Toward Metrics and Model Validation in Web-site QEM

Luis **Olsina** [1 2] Claudia **Pons** [2] Gustavo **Rossi** [2 3]
[1] *GIDIS, Department of Computer Science, Engineering School, at UNLPam*
Calle 9 y 110 - (6360) General Pico, LP - Argentina
E-mail olsinal@ing.unlpam.edu.ar
[2] *LIFIA, Informatics School at UNLP (*[3] *also at CONICET)*
Calle 50 y 115 ,1P - (1900) La Plata - *Argentina*
E-mail [grossi, cpons] @info.unlp.edu.ar

Abstract. In this work, a conceptual framework and the associated strategies for metrics and model validation are analyzed regarding website measurement and evaluation. Particularly, we have conducted three case studies in different Web domains in order to evaluate and compare the quality of sites. For such an end the quantitative, model-based methodology, so-called Web-site QEM (Quality Evaluation Methodology), was utilized. In the assessment process of sites, definition of attributes and measurements, preference criteria for elementary evaluation, and an aggregation model of attributes and characteristics for global evaluation have intervened. Hence, in the present paper, the validation framework and the theoretical validation of some used Web metrics and model for assessment purpose are discussed considering the representational theory of measurement.

Keywords. Web Metrics, Models, Web-site QEM, Theory of Measurement, Validation.

1. Introduction

Validation of software metrics is a very important process, but not an easy one. Measures must represent accurately those attributes they intend to quantify. So validation is key to the success of assessment and prediction processes. Although in several traditional disciplines like physic and classic engineering, the evolution, employment and validation of metrics have been developed along decades and centuries so that many of the metrics are today incorporated in daily life as measures of temperature, speed, distance, among others, without nobody doubting about the validity of them, the same thing doesn't happen in Software Engineering where it is still debated if there exists enough understanding on some of those popularly used as for example, Albrecht function points. Kitchenham *et al*. [5] refute the mentioned validity of this metric since it doesn't fulfill the basic principles of the representation condition, arguing that it violates the arithmetic operations imposed by the scale type constrains. Nevertheless, direct metric to measure the size of a source program (LOC Metric), metrics to measure density of defects, among others, have already been validated theoretical and empirically [3, 12]. In addition, predictive models like COCOMO, used in the determination of the development effort (and cost) of a software system, have already been validated.

On the other hand, the processes and products in the fields of Hypermedia and Software Engineering in the Web are rather recent, so that it is necessary a lot of hard research to understand, evaluate and validate. Web metrics validation has often been neglected in the metric community. In fact, there is no doubt that the activity of validation of attributes and metrics in the Web, as those employed in our task of site evaluation by means of Web-site QEM [7, 8, 9], is an important process. Thus, we will try to perform initial contributions based on the representational theory of

measurement [3, 5, 10, 11, 12].

In general terms, validation can be defined as the process of assuring that the measure is an appropriate numeric (or symbolic) characterization of the attribute of an entity, showing that the representation condition is satisfied. This is to say, the mapping between the empirical domain (the empirical world) and the new numerical domain (the formal world) preserves the relationship so that studying and analyzing the numbers, the entity of the empirical (and real) world can be explained or surmised. Hence, the homomorphism condition must be preserved [11, 12]. In the present paper, we will focus particularly on the theoretical validation of metrics useful to evaluate attributes of existent entities (sites) and we won't deal with the validation of predictive models for the Web. Indeed, this is an open line for future investigations.

The structure of this paper is as follows: in the next section, we describe a conceptual framework for metrics and models validation grounded in the representational theory of measurement. In Section 3, the theoretical approach is used in order to validate some Web metrics utilized in case studies. Besides an aggregation model used in Web-site QEM is validated as a ratio scale. Finally, Section 4 summarizes the paper and draws our conclusions.

2. A Conceptual Framework for Metrics and Models Validation

2.1 An Introductory Example.

As Fenton *et al*. [3] said, "*the representation condition asserts that a measurement mapping m must map entities into numbers and empirical relations into numerical relations in such a way that the empirical relations preserve and are preserved by the numerical relations*"..."*For the (binary) empirical relation 'taller than' we can have the numerical relation* $x > y$ *. Then, the representation condition requires that for any measure m, A is taller than B if and only if* $m(A) > m(B)$ *"* (pp. 31-32).

