Experimental Design of Metrics for Domain Usability

Michaela Bačíková, Lukáš Galko and Eva Hvizdová
Department of Computers and Informatics, Technical University of Košice, Letná 9, Slovakia

Keywords: Domain Usability, User Experience, Usability, Domain Engineering, User Interface, Domain Modeling, Domain Terminology.

Abstract: Usability and user experience gain more and more attention each year. Many large software companies consider it a priority. However, domain usability issues are still present in many common user interfaces. To improve the situation, in this paper we present our design of domain usability metric. In order to design the metrics, we performed an experiment with two surveys, one in a general domain and one in a specific domain of gospel music. The results confirm our previous experimental results and indicate, that five aspects of domain usability are not equal, but have different effect on overall usability and user experience. The results of the latter survey was used to design weights of domain usability aspects and these weights were used to calculate the overall domain usability of a user interface. Given that we know the number of all components in the analysed user interface, the designed metrics measures the domain usability in percentage. The designed metrics can be used to formally measure the domain usability of user interfaces in manual or automatized techniques and this way, we believe, improve the situation regarding interfaces that do not consider the domain dictionary of their users.

1 INTRODUCTION

Usability (Nielsen, 1993) and its evaluation is nowadays a common practice in a lot of large and successful IT companies such as Apple, Amazon or Google. A huge amount of effort is invested to satisfy customers and to meet all their requirements, since the success of software heavily depends on its users. The goal is to deliver a great software with the highest quality of user experience. If the product is pleasant to use, satisfying and useful, the customer will always use it with pleasure and (possibly) buy more products from the company. As Philip Kotler[1] said, *"The Best Advertising Is Done By Satisfied Customers"*.

Still, from our experience, *domain usability* issues are very common. Users get confused due to wrong terminology, misspelling, term inconsistency and low domain specificity. Furthermore, these usability failures limit their effectiveness and performance when using the software. Even when the focus of medium and small IT companies is on improving general usability, the time and money requirements drives them to fast development and, as a result, proper effort is not invested to get familiar with the work domain and its users. Multiple researchers, e.g. Chilana et al. (Chilana et al., 2010) and Lanthaler and Gütl (Lanthaler and Gütl, 2013), already recognized this issue.

To illustrate our definition (Bačíková and Porubän, 2013; Bačíková and Porubän, 2014) we introduced the concept of *domain usability* and examples. Recently we designed multiple manual domain usability evaluation techniques (Bačíková et al., 2017) and suggested a domain usability metrics design. However, in light of our recent experiments, the design needed further research.

Note: Because of lack of space, we will not present our domain usability definition in this paper. However, for better understanding of the rest of this paper, we strongly encourage the reader to familiarize with the definition (Bačíková and Porubän, 2014).

1.1 Research Questions and Tasks

To our knowledge, no formal metrics exist for measuring domain usability of existing user interfaces. Thus the *main goal of this paper* is the design of novel, formal *domain usability metrics*. The metrics could be used in manual (Bačíková et al., 2017) or automatized (Bačíková and Porubän, 2014) evalua-

[1]The founding father of the famous marketing management theories: Decision Making Unit (DMU) and the Five Product Levels.

BaÄ■Ã■kovÃą M., Galko L. and HvizdovÃą E.
Experimental Design of Metrics for Domain Usability.
DOI: 10.5220/0006502501180125
In *Proceedings of the International Conference on Computer-Human Interaction Research and Applications (CHIRA 2017)*, pages 118-125
ISBN: 978-989-758-267-7

tion of an existing user interface to represent *formal measurement of the target user interface's domain usability*. Combining with our design of manual and automatized evaluation techniques, we hope that our metrics would aid developers and improve the situation related to domain usability of user interfaces.

In our recent research (Bačíková et al., 2017) we tried to confirm or disprove the following hypothesis:

H: All domain usability aspects have equal impact on the overall usability and user experience.

We performed two experimental usability tests in the domain of gospel music with a mobile application from the same domain. We focused on consistency errors and language barriers. The results were inconclusive, but suggested the invalidity of the hypothesis H. Thus the following research question was raised:

RQ1: Are the five domain usability aspects equal? Do they have the same effect on overall usability and user experience?

Based on *RQ1*, the task of designing domain usability metric further raises the next research question:

RQ2: How to design a domain usability metric, given the hypothesis H proves to be invalid?

