QUALITY ASSESMENT OF CULTURAL WEB SITES WITH FUZZY OPERATORS

PAOLO DAVOLI
University of Modena e Reggio Emilia
Modena, Italy

FRANCESCA MAZZONI
University of Modena e Reggio Emilia
Modena, Italy

ELENA CORRADINI
Italian Ministry for Cultural Heritage
Bologna, Italy

ABSTRACT

We present FQT4Web, a quantitative inspector-based methodology for Web site evaluation, with a hierarchical structure. A new approach, based on fuzzy operators, permits a sophisticated aggregation of measured atomic quality values, using linguistic criteria to express human experts' evaluations. A wide reference to standards and the limited number of subjective items increase the reliability of site analysis. This methodology has been fine-tuned for cultural sites; however, the assessment method is general-purpose. It is possible to obtain quality charts, which allow decision-makers to validate site quality on the basis of design goals.

Keywords: Web quality, Web cultural applications, Web standards, fuzzy sets.

WHY WEB QUALITY INSPECTION? THREE CASE STUDIES

The age of simple "shop-window" Web sites is getting over. Nowadays, a Web site is more and more expected to provide users with quality services (updated and complete information, rich interactions, use of specific on line applications, access to on-line services, community experience in the Internet, …). A Web site is both a software project and a communication and organizational system. Hence, the design of an organization Web site is becoming a complex process, as it is based for instance on the analysis of the communication goals of the organization; on the investigation of needs and habits of target users (maybe profiled in different categories); on the definition of specific software services to be provided as well as of sector specific regulations and security features. Investigating the quality of a Web site is a fundamental step in the site life cycle, in order to ensure success and return of investments.

As far as e-commerce Web sites quality is concerned, for example, Scheffelmaier and Vinsonhaler (18) use a narrative and qualitative approach to synthesize almost 60 studies on *properties characterizing successful commerce sites*, facing a simple and crucial question: "Why should a Web site with 'good' properties sell more products than a site with 'poor' properties?" Even if not explicitly mentioned, the underlying issue is just that of site quality.

Checking site quality is not a trivial matter: it consists in finding if the site fits design and technological requirements, user needs, organization communications aims, etc… within given constraints (such as budget ranges).

The aim of this paper is to propose a robust and efficient methodology for Web site quality inspection (*FQT4Web, Fuzzy Quality Tree for Web Inspection*), that faces some open problems that still remain in recent literature, as it will be discussed in the next section. The methodology was developed for cultural sites (4, 3), but it is not limited to them. Its application can help decision-makers in the identification of specific quality lacks, in comparison with site goals, and in repairing such lacks.

Let us start from an example. Figure 1 shows the screenshots of the home pages taken from the Web sites of three different cultural institutions. Each row is about one single site, on the left side the site first release is represented (early 2004), while on the right side the current release (late 2004). As a consequence, each row is a comparison between the home pages before and after a site redesign occurred during 2004 (identification data reported in Table 5).

The external differences between *before* and *after* are quite evident, and the right side pages look more "polite" than the left side ones. Is it a matter of restyling or do they actually correspond to an increase in site quality? Which specific quality sides increased during the redesign, and how much? Did the investment in site redesign gain the expected results? Our answers to these questions will be given in Section 6, resulting from the application of the proposed methodology for Web site quality inspection.

This paper is organized as follows. In Section 2 we highlight some open matters on site quality inspection, with reference to the existing literature. The theoretical bases for the proposed methodology, with reference to research literature, are presented in Section 3 (standards and methodologies we refer to) and in the Appendix A (a dossier of the fuzzy operators which we use in the site assessment process). In Sections 4 and 5 the framework of our methodology is presented and in Section 6 it is applied to many case studies.

RELATED WORKS AND OPEN ISSUES

While usability studies are widespread, the issue of Web site quality assessment remains an unsettled matter. Olsina and Rossi (15) base their WebQEM model on ISO/IEC 9126-1 hierarchy (8), and on a summarization model called Logic Scoring of Preference. Atzeni et al. (1) propose a methodology which makes reference to four features – Objectives, Structure, Services and Effectiveness. Ramler et al. (16) (and following them also Ruiz et al. (17)) propose a Web testing methodology based on three orthogonal views (quality aspects, features, phase/lifecycle; the last dimension brings in a timely view). Barnes and Vidgen (2) base the WebQual 4.0 system on three quality characteristics – Usability, Information, Service Interaction. Mich et al. (13) founded their proposal on seven "dimensions" of quality derived form classical Ciceronian rhetoric rules, basically Kipling's *six honest serving-men* What-Why-When-How-Where-Who plus an examination of resources.

FIGURE 1
Comparison between the Home Pages of Three Sites Before and After a Site Re-design

