Trust Metrics to Measure Website User Experience

Andréia Casare, Tania Basso, Regina Moraes
University of Campinas - UNICAMP
Campinas, Brazil
email: casareandreia@gmail.com, {taniabasso, regina}@ft.unicamp.br

***Abstract*—Trust in computational systems and online applications depends on technical, social and personal aspects. The technical ones, such as a strong computational infrastructure, adequate network bandwidth, sufficient storage space, among others, have been largely studied. Social issues, such as runtime security and data privacy are important for users to be protected from attacks by malicious people. However, there are personal aspects that impact the sense of trust, which depends on the user experience when interacting with those systems. This paper tackles this issue by proposing a study on measures that can determine the user experience when using an online system or application. The approach relies on a quality model to combine these metrics and compose a trustworthiness score. Seven websites are used in the experiments in two different contexts and, based on the set of measures that composes the quality model, the approach suggests the one that presents the highest user perception of trust in each context, that is the one with the highest score.**

Keywords–Trustworthiness; User experience; Quality model.

I. INTRODUCTION

The increasing use of online systems and applications by individuals in a globalized market has introduced new challenges in software development, including issues related to human-computer interaction. Nonetheless, challenges regarding nonfunctional requirements, such as security, privacy and trust can arise from the value of a business or even the need to improve the relationship with the consumer.

Trust is defined differently in distinct areas [1] and, inspired by the existing definitions, we can define it as the reliance of a client on a service, that it will exhibit some expected behaviour. Then, trustworthiness can be defined as the level in which a service meets a set of those requirements, i.e., the worthiness of services for being trusted. In some cases, trust is subject to individual interpretation and context of use.

Trust does not only involve technical aspects, such as data security, fault tolerance, but also human interaction aspects that should consider attributes of usability, accessibility and user experience. If the system presents a pleasant interface, good performance, easy to use and learn, providing in the tasks execution a good experience of use, it will have great chances of being reused and trusted.

In this context, the goal of this work is to define a set of metrics and to propose a way to combine several metrics to get a score for trustworthiness focused on the user perception. This work is part of a wider proposal, in which several metrics should be defined, validated and combined in order to translate the importance of each metric toward trustworthiness score, being able to translate the user's perception and allowing to evaluate and determine the user experience when using online systems or applications. Based on it, users can compare and choose systems that present higher level of trust from the perspective of the system user.

In order to obtain a trustworthiness score composed of heterogeneous metrics, a quality model is used in this work. The quality model was proposed in the ISO/IEC 25000 (SQuaRE) standard [2] as a way to formalize the interpretation of measures and the relationship among them. These models are built by a user / analyst, who knows in advance the context, the final scores, their units and scales. This way, it is possible to define how the measures should be aggregated in the analysis, and what procedures have to be used to homogenize their values, so they can be aggregated. It is possible to define one quality model for each considered property, and then, these different perspectives can be aggregated following a hierarchical structure.

The contributions of this work are: (i) the selection of a set of properties that can be used to measure a user experience; (ii) a user experience quality model to compose the properties' metrics and their relationship. The model was evaluated using seven real systems (websites) to demonstrate its usefulness.

The idea followed by this work is aligned with the interest of the Adaptive, Trustworthy, Manageable, Orchestrated, Secure, Privacy-assuring Hybrid, Ecosystem for Resilient Cloud Computing (ATMOSPHERE) project [3]. ATMOSPHERE is an Europe-Brazil collaborative project that exchange experiences and results with its members. By defining a user experience quality model, the resulted model can easily be integrated with other quality models defined in the ATMOSPHERE project and complement the trustworthiness score with a user experience measurement.

The paper is organized as follows: Sections II and III present, respectively, relevant concepts and related work that guided our study. Section IV presents the proposed user experience quality model and the methodology used to get the metrics and final trustworthiness score. Section V shows the results of experiments applying the quality model to calculate the trustworthiness score of two categories of e-commerce websites. Finally, Section VI presents the conclusions and future work.

II. BACKGROUND

This section addresses, briefly, the issues that underpin this work. It discusses trust, user experience and quality model.

A. Trust

Trust and trustworthiness concepts have been studied in different areas, such as people social relationship and business environments. For example, Mayer et al. [4] proposed a model for defining trust including characteristics of the trustor, the trustee, and the role of risk. Their model is focused on trust in an organizational relationship. Venkatesh et al. [5] proposed a conceptual framework of online trust based on different views and requirements of different stakeholders (such as customers,

 ISBN: 978-1-61208-761-0

suppliers, employees, partners, etc.). In a broader context, McKinght et al. [6] proposed a multidimensional model of trust in e-commerce. The model includes four high-level constructs (disposition to trust, institution-based trust, trusting beliefs, and trusting intentions), which are further delineated into 16 measurable subconstructs.

