Data Triangulation in a User Evaluation of the Sealife Semantic Web Browsers

Helen Oliver, Patty Kostkova, and Ed de Quincey

City eHealth Research Centre (CeRC), City University London,
Northampton Square, London, UK
{helen.oliver.1,ed.de.quincey}@city.ac.uk, patty@soi.city.ac.uk

Abstract. There is a need for greater attention to triangulation of data in user-centred evaluation of Semantic Web Browsers. This paper discusses triangulation of data gathered during development of a novel framework for user-centred evaluation of Semantic Web Browsers. The data was triangulated from three sources: quantitative data from web server logs and questionnaire results, and qualitative data from semi-structured interviews. This paper shows how triangulation was essential in validation and completeness of the results, and was indispensable in ensuring accurate interpretation of the results in determining user satisfaction.

Keywords: Semantic Web Browsers, User Evaluation, Data Triangulation, Healthcare Ontologies, Sealife.

1 Introduction

The Semantic Web (SW), as a realisation-in-progress of the original vision of the World Wide Web [1], aims to increase findability of specific information among the many results returned by a Web search. Semantic Web Browsers (SWBs) are emerg- ing as a potential solution, but little attention has been paid to evaluating these brows- ers to assess real-world user satisfaction.

In the course of the EU-funded Sealife project [2], we addressed this lack by developing an innovative framework for user-centred evaluation of Semantic Web Browsers [3] for the life sciences using data from 3 sources: web server logs, questionnaire results, and semi-structured interviews. The data provided invaluable insight into user thought processes and satisfaction with the SWBs.

However, it is essential to bring together the quantitative and qualitative results in order to draw the appropriate conclusions about user satisfaction. In this paper we discuss an adaptation of a triangulation method, and the triangulated results of the Sealife SWB evaluation, demonstrating the necessity of data triangulation to ensure accurate interpretation of the collected data. We show how the impression received from one type of data can be dramatically altered by another type of data.

P. Kostkova (Ed.): eHealth 2009, LNICST 27, pp. 80–87, 2010.

2 Background

Most evaluations of web portals combine qualitative data (e.g. from interviews and focus groups) with quantitative data (from weblog analysis, standard questionnaires, etc.). As each source accumulates data in answer to a particular question, combining data sets is essential to paint a more complete picture of user acceptance. Triangulation has been investigated in evaluations of the impact of digital libraries (DLs) of medicine on clinical practice [4]; of electronic transmission of medical findings [5]; and of nursing documentation systems [6]. Given the ever-increasing interest in Se- mantic Web (SW) technologies in the life sciences, user-centred evaluation making use of triangulation is indispensable in producing much-needed results.

Not only has little attention thus far been paid to user-centred evaluation of SWBs, Ammenwerth [6][7] has pinpointed a lack of attention to data triangulation as a major weakness in user evaluations of health information systems. Despite substantial contributions [4][5][6] this need still has not been addressed in such comparable user evaluations as have been done on SWBs [8][9]. We investigated triangulation methodology in our evaluation of SWBs for the Sealife project.

2.1 Sealife SWB Evaluation

The framework we created for the Sealife evaluation was tested in the first user- centred evaluation of SWBs of its kind. It was a within-subjects [10] evaluation of three SWBs for the life sciences, using live, real-world systems with established user bases as control platforms, and recruiting study participants from the real-world target audiences of the SWBs.

The SWBs that were evaluated were the three Sealife browsers: COHSE [11], a CORESE-based SWB [12], and GoPubMed [13] and its related system GoGene as well as an extended version of GoPubMed. The control platform for COHSE and the CORESE-based SWB was the NeLI Digital Library (DL) [14], which has infectious disease professionals as its target audience. The control platform for GoPubMed/GoGene, which have microbiologists as their target audience, was PubMed [15]. A detailed breakdown of methods and results is beyond the scope of this paper, but more information can be found in [3]. This paper will focus on the aspects of the study relevant to triangulation.