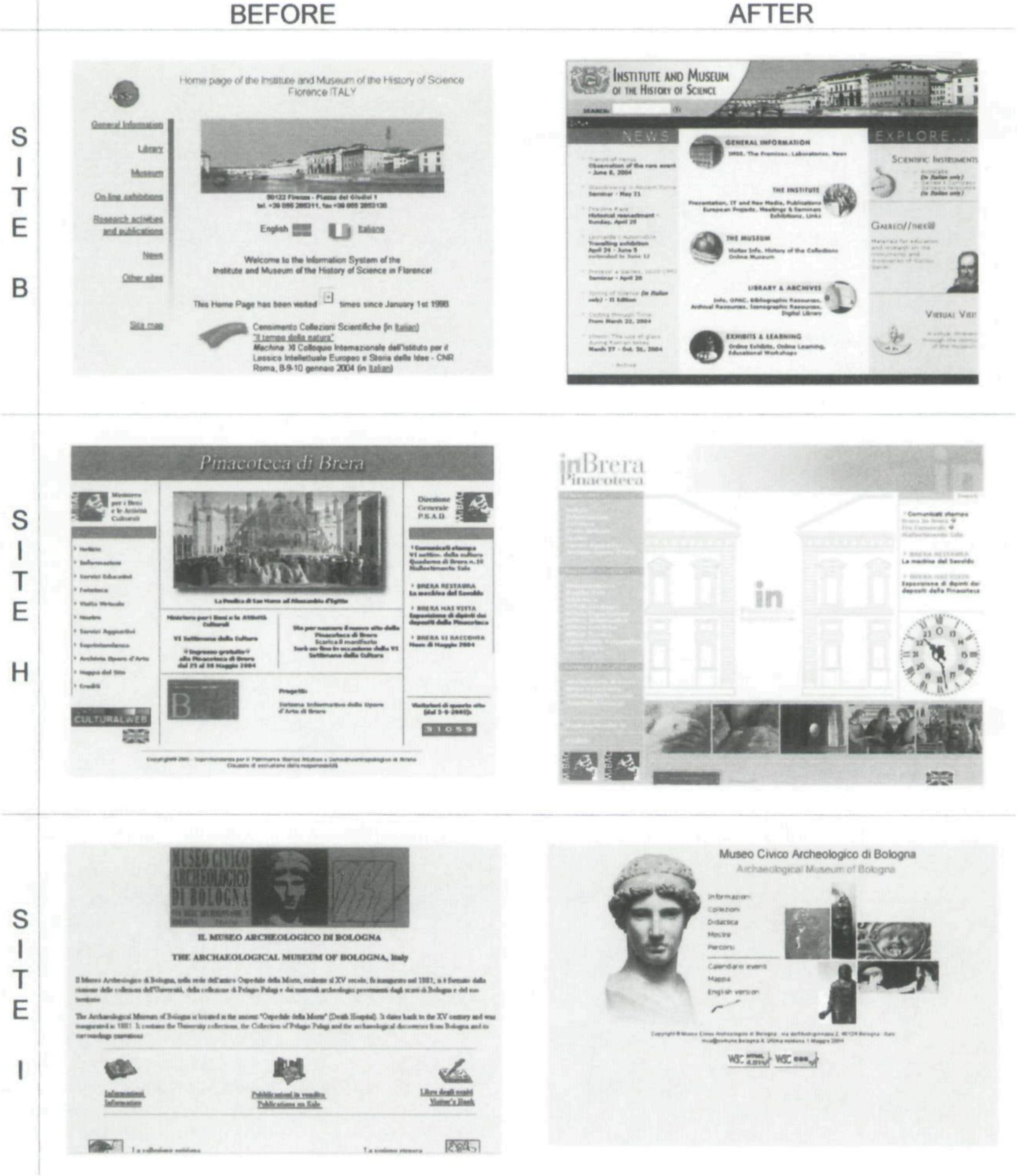

BEFORE
AFTER
SITE B
Home page of the Institute and Museum of the History of Science Florence ITALY
Welcome to the Information System of the Institute and Museum of the History of Science in Florence!
INSTITUTE AND MUSEUM OF THE HISTORY OF SCIENCE
NEWS
EXPLORE...
GENERAL INFORMATION
THE INSTITUTE
THE MUSEUM
LIBRARY & ARCHIVES
EXHIBITS & LEARNING
SITE H
Pinacoteca di Brera
inBrera Pinacoteca
SITE I
MUSEO CIVICO ARCHEOLOGICO DI BOLOGNA
IL MUSEO ARCHEOLOGICO DI BOLOGNA
THE ARCHAEOLOGICAL MUSEUM OF BOLOGNA, Italy
Museo Civico Archeologico di Bologna
Archaeological Museum of Bologna

Table 1 compares some characteristics of these papers that will be discussed in the following.

Quantitative vs. Qualitative Approach

The quality of a Web site has various facets, and each one of them is rather independent on the others. As far as site-human interaction is concerned, every quantitative evaluation has some limitations. On the other hand, qualitative evaluations, by their own nature, give ambiguous results, and do not permit to compare expected and actual site behaviors, nor to evaluate whether a given investment resulted in an appreciable (i.e. numerically comparable) increase in quality. Hence, we think that, even though quantitative evaluations cannot completely capture some aspects related to the "soundness" of a site, they are an indispensable component of any quality check. The methodology here proposed, therefore, will focus on them.

TABLE 1
Some Topics from Reported Literature

	PAPER					
	(15)	(1)	(16)	(17)	(2)	(13)
Quantitative/Qualitative	Quant.	Both	Quant.	Quant.	Qual.	Both
Inspector based/user based	Insp.	Insp.	N.a. (1)	n.a. (1)	User	Both
No. of atomic features	~100	~50	~150	n.a.	22	n.a.
Is a site scoring method described?	Yes	No	No	No	Yes	No
Is subjectivity discussed?	Yes	Yes	No	No	Yes	No (2)

n.a.=not applicable or not available
(1) inspector, likely (2) mentioned but not discussed

Inspector-based vs. User-based Evaluation

It is quite common, in usability and interface studies (11), to distinguish between user-based analyses (empirical methods) and inspector-based analyses (inspection methods). The former are based on groups of final users: e.g. they are asked to access the site performing tasks, and their behavior is observed and/or their opinions are gathered. The latter are performed by an "expert" inspector. In inspection methods the user has a central role once again, but in a "mediate" way – both the researchers who set up the evaluation methodology and the inspectors who perform the evaluation of a specific site should have a careful attention to the user's point of view.

Either such categories of analyses have well-known advantages and disadvantages, and gather quite different information. In Web quality studies, both of them can be performed as well. As far as the overall site quality is concerned, we think that a great variety of aspects is involved, often with technical implications that a common user could hardly manage. In such context, our opinion is that inspector-based methods can be managed in a simple and effective way. The methodology here proposed, therefore, is inspector based. This does not mean that the quality criteria should not be user-centered, but simply that the quality analysis is not made by the users themselves.

Stakeholders and Quality

Web quality represents a complex matter, which involves a variety of "stakeholders." They naturally include users (generic visitors, specific users, task-oriented users, etc.), as well as people involved in site design and implementation (such as software developers and designers, graphic professionals, communication experts, marketing teams), and, last but not least, the site purchasers/owners. Measuring only user-perceived quality may result in a partial view of the matter.

The user's perspective is the main point of reference, but the user himself could not notice important quality aspects related to other stakeholders. For instance, the site maintenance is greatly advantaged by standard adherence - it allows the use of budgetary resources to improve site services instead of to rearrange ill-coded pages. This is an important quality factor for a site, but a common user would hardly rank it.

A Web quality model should be able to account for this complexity and to give information about the quality related to each of these stakeholders categories. The methodology here proposed permits to examine quality features regarding each group of stakeholders.

Open Issues

According to the above reported literature, two unsolved problems should be highlighted:

1) Only few of the reported papers clearly state a straight method for assigning a global assessment to the site and to its quality dimensions, i.e. a method for obtaining in a transparent way a significant score for the site quality, starting from disorganized and inhomogeneous experimental data. Such assessment should not be used to draw up the "top-ten" sites classification (perhaps it could be, but not as a first goal), rather to compare real site quality and design goals, in order to find in which part of the site improvements are needed. Sometimes a simple weighted average is used to summarize the various quality measurements into a single score. As a significant improvement, the Logic Scoring of Preference (LSP) method (5), used by Olsina and Rossi (15), permits a more sophisticated math treatment of the experimental data. LSP is quite simple from the point of view of the mathematical concept involved (it is based on root-weighted-mean-power). Unfortunately, its parameters and calculations, even though rigorous as for the mathematical procedures, are hardly to be intuitively understood by non-math oriented professionals - as most of the stakeholders dealing with site quality are.
2) Another major problem affecting this kind of quality evaluation methodologies (actually including the one we present) is that of arbitrariness and subjectivity.

- *Researcher and domain expert arbitrariness* - There is a lack of commonly accepted references when the evaluation model is designed. Authors point out various combinations of quality aspects, perhaps starting with a personal selection from existing literature of "what is important" for the evaluation process. The risk that the proposed methodology is based on researcher's personal preferences is high.
- *Inspector subjectivity* – The personal preferences of the expert inspector, who performs the actual analysis on a site, may play a significant role during the assignment of scores to the single quality features.

A certain amount of researcher's arbitrariness and inspector's subjectivity is unavoidable in these evaluation processes. Hence, the identification of the sources of arbitrariness and of subjectivity is an important prerequisite of every quantitative methodology, so that it becomes possible to highlight (and, if possible, to limit) their impact on the final quality score.

In facing these matters, the major novelties of the *FQT4Web* methodology here presented are the following:

1. Fuzzy-sets based Ordered Weighted Averaging (OWA) operators are used in the assessment aggregation process, to collect disperse and inhomogeneous experimental data into a single site score (sect. 5 and Appendix A). They are simple to understand in practice and simple to be used, even for a non-math-oriented professional. OWA operators permit a translatiron of the informal quality criteria expressed by a human domain expert into a clean mathematical way, through linguistic quantifiers. Their use in this context is innovative. Recent papers (6, 7) propose applications of fuzzy operators in theoretical models to characterize the *user-perceived informative quality* of Web documents in XML format and of Web sites that provide information stored *in XML documents*. In those papers a model is proposed to obtain *qualitative recommendations* on Web document/site informative quality, through user-perceptible Web evaluation indicators. The approach is therefore qualitative, subjective, user-perception based, and it mainly focuses on information quality. To the best of our knowledge, OWA operators were not proposed nor applied to real cases for a *quantitative* evaluation of the *overall* quality for actual (X)HTML sites until now.
2. As the sources of arbitrariness and subjectivity should be identified (and limited if possible) the proposed methodology strongly refers to standards and specifications that achieved shared consensus by the scientific community (Section 3). Moreover, the number of subjective items is limited and the focus is set on inspector-independent ones (Section 4). As for the point 1 above, it is important to notice that the use of OWA operators does not reduce the role of researcher personal preferences in aggregating data, but it highlights them in a transparent way, so that their influence on the final quality judgment can be clearly debated.

