User Satisfaction of a Hybrid Ontology-engineering Tool

Christophe Debruyne[1] and Ioana Ciuciu[1,2]

[1] Vrije Universiteit Brussel – STARLab
chrdebru@vub.ac.be
[2]Joseph Fourier University – SIGMA / LIG
ioana-georgiana.ciuciu@imag.fr

Abstract. In an effort to continuously improve a research prototype for collaborative ontology engineering, we report on the reapplication of a usability test within an ontology-engineering experiment involving 36 users. The tool offers additional functionalities and measures were taken to address the problems identified in a previous study. The evaluation criteria proposed in the study were developed by taking into account the people involved, the processes and their outcomes, focusing on the user experience, in an approach that goes beyond usability; users were asked if the tool helped them in achieving their goals. We identify the problems the users encountered while using the system and also investigate whether the measures did tackle the problems observed in the first study. A set of recommendations is proposed in order to overcome these new problems and to improve the user experience with the system.

Key words: Usability testing, user satisfaction, socio-technical systems theory, ontology engineering, social dynamics, community, HCI

1 Introduction

We present the result of an ongoing study on the usability and user satisfaction of a collaborative ontology-engineering tool developed in the context of the Open Semantic Cloud for Brussels project[1]. The tool – which we will describe later in this paper – offers additional functionalities and measures were taken to address the problems identified in a previous study, reported in [2]. This tool was then used in an experiment similar in size and complexity as the experiment mentioned in [2], of which the user satisfaction will be presented in this paper. Next to identifying the main (usability) problems to draw conclusions and recommendations for future improvement, we also examine whether the problems reported in the previous study were properly addressed.

The rest of the paper is organized as follows: Section 2 provides the background of the paper. The usability test design is described in Section 3 and Section 4 reports on the results and presents some recommendations for improvement. Section 5 concludes and presents the future work of this research.

[1] http://www.oscb.be/

2 Background

Usability is defined by the ISO-9241 standard [7] as *the effectiveness, efficiency and satisfaction with which specified users can achieve specified goals in particular environments.* Usability is a key factor in making the systems easy to learn and to use. Usability testing has been extensively studied and applied by Lewis [10] at IBM Software Group. Classically, usability tests gather both subjective and objective data coming from realistic use case scenarios, as well as descriptions of the most common problems encountered by the test participants [8]. Subjective data reflect the participants' opinions regarding the evaluated system, while objective data reflect the participants' observed performance when using the system.

The focus of this study is the *user satisfaction* of an ontology-engineering system based on dynamic social processes. A classic way to measure the user satisfaction is via questionnaires (e.g., After-Scenario Questionnaire - ASQ, Computer System Usability Questionnaire - CSUQ [8], System Usability Scale - SUS [1]). However, a common mistake is to rely on questionnaires only while evaluating the user satisfaction. There are alternatives to measuring satisfaction with a questionnaire, e.g., the Microsoft "Desirability Toolkit". However, while the questionnaires are often biased towards positive responding, this tool helps elicit negative comments from participants [13].

A proven standard and effective instrument to assess the user satisfaction is the Post-Study System Usability Questionnaire (PSSUQ). PSSUQ was developed for scenario-based usability evaluation at IBM [8]. The environment used was an enterprise-wide and networked office application suite. A follow up study by IBM [9] was performed in the domain of speech recognition [9] using data from five years of usability studies. The follow up produced similar psychometric properties as the original survey. Fruhling and Lee [6] validate these results of the PSSUQ instrument and assess its adaptability to other domains, such as telemedicine. The reason for choosing PSSUQ for this study is mainly the richness of the provided information, with little effort from the user, and the extensive IBM documentation and experience for the statistics it can provide.

2.1 The Post-study System Usability Questionnaire (PSSUQ)

In this study, the user satisfaction was measured using a standard instrument, namely the Post-Study System Usability Questionnaire (PSSUQ) [8,9] developed by IBM. PSSUQ originally consisted of 19 questions, each question being a statement about the usability of the system. Participants need to answer each statement using a Likert scale of 7 points, where 1 indicates that the user "strongly agrees" with the statement whilst 7 indicates that the user "strongly disagrees" with it. PSSUQ is based on a comprehensive psychometric analysis, providing scales for three sub-factors, namely: (1) system usefulness; (2) information quality; and (3) interface quality. The short version of PSSUQ (and the most recent one, see Table 1) was used with the purpose of saving study time.