The evaluations were presented in web format and began with a pre-questionnaire regarding demographics and previous experience of the control platform. There followed a number of information-seeking tasks tailored to the SWB. After each task was a post-task questionnaire with two questions regarding findability and ease of use, and the evaluation ended with a post-questionnaire asking for users‘ ratings of both the control platform and the intervention SWB. The evaluation was conducted both online and in workshops; workshop participants were asked to give semi- structured interviews.

3 Use of Triangulation for Semantic Web

The triangulation in this study combined qualitative and quantitative data from three sources [16].

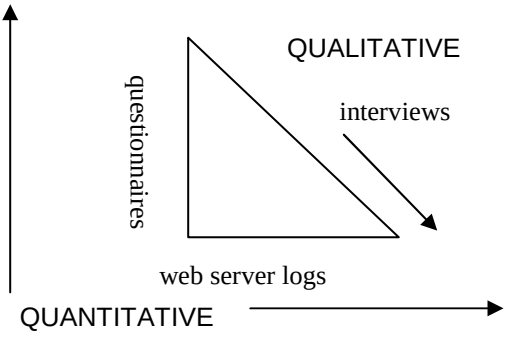

Fig. 1. Triangulation of data: qualitative and quantitative dimensions

The quantitative data was provided by the web server logs and questionnaires eliciting ratings of both the control and intervention systems; the qualitative data, by the semi-structured interviews, conducted during workshops with those participants who had the time to be interviewed. Because not all of the participants were interviewed, this evaluation was not fully triangulated; further study with full triangulation or sampling is needed and we envisage future work to develop a framework for triangulation of data in user-centred evaluation of SWBs.

3.1 Value of Data Triangulation in Interpreting the Results

The web server logs provided measures of time taken by each user to complete the tasks, and of usage of semantic links against non-semantic links, number of external pages viewed; and whether or not users viewed the target documents containing the answers to the tasks. The post-task and post-evaluation questionnaires gathered users' ratings of both the control platform and the intervention SWB in terms of information *findability, usability,* overall system *speed, relevance* of results, and overall *likeability* of the system. However, questionnaires could only elicit answers to the questions the evaluators thought to ask; the semi-structured interviews were essential for completeness [17] of results; in particular, observation of workshop users would tell us how intuitive they found the SWBs, complementing the questionnaire responses regarding usability. In the next section, we will illustrate how triangulation's core benefits of validation and completeness were demonstrated in our study. [6][17]

4 Sealife Results

The validity and completeness of the Sealife results were attained with triangulation of the web server log and questionnaire data with semi-structured interviews of some of the participants. COHSE was evaluated by 67 participants, 39 online and 28 in workshops. The CORESE-based SWB was evaluated by 14 participants; 2 online (only one of whom completed the evaluation) and 12 in workshops. GoPubMed was evaluated by 137 participants online and 4 in an informal workshop setting where full interviews were not conducted. GoGene and an extended version of GoPubMed were evaluated by 14 participants in a single workshop. The triangulated results are

not statistically significant because of the low numbers of interviewees per workshop, and so although some of our triangulated data is quantitative, our *interpretation* of the triangulated data is entirely qualitative. In this section, we will discuss the data gathered from the three sources of logs, questionnaires, and interviews in order to demonstrate that it is indispensable to combine them by triangulation.

4.1 Web Server Logs

Behind the gathering of data were a number of implicit assumptions. In considering time taken to complete tasks, we assumed that faster completion of tasks was better and that greater use of semantic links was better, where "better" equates to greater user satisfaction. The assumptions become risky if applied uncritically to a single dimension of the data. Ammenwerth has explained the thinking behind the quantitative approach thus: "The results of a measurement are clearly interpretable. Any subjec- tive interpretation is not helpful, and therefore, has to be avoided." [6] On the other hand, Brown [18] argues: "Anyone expecting to arrive at a picture of user-behaviour from web-log analysis is likely to be disappointed." Table 1 shows the average times spent using each system, gathered from the web server logs:

Table 1. Average time for all tasks by all users on each system in seconds

GoGene	GoPubMed	COHSE	CORESE	PubMed	NeLI
229	126	478	266	194	387

This shows that tasks completed using PubMed were completed more quickly than the GoGene tasks. If, as we assumed, faster is better, did users prefer the control platform of PubMed to the slower, implicitly "worse" intervention SWB of GoPubMed? The logs only tell us the speed at which users worked; it does not tell us how that speed affected, or was affected by, user satisfaction. The significance of the web log data could only be determined by asking the users.