A General-purpose Assessment Method

We designed our quality evaluation methodology orienting it to cultural Web sites. They represent very interesting case studies, as they have to refer to a variety of site behavior models and ask the evaluation methodology to deal with a wide range of quality factors. In fact, cultural Web sites have institutional communication aims, and they generally use information retrieval technologies to present collections and data, sometimes with innovative and user-oriented search mechanisms. In addition, they often host users-communities and deal with "edutainment" (education + entertainment) aspects and multimedia techniques. It is not unusual to find on-line services such as e-shopping activities or ticket reservation.

Therefore, focusing on cultural Web sites does not necessarily mean a lower generality. As it will be discussed in the next section, every evaluation model must be made suitable to domain specific features – the concept of quality for an e-commerce site is different from the one of an institutional site. An evaluation model is not expected to be generic. On the one hand it is expected to be able to capture domain specific quality features, and on the other hand to be sufficiently general so that it is adaptable to different site models. In particular, the assessment method in *FQT4Web* methodology is completely general-purpose.

WEB QUALITY, A MATTER OF STANDARDS

In this section the main standards used in *FQT4Web* methodology are presented, and the reasons why we decided to take into account such standards are discussed.

Software Quality Standards

On the one hand, a Web site is basically a software application; consequently, it is important to refer to the mature techniques of software quality. In particular, the International Standard Organization has set the ISO/IEC 9126-1 standard (8), defining a software quality model which consists of a hierarchy of characteristics and sub-characteristics. The six main characteristics are *functionality*, *reliability*, *usability*, *efficiency*, *maintainability*, and *portability*, and these are then divided into subcategories, such as accuracy, security, fault tolerance, understandability, attractiveness, resource utilization, etc.

On the other hand, a Web site involves many aspects which are not present in common software products. For example, the design team has to care about domain contents (common software does not have a "content"), security issues play an important role, which generally has a lower weight in common stand-alone software; network behavior between the Web server and the user client node (such as for routing or bandwidth) is generally not within the site control. On this basis, following (15, 16), we feel that the ISO/IEC 9126-1 standard (8) should be used as a general framework, with some adaptations to fit the specific nature of Web sites (see sect. 4).

Web Standards

A Web application has to take into account Web standards, both *de iure* standards (protocols such as SSL, and markup languages such as (X)HTML or XML) and *de facto* standards, such as multimedia widespread formats. Traditional Web browsers work with ill-made pages as well, and consequently many Web designers have a loose adherence to standard recommendations (23). Nonetheless, the Web is an heterogeneous environment where users access contents with a variety of capabilities and devices - which are still unknown even today – and where sites interact not only with users, but also with unknown software agents on the Net (such as search engine robots). Therefore, standard compliance greatly increases site interoperability and it represents an important quality factor, both for the final user and the site purchaser: it helps facing problems with client side software configuration (such as operating system, browser type and version); it facilitates site maintainability and reusability; it increases device independence and accessibility.

The *FQT4Web* methodology here presented deals with several technical W3C recommendation, referring for instance to (X)HTML and CSS coding correctness, proper use of META tags, accessibility requirements, URI quality, etc. (Section 4).

Domain Specific Standards

Every site quality analysis has to refer to a specific domain, to its typical contents, regulations, accepted customs, users' habits and expectations – for instance, see (15, 18, 19, 22). To limit his own arbitrariness, a researcher who sets up an evaluation methodology should look for (and refer to) existing standards and authoritative recommendation in the specific domain. Reciprocally, every active stakeholders' community on the Web should promote the growth of such accepted recommendations.

As far as cultural sites are concerned, the Minerva Project proposes an authoritative approach to quality. Minerva is a network of European States' Ministries for Cultural Heritage. The Minerva's Fifth Working Group has recently released the *Handbook for quality in cultural Web sites – improving quality for citizens* (14), whose scope is to provide the principal guidelines in projecting and implementing quality cultural Web applications, with the attempt of requirement standardization. The Handbook has a specific section which reports 12 goals for the Web site of a Cultural Entity, among whom, for instance, we find "Transparency on the activities of the Cultural Entity," "Presentation of standards and regulations of the sector," "Offer of educational services," "Promotion of Web communities," etc. In this paper Minerva Handbook quality models are closely adopted as authoritative recommendations (see Section 4).

FQT4WEB – A FUZZY QUALITY TREE FOR WEB SITE INSPECTION

The Main Framework

Table 2 lists the main framework of *Fuzzy Quality Tree for Web Inspection, FQT4Web* (4, 3). It is based on a hierarchical tree, adapted from the ISO/IEC 9126-1 software quality model. In reference to ISO/IEC model, six quality characteristics were maintained. However, Maintainability, Reliability and Portability were grouped together, while Functionality was broken into two sections, as a result of the necessary adaptations for specific nature of a cultural Web sites discussed in sect. 3. In particular, a cultural Web site may offer very different types of services - some of them are to be considered almost compulsory for any site ("basic" functionality), while others might be considered valuable but not indispensable ("advanced" functionality). Accessibility and Usability are divided into two different characteristics, due to Accessibility increasing importance in W3C recommendations. In the following, we are referring to these "characteristics" as "quality dimensions," according to the agreed idea that Web quality is a complex matter, which requires a multidimensional approach.

Some of the W3C guidelines were taken as reference for points 4 to 6, while general recommendations from the Minerva Handbook (14) represent the basis of the points from 1 to 3 and of some items in point 5. Each line in Table 2 represents an internally structured sub-tree. Sub-trees deeply affected by domain specific features are tagged with an asterisk – they should be strongly modified when a different content domain is dealt with, for example educational or government sites.

The hierarchy is organized into 6 main quality dimensions and 34 first-level sub-characteristics. For the sake of simplicity, Table 2 only reports the higher level nodes. Each sub-characteristic is recursively composed by a set of atomic quality properties (leaves of the tree), or by other non-atomic characteristics (intermediate nodes), which in turn are structured as a grouping of lower level nodes. The actual depth of the tree used for the analyses here reported consists of six levels. The number of atomic questions in each sub-tree is reported on the right side of Table 2, about 160 in all – each one of them has to be empirically answered or measured by an expert evaluator (not a final user), obtaining a value (numerical, Y/N, 1-4; see below). In the quality tree, various sub-trees could be identified, to take into account different users needs: generic user; people interested in preparing a visit to the museum; people interested in museum themes; educational user (children, students, teachers); researcher and museum professionals.

Example Atomic Questions

A few examples of atomic questions regarding interoperability and long term maintainability are listed.

- In Group 4.1 the percentage of pages which do not use the ALT attribute for images is measured, and it is verified if image maps are organized with MAP tags and proper text for hotspots – or (improperly) with tables. In group 4.2 it is verified if an alternative content is provided when scripts, applets, and plug-ins active features are inaccessible or unsupported.
- In Group 5.4 URIs quality and significance is checked from an objective point of view (www.louvre.fr is better than www.paris.fr/culture/musees/louvre/), and from the subjective point of view of the inspector (do we prefer www.nationalgallery.org.uk or www.ng.ac.uk?). Moreover, URI readability and meaningfulness is examined. In (9), it is suggested that a URI should avoid opacity, and should be persistent. A URI like http://www.foo.org/FOO/fooNews/HTML.NSF/By+Filename/mosimple+index?OpenDocument is difficult to memorize, and it is very likely to change the next time that the server technology is updated, thus becoming a broken link somewhere on the Net. (The link is a real one, where the institution name is simply masked with foo.)
- In Group 6.2, we check the use of frames (whose use is controversial, and unsupported in XHTML Strict), the presence of DOCTYPE and char-encoding definitions, the number and severity of (X)HTML coding non-conformities, the use and validity of external style sheets, the inappropriate presence of proprietary non-standard navigation tools.