Although these concepts are differently defined in distinct areas, one of the common main goals in all definitions is to accurately assess the trust level as a robust basis for decision making (e.g., system adaptation), which turns out to be a very complex problem. A key problem is that the trust level is uncertain and may dynamically change. Mainly, it can be strongly dependent on the feeling of a user, when she / he is interacting with the system, i.e., the quality of the interaction between the human and the system. So, the user experience should be included among the properties that are used to compose the trustworthiness score of a system. Thus, establishing trust and building trustworthy services is a challenge and can benefit from research on the quality of the system interface and user experience.

B. User Experience

ISO 9241-210 [7] defines user experience as *the user's perceptions and reactions resulting from the use of a software product, system or service*. The user experience includes all of the user's emotions, perceptions, preferences, physical and psychological responses, behaviors, and achievements that occur before, during, and after the use. Therefore, user experience is a consequence of the features, performance, system interactivity or products that the user has had as a result of previous experiences, abilities and context of use.

C. Quality Model

Trustworthiness can be understood as a multi-dimensional construct combining specific attributes, properties and characteristics (for example, security, privacy, fairness, transparency, dependability, among others). All of them have other sub-attributes that increase the number of possibilities to be addressed.

Since several conflicting properties may be involved in the analysis, a Multi-Criteria Decision-Making (MCDM) based technique can be useful to define how to compute the global score of a service. In this work, Logic Score of Preferences (LSP) [8] was chosen due to its previous use in the dependability field. It comprises multiple aggregation blocks to define how the different elements should be used to produce a final score.

Usually, measures of services present distinct scales and dimensions. In order to apply LSP, the measures should be brought to the same scale before the aggregation. To do this, we used the normalization functions proposed in [9].

To use the LSP technique, it is necessary to first define a Quality Model [10], which is essentially a conceptual representation of attributes, weights, thresholds and operators that should express the requirements that the system should meet (for example, the tree structure in Figure 2). The blocks, in this work, represent (leaf or composite) attributes, which are aggregated (by the operators). Values at the bottom level (leaf attributes) are aggregated to calculate upper level values (composite attributes), towards the calculation of the final score of the system through a single 0-to-100 score. *Thresholds* are elements used in the normalization function to specify the range of acceptable input values of leaf-level attribute. *Weight* is an adjustable element which specifies a preference over one or more characteristics of the system (e.g., in certain contexts memory usage might be more important than throughput).

III. Related Work

The Human-Computer Interface (HCI) literature presents some works that consider usability, accessibility and quality of product as attributes / characteristics that influence the users' trust perception when using a website or application, as well as works that propose quality models for software measurement.

Few works address user experience related to trust. Most of them use the eye-track technology (e.g., the work of Djamasbi et al. [11], which examines the effect of images of faces on the visual appeal, efficiency, and trustworthiness of a page). The work most related to ours is the one from Ramadhan et al. [12]. The authors evaluate the user experience regarding factors that influence user trust through the design of website interface. They evaluated the three cryptocurrency websites most frequently accessed from Indonesia using methods, such as Performance Metrics, Post-Task Rating, Post-Session Rating, Experiential Overview and eye-tracking device. However, they did not use quality models to represent the multi-dimensional attributes nor calculate trustworthiness scores to help define the more trusted website.

Regarding quality models, Seffah et al. [13] proposed an hierarchical model of usability measurement, called Quality in Use Integrated Measurement (QUIM). It has 10 factors, which are decomposed into 26 sub-factors that are further decomposed into 127 specific metrics. Some factors are: efficiency, effectiveness, productivity, accessibility, trustfulness, among others. The metrics may be extracted from log files, video observations, interviews, or surveys. Lew et al. [14] proposed, based on ISO 25010, a framework for modeling requirements for quality, usability and user experience. The goal is to evaluate quality attributes of software and Web applications. Some attributes are accuracy, suitability, accessibility, and legal compliance. Hendradjaya and Praptini [15] proposed a quality model with attributes to evaluate e-government websites. The attributes are: functionality, reliability, usability, efficiency, portability and productivity. The measures were obtained through some specific Web tools and questionnaires.

Although these previous works [13]-[15] presented quality models related to usability and user experience and even presented some quality attributes related to trustworthiness (e.g., reliability, accessibility), they are not focused on this characteristic. Furthermore, although Seffah et al. [13] defined some metrics, their focus is only on usability. Hendradjaya and Praptini [15] also defined some metrics that go beyond usability, but they are specific for e-government and may not be generalized.

IV. User Experience Quality Model and User Trustworthiness Score Calculation

This section presents the quality model we defined, as well as the methodology used to get the metrics and the final user experience quality score. Also, it presents the results of applying the quality model on seven real websites of two different contexts - optics sellers and airlines.

