

ment of the ontology schema designed for obtaining thematic annotations in Figure 1. Explicit assertions are amenable to be obtained through text mining and NLP techniques, such as, the direct occurrence of the word or a phrase that directly indicates the *manifestation* of a particular theme. However, even when a word or a phrase explicitly appears in a verse, in the Arabic language, it may manifest different meanings in different contextual settings. When modelling such thematic assertions, at sub-verse level, disambiguation by a human expert becomes indispensable. To add to the challenge, themes may also exhibit implicitly; whereby, the same theme may be manifested using not just a mere difference of word morphology, rather, a difference in expression or rhetoric. It is extremely difficult to extract such implicit thematic assertions via automated techniques, therefore, it becomes imperative that human contribution be sought.

3. APPROACH: CROWDSOURCING WORKFLOW

We devised the generic workflow that a typical crowdsourcing driven method for obtaining semantic annotations will entail. This is shown in Figure 2. There are several key stages and components. *Ontology Design*: An ontology schema such as the one given in Figure 1 guides the semantic annotation process. This serves as input to the *Task Generation and Design* stage, which creates an annotation or disambiguation relevant task to be crowdsourced based on the nature of the entity, relation or both, as specified in the *task specification*. The relevant task input is generated by retrieving relevant candidate verses from the available data sources such as the Semantic Qur’an [2] dataset or the quranontology². The tasks are published on the Amazon Mechanical Turk (AMT)³ platform. A complete workflow management system is implemented (a derivative of a workflow model for Linked Data Management presented in [1]), which includes means for generating dynamic tasks from a range of task profiles. The task generation module creates the required input, question and parameter files needed by the AMT API for publishing the task. The AMT crowd performs the *disambiguation and annotation tasks*. Both tasks are based on the Arabic script of the Qur’an, therefore, requiring the crowd workers to be familiar with the Arabic language of the Qur’an. For the disambiguation task, a question is presented to the crowd, which includes a verse, along with a highlighted, candidate explicit assertion for the given theme, and the crowd responds by declaring this assertion as either a positive or negative by determining if the occurrence is a true occurrence of the given theme. The annotation tasks require deeper knowledge and understanding of the Arabic text. The crowd determines whether the given verse contains any implicit reference to the given theme. If their response is positive, then they are also required to provide the portion of the verse (a meaningful phrase or a word) that implies the presence of the theme. Each task is required to have at least five responses for it to be marked as complete. As a form of a quality measure, the crowd is also required to provide a confidence level (ranging from Very High to Very low), to indicate their confidence in their response. There is a *response collection and aggregation* module that collects and aggregates the responses based on statistical measures

²<http://quranontology.com>

³<http://www.mturk.com>

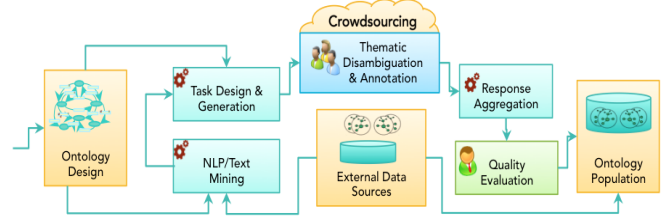


Figure 2: Generic workflow for Crowdsourcing semantic annotations for ontology population

of aggregation. Weighted confidence measures and thresholds are applied. Based on this aggregation, the completed tasks are marked as either *Approved* or *Reviewable*. A high confidence and aggregation threshold is applied for the approved tasks. An *expert review and validation* is conducted as a followup for the reviewable tasks that have a degree of disagreement more than the baseline threshold. The approved and validated annotations are passed on for *Ontology Population* and linked with existing data sources.

4. RESULTS AND DISCUSSION

For the initial pilot study, we obtained 12,000 (appx) submissions for the two types of tasks, for 70 distinct themes for each type. Of these, 1300 key verse-theme pairs were disambiguated, with explicit assertions, while 2200 pairs were disambiguated/annotated with implicit ones. Of the completed submissions, 96% tasks were approved for explicit assertions, while, for the implicit ones, 81% were approved, without applying any expert validation. While the results are promising, a considerable number of tasks remained incomplete for the implicit annotations. While these were not included in the results compilation, this indicated lower crowd engagement, which may have been due to that fact that the task statements were presented only in the Arabic language or since only the AMT sandbox was employed for the pilot study.

The results of the ongoing study suggest that the crowd can significantly assist in scaling the knowledge engineering activities such as knowledge formalization, and semantic annotation with reasonable reliability. Our ongoing efforts are focused towards augmenting this workflow to obtain expert validations for the reviewable tasks (through our own custom web framework) so that high quality annotations may be obtained. In future, we plan to increase the size of the study, and experiment with a range of other task designs of varying complexity. The study clearly shows the potential benefit of crowdsourcing that can be harnessed by knowledge intensive and expertise driven domains.

5. REFERENCES

- [1] A. Basharat, I. B. Arpinar, S. Dastgheib, U. Kursuncu, K. Kochut, and E. Dogdu. Semantically enriched task and workflow automation in crowdsourcing for linked data management. *International Journal of Semantic Computing*, 8(04):415–439, 2014.
- [2] M. A. Sherif and A.-C. N. Ngomo. Semantic Quran - a multilingual resource for natural-language processing. *Semantic Web*, 6(4):339–345, 2015.