Let us consider the *Scoped Search* attribute (for museum collections) used in [7]. Observing and using a Web site, we can understand that categories exist for the search mechanism that facilitates users with more or less functionality in the search process. Therefore, our initial understanding of the attribute from the empirical domain (experiencing with sites), can take us naturally to categorize the functionality in three unary relations. Namely, SF1, (no scoped search mechanism available -m1, for short); SF2, (basic scoped search functionality available, i.e., by author or title -m2), and SF3, (expanded scoped search functionality available, i.e., by author or title, by school, and/or style, and/or century (or date), and/or painting, and/or medium -m3). Then, we can assume that each observed mechanism is either SF1, or SF2, or SF3. Also, let us suppose by the time being that we don't consider the level of functionality that members of the classes imply, but we want to map members of the classes of the empirical domain, in three different real numbers (or labels) in the numerical (or symbolic) domain. For example, it can assign the following mapping:

$m(m1) = 5;\ m(m2) = 1;\ m(m3) = 2;$ (1)

This assignment is a representation (in the nominal scale type), because we have a one-to-one equivalence corresponding to SF1, SF2, and SF3 respectively. This is, the numerical representation corresponding to SF1 is the relationship "it is 5", and so on. However, if we want to indicate to the previous representation so that it expresses the binary empirical relationship "it is more functional than" in order to establish a ranking, knowing that m3 is more functional than m2, and that m2 is more functional than m1, then, we have to look for a more appropriate mapping. A more appropriate numerical representation that (1) is the following one (in the ordinal scale type):

$m(m1) = 0;\ m(m2) = 1;\ m(m3) = 2;$ (2)

This numerical representation expresses in the formal (mathematical) domain an order relationship starting from the binary empirical relationship "x is more functional than z ".

2.2 The Validation Framework.

In this section, we will discuss a conceptual framework, some properties, and axioms useful to validate metrics, according to previous investigations [1, 3, 5, 10, 12]. Fig. 1, depicts a schema of the conceptual model for direct metrics, representing the main classes and relationships. Their contribution to the validation process will be highlighted.

