We thank reviewers R1,R2,R3 for their constructive feedback.

Concerning the main reasons for rejection (1) since the CFP mentions “system papers” and “tools”, we are positive that our paper fits the research track (R1); (2) we wish to clarify that the “experimental evaluation of crowd sourcing” (already addressed in [9, 19]) was not the sole focus/contribution (R2) – our focus was rather on understanding and evaluating how crowdsourcing could be embedded into and impact ontology engineering (OE) workflows. Therefore understanding the types of often crowdsourced tasks and building a tool for embedding such tasks into OE scenarios were necessary, and we regard them as important, not yet available contributions to the field (as R1/R3 agree); (3) Plugin support for all mentioned tasks is under development (R3).

We agree that the evaluation can be extended with details and discussions of design decisions (R1/R2), with all requested additions and changes doable for a potential CR version as sketched next.

Setup: An HC task is complete when all requested judgments have been collected (R2). Although one can predict that plugin use shortens OE time, we saw the confirmation of this aspect and an understanding of the ratio reduction in OE time important (R2). In fact, experiments will be re-run to get detailed timing data for all stages of Figure 1 as R3 suggests.

Evaluators:
Four of the evaluators were experienced Protege users, the other four work in the Semantic Web area but have limited knowledge of Protégé and were shortly trained in Protege (R2). Since Setting 1 is more complex on the ontology engineers’ end, we gathered 8 experimental data settings to derive meaningful average values. Setting 2 is simpler, so we judged that four runs would suffice but would be necessary to get timing data and crowdsourced data for measuring data quality (R2).

Data:
We chose data generated by ontology learning [23] because 1) bootstrapping OE by extracting an initial ontology automatically is a feasible (and probably frequent) OE approach and 2) automatically generated ontologies present errors that can best be solved through human intervention (R2). Ontologies were used as generated; larger ontologies would have made the manual evaluation stage unfeasible. Our focus was not on scalability, but as the underlying framework automatically distributes tasks for parallel processing we don't expect major issues (R1). Domain choice: we perceive finance as a general knowledge domain while climate change requires domain familiarity or interest (R2). More specialised domains could be tested (as future work), but earlier work has already [9] investigated crowd-worker performance across ontologies of different domains/generality. Evaluation data, results and instructions are now online http://tinyurl.com/ucomp (R2/R3).

Results Interpretation:
We can add usability study details (R2) and/or run the SUS questionnaire with our evaluator base (R3).

The agreement between the crowd and experts is higher than among experts, possibly because crowdsourcing data is the majority view derived from 5 judgements as compared to a single expert judgement.
(R2/R3). An analysis of the results per domain ontology can be added (R2).