Table 1. PSSUQ - short version [10] . The questions correspond with the three sub-factors as follows: (1) System usefulness: the avg. of items 1 through 6; (2) Information quality: the avg. of items 7 through 12; (3) Interface quality: the avg. of items 13 through 15; (4) Overall: the avg. of items 1 through 16.

Item	**Item Text**
Q01	Overall, I am satisfied with how easy it is to use this system.
Q02	It was simple to use this system.
Q03	I was able to complete the tasks and scenarios quickly using this system.
Q04	I felt comfortable using this system.
Q05	It was easy to learn to use this system.
Q06	I believe I could become productive quickly using this system.
Q07	The system gave error messages that clearly told me how to fix problems.
Q08	Whenever I made a mistake using the system, I could recover easily & quickly.
Q09	The information provided with this system was clear.
Q10	It was easy to find the information I needed.
Q11	The information was effective in helping me complete the tasks and scenarios.
Q12	The organization of information on the system screens was clear.
Q13	The interface of this system was pleasant.
Q14	I liked using the interface of this system.
Q15	This system has all the functions and capabilities I expect it to have.
Q16	Overall, I am satisfied with this system.

PSSUQ is used in order to measure the user satisfaction when dealing with GOSPL (the collaborative ontology-engineering method and tool described in the next section). An advantage is that besides the 16 items in the test, the test participants can make comments and elaborate on their answers. Based on these comments, conclusions are drawn and recommendations for improving the human-system interaction provided.

2.2 GOSPL

GOSPL [4] is a method and tool for collaborative hybrid ontology engineering. A hybrid ontology is an ontology in which the community is promoted to first-class-citizen and all ontology evolution operators are grounded with the community agreements in which information between human stakeholders are exchanged in natural language [11]. It supports communities of stakeholder in collaboratively achieving an approximation of the domain to support their semantic interoperability requirements. Hybrid ontologies are ontologies where concepts are both described informally in natural language by means of glosses for high level reasoning between the community members and formally suitable for machine reasoning and data annotation.

Starting from co-evolving communities and requirements, the informal descriptions of key terms have to be gathered before formally describing those concepts. Communities define the semantic interoperability requirements, out of which a set of key terms is identified. Those terms need to be informally described before the formal description can be added. Concept are represented formally by means of lexons [12], which depicts a relation between to terms (referring two concepts) that hold in a particular context and in which the two roles of that relation are made explicit. In order for a lexon to be entered, at least one of the terms needs to be articulated. The terms and roles in lexons can be constrained and the community can then commit to the hybrid ontology by annotating an individual application symbols with a constrained subset of the lexons. At the same time, communities can interact to agree on the equivalence of glosses and the synonymy of terms. Synonyms are agreements that two terms refer to the same concept, and gloss-equivalences are agreements that two descriptions refer to the same concept. Committing to the ontology allows for the data to be explored by other agents via that ontology. Commitments also enable the community to re-interpret the ontology with its extension (i.e. the instances in each annotated system). This will trigger new social processes that lead to a better approximation of the domain, as the community is able to explore the increasingly annotated data, e.g., by formulating queries.

Ever since the publication of [4], the following functionalities have been added to the GOSPL prototype:

- Explicit social processes for defining the key terms and goals that constitute the semantic interoperability requirements of a community.
- Social processes for communities to agree that terms and roles in formal descriptions refer to the same concepts as classes and properties in other ontologies stored somewhere on the Web (e.g., OWL ontologies).
- Tool support for collaboratively managing application commitments.
- An RSS feed such that users did not need to check the platform for any new discussions and observe the activity by means of an RSS reader. The RSS feed has for each community commitment a separate channel, allowing one to filter to the community of interest.
- A reputation framework[2]. The reputation framework provides users how well he or she performs with respect to following the method. The reputation framework also took into account an evaluation of other users on a user's action. The users were presented "scores" as to encourage them to do better.