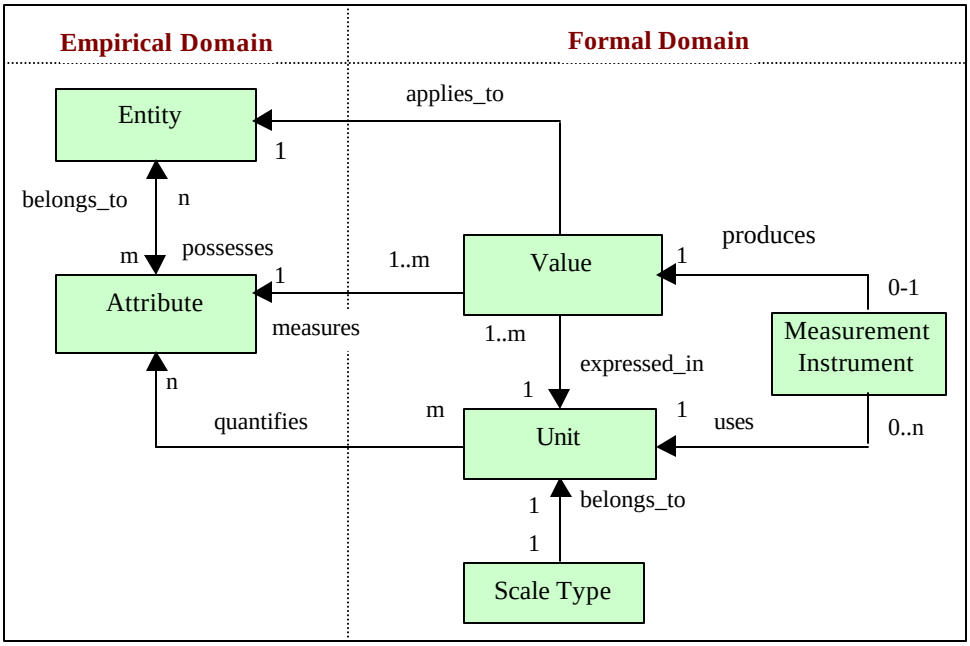

Fig. 1 *Conceptual model for direct metrics.*

2.2.1 Empirical and Formal Domain and Relational Systems. As aforementioned, there exist a theory, denominated theory of measurement that declares how to combine empirical conditions with numerical conditions under the assumption of logical equivalence or homomorphism. Extensions of the foundations of this theory with implications for software engineering are well analyzed and documented in the Zuse book [12].

For instance, lets S be a set of entities (Web sites) where X is an observable attribute so that x1, x2 belong to S1, S2 respectively. Thus, the binary empirical relationship x1 •> x2 holds if and only if it is judged that x1 is more functional than x2. Therefore, it is wanted to assign a real number m(x1) and m(x2) to each x1, x2 such that for all pair belonging to S, it is fulfilled that,

x1 •> x2 <=> m(x1) > m(x2) (3)

This statement is the base of measurement, and it can be read in two ways: first, if x1 is more functional than x2, it implies (=>) in the formal domain that m(x1) is greater than m(x2); and vice versa, if m(x1) is greater than m(x2), then it implies (<=) that x1 is more functional than x2 (in the empirical domain). This double implication is called homomorphism. Besides, in (3) the empirical statement is: x1 •> x2; and the numerical statement is: m(x1) > m(x2), being •> and > the respective relational operators. It is important to notice that empirical statements are not true *per se*; they can be falsified by means of observations. On the other hand, for the empirical relationship •> (or •>=) there can be several interpretations, according to the case. Examples are: equal or more difficult of understand; equal or higher level of availability, equal or higher level of errors, etc.

Moreover, the concepts of empirical and formal relational systems, and the concept of metric (or measure) can be introduced. For the ranking order, the empirical relational system is defined as: **S** = (S, •>=); and the numerical relational system is defined as: **N** = (R, >=), where R are the real

numbers. Then, we can write for the metric m, the following expression:

(**S**, **N**, **m**) = ((S, •>=), (R, >=), m) , (4)

where a metric is a correspondence (or mapping) m: S -> R in which the (3) expression is fulfilled.

2.2.2. Entities, Attributes and their Relationships. From the evaluation standpoint, in the empirical domain, we have Entity and Attribute classes. The Entity can be decomposed basically in three main classes of interest for evaluators, that is: a) *Process*: it is the entity possibly compound of other sub-processes and activities, used to produce artifacts; b) *Artifact*: it is the temporary or persistent entity representing the product of performing a process, and c) *Resource*: it is an entity required by a process as input to produce some specified output (resources of a project are: human, monetary, materials, technological, temporal).

The Attribute class, represents what is observed and attributed regarding what is own of an entity of the real world, being an object of interest to be evaluated. Attributes can be measured by direct or indirect metrics. For a given attribute, there is always at least an empirical relationship of interest that can be captured and represented in the numerical domain, enabling us to explore the relationship mathematically. Fig. 1 shows a many-to-many relationship between Entity and Attribute classes. That is, an entity can possess several attributes as long as an attribute can belong to several entities.

2.2.3. Value, Unit, Scale Type and their Relationships. When we measure a specific Attribute of a particular Entity, we consider a Scale Type and Unit in order to obtain magnitudes of type Value. A Unit of measure determines how it should be quantified the attribute of an entity. Therefore, the measured value can not be interpreted unless we know to what entity is applied, to what attribute is measured and in what unit is expressed (i.e., it should be clearly specified the empirical and numeric relational systems). Value (or scale) and scale type are two different entities that frequently are confused. The concept of scale is defined by the triple (**S**, **N**, **m**) - see (4). So the scale is defined by homomorphism. Clearly, it is appreciated that the empirical relational system **S**, the numerical relational system **N**, and the metric **m** are needed in order to obtain a value or scale. On the other hand, a scale type is defined by admissible transformations. An admissible transformation is a conversion rule f in which given two measures m and m', it keeps that m' = f m. For example, admissible transformations are m' = a m + b, with a > 0 and b ∈ R; m' = a m, among others. The scale type doesn't change when is performed an admissible transformation.

