Why Evaluating Semantic Web Applications is Difficult

Jacco van Ossenbruggen, Alia Amin and Michiel Hildebrand

ABSTRACT

This position paper discusses our experience in evaluating our cultural search and annotation engine. We identify three aspects that determine the quality of a semantic web application as a whole, namely: the quality of data set, the quality of underlying search and inference software and the quality of the user interface. We argue that evaluation of semantic web applications is particularly difficult because of the strong interdependency between the three aspects.

INTRODUCTION

The authors of this paper participate in the MultimediaN E-Culture project. The goal of the project is to build end-user applications using state-of-the art Semantic Web technology in a domain that should be ideal for this purpose: cultural heritage. Cultural heritage is an ideal domain because it has a tradition of creating rich metadata that goes centuries back. As a result, cultural heritage institutions typically have many interesting cultural objects, which have typically been carefully annotated by hand, and these annotations have been rechecked on their quality by curators. In addition, many institutions use carefully crafted controlled vocabulary or thesauri that provides the terms that are using in the annotations, and for each of these terms the thesauri typically define the meaning of the term and its relationship with other terms. While most of the project's research team members are Semantic Web advocates, this setting has given the project also a healthy, skeptic undertone: "If we cannot even show that the Semantic Web is useful in this ideal domain, well . . . "

THE MULTIMEDIAN E-CULTURE DEMONSTRATOR

Our project's online demonstrator includes quite a number of Semantic Web features[1]. It even won the first prize in the ISWC Semantic Web Challenge 2006. It won before any formal user testing on the demo had been carried out. The fact that until that point no formal user testing had been applied, was *not* because we felt this was not necessary. It was because at that point in time, we could, even after long brainstorm sessions, not come up with a good evaluation design that would tell us the things we wanted to know. The main goal of this paper is to explain why we think our project's application is so hard to evaluate and why we think the same arguments apply to many other, if not almost every, semantic web application.

Issue 1: quality of the data set

Our application is all about meaningful interaction with complex data. We believe this is true for all realistic Semantic Web applications: they are by definition about complex data sets. If all your data is local and homogeneous, if it fits nicely into a single relational database and if all your users' information needs are best expressed as SQL queries, you would probably not have built the application on Semantic Web technology in the first place. From the data perspective, the use of Semantic Web technology only becomes interesting if the data is distributed, heterogeneous and bulky. And a usability study only becomes interesting if this distributed, heterogeneous and bulky data actually has interesting content the participants in the study can relate too. This implies that any usability study will need to make non-trivial assumptions on the quality of the RDF data set that is used.

First of all, for most domains it is hard to find existing RDF data sets that meet these criteria. In our case, the project explicitly targets Dutch cultural heritage professionals. Since there was no existing (public) RDF data set we could use that addresses the needs of these professionals, we needed to make one within the project. Creating such a data set is a non trivial task [2, 3]. It requires modeling and conversion of large amounts of heterogeneous instance data set to RDF, modeling and conversion of heterogeneous vocabularies and thesauri to SKOS and/or OWL, alignment and mapping of these vocabularies to one another, and mapping the often plain text metadata to the proper terms from the vocabularies. In principle, the quality of the resulting data of every step in this process needs to be evaluated. Because as long as the data does not make sense to our users, usability testing on interfaces built on top of this data will have too much noise to be useful.

Issue 2: Quality of the underlying search and inference software

Our data is too bulky and complex to expose users directly to the raw underlying data. For all UI components used, advanced middleware software is needed to allow these components to efficiently search, retrieve and navigate meaningful bits of data, and this software is exploiting the semantics in the data. Again, we believe this to be true for all realistic Semantic Web applications: if the software doesn't exploit the semantics in the data, it is by definition not a Semantic Web application. But again, this implies that a usability study will need to make non-trivial assumptions on the quality of the middleware software. We are constantly con-

[1] `http://e-culture.multimedian.nl`

fronted that algorithms and heuristics that work well with one particular data set, perform unacceptable on another; approaches that work fine for one ontological modeling choice, work less for another. Getting your middleware to a quality level where it can be effectively demonstrated is one thing, getting it to a level where its limitations no longer effect a realistic user study, is not trivial at all.