4.2 Questionnaires

Details of the questionnaires can be found in [3]; the results paint an informative picture of users' attitudes. GoPubMed/GoGene were rated the highest in the dimensions of likeability, information findability, relevance, and system speed. The one dimension in which GoPubMed/GoGene did not "win" was usability; the questionnaires seemed to portray COHSE as the most usable of the SWBs. Overall, the GoPubMed and GoGene semantic browsers received far more positive ratings than either COHSE or the CORESE-based SWB, with more and larger differences in mode ratings between the control system (PubMed) and the intervention system. In no case did GoPubMed/GoGene receive *worse* mode scores than the control platform, whereas COHSE and the CORESE-based SWB received several inferior mode scores.

4.3 Semi-structured Interviews

Examining the questionnaire results alone might lead us to believe that GoPubMed/GoGene, the overall "winner", was reasonably well liked but that COHSE was

considered the most usable. However, questionnaire respondents could only answer the questions that the evaluators chose to ask. GoPubMed included free text fields, which elicited important feedback about accessibility: “Looks great be careful with the colors as dyslexic people find some color difficult to read”. It was in the work- shops, however, that the most dramatic discrepancies between our assumptions and reality were revealed. It became apparent early that user interface (UI) maturity was a fundamental, rather than a superficial, concern, with the unpolished UIs of the univer- sity-developed research applications COHSE and CORESE a serious impediment to usability. The maturity of the much more abundantly-resourced GoPubMed/GoGene UI elicited critiques at a much higher level of functionality than the other SWBs, which were difficult for participants to use at all. To test intuitiveness, the early work- shops presented the SWBs with minimal introduction. Online evaluations had been running for some time, but observing user behaviour, and hearing interview feedback, immediately made it clear that the SWBs were not intuitive to their target audiences as they were to us as computing professionals. Later workshops were preceded by brief explanatory presentations, which reduced users’ confusion but were not (we were told) in-depth enough to eliminate it. Most startlingly, it was discovered at the earliest workshops that many users could not tell the difference between the control platform and the intervention SWB, and much of the feedback from the first set of in- terviews turned out to be critiques of the control platform, the NeLI DL. When one such user was asked her opinion of the COHSE link boxes, the participant replied: “Those awful little boxes? They were really distracting, I didn’t really understand what they were.” Explanatory presentations eliminated the problem, but users still ex- pressed difficulty: one cited the “busy-ness” of the NeLI home page as a source of confusion between the control and intervention platforms, and another remarked that there was “not much difference” between the NeLI DL alone and the NeLI DL en- hanced by the COHSE service.

5 Sealife Evaluation: Validation and Completeness of Results

The value of triangulation is that it provides *validation* and *completeness* to the results of a study. [17][6]. This was certainly the case with the Sealife evaluation.

5.1 Validation

We were somewhat expecting triangulation of user data to show discrepancy [17] between what users said and what they did, and between statements made in person and responses entered into web forms. This was certainly the case for COHSE’s find- ability ratings – at workshops where some users rated this as adequate or good, the logs showed that none of that session’s participants had actually found the answer, (which was very specific and contained in a single target document). Other than this, we found that individuals who were interviewed tended to be consistent in their inter- views, questionnaire responses, and logged actions. One user worked quickly through the COHSE tasks and was so unusually positive in her ratings and comments about it that we suspected her responses were not genuine and should be discounted. However, the weblogs showed that time spent on each task was between one and two minutes per task: fast, but two others were faster. Logs also showed that she activated

4 link boxes, which matched the median number for all respondents. She viewed only one external page, which one might seize upon as confirmation of duplicity, only to realise that some users did not visit any external pages, and among those who did, one page was the mode.