Besides, the scale type of a measure affects the sort of arithmetical and statistical operations that can be applied to values. As said Zuse, scale types as nominal, ordinal, interval, ratio and absolute scales are hierarchically ordered and can been seen as keywords describing certain empirical knowledge behind values. For example, the nominal scale type implies a very simple empirical condition: the equivalence relationship. Lets be the empiric relational system (S, ≈), and given an observable attribute so that x1, x2 belong to S1, S2, then a function exists m: S -> R, that:

x1 ≈ x2 <=> m(x1) = m(x2); and (**S**, **N**, m) = ((S, ≈), (R, =), m), is a nominal value (5)

This was exemplified in Section 2.1. In addition, for the ordinal scale type, the empirical relational system of the nominal one is extended to reflect the ordinal scale as it is expressed in (4). Zuse indicates that the weak order is a prerequisite for ranking order measurement, which is transitive and complete. So we can express these properties as:

x1 •>= x2, and x2 •>= x3 => x1 •>= x3, (6) Transitivity

x1 •>= x2, or x2 •>= x1 (7) Completeness

Reassuming with the *Scoped Search* attribute, the metric produces an ordinal value as expressed in (2) and, therefore, it satisfies the properties (6) and (7). So, for the transitivity property, combining both relational systems, the following expression is fulfilled:

x1 •> x2, and x2 •> x3 => x1 •> x3 <=> m(x1) > m(x2), and m(x2) > m(x3) => m(x1) > m(x3); where m(x1) = 2, m(x2) =1, m(x3) = 0; and the operator •> means "it is more functional than".

Finally, the ratio scale type is very well known in physics and traditional sciences; e.g., longitude, money measures, among others. Zuse says, "*we want to have something above poor ranking or comparing of objects. We want to be additive in the sense that the combination of two objects is the sum of their measurement values*". The idea of a ratio scale is linked to additive and non-additive properties. An additive ratio scale is represented by:

((S, •>=, o), (R, >=, +), m) where o is the concatenation operator

Therefore, (S, •>=, o) is a closed extensive structure iff there exists a function m on S such that for all x1, x2 belonging to S,

x1 •>= x2 <=> m(x1) >= m(x2); and m(x1 o x2) = m(x1) + m(x2)

Also, a function m' exists that is the admissible transformation, i.e.: m'(x) = a m(x) with a > 0. Furthermore, Zuse defines the modified extensive structure and the empirical conditions where specific axioms must be satisfied (see [12], chapter 5). Finally, for an absolute scale type the only admissible transformation is identity, as we will exemplify for two Web metrics, in Section 3.

2.2.4. Measurement Instrument. In order to obtain the measured value one can do it manually, or using partial o totally a Measurement Instrument. Fig 1 shows that an instrument can be optional.

***2.2.5. Indirect Metrics*.** A direct metric of an attribute of an entity involves no other attribute measure. However, we can obtain values from equations involving two or more attributes measures. In the empirical domain, we have an Attribute Association that is formalized by an Equation in the formal domain. The conceptual model for indirect metrics is not discussed here for space reasons.

***2.2.6. Some Implications of the Conceptual Model and Properties for Validation*.** The model and measurement properties above introduced have several implications for validation, among them:

A. Different entities (or sub-entities) can share the same attribute. In the practice, any conceptual framework that groups attributes as belonging to a single entity type, it should not imply, in general, a many-to-one relationship.
B. Since a measure makes correspond values to an attribute, then if the representation condition is satisfied, the behavior of the attribute in the empirical domain should be reflected in the behavior of the new formal domain. Thus, the logical equivalence and axioms [12] should be accordingly satisfied and observed and this can imply a theoretical and/or empirical validation.
C. Since an attribute can be measure in different ways (with different criteria, units and scale types), the definition of an attribute is independent of specific units and scale types [5].
D. A unit can be applied to different attributes that belong to different entities. However, a specific value that doesn't has associated a specific unit is a clear sign that the measure is meaningless. Moreover, a specific measure of an attribute, its unit and scale type, must be clearly defined, in a particular context of evaluation. Accordingly, the acceptable statistical and/or mathematical operations and admissible transformations should be taken into account.
E. Since a direct metric makes correspond values to an attribute, the characteristic of the domain and the range of possible values should be considered. For instance, the elementary measurement criteria and, specifically, if the variable is continuous or discrete should be determined.
F. For an indirect metric, it must intervene a model or equation to calculate the value. Thus, attributes relationships, units and scale types should also be considered.