To address *RQ1* and *RQ2* and to fulfil the main goal of this paper, we state the following research tasks of this paper:

T1: Perform a survey with a sufficient number of users to evaluate the effect of five domain usability aspects on domain usability.

T2: Design a metric for formal evaluation of domain usability by using the results from the performed survey.

2 EQUALITY OF DOMAIN USABILITY ASPECTS

In our previous research we proposed several techniques for evaluating domain usability of existing user interfaces (Bačíková et al., 2017). In many of them we assume that all aspects of domain usability are equally important and have the same effect on user experience and usage. However this might not be true. As showed the experiments with evaluating consistency and language errors and barriers (Bačíková et al., 2017), some aspects might be significantly more or significantly less important for user performance or user experience than others, which suggest that hypothesis H might be disproved.

Suppose that the target user interface's domain usability would be measured formally. To achieve that, we could count all components of the application that contain any textual information. Then we would analyse all components for any domain usability issues. Having the number of all application terms n and erroneous terms e, we could determine the percentage of user interface's correctness, while 100% would represent the highest domain usability and 0% would be the lowest. Any component might have multiple domain usability issues at once (e.g. an unsuitable term and and a typo). If this would be the case of all user interface components, then the result would be lower than zero, thus we have to limit the resulting value. Given that each domain usability aspect has a different weight, we would define the formula to measure domain usability as follows:

$$\max(0, 100 * (1 - \frac{e}{n})) \tag{1}$$

where e can be calculated as follows:

$$\begin{aligned} e = {} & w_{dc} * n_{dc} + w_{ds} * n_{ds} + w_c * n_c \\ & + w_{eb} * n_{eb} + w_l * n_l \end{aligned} \tag{2}$$

Coefficients w_x ($x \in dc, ds, c, eb, l$) would be weights of particular domain usability aspects as follows:

- n_{dc} - the number of domain content issues,
- n_{ds} - the number of domain specificity issues,
- n_c - the number of consistency issues,
- n_{eb} - the number of language errors and barriers,
- n_l - the number of world language issues,

The weights w_x will be determined by a survey with multiple users that rated domain usability aspects in scale 1-5. Given that the results of the survey will point to inequality of domain usability aspects (thus hypothesis H will be disproved), the rating will be used to calculate the above stated weights w_x.

3 THE SURVEY

The goal of the survey was to validate whether some domain usability aspects have a significantly stronger effect on general usability and user experience than others. At the same time we aimed to confirm the results of our previous experiments.

In other words, in the survey we aimed to confirm or disprove hypothesis H and the following new hypotheses:

H1: The results will confirm our previous experiments, i.e. that language errors and barriers have no or small effect on domain usability.

H2: The results will confirm our previous experiments, i.e. that consistency has a strong effect on domain usability.

The survey was designed with the aim to determine the weights of particular domain usability aspects using the rating of survey participants. It is harder to design a survey in a specific domain, since for the survey to be valid, domain users or experts are needed. Finding a sufficient number of domain users in a specific work domain is more problematic. Thus, to get as many responses as possible, we used a general participant sample. In order to ensure equality between respondents, we tried to select such examples for the questionnaire, which would be known to any user.

3.1 Questionnaire Design

The questionnaire is composed of two parts. The first part contains 5 questions, in which the task of the participants is to rate particular domain usability aspects based on examples. Each of the 5 questions is aimed at one domain usability aspect. Questions are in form of visual examples - screenshots from different user interfaces. Each example contains a particular domain usability issue (corresponding to the particular domain usability aspect) and, to be sure the participant understands the issue, a supplementary explanation is provided.

After studying the example, the participants are asked to mark their view of the importance of a particular aspect by a number from the 1-5 Likert scale (Tomoko and Beglar, 2014) where 1 is the least important and 5 is the most important aspect. For unipolar scale such as ours, 5 points are recommended (Garland, 2011). At the same time, this amount preserves the consistency with the number of domain usability aspects.

We assumed, that in these 5 questions, multiple aspects might seem equally important to some participants. Thus we added a second part to the questionnaire, encouraging the participants to sort domain usability aspects according to their importance with the least important first, to let them think about the order more from a retrospective view.

The questionnaire was designed using Google Forms [2].