The problems reported in [2] were addressed as follows:

- **"The (error) messages displayed by the system were often not clear to the user. There was in general no online help or documentation available"** While teaching the method to the participants, they were offered a document and slide set (available online) in which the method and tool were explained. A running example for the creation of an application commitment was also provided.
- **"There is no "undo" or "edit" option available"** The problem reported here was not so much to undo an action, but rather to edit mistakes such as

[2] Details on the reputation framework will be reported elsewhere

typos. Also, the outcome of a discussion sometimes differs from the initial proposition. For some social interactions, the users are now able to conclude with the final outcome.

- **"No (top menu) link to the current community in the discussion page"** The style of the prototype has been adapted and the link to the community is made more obvious.
- **"It took a while to understand how the system works"** The availability of online documentation should solve that problem.
- **"Sometimes, listing items in the dynamic tables did not go well when after returning to a page it displayed the first item again"** This remark basically boiled down to search parameters being stored in a session such that users did not have to enter the same filter every time they leave the page. This was easily solved by storing the filters in a cookie.
- **"There was no "delete" option for the communities who "died" during the process"** As explained in [2], we did not wish to provide such a feature, as one can never know when a particular community can have an uptake. We therefore did not provide such functionality for the next experiment.
- **"The user name is not clear (just email addresses appear)"** We did not request users to provide an additional username, instead we strongly encouraged the users to use their institution's address or an address containing their names.
- **"Sometimes, more clicking necessary that one would expect (e.g. when browsing through several discussions)"** In communities with many discussions, browsing through the different discussions could have been cumbersome. This has been partly tackled by storing the filters in a cookie.

3 Test Design

The user satisfaction when interacting with the GOSPL prototype was assessed within a larger ontology-engineering experiment with a group of MSc students of a course on ontology engineering. The goal of the test is to evaluate the usability of GOSPL in two dimensions: *formative* and *summative*, from a user satisfaction point of view. The formative usability testing aims at identifying the usability problems of the tool. The summative usability test consists of a series of measurements (e.g., effectiveness, efficiency, satisfaction) which are performed in order to compare the usability results against a set of predefined objectives.

The objective of the experiment is to create a prototype ontology capturing the (shared) concepts and relations of two applications involving cultural events (e.g. concerts, exhibitions). One information system (IS) is developed by the experiment participants and one application whose database schema and data is provided to them. Both applications are portals. The objectives identified in the test are thus: (1) the ontology creation and (2) the annotation (of the IS and the existing database) with the ontology, together with their subsequent subtasks:

- Propose discussion;

- Discuss and vote;
- Conclude (accept/reject discussion);
- Create and manage a community;
- Use the ontology to annotate existing information systems.

The annotation of the existing systems is not a social process across groups of stakeholders and only concerns the users “representing” their own information system. The annotation of these systems, however, can result in new discussions as insights are gained while annotating the systems.

The satisfaction test was undertaken by a group of 23 end-users. The answers on the survey are depicted in Table 2. These end-users were part of a wider experiment involving 36 MSc in Computer Science students. In this experiment, the students were formed groups up to 4 persons to: develop their own information system, create a prototype ontology to enable semantic interoperability between those systems, and annotate their systems with the ontology.

The purpose of this study is to assess the user satisfaction with the system. The results are reported in the following section. The overall usability testing was carried out both implicitly by analyzing the data logs and the user-system interactions and explicitly, by collecting the user feedback via several questionnaires. The outcome of the experiment highlights three aspects of the evaluation: 1) effectiveness; 2) efficiency and 3) satisfaction [10]. Following the recommendations in [5] we have developed the evaluation criteria looking at the people involved, the processes and their outcomes. Some groups were completely represented in the survey (one group of 2 persons, 4 groups of 4 persons). The groups are also found in Table 2. As groups obviously worked together to ensure their own systems were annotated and hence divided the work, we will analyze the results looking at the groups and all the participants as a whole. We will furthermore compare these results with the ones reported in [2].

