Response to reviewers of
ICBO 2014 - PAPER 29

Clinical Data Wrangling using Ontology and Referent Tracking

Werner Ceusters, Chiun Yu Hsu and Barry Smith

---------- REVIEW 1 ----------
OVERALL EVALUATION: 1 (weak accept)
REVIEWER'S CONFIDENCE: 4 (high)

This paper reports interesting work in the area of referent tracking and is useful as an illustration of how RT is implemented, how templates and RTTs are generated and as such makes a valuable contribution to RT in Biomedical Ontology. As a scientific paper, however, it could benefit from narrowing and focusing the questions and conclusions drawn from the work that is described.

1. The goals of the paper are not clear. The first three sentences of the abstract suggest that it will address making heterogeneous data sets comparable, but only one data set was used in the work described in the paper.

→ That is not correct. These sentences describe, as explicitly stated in the first sentence, the goals of ontological realism, not the goals of the work reported on in this paper. No action taken.

Another possibility suggested by the second and third sentences in the abstract is that the paper will address problems with data sets that contain data items that do not denote real entities (or relations), …

→ That is a too narrow interpretation of these two sentences. We don't see how it even would be possible to have research data about something that doesn't exist. What will be the case, however, is that the data are presented in a structure that does not reflect the structure of reality as conceived through ontological realism. The task at hand is thus to reformulate such datasets in a way that does reflect the structure of reality.

… but while the importance of realism to the project is stressed, it is not clear that data that defy realism were addressed (nor is it clear whether they were intended to be addressed). The following four goals are discussed in the latter half of the abstract:

a) describe particulars that are implicitly referred to.

→ Data that reference particulars implicitly is an example of such 'defiance'

b) provide information about correspondences between data-items in a data set.

→ That is correct

c) describe which data items are "unjustifiably and redundantly present or absent"

→ this also is an example of such defiance

d) provide detailed statistics about the occurrence in specific data sets of each of these issues.

→ That is correct

The described work handles each of these four goals deftly.

→ That is correct

However, on the second page four questions are set up as being addressed in the paper. These questions do not match the four above stated goals from the abstract …

→ They do, but we agree it was not made so explicit. That has been corrected.

and it is not clear how 1, 3, and 4 are addressed by the described work. This could be improved by explicitly addressing each of these in the discussion or conclusion sections or by bringing the questions of the paper in line with the goals stated in the abstract.

→ This is a good suggestion; we followed the latter approach

2. The results section need further explication. In particular, the statistics in Table two need some interpretation and discussion of their relevance to the questions posed in the paper.

→ We did so.

3. The discussion section is interesting and informative. However, the first paragraph seems to commit a category mistake. In particular, this reviewer reads the definition of 'self-explanatory' as defining a property of data repositories that depends on a comparison of two or more datasets within the repository. However, the last sentence refers to "individual datasets which are themselves ... self-explanatory." This raises the question of how an individual dataset can self-explanatory (or free of idiosyncracies) outside of the context of comparison with other datasets in a repository.

→ A dataset can be considered self-explanatory when it comes fully described in terms of realism-based ontologies as explained in this paper. Being 'free of idiosyncrasies' is a separate issue which is not guaranteed by solely being self-explanatory, but by being as we defined 'maximally self-explanatory'. We clarified this in the discussion.

4. Conclusion - Rather than addressing the four questions from page 2 the conclusion section mentions that RTTs might replace ETLs. This is the first mention of ETLs in the paper, and as such this comparison seems more appropriate for a "Future Research" section that as a conclusion.

→ The questions and goals were addressed in the results section. We do not see any problem in making suggestions, specifically obvious ones as the suggestion here, in a conclusion.

---------- REVIEW 2 ----------
OVERALL EVALUATION: 2 (accept)
REVIEWER'S CONFIDENCE: 5 (expert)

This paper describes a method for converting a research dataset into a maximally explicit and self describing representation of reality according to the principles of Referent Tracking. As opposed to describing a new ontology, a new approach to ontology, or a new application of ontological principles to some existing terminology, thesaurus, vocabulary, or dictionary, it addresses the critically important, but often ignored, problem of how do we put ontologies into practice once they are built? That is to say, what should data look like that are collected according to the ontology as opposed to how do we retrofit existing data to ontologies.

