Semantics from the Questionnaire Up

Joseph R. Utecht, B.A.[1], Firstname B. Lastname, Degrees[2]
[1]University of Arkansas for Medical Science, Little Rock, AR, USA; [2]Institution, City, State, Country (if applicable)

Abstract

This abstract should eventually be between 125-150 words long and the paper itself must be between 5 and 10 pages long.

Background

Often at the core of clinical research is a questionnaire, in which a subject or researcher will enter data. The questions on this questionnaire will attempt to ascertain whatever information would be useful in the current research. There is much writing and research into how to properly word questions and design questionnaires so that the research can more accurately capture the information they desire. However, there is less to be said about how to represent the answers to said questions. A commonly used method is to simply record the exact answers to the question. For example on a questionnaire related to smoking, if the question was worded "How many cigarettes a day do you smoke?", and the subject answered twelve the number '12' would be recorded in whatever form was being used to track answers. This shouldn't cause a problem for the original researcher as they know that the answer '12' is to the question related to number of cigarettes smoked. A problem arises when this information needs to be compared to another source of data, either a later wording of the question or a separate study.

Imagine a second study where they were only interested in studying "heavy smokers". The questionnaire in this study asked a question to find heavy smokers "Is the subject a heavy smoker?", and this would have been represented with Yes/No or True/False. Now later when comparing the data from these two studies in an attempt to make a larger research cohort you come to the problem of how to compare the answer of 'Yes' to the number '12'. There is the obvious problem of what the definition of a "heavy smoker" in the second question is, but assuming that this is a well known number say greater than 10 cigarettes a day, it will still require human intervention to map the data in to form or the other. You cannot accurately map the "Yes" of heavy smoker into a specific number of cigarettes a day. This means the data can only be mapped to the less specific form of heavy smoker yes/no. In the new dataset we would lose the information of exactly how many cigarettes the subject smoked per day. This does not even address the issue of what happens if the definition of "heavy smoker" changes over time, when comparing two studies using the same question.

We propose a different way of recording the answers to the questions, which more accurately represents what the question is asking.

Another Way Forward

Outline the basic premise of representing the answers to questions with RDF directly.

What level of familiarity with the semantic web should we assume? JB: The AMIA 2016 proceedings are over 2000 pages long, and have only four papers that mention "semantic web." So it's not all that common. But a paragraph should be enough.

Explain the basics of the semantic web, hopefully inside of a single paragraph. Maybe this paragraph should go before we recommend using the semantic web to represent questions and answers.

How to Represent Questions

Mathias should write this section about the *proper* method of representing the answer to a question in RDF. We can use a few examples here from CAFE or DIDEO.

Advantages Over Previous Methods

Expand upon the problems with previous methods and extol the virtues of our method.

Talk about the idea of representing what a question is actually asking as opposed to just recording that the respondent answered Yes to question 1 or False to question 4.

Having to codify in RDF what a question is asking can also help you correctly word your questions by really thinking about what the question is asking.

Problems and Limitations

Does this merit an entire section, or just a few sentences in the conclusion.

Our Implementations

Lots of screen shots and example of use from CAFE and DIDEO. Show off a few examples of the output from DIDEO or the full RDF graph from a single run of the CAFE trauma center questionnaire.

Maybe mention semantic survey here, or would that be better for future work?

Conclusion

The way *you've* been doing things is wrong, our way is right.

References

1. Pryor TA, Gardner RM, Clayton RD, Warner HR. The HELP system. J Med Sys. 1983;7:87-101.
2. Gardner RM, Golubjatnikov OK, Laub RM, Jacobson JT, Evans RS. Computer-critiqued blood ordering using the HELP system. Comput Biomed Res 1990;23:514-28.