 ISBN: 978-1-61208-761-0

The research methodology consisted of the following steps: (i) HCI literature review to select the software quality attributes that influence the user experience when interacting with a system; (ii) a Quality Model development with measures hierarchically represented by quality attributes (such as usability, accessibility, performance, among others) that will compose the confidence perception score; (iii) the selection of tools that are able to collect the selected metrics; (iv) the experiments performed on real websites relying on automatic tools that return values of performance, accessibility, among other metrics; (v) based on the Quality Model and the experiments results, a website reliable score is computed; (vi) the analysis and discussion of the experiments results.

In step (i), we identified the quality attributes that impact user confidence during her/his interaction with the website or Web application. The complete set of metrics identified so far can be seen in Figure 1.

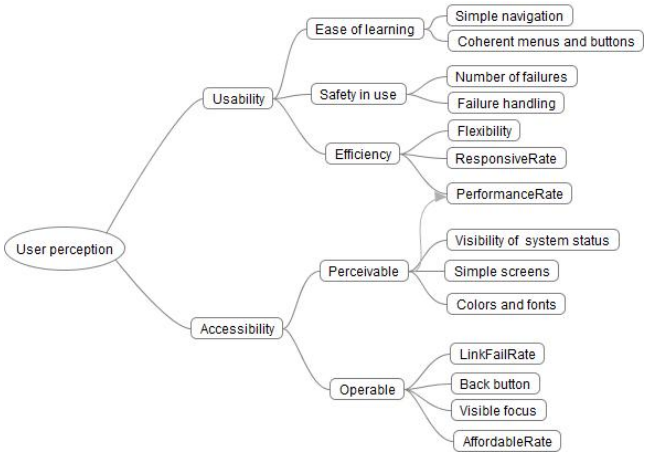

Figure 1. Quality attributes that influence trust

The attributes were grouped in two main categories: *Usability* and *Accessibility*. The *Usability* group is composed of *Ease of Learning* (regarding website navigation, menus and buttons coherence), *Safety in use* (regarding website failures), *Efficiency* (regarding website response, performance and flexibility).

The *Accessibility* group is composed of attributes *Perceivable* (regarding website design, as colors, screens and system status) and *Operable* (regarding website functions as buttons, links, focus, etc.).

From these quality attributes, we selected a subset to tackle in the present work, that is *PerformanceRate* (also called, in the quality model, *PerformancePageUp*, which refers to the time/rate to load the website), *AffordableRate*, *ResponsiveRate* and *LinkFailRate*. We first choose these metrics because they are objective measures and can be measured automatically by tools. The remaining attributes in Figure 1 will be assessed in future work, mainly to be dependent on user personal evaluation.

It is important to clarify that we decided, as a first stage, to consider only the attributes that can be measured by automatic tools because our main goal is to perform experiments that allow evaluating the quality model, the metrics and the scores calculation. It is obvious that, once we are interested in evaluating user perception, experiments with human users (for example, comparing whether the scores produced by the tools and human users match or not) would produce more solid results. However, we intend to extend the quality model and to perform experiments with users in a second stage.

Following the methodology, step (ii) defined a quality model to aggregate the several identified metrics. For now on, only the metrics to be tackled in this work were placed in the current version of the Quality Model, which is presented in Figure 2.

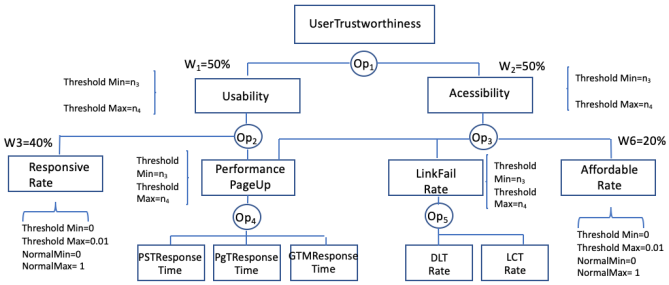

Figure 2. User Experience Quality Model

As mentioned in Section II-C, *UserTrustworthiness* is decomposed in composite and leaf attributes, which were aggregated based on operators. Some weights were defined for each attribute denoting the importance to represent *UserTrustworthiness* as a whole. For example, *Usability* and *Accessibility* collaborate both with fifty percent to compose *UserTrustWorthiness*. *Usability*, in turn, receives the collaboration of *ResponsiveRate* (40%) and *PerformancePageUP* (60%), and so on. This last attribute (*PerformancePageUp*) also collaborates with *Accessibility*, but, in this case, its importance is determined to be 30% (see Table III).