[2]The questionnaire can be found at: http://hornad.fei.tuke.sk/bacikova/domain-usability/surveys

3.2 Sample Selection

To ensure maximal coverage, we targeted the sample at the age between 14 and 44 years with at least minimal experience with using mobile or web applications. According to statistics (The Statistics Portal, 2014), internet and web pages are used by 73.6% of users in this age category.

3.3 Sample Description

The questionnaire was filled by 73 respondents of age between 17-44 years (with average of 24). Google Forms does not directly support sorting questions. Therefore, the second part of the questionnaire was, created by using five questions with selection boxes (one for each aspect). However, this type of form enables to input multiple duplicate aspects into the form. We excluded 4 responds with such duplicate answers in the second part of the questionnaire, which left us with 69 answers in overall.

3.4 Results

Ratings in the first questionnaire part were converted into the 0-4 range by subtracting 1. Answers related to aspects sorting were converted according to their order (less important aspect = 0, most important aspect = 4).

Results of the questionnaire can be seen in tables 1 and 2. In both tables the sum of rating for each aspect is noted in the third column and average rating in fourth column. Percentage ratio of gained and maximal possible rating is noted in the last column in both tables. The results are sorted according to the percentage ratio.

Table 1: First part of the general survey (control group).

DU Aspect	$\sum$	$w_x = \varnothing$	$\frac{max\_rating}{w_x}$
Language barriers	252	3,65	91,30%
Consistency	213	3,09	77,17%
World language	196	2,84	71,01%
Domain content	194	2,81	70,29%
Domain specificity	173	2,51	62,68%

Table 2: Second part of the general survey (control group).

DU Aspect	$\sum$	$w_x = \varnothing$	$\frac{max\_rating}{w_x}$
Domain content	164	2,38	59,42%
Language barriers	159	2,30	57,61%
Consistency	138	2,00	50,00%
Domain specificity	118	1,71	42,75%
World language	111	1,61	40,22%

The results between the first part of the questionnaire are quite different to the second part. In the first part of the questionnaire, the percentage ratio was quite similar in case of consistency, world language and domain content. In the second part, the rating of all aspects are relatively similar. Moreover, the answers of the respondents did not really match between each other.

The numbers in Tab. 2 are significantly lower than in Tab. 1, which we assumed was because the users were forced to explicitly order the domain usability aspects in the second questionnaire part. After looking at the results closely, we found out that for many participants, all aspects seemed highly important.

The results of this survey are quite different from our previous experimentation. Since language errors and barriers were previously shown to be less significant for the application's usage, we were surprised that it was rated so important by all participants. On the other hand, consistency, which in our previous experimentation showed as strongly important to abide, was behind language errors and barriers in both cases with ratings slightly above average.

I.e., the results (surprisingly) indicate the confirmation of hypotheses *H1* and *H2*. As for the hypothesis *H*, we cannot confirm nor disprove it. Although language barriers and errors have a significantly higher rating in the first part of the questionnaire, the ratings in the second part are quite similar for all aspects.

3.5 Discussion and Next Steps

We assume that there are two reasons for such results. First, that our focus was on a general domain. In general domains, the importance of domain usability aspects might not be so explicit as in specific domains, where using a general term might cause unexpected problems during usage.

The second reason is what we called *"imagined usage"* or *"imagined user experience"*. Despite that we tried to design the examples as understandable as possible and as general as possible, so that every participant could have experience with the user interfaces presented in the examples, the participants have not actually used the presented applications. They had to *imagine* the potential usage and based on that (not based on a real experience), they tried to answer the questionnaire. That might be the result that all domain usability aspects seemed the most important to them and also for the disunity of answers.

Based on the inconsistency of the results with our previous experimentation we decided to modify the survey by targeting at domain-specific users with the experience with a domain-specific application containing domain usability issues. At the same time, performing the survey in a specific domain will enable us to compare the results with the first attempt.

4 THE SECOND (DOMAIN SPECIFIC) SURVEY

The modified domain-specific survey was aimed at the specific domain of *gospel music*. For the purposes of the survey we developed a specific application called *Worshipper*[3] designed for gospel singers and musicians. All participants of the domain specific survey come from the domain of gospel music and all of them had a previous experience with this application, thus we assumed that the answers will be more relevant and the results will differ from the general survey.