3. Approches for Web Metrics and Model Validation

Basically, there are two strategies to corroborate or falsify the validity of metrics: the theoretical and the empirical. In turn, two common approaches for validation correspond to internal attributes or external attributes or characteristics of entities. For the former, it can be said that a measure with assessment propose is valid internally or "*valid in the narrow sense*" [3]. Or for the latter, for higher level characteristics (e.g., cost, quality, maintainability, etc.), it can be said that a measure is valid externally, or "*valid in the wide sense*". In this last category, the metric can mainly play a dual role: it can be used for assessment purpose or can be part of a valid prediction model or system.

The theoretical validation allows to confirm that a measure doesn't violate the properties of the empirical and numerical relational systems and the definition models and criteria. The empirical validation allows to confirm a measure by the planning and observation of experiments to see, for example, whether users agree with the existence of some attribute, or whether a mapping to a value captures theirs understanding of the attribute, or if a metric of an internal attribute can be used to predict the value of an external characteristic, among other issues.

In a general sense, the Kitchenham *et at*. assumption is that in order for a measure to be valid these two conditions must be held: 1) the measure must not violate any necessary property of its elements; 2) each model used in the process must be valid. Moreover, according to the proposed conceptual framework in order to decide whether a metric is valid, it is necessary at least to confirm [5]:

- ✓ *Attribute validity,* i.e., whether the attribute is actually exhibited by the entity being measured.
- ✓ *Unit and Scale Type validity,* i.e., whether the measurement unit and scale type being used are an appropriate means of measuring the internal or external attribute.
- ✓ *Instrument validity,* i.e., whether any model underlying a measuring instrument is valid and the same one is properly calibrated.
- ✓ *Protocol validity,* i.e., whether an acceptable measurement protocol has been adopted in order to guarantee repeatability and reproducibility in the measurement process.

3.1 Towards the Theoretical Validation in Web-site QEM

Web-site QEM is a stepwise, quantitative, expert-driven methodology useful to evaluate and compare quality of sites. So far, we have conducted case studies in the museum [7], academic [8], and e-commerce [9] domains. For instance, in the latter study, over a hundred and forty characteristics and attributes were taken into account considering attributes direct or indirectly quantifiable. The steps of the methodology are grouped in the following major technical phases:

1. *Quality Requirement Definition and Specification,*
2. *Elementary Evaluation: Definition and Implementation,*
3. *Global Evaluation: Definition and Implementation,*
4. *Analysis and Recommendations*

In the *"defining the Web-site quality requirement tree"* step, in the first phase, the evaluators should agree and specify the quality characteristics and attributes, grouping them in a requirement tree. In the process, we use the same high-level quality characteristics like *Usability*, *Functionality*, *Reliability*, *Efficiency*, *Portability,* and *Maintainability* in order to follow a well-known standard [4]. From some or all of these characteristics, we derive sub-characteristics, and from these, we can specify attributes with minimal overlap. For each quantifiable direct or indirect attribute A_i, a variable X_i will be associated. It is important to notice that many qualitative and empirical statements and conditions exist in software engineering. For instance, quality characteristics in the ISO 9126 standard are defined by empirical declarations. Likewise, the sub-characteristics and attributes defined in the studies. Empirical statements and conditions are more intuitive for users than mathematical statements.

In the second phase, the evaluators should define the basis for the elementary evaluation criterion (for each attribute), and perform the measurement and mapping process. As said above, for each attribute a variable X_i is associated, which can take a real measured or calculated value. Besides, for each variable it is necessary to establish a criterion function, called the elementary criterion function. This function models a new mapping among the measured or calculated value resulting afterwards in an elementary quality preference. In this way, the scale type and unit become normalized [8]. The elementary indicator or preference EP_i is frequently interpreted as a percentage of satisfied requirements for a given attribute, and it is defined in the range between 0 and 100%.