It is important to mention that the metrics configuration values (weighs, thresholds, normalization values, periodicity, operators) were defined by experts who integrated the ATMOSPHERE project teams. They worked on the layers to which a quality model refers, performing experiments during the framework layer development and defining the values based on these experiments. To the best of our knowledge, there are no previous works that also define these values, otherwise we would consider them to perform a more robust study.

Table I, Table II and Table III provide more details about the User Quality Model. In Table I, a description of each leaf metric is provided to give a broader view of the metric. In addition, details are provided about how the metric is computed, which type of data is each one of them, how and when it is collected, how it can be considered to improve the user experience and if the metric acts as *Benefit* (the higher the metric value is better) or *Cost* (the lower metric value is better).

Related to the leaf attributes configuration, Table II provides information about minimum (NMin) and maximum (Nmax) values that the attribute can assume and these values are used for normalization purpose (normalized between 0 - 1 range). It also provides the thresholds minimum (TMin) and maximum (TMax), the weight (W) of the attribute in the composition to the subsequent level attribute, the periodicity of

 ISBN: 978-1-61208-761-0

TABLE I. LEAF ATTRIBUTES METRICS DESCRIPTION (PARTIAL VIEW)

Metric Name	Description	How is it Computed ?	Type of data	How & When is it sent?	How is it used for adaptation	B(en) / C(ost)	Property
Responsive Rate	Check if the website is able to adapt to mobile devices	Extracted relying on Mobile Friendly Test Tool	1 (yes) / 0 (no)	On demand	Improving components and software development	B	Usability
PST Responsive Time	Measure the website performance (time to be up)	Extracted relying on PageSpeed Insights tool	0 up to 100	On demand	Compose the mean to obtain the performance metric	B	Performance

the calculation (CP), given in seconds (and, in this case, all of them are on demand (*On dem.*), and the operator (OP), which can be *Average* (Avg), *Min, Max, Sum* (S), through which the metrics will be aggregated.

TABLE II. LEAF ATTRIBUTES METRICS CONFIGURATION (PARTIAL VIEW)

Metric Name	N Min	N Max	T Min	T Max	W(%)	CP	OP
Responsive Rate	0	1	0	1	40%	On dem.	OP2 Max
PST Response Time	0	100	0	100	33.3%	On dem.	OP4 Avg

TABLE III. COMPOSITE ATTRIBUTES CONFIGURATION

Metric Name	T Min	T Max	W(%)	CP	OP
Performance Page Up	0	100	60% 30%	On dem.	Op2-N Op3-N
LinkFail Rate	0	100	20%	On dem.	Op3-N
Usability	0	100	50%	On dem.	Op1-N
Accessibility	0	100	50%	On dem.	Op1-N

In addition to the leaf attributes configuration, the composite attributes also need to be configured. So, Table III presents the configuration of composite attributes where the thresholds (TMin and TMax), the weight (W), the periodicity (CP) and the operator (OP) are specified. It is important to notice that, in this case, the operator refers to different operations, that can be:

- *Neutrality (N)*. Refers to the arithmetic mean and represents the combination of simultaneous satisfaction requirements with replaceability capability.
- *Simultaneity (S)*. This operation means that all requirements must be satisfied; it refers to a conjunction - i.e., the logical operator AND.
- *Replaceability (R)*. Is used when one of the requirements of the system has a higher priority replacing the remaining requirements; it refers to a disjunction - i.e., a logical operator OR - to perform aggregation.

V. CALCULATING METRICS AND SCORES: EXPERIMENTS AND RESULTS

This section presents the experiments regarding the application of the User Experience Quality Model in order to calculate user trustworthiness scores for some e-commerce websites.

A. Experimental setup

Step (iii) finds automatic tools to collect the metrics that were selected to compose the quality model. We found some freeware or open source and proprietary tools. This latter (proprietary) set was not considered in this work. Based on preliminary tests, the following tools were selected: *PageSpeed Insights* [16], *Pingdom Website Speed Test* [17] and *GTMetrix* [18] evaluate the performance of the website to be up; *Dead Link Checker* [19], *Xenu's link* [20] and *Screaming Frog* [21] inspect the links and count the broken ones; *Mobile Friendly Test* [22] verifies if the website is a responsive one; *ASES (Accessibility Evaluator and Simulator)* [23], *Nibbler* [24] and *Access Monitor* [25] verify the website affordable rate. All these tools are stable and freeware.

Step (iv) is reserved to run the tools on the chosen websites. For the experiments, the website selection is based on the website type (e-commerce) of two business segment (optics sellers and airlines) and the website size with up to 10,000 URLs to allow the experiments control. We selected four optics websites and three airline websites. The companies names or respective websites are not mentioned because they have commercial license. So, we will refer to the optics websites as Opt1, Opt2, Opt3, Opt4 and the airline websites as Air1, Air2 and Air3, without any special order.