Issue 3: Quality of the user interface

It may very well be that there are applications where all the semantic web technology remains hidden under the hood, and that for these applications, a traditional interface works fine. From a research perspective, however, we are interested in user interfaces in which semantic web issues play a key role. In our research, we focus on two categories: unified interfaces on heterogeneous and distributed data that used to be only accessible by using several different and isolated applications; and interfaces for tasks that are currently hard, or impossible without deploying semantic web technology. We give an example for each category.

Our keyword search is basically a traditional search interface on data that searches within integrated data, and is a good example of the first category. We assume that for this application, the first two issues are solved, that is, our RDF data and search software is of sufficient quality to make sense our users (i.e. cultural heritage experts). During an informal exposure to the most simple component of our interface, we found that these experts already distrust almost every aspect of our application just for the very fact that it combines information from various sources into one single application [1]. Almost all search tasks of our experts require them to rely on data from reliable sources. Because of this, they are currently used to go to a specific website or a specific database application of an authoritative organization in the area relevant for their search. After explaining the goal of our application, they appreciate the fact that we include the same information from the same authoritative source. They will, however, only take it seriously if they can clearly see the original source and verify the provenance of each data unit. The conclusion was clear: we would be unable to test any realistic search task if this issue was not solved first. Note that this did not come up during earlier interviews with the users, because in their current applications it is not an interface issue at all: there is no need to convey the source of information from organization X inside an application that only contains data from organization X.

Our relation search interface is a good example of the second category, an interface for a task that was until now not supported. It allows users to search for (semantic) relations between any two items in the data[2]. Here, the quality of the underlying search proved to be an hurdle that was too hard to overcome. Given a sufficiently dense RDF graph, any node in the graph has many, many relations to any other node. Finding relationships is not difficult, but finding which relations make sense to the user is. We have developed quite a number of relation search strategies and algorithms, each with their own pros and cons. We clearly need user testing to see which one performs better in a given situation. But the quality of non of the current algorithms comes even close to something what is needed for a user evaluation that says anything about how useful users find relationship search. The best we can do is an A-B comparison test, and conclude that for a given task and data set, search algorithm A seem to performs less worse that algorithm B.

CONCLUSIONS

We feel that the strong dependency between the user interface quality with data set quality, search and inference software quality is an important issue that needs to be raised in a forum such as the SWUI workshop. We would like to discuss ways to come to a proper research methodology to evaluate semantic web user interfaces that can minimize, or even isolate, the influence of data set and search algorithm quality. A start would be identification of a set of realistic but "representative" end user tasks for which there is currently a UI challenge, along with a suitable public RDF data set and necessary middleware technology. In this way, solutions to interface problems can be developed and compared without underlying software and data problems getting in the way. We are fully aware that the definition of "representative" will be major problem, because there seems little consensus in the Semantic Web community on what "typical" tasks for Semantic Web applications are. That may very well be an even bigger problem.

ACKNOWLEDGMENTS

This research was supported by the MultimediaN project funded through the BSIK programme of the Dutch Government and by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content K-Space.

REFERENCES

1. A. Amin, J. van Ossenbruggen, and L. Hardman. Searching in the cultural heritage domain: capturing cultural heritage expert information seeking needs. Technical report, Centrum voor Wiskunde en Informatica, 2007.

2. A. Tordai, B. Omelayenko, and G. Schreiber. Thesaurus and metadata alignment for a semantic e-culture application. In *K-CAP '07: Proceedings of the 4th international conference on Knowledge capture*, pages 199–200, New York, NY, USA, 2007. ACM.

3. M. van Assem, M. R. Menken, G. Schreiber, J. Wielemaker, and B. Wielinga. A method for converting thesauri to rdf/owl. In *Proceedings of the Third International Semantic Web Conference (ISWC'04)*, number 3298 in Lecture Notes in Computer Science, pages 17–31, Hiroshima and Japan, November 2004.

[2] `http://e-culture.multimedian.nl/demo/path`