.93	70	70	70
2.2.2 Global Navigability	80	80	80	80
2.2.3 Navigational Control Objects	34	61.6	86.8	52
2.2.3.1 Presentation Permanence and Stability	0	46	88	30
2.2.3.2 Level of Scrolling	85	85	85	85
2.2.4 Navigational Prediction	40	40	75	80
2.3 Domain Specific and Miscellaneous Functions	44.75	71.34	83.39	80.17
2.3.3 Electronic Commerce	0	39.43	93.95	93.95
2.3.3.1 Purchase Features	0	90	90	90
2.3.4 Image Features	60	100	80	60
3. Site Reliability	**89.67**	**82.97**	**53**	**89.67**
3.1 Nondeficiency	89.67	92.97	53	89.67
3.1.1 Link Errors	100	80	40	100
3.1.2 Miscellaneous Errors or Drawbacks	75	87.5	75	75
4. Efficiency	**62.44**	**62.44**	**64.39**	**80**
4.1 Information Accessibility	62.44	62.44	64.39	79.99
4.1.2 Readability by deactivating Browser Image Feature	64	64	66	82
Global Preferences	**51.74**	**68.40**	**50.95**	**79.26**

Table 3 *Quality characteristic outcomes and global preferences for Website museums*

Characteristics	Louvre Museum	Prado Museum	Met Museum	Gallery of Art
1. Usability	59.73	57.81	45.66	70.39
2. Functionality	27.94	72.67	49.19	80.41
3. Site Reliability	89.67	82.97	53	89.67
4. Efficiency	62.44	62.44	64.39	80
Global Preferences	**51.74**	**68.40**	**50.95**	**79.26**

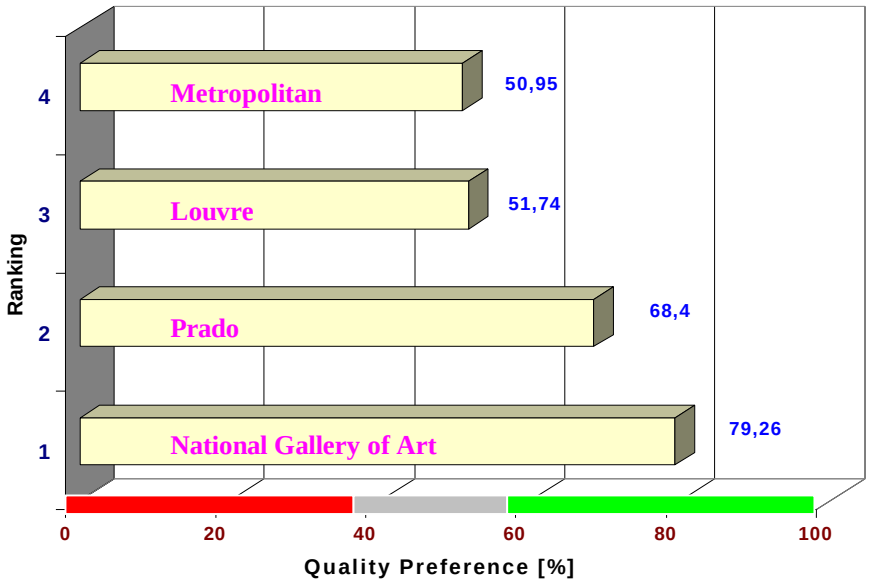

Figure 6. *Ranking for National Gallery of Art, Prado, Louvre, and Metropolitan museums*

Table 3 shows outcomes for the high-level characteristics and global quality preferences for each competitive museum, and figure 6, represents the final ranking. The colored quality bars at the bottom of fig. 6, as well as the colored numbers of table 3, indicate the rating levels as previously defined: satisfactory (green), marginal (gray), and unsatisfactory (red).

According to the results of the evaluation process of Web sites the National Gallery of Art has ranked first, falling into a satisfactory level; the Prado museum has got 68.40 % of the global quality preference; Louvre museum has ranked third, in a marginal level likewise Met museum, which has got 50.95 % of the preference.

Regarding a global indicator of preference, a scoring within a gray bar can be interpreted as improvement actions should be considered (this is the case for global preferences both to Met and Louvre sites), as long as an unsatisfactory rating level can be interpreted as necessary change actions must be taken (as observed in table 3 for the Louvre *functionality* characteristic, scored 27.94 %). Particularly, we can go back and refine the analysis to see why the Louvre, *functionality* characteristic resulted in that score. By observing table 2, the evaluators (and stakeholders in general) can argue that *Searching* functionality is lacking in the site (both scoped and global search). This gives a result of zero to *Searching* sub-characteristic (which is input to the non-mandatory C-- quasi-conjunctive operator, 0.25 weighted). The same outcome was observed for Met site by the time of evaluation (take into account the consideration made in Section 3).

Thus, final results show that Gallery of Art received satisfactory scores in the four main characteristics (between 70 and 90 % in the green quality bar), while Louvre site drew uneven quality characteristics (ranging from 27 to 90%, on the red, gray, and green bars). We can conclude that Louvre museum should improve the *functionality*, by adding a search feature and improve some *navigability* issues like *orientation* and *navigational control objects* characteristics. In addition, among the *Domain Specific and Miscellaneous Features*, we can say that the others museums (Prado, Gallery of Art, and Met) provide some *electronic commerce* features, so appealing regarding a broad audience.