The comparison of both results will be performed as an experiment, thus, from this point, we will declare the sample of the general survey as the *control group* and the sample of the modified (domain specific) survey will be the *experimental group*. We will formulate the hypothesis of the experiment as follows:

H3: Results of the general survey will be in correspondence with the results of the domain specific survey.

We expect the hypothesis *H3* to be disproved by the experiment.

4.1 Questionnaire Design

The domain-specific questionnaire was designed in the same manner as its general variant. Only in this case the particular pictures contain use cases focused on specific parts of the domain application *Worshipper*.

4.2 Sample Selection

Target sample is a group of people of age between 15-44 years and average age of 23 years consisting of singers or guitarists from the domain of gospel music, who participated in previous experiments targeted at domain usability issues. In these experiments, the *Worshipper* application was used with manually created domain usability errors in its user interface. Since they "lived through the experience" of using a

[3]https://play.google.com/store/apps/details?id=com.evey.Worshipper

faulty application with domain usability issues, the results will be more relevant than in the general survey.

4.3 Sample Description

The *Worshipper* questionnaire was filled by 26 gospel singers and guitarists, who have been users of this application. 11 of them participated in our previous experimentation with manual domain usability evaluation methods (Bačíková et al., 2017), so they experienced domain usability issues first-hand. To the rest of the participants we explained the meaning of domain usability explicitly and let them use a version of the *Worshipper* application with consistency issues. By this we ensured that all participants were equally experienced with domain usability issues in a domain-specific application known to them. To ensure equal knowledge of domain usability aspects to all participants, we shortly explained the idea to each of them.

4.4 Results Evaluation

The results of the second survey can be seen in Tab. 3 and 4. From the tables it is possible to see that the results of the control and experimental sample are clearly different by the order of the aspects, their average rating and percentage ratio, thus disproving the hypothesis *H3*. We assume that for the same reason, the results of both the first and second part of the domain specific questionnaire are more similar and preserve almost the same order of aspects. Also, the answers of the experimental group were more unite.

Language barriers and errors are on a significantly lower position and in the second part of the questionnaire they have percentage ratio under average. Summing with the results of the first questionnaire part (50%), the overall value is still below average, which confirms hypothesis *H1* for specific domains.

Consistency achieves a 10.58% above-average significance, which confirms the hypothesis *H2* for specific domains.

The results of the domain-specific survey are, unlike its general variant, in correspondence with our experimental results. The reason clearly is the domain-specific application, participants' first-hand contact with it and "live" experience of domain usability issues.

We can conclude that domain usability really has a significant impact on usability and user experience. All aspects significantly differ from each other, which disproves the original hypothesis *H*.

Table 3: First part of the domain-specific questionnaire (experimental group).

DU Aspect	$\sum$	$w_x = \varnothing$	$\frac{max\_rating}{w_x}$
Domain content	85	3,27	81,73%
Domain specificity	79	3,04	75,96%
Consistency	70	2,69	67,31%
Language barriers	52	2,00	50,00%
World language	44	1,69	42,31%

Table 4: Second part of the domain-specific questionnaire (experimental group).

DU Aspect	$\sum$	$w_x = \varnothing$	$\frac{max\_rating}{w_x}$
Domain content	66	2,54	63,46%
Consistency	63	2,42	60,58%
Domain specificity	57	2,19	54,81%
Language barriers	38	1,46	36,54%
World language	36	1,38	34,62%

4.5 Threads to Validity

Because of the lack of time and human resources we were not able to find a sufficient number of participants in the domain-specific survey. Meaning, the sizes of control and experimental group were not equal, which could affect the results.

In both surveys, we used only qualitative approaches along with questionnaires.

In most of the questions, respondents in the control group expressed their opinions hypothetically. This means they only imagined how they would react in case of the presented situation, which might be slightly different from reality.

5 THE DOMAIN USABILITY METRIC

Because the domain specific survey confirmed the hypotheses formulated on previous experimental findings and the answers were more consistent this time, the results of this survey will be used to design the metric of domain usability. By merging the first and the second part of the domain specific questionnaire we gained overall weights of the particular aspects (summarized in Tab. 5).

When we substitute the weights w_x (where $x \in dc, ds, c, eb, l$) in the formula (2) as follows:

$$e = 2.9 * n_{dc} + 2.6 * n_{ds} + 2.6 * n_c + 1.7 * n_{eb} + 1.54 * l \quad (3)$$

then the *formula (1) represents the metric of domain*

Table 5: Overall ratings of the particular domain usability aspects based on the domain-specific survey.