B. Results and discussions

In the experiments, we applied the selected tools to extract the following metrics of the websites: performance of the page up (*PageSpeed, Pingdom* and *GTMetrix*); amount of broken links (*Dead Link Checker*, *Xenu's link* and *Screaming Frog*); responsiveness (*Mobile Friendly Test*); affordable rate (*ASES, Nibbler* and *Access Monitor*). The next subsections present the results for each metric, including the trustworthiness score calculation.

1) Performance: To measure the performance of each website, nine experiments were performed using different machines, Internet networks and Web tools. For each website, we executed three tests using a desktop machine, processor I5 and Windows 7, accessing the wired network of the university administrative sector; three tests using a notebook, processor I5 and Windows 10, accessing the university WiFi network, and three tests using a notebook, processor I5, Windows 10, accessing a 15-megabytes wireless network at home.

 ISBN: 978-1-61208-761-0

All the tools used to measure the performance returned a value between 0 up to 100, which is normalized to 0 - 1 range for the calculation. In case of a *Benefit* attribute, the higher the value, the better the performance contribution to the score. Table IV shows the measurements obtained in the nine tests on the first website (*Opt1*). For all the other websites, a similar table has been created but for the sake of space and better presentation they are not included in this paper and we opted to present only a summary table. However, all the tables from these experiments can be found in our institutional website [26].

TABLE IV. PERFORMANCE MEASUREMENTS - WEBSITE OPT1

Test #	PageSpeed	Pingdom	GTMetrix	Remarks
1	82	67	48	Desktop
2	85	67	48	Desktop
3	80	67	48	Desktop
4	86	67	48	Notebook wifi
5	87	67	48	Notebook wifi
6	85	67	48	Notebook wifi
7	86	67	48	Notebook wifi - home
8	85	67	48	Notebook wifi - home
9	81	67	48	Notebook wifi - home

Table V summarizes the measurements obtained in the nine tests on each optics website, using the same tools and configurations. Table VI summarizes the tests for airline websites.

The performance score is calculated using the normalized measurements average:

$$PerformancePageUp = (AVG(PageSpeed) * W + AVG(PingDom) * W + AVG(GTMetrix) * W) \quad (1)$$

For example, for *Opt1*, PerformancePageUp = 0.8411 * 0.3333 + 0.67 * 0.3333 + 0.48 * 0.3334 = 0.6637.

Considering only the performance attribute, the *Opt2* website is considered better than the other three optics websites and the best considering both e-commerce segments, as its score is the highest. Among the airline websites, *Air1* is better than the other two. It is important to notice that the metrics collected by a tool, in any case (considering the same tool and the same website), are very similar when one considered different computers and networks configurations. Higher differences are recorded by the *PageSpeed* tool (*Opt3, Air2* and *Air3* websites) pointing out that this tool is more sensitive to the computational resources used by the user.

2) Broken links: To check the number of website broken links, one experiment was carried out using a notebook, a processor I5, and Windows 10, accessing a 15-megabytes wireless home network. Three tools were used that returned the total number of the website links and the number of defective links. In this case, only one experiment was performed using each tool, because this measurement is not impacted by the computational environment or network. Table VII shows the scores returned from the test of the *Opt1, Opt2, Opt3* and *Opt4*.

The broken link rate is calculated based on the maximum rate obtained by any of the tools, i. e., it is calculated as:

$$LinkFailRate = MAX(DeadLInkChecker(broken\_links/total\_links), Xenu'sLink(broken\_links/total\_links)). \quad (2)$$

Using expression (2), the *Opt1* score for Link Fail Rate is 8.46%, the scores of *Opt3* is 5.45%, *Opt3* is 1.66% and *Opt4* 0.55%. As this metric is a *cost* one (i.e., lower value is better for the score), *Opt4* presented better score for this attribute (broken link) in the optics segment and also considering both segments. The metrics collected by all tools are very similar for *Opt3* website but differ significantly for the other websites. Investigating why this difference happens, we observe some restrictions. *Dead Link Checker* is limited to 2000 links, so this justifies the difference observed for the *Opt4* website. *Screaming Frog* was not able to inspect the *Opt3* website even after several attempts. We could not identify this problem.

Considering the airline websites, we can better observe the limitation of *Dead Link Checker* tool in Table VIII, which stops the analysis of all the websites when 2000 links are inspected. For this reason, we ignored its results and considered the maximum rate among the other two tools. The best airline website related to broken links is *Air1* with 2.9% of broken links, followed by *Air2* (4.47%) and *Air3* (4.96%) .