In the third phase, the *"aggregating elementary preferences to yield partial and global quality preferences"* step should be performed. The decision-makers should prepare and implement the global evaluation process in order to obtain a quality preference indicator for each website. In the process, the type of relationships among attributes, sub-characteristics, and characteristics and the relative weights must be considered. For this purpose, regarding the amount of intervening characteristics and attributes in the studies, it was agreed the use of a robust and sensible model such as the Logic Scoring of Preference (LSP) model [2]. However, in simpler cases a merely additive scoring model can be used. The strength of LSP model over merely additives ones resides in the power to deal with different logical relationships and operators to reflect the evaluation needs. The basic relationships modeled are:

a) *replaceability*, when it is perceived that two or more input preferences can be alternated;
b) *simultaneity*, when it is perceived that two or more input must be present simultaneously;
c) *neutrality*, when it is perceived that two or more input preferences can be grouped independently (neither conjunctive nor disjunctive relationships).

Regarding the aggregation process, it follows the hierarchical structure of the requirement tree, from bottom to top. Applying a stepwise aggregation mechanism, the elementary preferences can be partially structured; in turn, repeating the aggregation process at the end a global schema can be obtained. This aggregation model allows computing partial and global preferences. The global preference represents the global degree of satisfaction of all involved requirements. LSP is based in a weighted power means mathematical model. Lets suppose we have to produce the partial or global preference GP, starting from m elementary preferences. The aggregation function should satisfy that:

1) each elementary indicator EP_i should have an associated weight W_i ; and
2) the resulting preference has a value between $\text{Min}(EP_1, ..., EP_m) <= GP_i <= \text{Max}(EP_1, ..., EP_m)$.

These properties can be satisfied by the following model:

$$GP(r) = (W_1 \ EP^{\,r}_1 + W_2 \ EP^{\,r}_2 + ... + W_m \ EP^{\,r}_m)^{1/r} \ ; \qquad (8)$$

$$-\infty <= r <= +\infty \ ; \quad 0 <= EP_i <= 1 \ ;$$

$$(W_1 + W_2 + ... + W_m) = 1; \quad W_i > 0 \ ; \quad i = 1 \ ... \ m;$$

$$GP(-\infty) = \text{Min}(EP_1 , EP_2 , ... , EP_m);$$

$$GP(+\infty) = \text{Max}(EP_1 , EP_2 , ... , EP_m);$$

The power r is a real number selected so to achieve the desired logical relationship of the aggregation function. If GP(r) is closer to the minimum then such a criterion specifies the requirement for the simultaneity of inputs. Conversely, if GP(r) is closer to the maximum then such a criterion specifies the requirement for the replaceablity of inputs. The model will be validated as a non-additive ratio scale in sub-section 3.1.2 for the above relations. However, the model is additive for neutrality relationship (when r = 1).

3.1.1 Some Web Metrics Validation. From the theoretical validation standpoint, it would be necessary to confirm for each attribute the previous properties and criteria.