4 User Satisfaction: Results and Recommendations

The results delivered by the PSSUQ questionnaire are as follows (cfr. Table 3): **SysUse**: the average of all groups remained the same as in the previous study. The overall average, however, had a small decline in satisfaction with 0.1 points. **InfoQual**: compared with the previous study, the system performed better in terms of information quality with 0.3 points for the groups and 0.4 points for the overall average. **IntQual**: for both the group average and the overall average, the interface quality was deemed more satisfying with 0.1 points compared to the previous study. **Overall**: the system performed slightly better in terms of user satisfaction.

4.1 Formative User Satisfaction

Of the 23 respondents, 17 have left comments. These comments were analyzed to pinpoint the problems of the tool. When indicating an occurrence, this corresponds with a respondent making a remark about the issue at least once.

Table 2. Respondents. On the left the answers of each of the respondents on questions Q01 to Q16. On the right we have the number of communities created (Con), discussions started (Pro), interactions in a discussion (Dis), number of votes (Vot), the number of discussions concluded (Con) and total (Tot) of all aforementioned numbers for each respondent.

User	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Com	Pro	Dis	Vot	Con	Tot	Group Tot.
Group A																							1051
x1	3	2	3	2	3	2	5	3	3	2	3	2	2	2	2	5	0	24	82	156	55	317	
x2	5	5	5	4	2	4	6	5	3	3	3	5	4	4	3	4	0	16	45	106	61	228	
x3	2	2	4	1	3	4	3	3	2	5	2	1	4	3	2	2	1	41	39	151	24	256	
x4	3	2	3	2	2	2	2	4	3	2	3	4	3	4	4	2	1	74	16	132	27	250	
Group B																							1693
x5	3	3	3	2	4	3	2	3	5	3	3	2	2	2	2	2	3	84	81	187	109	464	
x6	2	2	2	3	3	3	2	3	4	4	4	4	3	3	5	3	2	85	90	365	105	647	
x7	3	3	4	3	3	4	2	3	5	3	3	2	3	2	2	2	3	82	76	136	76	373	
x8	2	3	3	3	2	3	6	5	4	4	3	4	3	3	4	3	2	46	7	103	51	209	
Group C																							701
x9	5	4	2	2	4	6	6	6	5	3	3	3	4	6	6	6	2	71	35	46	38	192	
x10	4	3	5	5	5	5	4	6	4	6	4	3	6	5	5	4	0	12	51	163	75	301	
x11	4	4	5	6	3	5	4	5	4	4	4	4	5	6	4	6	0	1	4	18	2	25	
x12	5	3	3	3	2	4	6	4	3	4	3	5	5	3	5	3	0	2	11	161	9	183	
Group D																							354
x13	5	6	5	4	7	5	5	6	4	4	5	6	3	3	6	5	0	7	12	58	9	86	
x14	3	3	4	3	2	4	2	4	3	4	3	3	4	4	2	3	0	11	11	59	5	86	
x15	3	2	2	2	2	3	4	2	2	2	2	1	3	3	3	3	0	3	12	52	2	69	
x16	3	3	3	2	2	3	5	2	5	2	2	2	2	2	2	2	1	34	24	38	16	113	
Group E																							448
x17	3	3	2	3	2	3	7	5	5	6	5	3	2	2	5	3	1	31	8	83	20	143	
x18	2	2	1	1	4	1	2	1	5	3	1	2	1	1	3	1	0	58	57	152	38	305	
Remainder																							N/A
x19	3	3	2	2	2	3	3	2	3	2	2	1	1	1	2	2	1	23	13	36	6	79	
x20	4	2	6	5	5	2	3	6	2	6	1	1	2	2	6	3	0	49	23	35	32	139	
x21	3	2	2	4	2	3	3	2	2	1	4	1	5	5	2	3	0	13	5	23	0	41	
x22	3	3	5	4	2	6	4	5	3	5	2	3	2	3	5	4	0	38	8	127	37	210	
x23	5	4	5	3	2	4	5	5	4	4	4	6	3	4	3	4	0	13	29	139	7	188	

Some respondents commented about a certain issue multiple times in different comment sections of the survey. The problems identified by the users in the comments section of each item are as follows:

- Keeping an overview of the discussions (10 occurrences). Some proposals have been made to tackle this problem: 2 respondents proposed a central notification system, one respondent proposed the ability to follow the actions of a particular user, another proposed an RSS feed per community complementing the overall feed[3]. Other proposals were: identifying the "hottest" discussions, offering the changes after last login and even a search function over the whole system.
- The new version of the prototype offered possibilities for creating and managing application commitments, which are built according to a particular grammar. Users noted that the errors while parsing application commit-

[3] Even though one could filter on channel

Table 3. Summative user satisfaction: results looking at the groups, average of all complete groups, average all respondents and results of the previous study.

Metric	A	B	C	D	E	Average groups	Average all users	Average reported in [2]
SysUse ($Q01 \rightarrow Q06$)	2.9	2.9	4.0	3.4	2.3	3.1	3.2	3.1
InfoQual ($Q07 \rightarrow Q12$)	3.2	3.5	4.3	3.3	3.8	3.6	3.5	3.9
IntQual ($Q13 \rightarrow Q15$)	3.1	2.8	5.0	3.1	2.3	3.3	3.3	3.4
Overall ($Q01 \rightarrow Q16$)	3.1	3.1	4.4	3.3	2.8	3.3	3.3	3.4

ments were often too obscure to be practical and had to rely too much on our help (5 occurrences).

- Correcting mistakes (5 occurrences). The changes made to the tool to cope with mistakes or changes in a discussion proved inadequate to improve the user's satisfaction. They wished the ability to "undo" or "cancel" an interaction. One respondent also suggested the possibility to alter a comment.
- Problems creating constraints (4 occurrences). Surprisingly, the functionality of building constraints has not changed and yet 4 people reported that the construction of constraints was confusing. 2 of these 4 respondents also mentioned that the verbalization of these constraints were not clear.
- The voting mechanism (4 occurrences). 3 respondents wished the system would require a justification when one is against a proposal. With an additional respondent stating the voting system to be inadequate for stimulating the discussion.
- Imposing the GOSPL Method (2 occurrences). The tool has been developed for the GOSPL method, yet some freedom is allowed as to offer possibilities to deviate from the method or adopt other methods such as Business Semantics Management [3] that prescribe similar activities in a different order.
- Even though documentation was available as well as a running example in the slide set, participants wished for documentation within the tool next to the material offered (5 occurrences) and more examples (4 occurrences). Three participants merely noted there was not enough documentation without providing further details.
- Availability of concrete (worked out) examples and tutorials next to the documentation (3 occurrences).
- Availability of help functionality within the tool rather than online in separate documents (2 occurrences).
- Lastly, there were 4 complaints of the back button resulting in a warning on a discussion page. This needed to be added as the reputation framework kept track of the discussions visited by a user and one of the popular browsers not capturing the event of clicking the back-button properly. To this end, we asked the users whether they "wished to leave the page", whereupon a click on the button "Yes" called the method for logging.

4.2 Recommendations for Improvement

Taking the satisfaction results obtained from PSSUQ and the user comments, we drive the following conclusions: out of the three sub-factors identified by PSSUQ the system usefulness performed best (3.2). The users of this study seemed to be less satisfied than in the previous study. Information quality had a fairly important improvement in terms of satisfaction with respect to the previous study. Taking into account the complaints on error handling of the commitment manager (which was added to the prototype), we can conclude this is a very positive evolution. Both the interface quality and overall satisfaction evolved positively. The following steps will be taken to improve the system:

1. Investigating how the overview of all the discussions can be improved.
2. Improving the interface for managing the application commitments.
3. Allowing actions to be undone (i.e., "cancel" or "undo") in case of error.
4. Improving the verbalization and forms for constructing constraints.
5. Participants complained that the voting system did not require users not agreeing to a proposal to justify their opinion. The goal of the voting system was to allow users to participate to discussions in a "lightweight" manner. After the experiment, however, we feel that the voting mechanisms did not contribute to the discussion and sometimes lead to confusion. We therefore will most remove the voting mechanism.
6. Imposing the method. In other words, add pre-conditions to the social interactions such that the tool is completely compliant with the GOSPL method, thereby loosing the possibility to use the tool with other methods.