3) Responsiveness: To verify if the website is ready to run on mobile devices, one experiment was carried out using a notebook, processor I5, Windows 10, accessing a 15-megabytes wireless network. The *Mobile Friendly Test* tool is the only stable tool identified to collect this metric. All the websites used in the experiment (both segments) are ready for mobile devices (i.e., they are responsive). In this case, this metric should be 0 (non responsive) or 1 (responsive).

4) Affordable rate: The affordable rate score is given in percentage and calculated using the normalized measurements average:

$$AffordableRate = (AVG(ASES) * W + AVG(Nibbler) * W + AVG(AccessMonitor) * W) \quad (3)$$

Table IX and Table X present the results for applying the tools to the optics sellers and airline websites, respectively. In Table IX, the highest score is computed for *Opt2*, with approximately 0.7% and the lowest one for *Opt1* with approximately 0.5%. In Table X, the scores are very close, however, *Air1* presented the highest one (approximately 0.75%).

5) Trustworthiness score calculation: Considering the measurements that were obtained for the attributes of the Quality Model, the score of the next level (i.e., Usability and Accessibility) can be calculated. *Usability* is composed of *Responsive Rate* and *Performance Page UP* using the operator *Neutrality* (arithmetic mean), applying the *weight* (W) of each one of them (0.4 and 0.6, respectively) as follows:

$$ScoreUsability = (MeanResponsiveRate * W + MeanPerformancePageUp * W) \quad (4)$$

For example, the usability score for *Opt1* website = (1 * 0.4 + 0.6637 * 0.6) = 0.79822. Table XI presents the score

 ISBN: 978-1-61208-761-0

TABLE V. PERFORMANCE MEASUREMENTS - OPTICS SELLERS WEBSITES

Website Name	PageSpeed			Pingdom			GTMetrix			Score
	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	
Opt1	81	87	84.11	67	67	67	48	48	48	0.6637
Opt2	100	100	100	81	81	81	73	74	73.33	0.8477
Opt3	61	75	68.66	68	68	68	63	65	64.55	0.6707
Opt4	43	48	45.11	69	70	69.22	56	59	57.33	0.5722

TABLE VI. PERFORMANCE MEASUREMENTS - AIRLINE WEBSITES

Website Name	PageSpeed			Pingdom			GTMetrix			Score
	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	
Air1	82	94	87.77	67	67	67	59	59	59	0.7126
Air2	19	23	20.11	63	63	63	48	50	48.66	0.4392
Air3	9	12	10.11	63	63	63	52	53	52.77	0.4196

TABLE VII. AMOUNT OF TOTAL AND BROKEN LINKS - OPTICS WEBSITES

Tool	Opt1 Total/ Broken links	Opt2 Total/ Broken links	Opt3 Total / Broken links	Opt4 Total / Broken links
Dead Link Checker	1830 / 155	383 / 6	1785 / 15	2000 / 2
Xenu's link	552 / 39	403 / 22	1748 / 29	8593 / 47
Screaming Frog	1168 / 23	2 / 0	1697 / 2	7866 / 3
Score	0.0846	0.0545	0.0166	0.0055

TABLE VIII. AMOUNT OF TOTAL AND BROKEN LINKS - AIRLINE WEBSITES

Tool	Air1 Total / Broken links	Air2 Total / Broken links	Air3 Total / Broken links
Dead Link Checker	2000 / 58	2000 / 53	2000 / 22
Xenu's link	69482 / 262	2974 / 133	9748 / 484
Screaming Frog	93995 / 340	2498 / 61	7738 / 126
Score	0.029	0.0447	0.0496

of all optics sellers websites. The airline websites have their usability score presented in Table XII.

Accessibility score is composed for *Performance Page Up* (but now its weight is 0.3), *Link Fail Rate* and *Affordable Rate*. It is important to note that, in the composition of accessibility score, one of the measures, the *Link Fail Rate*, is a *Cost* attribute, so we use its complement in the calculation expression and *Performance Page Up* is now weighted as 30% to compose the accessibility score, since its importance for accessibility is lower. The expression to calculate the

TABLE IX. AFFORDABLE RATE - OPTICS SELLERS WEBSITES

Tool	Opt1 %	Opt2 %	Opt3 %	Opt4 %
ASES	61.19	76.42	79.54	64.17
Nibbler	62	78	67	78
Access Monitor	33.8	57.8	56.4	50.2
Score	0.5233	0.7074	0.6765	0.6412

TABLE X. AFFORDABLE RATE - AIRLINE WEBSITES

Tool	Air1 %	Air2 %	Air3 %
ASES	85.6	83.34	78.31
Nibbler	87	87	89
Access Monitor	52.8	44.4	48
Score	0.7513	0.7158	0.7177

accessibility score is:

$$ScoreAccessibility = MeanPerformancePageUp * W + (1 - MeanLinkFailRate) * W + MeanAffordableRate * W \quad (5)$$

For example, *Opt1* website accessibility score = 0.6637 * 0.3 + (1 - 0.0846) * 0.2 + 0.5233 * 0.5 = 0.64384. Table XI and Table XII present the accessibility scores for the optics sellers and airline websites, respectively.