Table 1. *Examples of Web metrics and criteria for theoretical validation*

Attribute	Scale Type	Unit	Criteria and Properties that Apply
Scoped Search (for collections, academic courses, personnel, etc.). High-level characteristic: *Functionality*	Ordinal	Functionality ranking	✓ The internal attribute is exhibited in sites. It is a direct metric and was measure by observation. ✓ It fulfills the representation condition (according to section 2.1), i.e., it accomplish the transitivity (6) and completeness (7) properties ✓ Different sites may have different functionality for the attribute. Conversely, different sites may have the same functionality. ✓ The unit and scale type are defined and confirmed.
Broken Links. (It intervenes in the three case studies) High-level characteristic: *Reliability*	Absolute	% of broken links.	✓ It is an indirect metric. The equation is X = BL*100/TL; where BL, represents the number of broken links found; and TL represents the total number of links of the site. ✓ It fulfills the representation condition (That is, greater number of broken links with regard to the total amount of links leads to more degree of deficiency for the attribute –the site is less reliable). ✓ Different sites may have different percentages of broken links. Conversely, different sites may have the same percentage. ✓ The unit and scale types are defined and are appropriate for variables and for the equation. The count for each variable is an absolute scale type. It produces a final absolute scale type (see demonstration in this section). ✓ It was measure automatically by a measurement instrument (the SiteSweeper 2.0 tool).
Image Title. (It intervenes in the three case studies) High-level characteristic: *Efficiency*	Absolute	% of absence of ALT property	✓ It is an indirect metric. The equation is X = AAR * 100/TAR; where AAR represents the number of objects without image title, and TAR represents the total number of objects (images) that should reference the ALT property (in the HTML code). ✓ It fulfills the representation condition (That is, greater number of absence of image title leads to lesser accessibility in the reading of graphic objects when users do not use the browser's image feature) ✓ The unit and scale types are defined and are appropriate for variables and for the equation. The count for each variable is an absolute scale type. It produces a final absolute scale type (see demonstration in this section). ✓ It was measure automatically by a measurement instrument (the SiteSweeper 2.0 tool).

Table 1, shows descriptions of the theoretical validity for a small group of metric used in the case studies. The target entity is a Web site in the operative phase of a product lifecycle. The instrument validity is not applicable when data gathering was carried out manually. When an instrument was utilized, evaluators confirmed the validity of units and parameters that such tool

allowed to configure. Besides, for each measure, the same one was repeated several times in order to see the error tolerance of the tool.

We can demonstrate for the last two attributes of table 1 that the scale type is absolute. Lets be the metric: m=a A/B ; with a > 0 and the absolute value of A, B (9)
where a is, in our case, the constant 100; A, for example, represents the number of broken links found (a count); and B represents the total number of links of the site (a count).

It is always satisfied that A <= B, and therefore it holds that A $\subseteq$ B. The relationship between A and B can be described by:
A = b B; with b>0. (10)

Replacing (10) in (9), we obtain: m = a b B/B = a b = c.

The resulting m is an absolute scale. Hence, the two quoted metrics are in the absolute scale type. Percentage measures, says Zuse, can be used as an absolute scale, but they do not assume an extensive structure.

3.1.2 The Validation of the LSP Aggregation Model. The following combination rule is meaningful for the ratio scale (see [12] pp. 219-222):
$m(x1 \circ x2) = (m(x1)^r + m(x2)^r)^{1/r}$

The LSP model -see formula (8) in section 3.1.1-, fulfills the above combination rule. It is additive to r = 1; it is supra-additive to r > 0, and it is sub-additive to r < 0. Therefore, supra-additivity (wholeness) and sub-additivity are meaningful properties to a non-additive ratio scale. It implies that the combination rule can be modified by the constant r without changing the empirical meaning of the measure m. In (8), the measures EP_1 ... EP_m represent the elementary quality preference as introduced in 3.1. The elementary quality preference EP_i satisfies the following criteria for a ratio scale type according to Fenton et al. ([3] pp. 51):

a) it is a measurement mapping that preserves ranking order, the size of intervals and ratios among entities;
b) there is a zero element, representing total lack of the attribute;
c) the measurement mapping must start at zero and increase at equal intervals known as units.

In addition, the admissible transformation for a ratio scale type is m' = a m, where a = W in (8).

4. Final Remarks

We should be keenly aware of the importance of the validation process in Software Engineering in general, and in the Web in particular, and also of the need of a more rigorous approach in measurement and evaluation for understanding and improvement purposes. Nevertheless, contrary to other traditional sciences there is no yet a clear consensus in the software community which approaches actually lead to a widely accepted view of validity. Although many progresses have been made in property-based software measurement and validation frameworks [1, 3, 5, 11, 12], among others, there are different views and disagreement points. For instance, Zuse only considers a mapping in R for a nominal scale, but this could not be necessary because a symbolic representation for each class (e.g., labels) could be enough. Furthermore, according to the same author, for a nominal and ordinal value a unit can not be assigned. His view is that units only can be assigned to absolute, ratio or interval scale measures. This position disagrees with the one of Kitchenham *et at*. [5] who consider that the use of units should be extended for values of nominal and ordinal scale types. Ultimately, axiomatic-based frameworks are not always well accepted in some software community of metrics [6]. (The above are only a reduced list of discrepancies).