Answer types and metrics

Each atomic question can be answered in one of the following ways:

- With Yes/No answers (e.g., does the site use style sheets?),
- Using defined metrics for measuring a certain property (e.g. what is the percentage of sample pages with DOCTYPE declaration?).
- With a four level scale (e.g. in Maintainability, how severe are HTML non-conformities?). In the Usability section, most questions belong to this group. In a few cases a fifth level (=zero) is used to take into account the absence of a specific feature. Even if the matter is controversial, we think that an even level scale is preferable to an odd level scale, because the inspector is forced to make a clear choice, avoiding a "low-responsibility" intermediate choice: 1 or 2 means (partially) insufficient, 3 or 4 means (very) good. We experienced that four levels (instead of 6 o 8) suited well in this evaluation context.

  The subjective questions are almost all in the third group. In order to limit the influence of inspector's subjectivity on the final quality evaluation, some key choices were made:
- the number of subjective answers was limited to about one fourth of the total ones, and objective inspector-independent questions were preferred
- each single question is related to a very specific characteristic (generic questions were avoided)
- when the inspection referred to sample pages, the kind and

the depth of the selected pages were precisely defined.

For some questions in the group of the about 40 "technological" ones, software inspection tools were used, such as code validators, site watchers, link checkers, image optimizers, and some informational features of common browsers.

TABLE 2
The Main Framework of *FQT4Web*

	No. of Items
1 – BASIC FUNCTIONALITY	
1.1 – Basic information	
1.1.1 – Information about visiting the institution (*)	4
1.1.2 – Transparency on the institution activity	8
1.1.3 – Spread of cultural content (*)	9
1.2 – Site management	
1.2.1 – Web site identity and responsibility	4
1.2.2. – Evidence of maintenance strategy and content currency	5
1.2.3 – Organization identity and internal functions	6
1.2.4 – Multilingualism	5
1.2.5 – Multimedia features	4
2 – ADVANCED FUNCTIONALITY	
2.1 – Services for common users	
2.1.1 – Offer of Web educational services (*)	3
2.1.2 – Offer of reservation and acquisition services (8)	7
2.1.3 – Privacy policies and transaction management	4
2.1.4 – Support to cultural tourism (*)	1
2.1.5 – Promotion of web communities	5
2.2 – Scientific services and networks	
2.2.1 – Services for scientific research (*)	2
2.2.2 – Services for specialists in the sector (*)	5
2.2.3 – Information on standards and regulation	2
2.2.4 – Evidence of sector network appurtenance	3
3 – USABILITY	
3.1 – Usability basics	
3.1.1 – Quality content and web writing	3
3.1.2 – User interface and metaphors	7
3.1.3 – Site structure	5
3.1.4 – Navigation characteristics	12
3.2 – Support and multimedia	
3.2.1 – Navigation support	6
3.2.2 – User Help	3
3.2.3 – Multimedia usability	5
4 – ACCESSIBILITY	
4.1 – Images, maps, multimedia features	4
4.2 – Client side programming features	2
4.3 – Screen and visual behavior	4
5 – EFFICIENCY	
5.1 – Connectivity	5
5.2 – Visibility on search engines	4
5.3 – Proper use of TITLE and META tags	4
5.4 – URIs quality	3
6 – MAINTAINABILITY & COMPLIANCE	
6.1 – Code quality and standard suitability	8
6.2 – Compliance	4
6.3 – Reliability	4

ASSESSING THE SITE

A Decision-making Problem

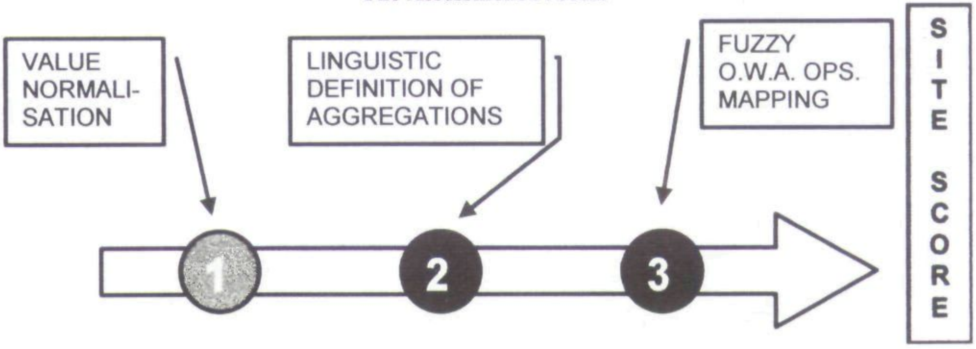
Once the 160 quality properties were measured for a given site, obtaining a general score for the site is an engaging challenge. A structured mathematical method has to be used to take into account all the measured values.

The problem of extracting a global score for the site can be qualified as a *Multi-Criteria Decision Making* problem, and fuzzy systems are well suited to this kind of problems. Actually, we do not have incomplete or uncertain data, but we rather have many atomic data of various semantics, which have to be aggregated in a complex way to produce a single final score value. The fundamental variables that are to be managed to produce the "decision" are, therefore, the aforementioned 160 atomic quality measures. If several sites are to be compared, the "decision" is to obtain a classification based on the calculated score for each site. When a single site is concerned, it can be said that the "decision" consists of extracting a meaningful score from a large amount of non-comparable data – or if preferred, in extracting six scores for the six main quality dimensions. These scores will permit the decision-makers to choose the preferable directions for improving the quality of the site, as it will be shown in the next section.

In order to achieve this goal, the assessment process follows 3 steps (Figure 2)

1. experimental values normalization
2. definition of aggregation criteria – this step may require to refer to qualitative criteria proposed by domain specific experts through linguistic expressions
3. and mapping with fuzzy Ordered Weighted Averaging (OWA) operators, which are part of a so-called mean operators class. We found them suitable for the assessment process, as they are both powerful and easy to use, and permit a simple modeling of various linguistic quantifiers (21). In Appendix A the reader can find a definition and a brief summary of the main characteristics of the OWA operators, which are relevant for this paper.

FIGURE 2
The Assessment Process


VALUE NORMALI-SATION
LINGUISTIC DEFINITION OF AGGREGATIONS
FUZZY O.W.A. OPS. MAPPING
1
2
3
SITE SCORE

Normalization

The answer to each atomic question corresponds to a number, which is the objective or subjective measured value for that specific atomic property. If we want to put them together, we first have to normalize each measured value to the same range, typically from 0 to 1 (or from 0 to 100 percent – is the same). A normalized value states the degree of satisfaction for that specific quality measurement.

In the cases of Yes/No answers and of answers with a score ranging 1-4, normalization is trivial. Other answers must be processed with a transformation function. For example, an atomic question in the group 5.1 Efficiency-Connectivity asks "What is the percentage of pages with images without WIDTH/HEIGHT attributes?" Suppose that the answer for a site is 20%. In theory, this percentage should be zero, but actually no site is (X)HTML compliant. Hence, a tolerant behavior can be adopted, stating that a percentage up to 10% is acceptable (score=1), while a percentage of 50% or higher is completely unacceptable (score=0). Intermediate percentages are converted into intermediate values (e.g. the percentage of 20% gives a normalized score of 0.75). This process is a source of "researcher arbitrariness" (see sect. 2 above), that should be properly evidenced.

Setting of Aggregation Process

After normalization, a set of 160 atomic quality values ranging from 0 to 1 is obtained. The normalized quality values for each group in the lower levels are to be summarized to obtain a group score. Each sub-group sums up a score to the upper level, contributing to the upper level scoring. Therefore, the overall score calculation requires an aggregation function, recursively applied to each group at each tree level:

$$S = \overset{n}{\underset{i=1}{AGGR}}(nqv_i)$$

S represents the final score, *nqv* $_i$ is the *i-th normalized quality value* of each group, *AGGR()* is an aggregation function – the trivial aggregation function is the arithmetic mean, in *FQT4Web* OWA operators were used. In the higher level aggregation step, a simple weighted average was used to summarize the six main quality dimensions (Basic and Advanced Functionality, Usability, Accessibility, Efficiency, Maintainability) into a global site quality score S.

