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

The more traditional way of recording data through a questionnaire or survey involves some work after the data has been collected to translate the results of the data to fit into whatever larger datastore contains the information the collected data is being compared. What we are proposing is roughly the same as that translation step except from a semantic web point of view.

This has a few advantages, especially when done while still constructing the questions. If the RDF representation of the answers are in place data will immediately be in its final state after the user answers a question and can either be used by the researcher right away or even used in the questionnaire application to provide feedback to the user.

It is important to note that the RDF representations are not set in stone and could be modified at any time, even after generated triples have been inserted into another store. One of the benefits of many triplestores and other types of graph databases is that each piece of information has an original context which enables a way to remove or update records from any source.

Also we have found that the step of codifying in RDF what a question is asking can also help correctly word questions by really thinking about what the question is asking.

Also mention ability to share "original data" for the purpose of reproducability and better transparency in science

Problems and Limitations

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

Poor suitability of RDF for likert scales and things that don't resolve to yes/no?

Our Implementations

The application of this process is the focus of the CAFE project. To those means we have built a few tools for this purpose. The main tool is the CAFE questionnaire, which is a web questionnaire to be filled out by trauma center administrators. The tool records all answers directly into RDF and will eventually allow users to perform comparisons to other organizations through semantically enriched queries, while also building a semantic representation of their organization for public health researchers.

The CAFE web application is a custom built site made with Angular2 and Django the source of which is avaliable on Github (https://github.com/cafe-trauma/). The questionnaire currently consists of just over 150 questions with various types of possible answers. Currently supported are yes/no, number field, checkbox selection, and drop downs. These questions have been entered into an administrative portion of the tool by a non-developer (figure out admin question interface) The answers to these questions have RDF representations which we have created in the tool (figure of rdf entry page), this can range from a single RDF triple (figure of question 161) to a complex set of triples (figure of large rdf graph). When a user answers a question in the positive, either yes on a yes/no or any selction on the other question types, the configured RDF for that question will be inserted into a triplestore on the server. This RDF can also be generated later at anytime as the tool also records all answers to the questionnaire in the same relational database which holds the questions and potential triples. From the users point of view the questionnaire does not differ in any way from a more typical questionnaire backed by a relational database.

Another application we have built using this framework is the DIDEO(?) - evidence of potential drug-drug interaction classification tool. In this tool trained users will enter properties from reported potential drug-drug interactions through a small number of questions. These questions will generate RDF in the same way as the CAFE application. After the questions are answered a logical reasoner is run over the generated RDF which can infer additional properties thus minimizing the number of questions the user will need to fill out. In this example we are asking the user to answer 10 questions which using this inference can record the same amount of information that it would normally take many more questions to specify.

In both of these example applications the user is presented with a familiar looking survey presentation, but their answers will produce a semantically rich representation of what the questions are asking. The generated data is

immediately ready to be used with any existing RDF data or even translated back into a relational database format with all of the logical inferences intact.

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

Conclusion

What we have set forth here is methodology for representing the answers to questionnaires directly into RDF.

Advantages of semantic web, sharability/usability of generated data, question framing

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.