We think that a practical and robust validation framework should be based on identifying the elements of measurement and theirs properties regarding also axiomatic conditions, identifying how

these elements and definition models are considered when we construct a measure in a particular context, and specifying robust and flexible theoretical and empirical methods of validating those intervening elements and models. On the other hand, software measurement should focus much more on empirical and qualitative conditions and models than measurement in traditional sciences due to the particular features of the software as a product.

Therefore, based on these previous foundations, we have described a conceptual framework regarding its relational systems, conceptual classes and relationships, some axioms, properties and criteria, useful to validate metrics for assessment and prediction purposes. Particularly, we have targeted the validation of some direct and indirect Web metrics for assessment purpose and the validation of an aggregation model utilized in Web-site QEM.

Currently, we are developing an integrated environment, called WebQEM_Tool, to support metric automation as well as the edition, calculation, and hyper-documentation in the whole evaluation process. On the other hand, the field studies and the validation process carried out allow us nowadays to get an insight into the elements for a predictive model in order to estimate the effort of Web developments. Indeed, this is an open line for our future investigations.

Acknowledgment

This research is supported by the "*Programa de Incentivos, del Ministerio de Cultura y Educación de la Nación*", and "*Facultad de Ingeniería*" in the UNLPam-09/F013 research project.

References

1. Briand, L., Morasca, S., Basili, V.; 1996, Property-Based Software Engineering Measurement, *IEEE Transaction on Software Engineering*, 22(1), pp. 68-85.
2. Dujmovic, J; Bayucan, A;1997, Quantitative Method for Software Evaluation and its Application in Evaluating Windowed Environments, *IASTED Software Engineering Conference*, SF, US.
3. Fenton, N.E.; Pfleeger, S.L., 1997, Software Metrics: a Rigorous and Practical Approach, 2nd Ed., PWS Publishing Company.
4. ISO/IEC 9126-1991(E) International Standard, *Information technology – Software product evaluation – Quality characteristics and guidelines for their use*, Geneva, Switzerland, 1991.
5. Kitchenham, B., Pfleeger, S. L., Fenton, N., 1996, Towards a Framework for Software Measurement Validation. *IEEE Transactions on Software Engineering*, 21(12), pp. 929-944
6. Kitchenham, B., Stell, J., 1997, The danger of using Axioms in Software Metics, *IEEE Proceedings on Software Engineering*, Vol. 144, Nº 5-6, pp. 279-285.
7. Olsina, L., 1999, Web-site Quantitative Evaluation and Comparison: a Case Study on Museums, *Workshop on Software Engineering over the Internet, at Int'l Conference on Software Engineering*, LA, US, http://sern.cpsc.ucalgary.ca/ ~maurer/ICSE99WS/ICSE99WS.html.
8. Olsina, L., Godoy, D; Lafuente, G.J; Rossi, G.; 1999, Assessing the Quality of Academic Websites: a Case Study, New Review of Hypermedia and Multimedia (NRHM) Journal, Taylor Graham Publishers, UK/USA, Vol. 5, pp. 81-103
9. Olsina, L.; Lafuente, G.J; Rossi, G.; 2000, E-commerce Site Evaluation: a Case Study. To appear in LNCS of Springer-Verlag, 1st International Conference on Electronic Commerce and Web Technology, London-Greenwich, UK.
10. Pons, C., Olsina, L., Prieto, M., 2000, A Formal Mechanism for Assessing Polymorphism in Object-Oriented Systems, To appear in First Asia-Pacific Conference on Quality Software (APAQS), HK.
11. Roberts, F., 1979, Measurements Theory with Applications to Decision-Making, Utility, and the Social Sciences , Encyclopedia of Mathematics and its Applications Addison-Wesley Pub. Co.
12. Zuse, H., 1998, A Framework of Software Measurement, Walter de Gruyter, Berlín-NY.