Linguistic-Quantifier Guided Aggregation and OWA Mapping

To define the aggregation function AGRR() above, it is advisable to collect opinions from quality experts in the fields involved by the quality dimensions under examination – that means both technological experts and domain specific experts (in the current case, for cultural institutions). Their opinions are likely not to be expressed in a direct mathematical way, but rather with more generic sentences like "if *at least some* of the values to be aggregated are satisfactory, the aggregation score is satisfactory." A relevant feature of OWA operators (see Appendix A) is that they permit to implement the so-called *Linguistic Quantifiers* such as *many, most, at least, about, ...* in the aggregation process. That is, they permit to express, in a mathematically transparent way, sentences like the one above (21).

Therefore, sub-trees in Table 2 were carefully examined, and for each one of them an aggregation criterion was fine-tuned according to its characteristics and to domain expert opinion, using a qualitative linguistic expression (see examples below). Then, the OWA operator corresponding to that criterion was defined and applied to the values of the group.

By tuning *orness* and *crispness* of OWA operators (see Appendix A), it is possible to design a large variety of operators and have a fine control of their behavior. In particular, both Regular Increasing (Decreasing) Monotone Quantifiers and Regular Unimodal Quantifiers (21) were used – for RIM and RDM Quantifiers it was found that suitable functions were $Q(x)=x^{1.75}$ and $Q(x)=1-(1-x)^{1.75}$

This step requires managing human expert opinions, and it is therefore a further source of *researcher arbitrariness* (see sect. 2 above). OWA operators permit to manage this arbitrariness highlighting the role of human domain expert choices in a transparent way, so that their influence on the final quality judgment can be clearly debated.

Some Examples

Let us exemplify some of the OWA operators that were used. For the sake of clearness, here we report only examples of quality score input vectors with binary values, such as A = (0,1,1,0,0). Nevertheless, OWA operators work properly on input values <1 as well (as most of the intermediate calculated group-values are).

OPERATOR 1

Linguistic quantifier → **Two or three of five**

It means that one positive value is encouraging, two positive values are sufficient, three positive values represent full satisfaction, further positive values are not relevant.

OWA op → W = (0.4, 03, 0.3, 0, 0)

Orness(W) → 0.78

Character → Very tolerant

e.g. applied to → Group 1.1.3 – Spread of cultural content and collection presentation

In this group we check five different ways of presenting cultural institution content – rooms' clickable maps, navigation based on collections, on fixed arguments, on authors, or on different criteria. The presence of multiple ways of navigating collection content is an added value, but it is a non-sense to ask for the presence of all possible criteria, in order to obtain a full satisfaction.

Example A=(1,0,0,1,1) – The site provides a collections navigation apparatus based on clickable maps, on authors and on another criterion not foreseen in our questions.

Score=1 (full satisfaction, even if only three criteria are met)

OPERATOR 2

Linguistic quantifier → **Some items**

It means that the presence of some positive values gives a sufficient relevant contribute, but the full satisfaction is obtained when all quality characteristics are fit. We believe that it is sufficient to fit some of the input criteria.

OWA op → W = (0.323, 0.268, 0.208, 0.141, 0.060)

Orness(W) → 0.66

Character → Tolerant

e.g. applied to → Group 2.1 – Services for common users

In this high level sub-free group, we group the scores for on-line educational services, on-line reservation services, transaction management, support to cultural tourism and Web communities. They are advanced and sometimes expensive services; they are a promising direction for increasing museum Web quality, but still in progress. The presence of some of them represents a good direction for its growth..

Example A=(1,0,0,0,1) – The site proposes good on-line educational services and support to Web communities.

Score=0.591 (sufficient, even if only two criteria are met)

OPERATOR 3

Linguistic quantifier → **Most of items**

It means that a good score is obtained only when most of the input data have positive values. We believe that most of the input criteria have to be satisfied.

OWA op → W = (0.088, 0.209, 0.307, 0.396)

Orness(W) → 0.33

Character → Intolerant

e.g. applied to → Group 4.3 – Screen and visual accessibility

*In this group we check if text characters are zoomable, if links are distinguishable from non-clickable text, if the site is still usable at 800*600 resolution, and if color is not indispensable. We require most of the criteria to be satisfied to obtain a sufficient score.*

Example A=(1,1,0,1) – Three criteria are satisfied but the site is strongly dependent on 1024*768 screen resolution

Score=0.604 (just sufficient, even if almost all criteria are met)

e.g. applied to → group 5.3 – Proper use of tags TITLE and META

In this group we check the proper use of TITLE, META/keywords, META/content-type, META/description. A good developing team should have strict policies about this use.

Example A=(1,1,0,0) TITLE and META/keywords are used, but not the others.

Score = 0.297 (not sufficient, even if half of criteria are met)

OPERATOR 4

Linguistic quantifier → **About three of four (or at least three of four)**

It means that one or two positive input values give an irrelevant contribute, three positive values represent good satisfaction, one more positive value represents full satisfaction. We believe that the presence of at least three good quality values is indispensable.

OWA op → W = (0, 0.2, 0.6, 0.2)

Orness(W) → 0.33

Character → Intolerant and quite crisp (in fact, note that the operator is an Unimodal one, and shows a peak around the third position – it has a low *dispersion)*

e.g. applied to → Group 1.1.1 – Information about visiting the cultural institution

In this group we check the presence of information on opening days and hours, tariffs, use of public and private means of transport. They are indispensable pieces of information, thus, we ask for the presence of at least three of four Y/N answers.

Example A=(1,1,0,0) – Only information on opening days and tariffs is given

Score=0.2 (not sufficient, even if half of criteria are met)

e.g. applied to → Group 6.1 – Code quality and standard suitability

In this group we ask if the site uses CSS, if it has a reduced number and type of HTML non-confirmities, if it doesn't use frames, if it doesn't use or proprietary technologies for the main navigation menus. We expect a good site to fit at least three of these four quality points. If not, we decide to penalize its score.

Example A=(1,1,0,1) The site uses frames but completely fits the other three items.

Score = 0.8 (good satisfaction)

FQT4WEB AT WORK

FQT4Web methodology was applied to 15 major European cultural institution sites, in particular museum sites – English, French, German, Italian, Spanish ones – examined in their native languages. Two different expert inspectors, one with technological expertise and one with humanistic background and Museum expertise, examined each site. Each inspector answered only the questions inside his/her own competencies, respectively about 40 and 120. The experimental investigations required some hours for each site (three to six hours, depending on site complexity).

The sites were evaluated in January - February 2004. Some sites (see below) were evaluated again in June - September 2004, after a site redesign - Table 3 below shows the last results.

Overall Site Score

Table 3 shows the total quality scores for the sites, reported from 1 to 100.

TABLE 3
Scores for Examined Sites, On Basis 100 Points

Site id	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Score	88	77	76	71	64	63	63	63	62	60	60	57	52	51	49

It can be noted that the scores are largely scattered between 50 and 90 points. This accounts for a good ability of *FQT4Web* in capturing different site quality behaviors – it is a selective tool.

It is impossible to fulfill completely the theoretical quality criteria captured in the 160 atomic questions – and maybe, this should not be included in the objectives of site staff. Therefore, it can be considered that a score of 85 points and over represents excellence. 70-85 points mark a very good site, 60-70 reveals a good site where some problems could be corrected, while under 60 points the site lacks of various quality aspects.