DU Aspect	$\sum$	$w_x = \varnothing$	$\frac{max\_rating}{w_x}$
Domain content	75,5	2,90	72,60%
Domain specificity	68	2,62	65,38%
Consistency	66,5	2,56	63,94%
Language barriers	45	1,73	43,27%
World language	40	1,54	38,46%

usability with the consideration of its aspects and with the result in percentage.

The average weights have been rounded to one decimal place considering that in this case, hundredths are negligible, since they do not remarkably change the overall result. The variables of domain content n_{dc}, domain specificity n_{ds}, consistency n_s, language errors and barriers n_{eb} and world language n_l again represent the number of errors found in the analysed user interface.

The sum of multiplies between weights and errors is divided by the number of all components n. All components can contain domain information such as terms, icons or tooltip descriptions. One domain component can have one or more domain usability issues (e.g. language error in its description and also an inaccurate description). The result is converted into percentage and in case of a negative number, the result will be 0%.

To interpret the results, evaluators can follow Tab. 6 where we can see the interpretation of the results achieved using the designed metric. The interpretation corresponds to the scale, in which the participants rated the particular aspects.

Table 6: Interpretation of the ratings achieved via the proposed domain usability metric.

Rating	Interpretation
100 - 90%	Excellent
90 - 80%	Very good
80 - 70%	Good
70 - 55%	Satisfactory
less than 55%	Insufficient

For a demonstration, we will use the *Worshipper* application, which contains 63 graphical components. We will use its older version that was created before our first user testing, during which we discovered 5 domain usability issues. As can be seen in Fig. 1, the issues were mainly connected with domain specificity (iv), domain content (ii, iii) and consistency (i and iii), as can be seen in Fig. 1. Errors of other aspects were not discovered.

If we substitute the values of n_x (where $x \in dc, ds, c, eb, l$) in the formula (3),

$$e = 2.9*2 + 2.6*1 + 2.6*1 + 1.7*0 + 1.5*0 = 11 \quad (4)$$

and the resulting error rating e into the formula (1):

$$\begin{aligned} m_{du} &= max(0, 100*(1-\frac{11}{63})) \\ &= max(0, 100*(1-0.1746)) = 82.5\% \end{aligned} \quad (5)$$

we will get the resulting domain usability rating of the application as 82.5%, which according to the Tab 6 can be interpreted as very good. At the same time the result roughly corresponds with our previous results using System Usability Scale questionnaire (Brooke, 2013), which was 86,5%.

In our further research we will use the designed metric in the DEAL tool (mentioned in the introduction) to automatically evaluate the domain usability of user interfaces. DEAL is able to extract terminology from an existing user interfaces and based on that it is able to automatically discover issues of domain specificity, content, language barriers/errors and world language. However, since consistency errors are not possible to evaluate in an automatized manner, in this case the analysis needs to be performed manually by using the technique of consistency inspection (Bačíková et al., 2017).

6 RELATED WORK

The following paragraphs summarize the state of the art works directly referring to the aspects of domain usability. Their terminology might differ from our definition. According to our knowledge, there are no metrics of domain usability similar to ours in the current literature.

To begin with the *domain content* aspect, there is a lot of existing literature from different authors referring to this topic. Jacob Nielsen generally refers to the topic of domain content using the concept "*textual content*" of user interfaces. He stresses the important feature that the system should address the user's mental model (Nosáľ and Porubän, 2015) (Porubän and Chodarev, 2015) of the domain (Nielsen, 1993). Accordingly, Shneiderman (Shneiderman, 1984), Becker (Becker, 2004) and Kincaid (Kincaid et al., 1975) (Kincaid and W.C., 1974) say that the less complex the textual content is the more usable the application will be. Complexity is related to the aspect of domain content in ways of appropriate terminology for target users.

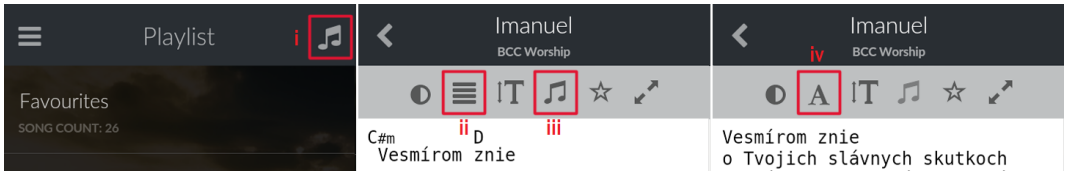

Figure 1: Domain usability issues in the *Worshipper* application.