Ultimately, with regard to the *Usability* factor, only Gallery of Art is in the green quality bar, the remainder falls in the gray bar, being the lower score to Met museum with the 45.66 % of the preference. For instance, Met site does not support *foreign languages* in no way, and has neither *table of content*, nor *global indexes*, nor *global site map* (amazingly, the same to Prado, and Louvre). The absence of these last attributes contributes to the lack of *Global Web-site Understandability*, when the visitor enters the first time mainly. On the other hand, Louvre is the unique site with major support to the *Foreign Language* feature (the elementary preference is 90 % to 1.4.1 attribute –see table 1; instead, for Prado is 60; and N. Gallery of Arts is 24 %)

7. CONCLUDING REMARKS

Complex Internet-centered developments are growing at a rapid pace, and among them Web site applications. However, this raises questions like how to design and produce for quality taking into account different audiences; or how to analyze, assess, and improve the quality of Web sites, among other issues. One effective strategy to face these, is product (and process) modeling using prescriptive and/or descriptive approaches [17]. Product modeling potentially allows us, the understanding, evaluation, prediction, and improvement of artifact quality.

In this work, we have presented a systematic and quantitative engineering-based approach to evaluate and compare Web-site quality characteristics and global preferences, regarding intended users. Particularly, we have shown the different Web-site QEM process steps by means of a case study for museums. Considering the general visitor view, we arrange in a hierarchy, quality characteristics and quantifiable attributes following well-known international standards and guidelines. The used requirement decomposition framework is ease to understand and also flexible allowing deletions, additions, and modification of its elements. Moreover, we are

classifying sub-characteristics and attributes to be as useful for most Web-site domains as possible regarding specific users [18].

Furthermore, to model simple and sophisticated aggregation criteria functions we use the LSP model grounded in the continuos preference logic. We can model simultaneity, replaceability, neutrality, and symmetric and asymmetric attribute relationships using logic aggregation operators based on weighted power means. At the end of the evaluation process, the model generate quality preference scores which indicate the level of satisfaction of previously defined user's needs. The use of the Web-site QEM should reduce subjectivity in the assessment process by providing a quantitative ground for the decision-making activity.

Currently, we are making a survey over thirty attributes regarding a sample of 24 museums. In addition, we are carrying out a case study for academic sites and, in next months, we will begin evaluating well-known e-commerce sites. Finally, data collection is an issue. In this direction, we are analyzing integrated tool to automate some suited metrics.

REFERENCES

1. Dujmovic, J.J., 1996, *"A Method for Evaluation and Selection of Complex Hardware and Software Systems"*, The 22nd Int´l Conference for the Resource Management and Performance Evaluation of Enterprise CS. CMG 96 Proceedings, Vol. 1, pp.368-378.
2. Dujmovic, J.J.; Elnicki, R., 1982, *"A DMS Cost/Benefit Decision Model: Mathematical Models for Data Management System Evaluation, Comparison, and Selection "*, National Bureau of Standards, Washington D.C. Nº GCR 82-374. NTIS Nº PB 82-170150 (150 pp).
3. Fenton, N.E.; Pfleeger, S.L., 1997, *"Software Metrics: a Rigorous and Practical Approach"*, 2nd Ed., PWS Publishing Company.
4. Furano, F.; Orsini, R.; Celentano, A., 1997, *"Museum-on-demand: dynamic management of resources in World Wide Web museums"*, Hypertext and Hypermedia: Products, Tools, Methods (H2PTM´97), V. 1, Nº 2-3-4/97 pp. 115,124, Hermes Editorial, Paris, Fr.
5. Garzotto, F.; Mainetti, L.; Paolini, P., 1995, *"Hypermedia Design, Analysis, and Evaluation Issues"*, CACM 38,8 (Ago-95); pp. 74-86.
6. Garzotto, F.; Mainetti, L.; Paolini, P., 1997, *"Designing Modal Hypermedia Applications"*, The Eighth ACM Conference on Hypertext, Southampton, England, pp. 38-47.
7. IEEE Web Publishing Guide, http://www.ieee.org/web/developers/style/
8. IEEE Std 1061-1992, *"IEEE Standard for a Software Quality Metrics Methodology"*
9. ISO/IEC 9126-1991 International Standard, *"Information technology – Software product evaluation – Quality characteristics and guidelines for their use"*.
10. Lohse, G.; Spiller, P., 1998, *"Electronic Shopping"*, CACM 41,7 (Jul-98); pp. 81-86.
11. Louvre Museum: http://www.louvre.fr
12. Metropolitan Museum: http://www.metmuseum.org
13. Museo del Prado: http://museoprado.mcu.es
14. National Gallery of Art Museum: http://www.nga.gov
15. Nielsen, Jakob; The Alertbox, http://www.useit.com/alertbox/
16. Olsina, L, 1998, *"Building a Web-based Information System applying the Hypermedia Flexible Process Modeling Strategy";* 1st International Workshop on Hypermedia Development, at ACM Hypertext 98, Pittsburgh, US, http://ise.ee.uts.edu.au/hypdev/
17. Olsina, L., 1998, *"Toward Improvements in Hypermedia Process Modeling"*, Proceed. at the Symposium of Software Technology, 27 JAIIO, Bs As., pp. 219-226.
18. Olsina, L., Rossi, G.; 1998, *"Toward Web-site Quantitative Evaluation: defining Quality Characteristics and measurable Attributes"*, Submitted paper.
19. Rosenfeld, L., Morville, P., 1998,*"Information Architecture for the WWW"*, O`Reilly.
20. Webby, R.; Lowe, D., 1998, *"The Impact Process Modelling Project"*, 1st International Workshop on Hypermedia Development, at ACM Hypertext 98, Pittsburgh, US, http://ise.ee.uts.edu.au/hypdev/
21. W3C, 1998, W3C Working Draft, "WAI Accessibility Guidelines: Page Authoring", http://www.w3c.org/TR/1998/WD-WAI-PAGEAUTH-19980918/