Quality Shapes and Site Goals

A deeper insight may be gained by examining each of the six quality dimensions that contributes in creating the overall score. This step can help site managers to identify the specific lacks of specific quality dimensions according to their particular goals, and, if needed, to fix them.

Data for four selected sites from Table 3 are reported in Table 4.

In Figure 3, four charts show the six main quality dimensions of FQT4Web in detail (Basic Functionality, Advanced Functionality, Usability, Accessibility, Efficiency, Maintenance-Compliance) for sites A, C, E, M – one chart for each site.

These charts can be considered as the *Quality Shapes* for the sites. The greater the coloured area of the chart is , the higher quality has the site. The more regular the hexagonal shape is, the more well-balanced are the quality features of the site.

- National Gallery (London, UK) has almost full scores in each of the six quality dimensions – the site has excellent and sound quality features.
- Louvre Museum site (Paris, France) proposes excellent basic services (such as information on visits and activities, multimedia presentation of museum collections, translation in various languages). It pays high attention to user-site interaction (usability, accessibility) and has good technological features (efficiency, maintainability). However, the advanced features are not equally good – for instance, there are no educational and interactive on-line proposals, nor promotion of users communities by using

newsletters or feedback forms. We can infer that the Louvre has a very good *traditional* site. The limited presence of advanced features is not a lack *per se* – it depends on the objectives of the Louvre managers. (Actually, it must be said that some of the latter considerations hold no more at the time of paper publication, as the Louvre site was enriched with some services, e.g. in educational field, after our analysis in January 2004).

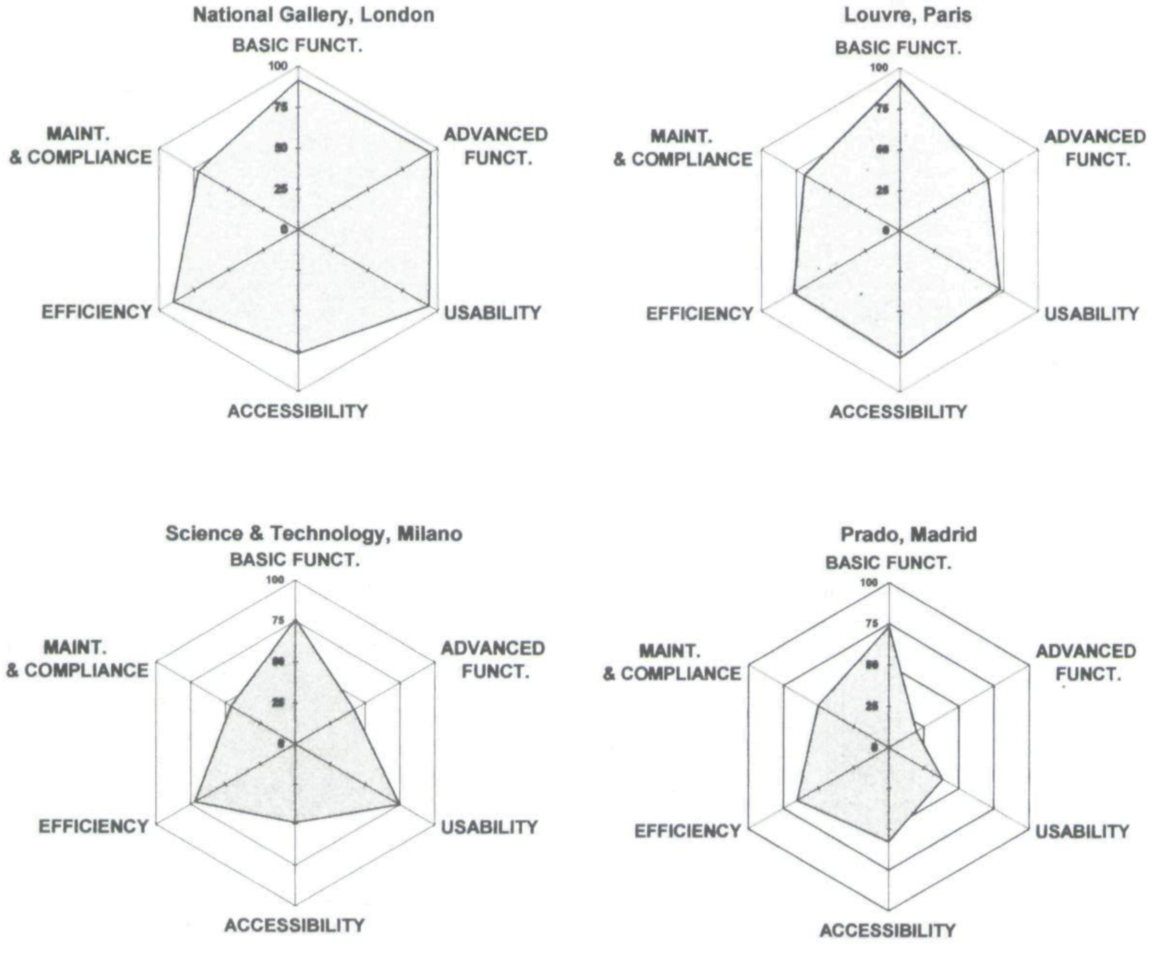
- The Quality Shape of Science & Technology Museum (Milano, Italy) is almost triangular, because the museum has very good scores for Basic Functionality, Efficiency, and Usability, but scarce features for the other quality dimensions. However, while the limited presence of Advanced Functionality features could be a strategic choice made by the site decision-makers, the scarce scores in Maintainability and Accessibility dimensions point out some problems that should be fixed.

The last one is the Prado's site (Madrid, Spain). Its poor overall score is obtained in spite of good Basic Functionality features and a relatively good Efficiency score. The site is surprisingly scarce from the Usability point of view (no navigation support, inconsistent use of colors and interaction styles) and Accessibility (dependence on client side scripts, missing ALT attributes). Moreover, the site has almost no Advanced Features.

TABLE 4
Identification Data for Selected Sites

Site ID	Quality Score	Cultural Institution	URL
A	88, excellent	National Gallery, London	http://www.nationalgallery.org.uk/
C	76, very good	Louvre, Paris	http://www.louvre.fr/
E	64, good	Science and Technology Museum, Milan	http://www.museoscienza.org/
M	52, quite poor	Prado, Madrid	http://museoprado.mcu.es/

FIGURE 3
The Six Main Quality Dimensions of FQT4Web as Calculated in Four Selected Sites



Some of the experimental findings are quite surprising, when we consider that major national sites were examined. In fact, while the lack of specific services could be derived from strategic choices by Web staff, or moreover, from budgetary constraints, poor features in usability and in technological standard adherence are less understandable. Probably, a more mature approach to the design of robust and standard applications is still in the growing phase.

Examining three sites before and after a redesign

Let us now examine the three case studies we started with, in section 1, Figure 1, and let us answer to the questions there proposed (save for budget question, obviously):

- Home page changes are a matter of restyling or do they actually correspond to an increase in site quality?
- Which specific quality sides increased during the redesign, and how much?

The three sites are Italian ones – with complete (site B) or extensive (sites H and I) English translation. They were examined in early 2004 and again in late 2004 after a site redesign.

Table 5 reports identification data for the sites, and their overall score *before* and *after* the site redesign. Figure 4 reports their Quality Shapes, each one showing the comparison between the results of the *FQT4Web* methodology *before* (yellow) and *after* (orange). These data permit a detailed evaluation of the redesign step results, as follows.