Following the Quality Model, the last calculation (the top of the Quality Model tree) is the user trustworthiness score. The aggregation is guided by *Operation 1* (OP1), which was configured as *Neutrality*. So, the user trustworthiness score is computed as follows:

$$UserTrustworthinessScore = (UsabilityScore * W) + (AccessibilityScore * W) \quad (6)$$

For example, *Opt1* website user trustworthiness score = 0.79822 * 0.5 + 0.64384 * 0.5 = 0.72103. The user trustworthiness score for the optics sellers websites are presented in Table XI and Table XII presents the same score for airline website.

Comparing the user trustworthiness scores obtained, we

observe that, considering the selected attributes as the ones which impact the user perception of trust, *Opt2* is the website that has the highest chance to please the users during their interaction, followed by *Air1* website. The worst website in this selection is *Air3* website which presents the smallest score among all.

TABLE XI. USABILITY, ACCESSIBILITY AND USER TRUSTWORTHINESS SCORES - OPTICS SELLERS WEBSITES

Attributes	Opt1	Opt2	Opt3	Opt4
Usability	0.79822	0.90862	0.80242	0.74332
Accessibility	0.64384	0.79711	0.73614	0.69116
User Trustworthiness	0.72103	0.85286	0.76928	0.71724

TABLE XII. USABILITY, ACCESSIBILITY AND USER TRUSTWORTHINESS SCORES - AIRLINE WEBSITES

Attributes	Air1	Air2	Air3
Usability	0.82756	0.66352	0.65176
Accessibility	0.78363	0.68072	0.67481
User Trustworthiness	0.80559	0.67212	0.66328

In general, considering both segments, the order for user trustworthiness score (from highest to lowest score) is: *Opt2, Air1, Opt3, Opt1, Opt4, Air2 and Air3*. We could observe that *Opt2* website presents the best trustworthiness score being almost 22% better than *Air3* website (the worst one). In the middle, *Opt3* presents a score 10.5% smaller and *Opt1* 15.2% smaller when compared to *Opt2*. Observing the airline websites, the best is *Air1*, but its score is almost 6% worse when compared to the *Opt2* (the best score).

Also, considering both segments, it is important to notice that *Opt2* is the best when we observe the attribute *Performance Page Up*. It is not the number one in the other three attributes, but also it is not the worst one in any of the attributes. Moreover, *Performance Page Up* is an important attribute in the current version of the Quality Model, since it is considered as component of the Usability attribute and the Accessibility attribute scores as well. *Air1* is the best in Affordable Rate and *Opt4* is the best in the broken link rate. Even being better in these attributes, neither *Air1* nor *Opt4* websites are able to overcome the *Opt2* website trustworthiness score due to the importance of *Performance Page Up* attribute for the context. *Opt1* presents the smallest score in the *Affordable Rate* and the highest *broken link* score.

VI. CONCLUSIONS

This work presented a definition of a set of metrics with the aim of obtaining a score for the user perception regarding trust. It is part of a wider proposal, in which several metrics should be defined, validated and combined following a methodology toward trustworthiness score calculation. The trustworthiness score should translate the user's perception when using online applications. This score will allow users to compare and choose systems that present a high level of trust.

Using the use case composed by four optical sellers and three airline websites, it was possible to calculate the trustworthiness scores and allow users to select the more trustworthy websites among the same business segment. Moreover, it was possible to observe the importance of the proposed mechanism to obtain the score (the quality model) as it balances the results based on the importance of the attributes and not only one attribute or another, neither the amount of attributes with the best scores only.

Future work will tackle more complex metrics, which are more subjective and will require some tests with the users, to evaluate usability and accessibility. It is our intention to perform usability tests using the think aloud technique and accessibility evaluation by an expert based on the World Wide Web Consortium (W3C) [27] recommendations among other evaluation techniques.

ACKNOWLEDGMENT

This work has been partially supported by the project ATMOSPHERE- Adaptive, Trustworthy, Manageable, Orchestrated, Secure, Privacy-assuring Hybrid, Ecosystem for Resilient Cloud Computing (https://www.atmosphere-eubrazil.eu/ - Horizon 2020 grant agreement No 777154 - MCTIC/RNP) and by the project ADVANCE - Addressing Verification & Validation Challenges in Future Cyber-Physical Systems (https://www.advance-rise.eu/ - call H2020-MSCA-RISE-2018, number 823788).

REFERENCES

[1] D. Artz and Y. Gil, "A survey of trust in computer science and the semantic web," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 5, no. 2, 2007, pp. 58–71.