→ That is correct

As a first step, the authors take a research dataset and convert it into Referent Tracking tuples. This approach thereby gives the reader an understanding of how these datasets would have looked had they been collected using a Referent Tracking System in the first place.

→ That is correct

At the same time, they encountered several issues that heretofore had not been identified, and thus advance the field towards solving problems in achieving interoperability with ontology.

My main concern with the paper is that it is difficult to understand at times. My major suggestion in that regard is to show and describe the 3 variables from the dataset that the authors decompose in Table 1 up front. It might also help to show examples of how data in those 3 variables in the source table end up looking as RT tuples (although I appreciate the space constraints of the proceedings format).

→ We improved the legend of table1 in the hope to make it easier to understand. We will also make sure that the presentation slides – which will be made publicly available – contain these examples.

Also, more concretely tying the limitations of IAO to the work described is important, as it represents one of the major contributions of the work (for both this and above comment, see specific comments for more information about how these issues might be addressed).

→ We tried to do so.

Specific comments:

At the end of page 2, the authors should make clear that #3 denotes the second polyp in the second scenario (the original benign polyp is not the malignant one found later).

→ We did so.

Page 3, paragraph beginning “Another goal of RT is to make explicit..." . The phrase “...some of which resulting from..." should be “...some of which result from..."

→ The phrase has been changed

The reader is referred to Table 1 before s/he has enough information to understand it. For example, codes in the RT column are explained later in the text, and the acronym IUI is not defined first. The reader might be helped if the authors at least say here that the text that follows will develop the information in Table 1, or an understanding of it.

→ We expanded the legend of Table 1, and also refer in it to the section where ample detail is available.

Calling column 1 in table 1 “record number” led to some confusion on the part of this reviewer, and thus could confuse the reader also.

→ Agreed. We changed ‘record number’ to ‘template line number’ or ‘line’ (L) for short

It was not immediately obvious at first whether each row in Table 1 is attached to a column in the source table, or a row. If column, then obviously table 1 shows a subset of the total number of variables in the source table (n=161). Also, it appears that RN=4 through 7 are dealing with one column or variable, so it does not seem to be the case that one row in Table 1 corresponds one-to-one with a column in the source table. The best solution perhaps is orient the reader by pointing out the Var column in table 1 as denoting one of the 161 variables in the source table.

→ We added this information to the legend.

Similarly, although I understand the space limitations of the proceedings format are constraining, it would be very helpful to see what q3 looks like in the variable codebook. It seems that the answer to the question is a coded value of 0 or 1, where the former means no pain the lower face, and 1 means pain in the lower face in the past month.

→ That is correct. It is described in section B of the discussion

What an_8_gcps_1 denotes is not clear. If the answer to q3 is 1, is there additional information required? For example is the information about "last 30 days" and the time q3 was answered (time_of_q3_concretization), only entered if the first part is answered with '1'?

→ This is explained in the discussion and we added a note to that effect in the legend of the table.

The paper would be improved if the authors could give examples from the work described that motivate their assertions in the limitations section. It is not immediately obvious how the present work motivates "a fully adequate set of relations for the various flavors of aboutness and a better theory of ICE..." For sure, one example is the "corresponds-to" relation. However, readers not familiar with the IAO and its theory of ICE will not understand that this relation is not present or defined in IAO (or that according to the method of realism followed by the authors, that no other ontology like IAO exists because the method calls for only one ontology per portion of reality).

→ We added this to the section

Some indication of the total number of tuples generated for the dataset, and tuples per record in the source table, and the overall size (in MB or GB) of the resulting RTT dataset would be interesting to know. Is there a one-to-one correspondence between RT tuples and the record types in table 1? From table 1, I am tempted to calculate ~4 RT tuples per patient per variable times 161 variables times 390 patients = ~251,160 RT tuples.

→ This estimate is correct.