5.2 Completeness

Interestingly, while COHSE interviewees who rated the SWB negatively often had spent substantial time on each task (more than the expected 5 minutes, and more than they spent on the control platform), several GoPubMed/GoGene users who spent more time on GoGene than on PubMed or the extended GoPubMed spoke of GoGene as their favourite and rated it highly in the questionnaires. One respondent spent just under 14 minutes on the four PubMed tasks, just under 10 minutes on the three ex- tended GoPubMed tasks, and just under 19 minutes on the four GoGene tasks. She gave the PubMed tasks a high rating (92% of the maximum score), the extended GoPubMed tasks 67% of the maximum score, but GoGene 100%. She stated in the in- terview that the SWBs were "very useful tools" but also mentioned difficulties using the extended GoPubMed. Two other users showed similar patterns in their triangu- lated data, spending the longest time on the GoGene tasks but rating and describing it as the best one. We therefore cannot jump to the conclusion that spending more time completing tasks implies that the SWB is worse (or better).

6 Discussion

The GoPubMed/GoGene workshop tended to confirm positive impressions; the CORESE-based SWB workshop confirmed the negative questionnaire results. How- ever, the GoPubMed/GoGene workshop also confirmed that the issues with this SWB were the most trivial and that the *somewhat* higher questionnaire ratings mask a *dra- matically* better user experience. While the other SWBs were rated rather negatively by the questionnaires, impressions of COHSE's greater usability were quashed by contact with the users in person; and the severity of users' problems would have gone undetected without interviews. We had hoped to gather observational data of user ac- tions in situ, and the use of eye tracking software was considered, but time constraints prevented implementation of this or other forms of recording such as video; this will inform planning for future work. While the study fell short of complete triangulation because not all participants were interviewed, recruitment of in-person participants, particularly of busy clinicians, was difficult and resource-intensive. Recruiting enough to attain statistical significance for all data sources would have been impracti- cal even had it been possible. In future work, careful sampling of a subset of individu- als for interview might be a better solution than trying to interview 100% of a large number of participants.

7 Conclusion

We have developed a method of triangulating quantitative and qualitative data in user centred evaluation of SWBs, addressing a need for greater attention to a technique

which is essential for accurate interpretation of data. Having previously applied the framework we developed for user-centred evaluation of SWBs, we triangulated quantitative data from the web server logs and from questionnaires eliciting ratings of users' satisfaction with a number of dimensions of the system, with qualitative data from semi-structured interviews eliciting users' opinions on matters which were im- portant to the users but which had not necessarily been considered by the evaluators. This triangulation was demonstrated to be essential in building up a true interpretation of the results, as impressions built up from one type of data changed dramatically in light of another type of data. Answers about system speed were provided by log data, but the meaning of the answers could only be found in the questionnaires and inter- views. Questions about usage of semantic links compared with non-semantic links, and whether or not users found the answers to tasks, could only be answered by log data; but questionnaires and interviews revealed discrepancies between users' reports and their actions. Questions about the intuitiveness of the system were partly an- swered by questionnaire results, but the full meaning and significance of the results was only discovered in the interviews.

The ultimate question about user satisfaction was only answerable by triangulating the data from all three sources. If any one of the three modes of data collection had been excluded, the evaluation results might have been severely misinterpreted.

Acknowledgments. This paper is a direct result of the work of Gawesh Jawaheer, Gemma Madle (CeRC); Dimitra Alexopoulou, Michael Schroeder (TU Dresden); Bianca Habermann (Scionics); Simon Jupp, Robert Stevens (University of Manchester); and Khaled Khelif (INRIA Sophia-Antipolis).