We observed a need for more worked out examples and the availability of help functionality within the tool rather than online in a separate document.

5 Conclusion and Future Work

In conclusion of this study, we provide a synthesis of the results in Fig. 1 illustrating the 3 sub-factors per group. Compared to the previous iteration of the user satisfaction testing, and as discussed in the previous section, we observe a overall improvement of the user satisfaction. In particular, positive evolutions have been obtained regarding the information quality and the interface quality.

Future work will consider testing the user satisfaction in particular and the usability testing in general from a socio-technical systems theory point of view with users from various domains, different than students.

Acknowledgements This study was supported by the INNOViris Open Semantic Cloud for Brussels project, financed by the Brussels Capital Region.

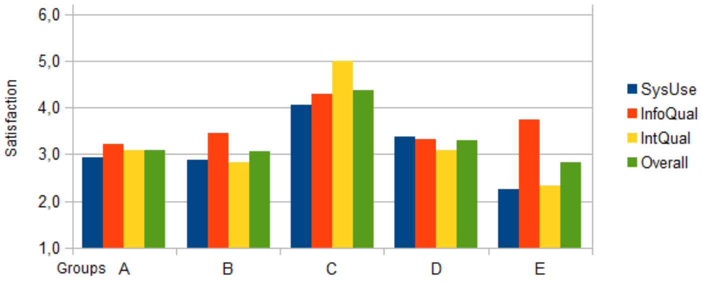

Fig. 1. Bar chart of user satisfaction results per sub-factor, per group.

References

1. Brooke, J.: Sus-a quick and dirty usability scale. Usability evaluation in industry (1996) pp. 189–194
2. Ciuciu, I., Debruyne, C.: Assessing the User Satisfaction with an Ontology Engineering Tool Based on Social Processes. In: OTM Workshops: Herrero, P., Panetto, H., Meersman, R., Dillon, T.S. (eds.). Vol. 7567 of LNCS, Springer (2012) pp. 242–251
3. De Leenheer, P., Christiaens, S., Meersman, R.: Business semantics management: A case study for competency-centric HRM. Computers in Industry **61**(8) (2010) pp. 760–775
4. Debruyne, C., Meersman, R.: GOSPL: A Method and Tool for Fact-Oriented Hybrid Ontology Engineering. In: ADBIS: Morzy, T., Härder, T., Wrembel, R. (eds.). Vol. 7503 of LNCS, Springer (2012) pp. 153–166
5. Dillon, A.: Beyond usability: process, outcome and affect in human-computer interactions. Canadian Journal of Library and Information Science **26**(4) (2008)
6. Fruhling, A.L., Lee, S.M.: Assessing the reliability, validity and adaptability of pssuq. In: AMCIS: Khazanchi, D., Zigurs, I. (eds.), Association for Information Systems (2005) pp. 378
7. ISO: ISO 9241-11: Ergonomic requirements for office work with visual display terminals (vdts) – part 11: Guidance on usability. Technical report, ISO (1998)
8. Lewis, J.R.: IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use (technical report 54.786). Technical report, Human Factors Group, IBM (1993)
9. Lewis, J.R.: Psychometric evaluation of the PSSUQ using data from five years of usability studies. International Journal of Human-Computer Interaction **14**(3-4) (2002) pp. 463–488
10. Lewis, J.R.: Usability testing. In: Handbook of Human Factors and Ergonomics. 4 edn. John Wiley (2012) pp. 1267–1312
11. Meersman, R., Debruyne, C.: Hybrid Ontologies and Social Semantics. In: Proceedings of 4th IEEE International Conference on Digital Ecosystems and Technologies (DEST 2010), IEEE Press (2010)
12. Meersman, R.: The Use of Lexicons and Other Computer-Linguistic Tools in Semantics, Design & Cooperation of Database Systems. In: CODAS. (1999) pp. 1–14
13. Travis, D.: Measuring satisfaction: Beyond the usability questionnaire. http://www.userfocus.co.uk/articles/satisfaction.html (2009)