TABLE 5
Identification Data for Three Redesigned Sites, and Overall Quality Score Before and After

Site ID	Quality Score Before/After	Cultural Institution	URL
B	63 (before) 77 (after)	Institute of the History of Science, Florence	http://www.imss.fi.it/
H	55 (before) 63 (after)	Brera Gallery, Milan	http://www.brera.beniculturali.it/
I	53 (before) 62 (after)	Archaeological Museum, Bologna	http://www.comune.bologna.it/ museoarcheologico/

History of Science Institute site has an evident increase in quality (an increased hexagonal area, from score 62 to score 77, a very good one) with a better balancing between the various quality dimensions as well (a more regular hexagon). Only one quality dimension remains underrated in both site versions, and slightly gets worse after redesign, i.e. Maintainability – the new site release has better (X)HTML coding and CSS support, but it has unexpected navigation bugs in some higher level pages and an increased URL *opacity* (see URL examples in Section 4).

In Brera Gallery site, the comparison of the Quality Shapes between *before* and *after* clearly shows that four quality dimensions remain practically unchanged, while there are significant increases in Accessibility and Maintenance/ Compliance (but slower pages, and a slightly lower Efficiency score). We can argue that the redesign is basically a technological one, and that site functionality and usability were not (or secondarily) involved in the process. The overall site score increases 8 points, from 55 to 63.

In Archaeological Museum, Bologna, both Basic and Advanced Functionality significantly improved, and there is a slight increase in Efficiency and Usability (some features improved, other worsened). Other technological features remain almost unchanged – on the contrary, Javascript has become necessary in various pages, corresponding to an Accessibility lower score.

In all of the three sites, open problems and quality improvements are clearly identified by *FWT4Web* methodology; therefore, the sites staffs are able to focus the attention on such matters.

CONCLUSIONS AND FUTURE WORK

FQT4Web (Fuzzy Quality Tree for Web Inspection), a quantitative inspector-based methodology, has been presented, which produces six measures of quality dimensions and an overall quality score for a Web site. It was shown how decision-makers may use this methodology to validate site behavior in comparison with their initial goals. Two main novelties of the methodology are a) a strong reference to existing standards in order to put in evidence and limit arbitrariness and subjectivity and b) an innovative use of fuzzy OWA operators in the assessment process to set aggregation criteria through qualitative linguistic quantifiers. OWA operators permit to manage the existing arbitrariness highlighting the role of human researcher choices in a transparent way, so that their influence on the final quality judgment can be evidenced.

The methodology was developed for cultural Web sites, but it is easily adaptable to other domains. In particular, the assessment method is completely general-purpose.

The use of fuzzy logic in Web quality assessment is in its early stage, and it promises interesting results. We are going to examine how analyses results are changed when the relative importance of the various quality criteria is taken into account, and to compare the suitability of different fuzzy aggregators and of rule-based fuzzy systems. A related system which seems to be promising in such context is Analytic Hierarchy Process – e.g. see (10).

One problem affecting evaluation methodologies, including *FQT4Web*, is the unknown level of trustworthiness of their results. Someone could wonder how much we can trust in the investigation results, i.e. how much the final judgment could have been changed if different evaluators with different preferences examined the site, or if a different summarization method was used. Preliminary quantitative investigations on *FQT4Web* indicate that reasonable changes in the summarization method and in inspector choices do not distort the general results. The final site score could change, but not radically – a good (or poor) site remains good (or poor) even when diverse inspectors examined the site or different summarization choices were made.

APPENDIX A – DOSSIER: OWA OPERATORS

In the following, the definition and the main characteristics of the OWA (Ordered Weighted Averaging) operators, which are relevant for our purposes, are briefly summarized.

Definition

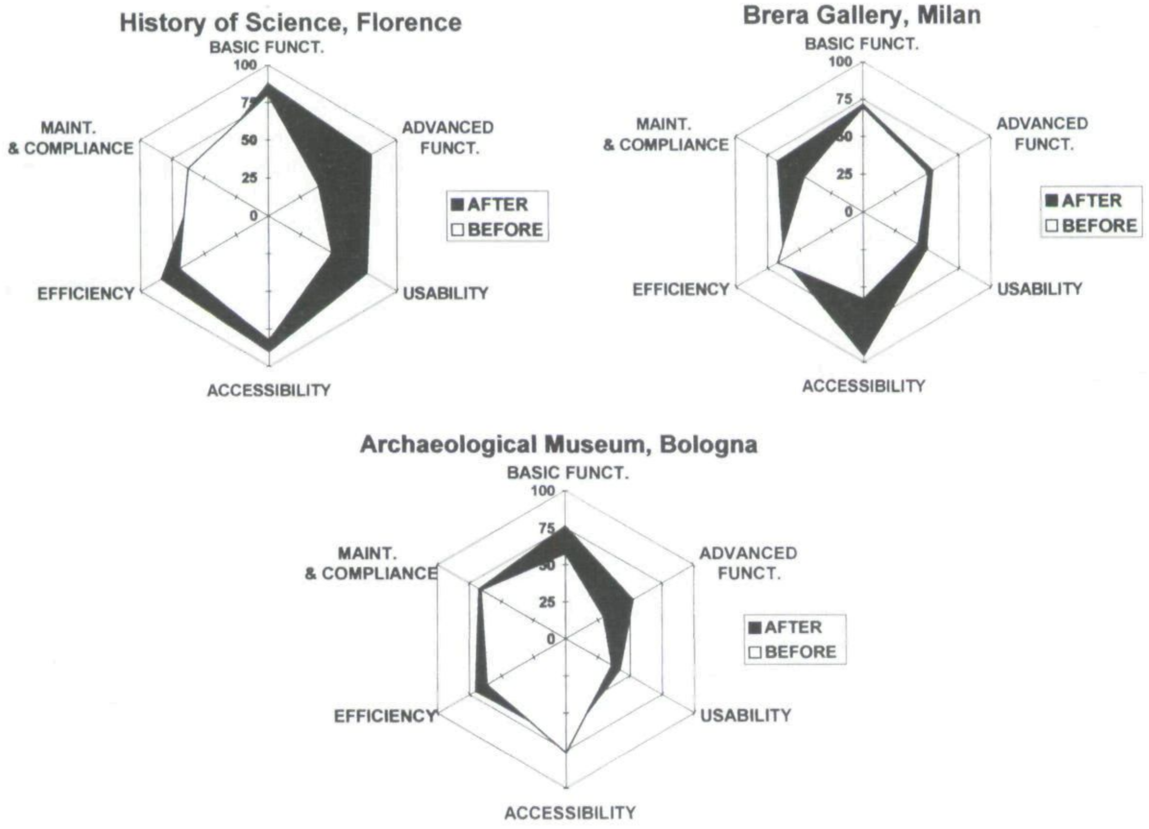
Let us have a vector of input values A = (a_1 , ..., a_n) in the range (0,1), which e.g. represents a set of normalized quality values nqv_i measured for a group of elementary properties. An OWA operator, as defined by Yager (20), is a mapping $R^n \rightarrow R$ that has an associated vector of weights W = (w_1 , ..., w_n) so that

- each w_i is in the range (0,1),
- the sum of all w_i is 1,
- and the mapping of the input values A in the range (0,1) is obtained according to

$$S(A,W) = \sum_i w_i * b_i$$

where b_i is the largest *i-th* element in the collection a_1 , ..., a_m, i.e. where b_1 , ..., b_n is substantially the collection of input values a_1 , ..., a_n after a descending ordering step. Even if the above summarization formula is the same as the weighted average, the reader should carefully note that the meaning of the weights is completely different. The key point is ***the reordering step*** – the weights w_i do not refer to input data relevance (as for the weighted average), because they are not associated with a particular input element a_i, but with a particular position *i*.

FIGURE 4
Quality Shapes Before (Yellow) and After (Orange) Site Redesign for the Three Sites



Tolerant/Intolerant Behavior

EXAMPLE 1

OWA operator	W=(0.4, 0.3, 0.2, 0.1)
Input data	A=(0.25, 1, 0, 0.75)
Result	The application of this OWA operator W to the input values A gives the result S=0.4*1+0.3*0.75+0.2*0.25+0.1*0=0.675
Comment	The reordering step for the input vector A guarantees the *higher weight* in the *first* position of weighting vector W to be associated to the *higher input value* in the vector A; the second higher weight in the second position of weighting vector W will be associated with the second higher input value in the vector A, and so on.