Kleshchev (Kleshchev, 2011), Artemieva (Artemieva, 2011) and Gribova (Gribova, 2007), whose works address to the importance of ontology and domain dictionary of user interfaces, presented a method for estimating a user interface usability using its domain models. A *general ontology* is also used as a core of *semantic user interfaces*, proposed by Tilly and Porkoláb (Tilly and Porkoláb, 2010) to solve the problem of ambiguous terminology of user interfaces. Despite of user interfaces's different arrangement and appearance, domain dictionary must remain the same.

In the experiment of Billman et al. (Billman et al., 2011) usability of an old and new application was compared. An improvement in NASA user's performance was shown in the new application that had used domain-specific terminology.

Badashian et al. (Badashian et al., 2008) and many of the above listed authors present the importance of designing user interfaces that match with the real world and correspond to their domain of use. Furthermore, Hilbert and Redmiles (Hilbert and Redmiles, 2000) stress that even event sequences of applications should correspond with the real world processes and domain dictionary.

In addition, *consistency* is another aspect that many authors refer to. The importance of this feature is presented by Badashian et al. (Badashian et al., 2008). Ivory and Hearst (Ivory and Hearst, 2001) analyse and compare multiple automatic usability methods and tools. Mahajan and Shneiderman (Mahajan and Shneiderman, 1997) designed a tool called Sherlock that can check the consistency of user interface terminology automatically. However, with this tool it is not possible to check whether the same functionality is described by the same term.

Aspects of *world language*, *language barriers and errors* are studied by Becker (Becker, 2004), who deals with user interfaces and their translations.

Authors Isohella and Nissila (Isohella and Nissila, 2015) address the *appropriateness* of user interface terminology and its evaluation by the users. According to the authors, the more appropriate the information system's terminology is, the higher is its quality. Chilana et al. (Chilana et al., 2010) stress that deeper study of the target domain is needed to provide a successful and usable product in difficult domains.

Formal languages (Kollár et al., 2013) (Tomášek, 2011) (Šimoňák, 2012) (mainly domain-specific languages) are closely related to domain usability. Multiple authors identified the relation between the system's model (Porubän and Chodarev, 2015) to user interface features (Nosál' and Porubän, 2015) (Szabó et al., 2012).

The number of works mentioned above indicates the importance of domain usability and a need for domain usability metrics to be able to formally evaluate user interfaces from this point of view.

7 CONCLUSIONS

In this paper we described our experimentation that led to the design of a formal domain usability metric. The experiment confirmed our previous experimentation, thus we believe that the metric can be used to correctly measure domain usability of user interfaces.

Our further research includes the use of the designed metric in our method for automatized domain usability evaluation implemented by the DEAL tool to formally measure the domain usability of automatically analysed user interfaces.

Some areas of research are still open considering the designed metrics. One of them is to consider the impact of the component positioning and domain specificity on domain usability.

ACKNOWLEDGEMENTS

This work was supported by the KEGA grant no. 047TUKE-4/2016: Integrating software processes into the teaching of programming

REFERENCES

Artemieva, I. L. (2011). Ontology development for domains with complicated structures. In *Proceedings of the First international conference on Knowledge processing and data analysis*, KONT'07/KPP'07, pages 184–202, Berlin, Heidelberg. Springer-Verlag.

Bačíková, M. and Porubän, J. (2014). *Domain Usability, User's Perception*, pages 15–26. Springer International Publishing, Cham.

Badashian, A. S., Mahdavi, M., Pourshirmohammadi, A., and nejad, M. M. (2008). Fundamental usability guidelines for user interface design. In *Proceedings of the 2008 International Conference on Computational Sciences and Its Applications*, ICCSA '08, pages 106–113, Washington, DC, USA. IEEE Computer Society.

Bačíková, M., Galko, L., and Hvizdová, E. (2017). Manual techniques for evaluating domain usability (in review process). In *2017 IEEE 14th International Scientific Conference on Informatics*. Technical University of Košice, Slovakia.