[2] International Organization for Standardization, "Systems and software engineering — systems and software quality requirements and evaluation (square) — guide to square," 2014, URL: https://www.iso.org/standard/64764.html [Last access on September, 2019].

[3] ATMOSPHERE, "Adaptive, trustworthy, manageable, orchestrated, secure, privacy-assuring hybrid, ecosystem for resilient cloud computing," 2018, URL: https://www.atmosphere-eubrazil.eu/ [Last access on January, 2020].

[4] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," Academy of management review, vol. 20, no. 3, 1995, pp. 709–734.

[5] V. Shankar, G. L. Urban, and F. Sultan, "Online trust: a stakeholder perspective, concepts, implications, and future directions," The Journal of strategic information systems, vol. 11, no. 3-4, 2002, pp. 325–344.

[6] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," Information systems research, vol. 13, no. 3, 2002, pp. 334–359.

[7] International Organization for Standardization, "Ergonomics of human-system interaction — part 210: Human-centred design for interactive systems," 2010, URL: https://www.iso.org/standard/52075.html [Last access on February, 2020].

[8] J. Dujmovic and R. Elnicki, "A DMS cost/benefit decision model: mathematical models for data management system evaluation, comparison, and selection," National Bureau of Standards, Washington DC, No. GCR, 1982, pp. 82–374.

[9] M. M. Friginal, Jesus, D. de Andres, and J.-C. Ruiz, "Multi-criteria analysis of measures in benchmarking: Dependability benchmarking as a case study," The Journal of Systems and Software, no. 111, 2016, pp. 105–118.

[10] I. IEC, "Software Product Quality Requirements and Evaluation - SQUARE," ISO/IEC, User Guide, 2005.

[11] S. Djamasbi, M. Siegel, T. Tullis, and R. Dai, "Efficiency, trust, and visual appeal: Usability testing through eye tracking," in 2010 43rd Hawaii International Conference on System Sciences. IEEE, 2010, pp. 1–10.

 ISBN: 978-1-61208-761-0

[12] B. A. Ramadhan and B. M. Iqbal, "User experience evaluation on the cryptocurrency website by trust aspect," in 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), vol. 3. IEEE, 2018, pp. 274–279.

[13] A. Seffah, M. Donyaee, R. B. Kline, and H. K. Padda, "Usability measurement and metrics: A consolidated model," Software quality journal, vol. 14, no. 2, 2006, pp. 159–178.

[14] P. Lew, L. Olsina, and L. Zhang, "Integrating quality, quality in use, actual usability and user experience," in 2010 6th Central and Eastern European Software Engineering Conference (CEE-SECR). IEEE, 2010, pp. 117–123.

[15] B. Hendradjaya and R. Praptini, "A proposal for a quality model for e-govemment website," in 2015 International Conference on Data and Software Engineering (ICoDSE). IEEE, 2015, pp. 19–24.

[16] PageSpeed Insights, "Increase the speed of your web pages on all devices," 2020, URL: https://developers.google.com/speed/pagespeed/insights/ [Last access on January, 2020].

[17] Solarwinds Pingdom, "Pingdom website speed test," 2018, URL: https://tools.pingdom.com/ [Last access on January, 2020].

[18] GTmetrix, "How fast does your website load? find out with gtmetrix," 2020, URL: https://gtmetrix.com/ [Last access on January, 2020].

[19] Dead Link Checker, "Free broken link checker," 2013, URL: https://www.deadlinkchecker.com/ [Last access on January, 2020].

[20] Xenu, "Xenu's link sleuth," 2020, URL: https://xenu-link-sleuth.br.softonic.com/ [Last access on January, 2020].

[21] Screaming Frog, "A website crawler and log file analyser tools," 2020, URL: https://www.screamingfrog.co.uk/ [Last access on January, 2020].

[22] Mobile Friendly Test, "Is your webpage mobile-friendly?" 2020, URL: https://search.google.com/test/mobile-friendly [Last access on January, 2020].

[23] ASES, "Site accessibility evaluator and simulator," 2020, URL: http://asesweb.governoeletronico.gov.br/ases/ [Last access on January, 2020].

[24] Nibbler, "Test any website," 2020, URL: https://nibbler.silktide.com/ [Last access on January, 2020].

[25] Tenon, "Access monitor," 2020, URL: https://wordpress.org/plugins/access-monitor/ [Last access on January, 2020].

[26] Faculty of Technology, "Software engineering department," 2020, URL: http://www.ft.unicamp.br/∼regina [Last access on January, 2020].

[27] World Wide Web Consortium, "Leading the web to its full potential," 2020, URL: https://www.w3.org/ [Last access on February, 2020].

 ISBN: 978-1-61208-761-0