References

1. Berners-Lee, T., Hendler, T., Lassila, O.: The Semantic Web. Scientific American 284, 34–43 (2001)
2. Schroeder, M., Burger, A., Kostkova, P., Stevens, R., Habermann, B., Dieng-Kuntz, R.: From a Service-based eScience Infrastructure to a Semantic Web for the Life Sciences: The SeaLife Project. In: Workshop on network tools and applications in biology, NETTAB 2006, Pula, Italy (2006)
3. Oliver, H., Diallo, G., de Quincey, E., Alexopoulou, D., Habermann, B., Kostkova, P., Schroeder, M., Jupp, S., Khelif, K., Stevens, R., Jawaheer, G., Madle, G.: A User-Centred Evaluation Framework For The Sealife Semantic Web Browsers. BMC Bioinformatics, Special Issue (in press, 2009)
4. Madle, G., Kostkova, P., Mani-Saada, J., Roy, A.: Lessons Learned From Evaluation of The Use of The National Electronic Library of Infection. Health Informatics Journal 12, 137–151 (2006)
5. Machan, C., Ammenwerth, E., Schabetsberger, T.: Evaluation of the Electronic Transmission of Medical Findings from Hospitals to Practitioners by Triangulation. Methods of Information in Medicine 44, 225–233 (2005)
6. Ammenwerth, E., Iller, C., Mansmann, U.: Can evaluation studies benefit from triangulation? A case study. International Journal of Medical Informatics 70, 237–248 (2003)

7. Ammenwerth, E., Gräber, S., Bürkle, T., Iller, C.: Evaluation of Health Information Systems: Challenges and Approaches. In: Spil, T.A.M., Schuring, R.W. (eds.) E-health systems diffusion and use, pp. 212–236. IGI (2005)
8. Reichert, M., Linckels, S., Meinel, C., Engel, T.: Student's Perception of A Semantic Search Engine. In: Proceedings of the IADIS Cognition And Exploratory Learning In Digital Age (CELDA 2005), Porto, Portugal, pp. 139–147 (2005)
9. Uren, V., Motta, E., Dzbor, M., Cimiano, P.: Browsing For Information By Highlighting Automatically Generated Annotations: A User Study And Evaluation. In: K-CAP 2005: Proceedings of the 3rd International Conference on Knowledge Capture, pp. 75–82. ACM Press, New York (2005)
10. Hoeber, O., Yang, X.D.: User-Oriented Evaluation Methods for Interactive Web Search Interfaces. In: 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology Workshops, pp. 239–243. IEEE CS Press, Los Alamitos (2007)
11. Yesilada, Y., Bechhofer, S., Horan, B.: Dynamic Linking of Web Resources: Customisation and Personalisation. In: Wallace, M., Angelides, M.C., Mylonas, P. (eds.) Advances in Semantic Media Adaptation And Personalization. Springer Series on Studies in Computational Intelligence, vol. 93, pp. 1–24. Springer, Berlin (2008)
12. Diallo, G., Khelif, K., Corby, O., Kostkova, P., Madle, G.: Semantic Browsing of a Do- main Specific Resources: The Corese-NeLI Framework. In: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent technology, WI-IAT 2008, pp. 50–54. IEEE, Sydney (2008)
13. Doms, A., Schroeder, M.: GoPubMed: Exploring Pub. Med. with the Gene. Ontology. Nucleic Acids Research 33 (Web Server Issue), W783–W786 (2005)
14. Diallo, G., Kostkova, P., Jawaheer, G., Jupp, S., Stevens, R.: Process of Building a Vocabulary for the Infection Domain. In: 21st IEEE International Symposium on Computer-Based Medical Systems, pp. 308–313. IEEE, Los Alamitos (2008)
15. Pub. Med., `http://www.ncbi.nlm.nih.gov/pubmed/`
16. Denzin, N.: Strategies of Multiple Triangulation. In: Denzin, N. (ed.) The Research Act. a theoretical introduction for sociological methods, pp. 297–331. Aldine, Chicago (1970)
17. Greene, J., McClintock, C.: Triangulation in Evaluation. Evaluation Review 9, 523–545 (2005)
18. Brown, S., Ross, R., Gerrard, D., Greengrass, M., Bryson, J.: RePAH: A User Requirements Analysis for Portals in the Arts and Humanities - Final Report. Arts & Humanities Research Council, UK (2006)