Reasoners derive the logical consequences from the ontologies we've defined and work with a scale of input and output infeasible for a human to derive by hand. Hence, we're dependent upon these reasoners functioning correctly to have confidence in what is derived. Unfortunately its known that reasoners don't function perfectly.

We have developed a methodology that verifies reasoner correctness. This method is based on the concept of N-version software development, using multiple pieces of software designed in parallel for the same task to help detect errors. We use justifications to verify reasoner correctness with respect to classification of ontologies. Briefly stated, any dissent between reasoners required the production of justifications for the disagreed upon entailment – these were then verified, sometimes by human expertise.

A trial of this with 4 reasoners and the bioportal corpus produced a variety of different cases in which reasoners exhibited buggy behaviour, as well as determined the degree to which they agreed on Ontology Classification. A consequence of these disagreements is that it is generally advisable to check classification and other reasoner task results with more than one reasoner. Reasoners failed for a variety of reasons. For instance, some simply failed to deal with data type restrictions. Other failures occurred for more complex reasons that were not easily diagnosable.

It's unknown at this point the degree of skill required to evaluate the putative justifications in the corpus, nor the degree that this task can be aided by current methods. We wish to evaluate the level of skill required to judge a sample of the corpus generated from the experiment. This will be investigated with a study in which participants are required to sort putative justifications into real or fake ones.