EXAMPLE 2

OWA operator	W=(0.1, 0.2, 0.3, 0.4)
Input data	A=(0.25, 1, 0, 0.75)
Result	The application of this OWA operator W to the input values A gives the result S=0.1*1+0.2*0.75+0.3*0.25+0.4*0=0.325
Comment	The reordering step for the input vector A guarantees that the *higher weight* in the *last* position of weighting vector W will be associated with the *lower input value* in the vector A; the second higher weight in the second-last position of weighting vector W will be associated with the second lower input value in the vector A, and so on.

With the OWA operator in Example 1, we overweight the best experimental results; hence, we can obtain a good aggregation score, even if we start with a set A, which is made of both good and poor values, i.e. even if only *some* of the input values are good. From an intuitive point of view, we adopt here a *tolerant behavior* – we accept that only some criteria are satisfied; but not completely tolerant – a single positive value does not represent full satisfaction.

With the OWA operator in Example 2 we underweight the best experimental results; therefore, we can obtain a good aggregation score only if the lower input values are good, i.e. only if *most* of the input values are good. From an intuitive point of view, we adopt an *intolerant behavior* – we demand that most criteria are satisfied; but not completely intolerant – the presence of some positive input values results in a limited degree of satisfaction. (The terms tolerant and intolerant are borrowed from Marichal (12), who applies them to the Choquet integral).

Orness and Crispness

This behavior is related to the concept of *orness* of the OWA operators (20), which is a real number in the range (0,1) that measures the OR-like behavior of the operator, as follows:

$$Orness(W) = (1/(n-1))\Sigma_i w_i * (n-i)$$

OWA operator in Example 1 has orness = 0.67 (>0.5, OR-like), while OWA operator in Example 2 has orness = 0.33 (<0.5, AND-like). The closer the weights in the OWA operator are in the left side of the vector, the closer the operator acts like a pure "OR" operator. Reciprocally, the closer the weights in the OWA operator are in the right side of the vector, the closer the operator acts like a pure "AND" operator. The operator W=(0.25, 0.25, 0.25, 0.25) acts as the usual arithmetic average for a set of four input values A, and has orness=0.5

The OWA operators for a set of input values are infinite, and the choice of a specific set of weights is related to the aims of the human decision-maker. It is relevant to note that the OWA operators with a given orness are infinite. E.g. the two weighting sets W1=(0.33, 0.33, 0.33) and W2=(0, 1, 0) have the same orness=0.5, but they show different behavior – the second is more "selective" and is a "crisp" operator. They have a different degree of dispersion around a specific element, a concept related to the Shannon information concept (20).

ACKNOWLEDGMENTS

We acknowledge the contribution of Du Tu Tran Anna, Faculty of Humanities, and Marcello Nuccio, Dept. of Information Engineering, University of Modena-Reggio Emilia. Their suggestions and efforts in the experiments were greatly appreciated. We also thank Michele Colajanni, Professor of Computer Networks at the University of Modena and Reggio Emilia, and Silvio Giove, Associate Professor of Decision Methods and Economy of Complex Systems at the University of Venice, for their useful suggestions on a preliminary version of the paper.

REFERENCES

1. Atzeni, P., P. Merialdo, and G. Sindoni. "Web Site Evaluation: Methodology and Case Study," **Lecture Notes in Computer Science 2465,** Springer-Verlag, 2002, pp. 253-263.
2. Barnes, S.J. and R.T. Vidgen. "An Integrative Approach to the Assessment of E-commerce Quality," **Journal of Electronic Commerce Research,** 3:3, 2002, pp. 114-123.
3. Corradini, E., P. Davoli, and A. Russo. "A Quality Tree for Cultural Web Sites Inspection," **8th International Cultural Heritage Informatics Meeting,** Berlin, August 2004.
4. Davoli, P., E. Corradini, E. Garzillo, M. Nuccio, and A. Russo. "Inspection of Museum Web Application Quality-Analysis of Selected European Sites." In Bearman, D, and J. Trant (Eds.). **Museums and the Web 2004: Proceedings CDROM.** Toronto: Archives & Museum Informatics, March 2004.
5. Dujmovic, J.J. "Extended Continuous Logic and the Theory of Complex Criteria," **Univ. Beograd. Publ. Elektrotechn. Fak,** 1975, pp. 197-216.
6. Herrera-Viedma, E. "Fuzzy Qualitative Models to Evaluate the Quality on the Web," **Lecture Notes in Artificial Intelligence 3131.** Springer-Verlag, 2004, pp. 15-27.
7. Herrera-Viedma, E., E. Peis, M.D. Olvera, J.C. Herrera, and Y.H. Montero. "Evaluating the Informative Quality of Web Sites by Fuzzy Computing with Words," **Lecture Notes in Artificial Intelligence 2663,** Springer-Verglag, 2003, pp. 62-72.
8. ISO/IEC 9126-1:2001. **Software Engineering-Product Quality-Part 1: Quality Model.** Intl. Org. for Standardization, Geneva, 2001.
9. Jacobs, J. (Ed.). **Architecture of the World Wide Web.** W3C Technical Architecture Group, W3C Working Draft, Dec. 2003. http://www.w3.org/TR/webarch/
10. Kim, G.M. and G.S. Lee. "E-catalog Evaluation Criteria and Their Relative Importance," **Journal of Computer Information Systems,** 43:4, 2003, pp. 55-62.
11. Lewis, C. and J. Rieman. **Task-centered User Interface Design: A Practical Introduction.** Boulder: University of Colorado, 1994.
12. Marichal, J.-L. "Tolerant or Intolerant Character of Interacting Criteria in Aggregation by the Choquet Integral," **European Journal of Operational Research,** 155:3, 2004, pp. 771-791.
13. Mich, L., M. Franch, and L. Gaio. "Evaluating and Designing Web Site Quality," **IEEE Multimedia,** 10:1, 2003, pp. 34-43.
14. Minerva. **Handbook for Quality in Cultural Web Sites,** Ver. 1.2., November 2003. http://www.minervaeurope.org.
15. Olsina, L. and G. rossi. "Measuring Web Application Quality with WebQEM," **IEEE Multimedia,** 10:1, 2003, pp. 34-43.
16. Ramler, R., W. Schwinger, and J. Altmann. "Testing Web Quality Aspects," Software Competence Center, Hagenberg, Austria, Technical Report SCCH-TR-0216,

2002.

17. Ruiz, J., C. Calero, and M. Piattini. "A Three Dimensional Web Quality Model," **Lecture Notes in Computer Science 2722,** Springer-Verlag, 2003, pp. 384-385.
18. Scheffelmaier, G.W. and J.S. Vinsonhaler. "A Synthesis on Research of the Properties of Effective Internet Commerce Web Sites," **Journal of Computer Information Systems,** 43:2, 2002, pp. 23-30.
19. Shchiglik, C. and S.J. Barnes. "Evaluating Website Quality in the Airline .Industry," **Journal of Computer Information Systrems,** 44:3, 2004, pp. 17-25.
20. Yager, R.R. "On Ordered Weighted Averaging Aggregation Operators in Multicriteria Decision Making," **IEEE Trans. Systems, Man Cybernet,** 18:1, 1988, pp. 183-190.
21. Yager, R.R. "Quantifer Guided Aggregation Using OWA Operators," **Internatl. J. Intel. Systems,** 11, 1966, pp. 49-73.
22. Yeung, W.L. and M.T. Lu. "Gaining Competitive Advantages Through a Functionality Grid for Web Site Evaluation," **Journal of Computer Information Systems,** 44:4, 2002, pp. 67-77.
23. Zeldman, J. **Designing with Web Standards.** Indianapolis, IN: New Riders Press, 2003.