Bačíková, M. and Porubän, J. (2013). Ergonomic vs. domain usability of user interfaces. In *2013 The 6th International Conference on Human System Interaction (HSI)*, pages 159–166.

Becker, S. A. (2004). A study of web usability for older adults seeking online health resources. *ACM Trans. Comput.-Hum. Interact.*, 11(4):387–406.

Billman, D., Arsintescucu, L., Feary, M., Lee, J., Smith, A., and Tiwary, R. (2011). Benefits of matching domain structure for planning software: the right stuff. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 2521–2530, New York, NY, USA. ACM.

Brooke, J. (2013). Sus: A retrospective. *J. Usability Studies*, 8(2):29–40.

Chilana, P. K., Wobbrock, J. O., and Ko, A. J. (2010). Understanding usability practices in complex domains. In *Proc. of the SIGCHI Conf. on Human Factors in Comp. Syst.*, CHI '10, pages 2337–2346, NY, USA. ACM.

Garland, P. (2011). Yes, there is a right and wrong way to number rating scales. https://www.surveymonkey.com/blog/2011/01/13/how-to-number-rating-scales/. cit. 2017-4-20.

Gribova, V. (2007). A method of estimating usability of a user interface based on its model. *International Journal "Information Theories & Applications"*, 14(1):43–47.

Hilbert, D. M. and Redmiles, D. F. (2000). Extracting usability information from user interface events. *ACM Comput. Surv.*, 32(4):384–421.

Isohella, S. and Nissila, N. (2015). Connecting usability with terminology: Achieving usability by using appropriate terms. In *IEEE IPCC'15*, pages 1–5.

Ivory, M. Y. and Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Comput. Surv.*, 33(4):470–516.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Technical report.

Kincaid, J. P. and W.C., M. (1974). An inexpensive automated way of calculating flesch reading ease scores. patient disclosure document 031350. Us Patient Office, Washington, DC.

Kleshchev, A. S. (2011). How can ontologies contribute to software development? In *Proceedings of the First international conference on Knowledge processing and data analysis*, KONT'07/KPP'07, pages 121–135, Berlin, Heidelberg. Springer-Verlag.

Kollár, J., Halupka, I., Chodarev, S., and Pietriková, E. (2013). Plero: Language for grammar refactoring patterns. In *2013 Federated Conference on Computer Science and Information Systems, FedCSIS 2013*, pages 1503–1510.

Lanthaler, M. and Gütl, C. (2013). Model your application domain, not your json structures. In *Proc. 22Nd Int. Conf. on WWW '13 Companion*, pages 1415–1420, NY, USA. ACM.

Mahajan, R. and Shneiderman, B. (1997). Visual and textual consistency checking tools for graphical user interfaces. *IEEE Trans. Softw. Eng.*, 23(11):722–735.

Nielsen, J. (1993). *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Nosáľ, M. and Porubän, J. (2015). Program comprehension with four-layered mental model. In *13th International Conference on Engineering of Modern Electric Systems, EMES 2015, art. no. 7158420*.

Porubän, J. and Chodarev, S. (2015). Model-aware language specification with java. In *13th International Conference on Engineering of Modern Electric Systems, EMES 2015, art. no. 7158424*.

Shneiderman, B. (1984). Response time and display rate in human performance with computers. *ACM Comput. Surv.*, 16(3):265–285.

Szabó, C., Korečko, v., and Sobota, B. (2012). Data processing for virtual reality. *Intelligent Systems Reference Library*, 26:333–361.

The Statistics Portal (2014). Distribution of internet users worldwide as of november 2014, by age group. https://www.statista.com/statistics/272365/age-distribution-of-internet-users -worldwide/. cit. 2017-4-20.

Tilly, K. and Porkoláb, Z. (2010). Automatic classification of semantic user interface services. In *Ontology-Driven Software Engineering*, ODiSE'10, pages 6:1–6:6, New York, NY, USA. ACM.

Tomášek, M. (2011). Language for a distributed system of mobile agents. *Acta Polytechnica Hungarica*, 8(2):61–79.

Tomoko, N. and Beglar, D. (2014). Developing likert-scale questionnaires. In *N. Sonda, A. Krause (Eds.), JALT Conference Proceedings*. JALT.

Šimoňák, S. (2012). Verification of communication protocols based on formal methods integration. *Acta Polytechnica Hungarica*, 9(4):117–128.