Multi-user Input in Determining Answer Sets (MIDAS)

Albert Kalim, Satrio Husodo, Jane
Huffman Hayes, Erin Combs, Jared Payne
Department of Computer Science
University of Kentucky
Lexington, Kentucky, USA

***Abstract* - Empirical validation is an important component of sound requirements engineering research. Many researchers develop a gold standard or answer set against which to compare techniques that they also developed in order to calculate common measures such as recall and precision. This poses threats to validity as the researchers developed the gold standard and the technique to be measured against it. To help address this and to help reduce bias, we introduce a prototype of Multi-user Input in Determining Answer Sets (MIDAS), a web-based tool to permit communities of researchers to jointly determine the gold standard for a given research data set. To date, the tool permits community members to add items to the answer set, vote on items in the answer set, comment on items, and view the latest status of community opinion on answer set items. It currently supports traceability data sets and classification data sets.**

***Index Terms* - Data set, gold standard, answer set, evaluation**

I. INTRODUCTION

There are many important problems in requirements engineering and software engineering whose attendant research are evaluated empirically. Examples of areas that are evaluated empirically include, but are not limited to, requirement classification, requirement prioritization, fault prone code prediction, and traceability. The aforementioned all require a gold standard for assessing the accuracy of research techniques.

Some research areas or contexts are fortunate to have open source projects or closed source projects that can be mined for gold standards or for which gold standards/answer sets are provided. Other areas are not as fortunate; traceability is a prime example. It is not a given that trace links exist in closed source projects let alone in open source projects. As a result, traceability researchers must develop answer sets for trace data sets. This leads to a number of problems: there is possible bias in that the researchers developed the answer set for their own technique's evaluation, the answer set may be wrong, the data set and answer set may only be used by the creating research group and hence not be vetted externally, to mention a few. To encourage researchers to share data sets and to also assist in evaluating the gold standards for these datasets, we present a prototype tool to support community vetting of answer sets. Community vetting permits analysis of the answer set for inter-rater agreement – with only "community accepted" entries being used as part of the official answer set, thus reducing threats to validity/bias[1].

The requirements for an answer set voting tool have been widely discussed in the traceability community as far back as 2006 and as recently as April 2017 by a large group of researchers gathered for the Grand Challenges of Traceability: The Next Ten Years. Based on these requirements, a prototype tool has been developed called Multi-user Input in Determining Answer Sets (MIDAS). While it was initially developed specifically for traceability research, it has been tailored to also support classification datasets. With tailoring for the format of the data sets and answer sets, it can be applied to any research area that needs to build answer sets. Note that our tool is an early prototype and is open for community collaboration via GitHub.

The paper is organized as follows: Section 2 presents some background information, Section 3 describes the tool's features, Section 4 describes related work, and Section 5 concludes the paper and presents future work.

II. BACKGROUND

Before detailing the features of our tool, we provide a short introduction to tracing and answer sets.

In tracing, one must form correct trace links, collectively called trace matrices, between pairs of project artifacts. Ideally, a correct trace matrix includes only pairs of artifacts that are related. Project artifacts can include, for example, requirements, source code, and test cases. Trace links can represent different artifact element relationships depending on the context. For example, an implementation link could be found between a requirement element and a code element - code file X implements requirement Y; a trace link could represent requirement satisfaction - requirement X is partly addressed or satisfied by design document Y.

[1] Group vote on gold standard for s/w engineering research appears as recently as MSR 2018: https://github.com/collab-uniba/EmotionDatasetMSR18

In the tracing field, researchers work to develop methods that, given a set of artifacts, will automatically generate trace links. The method accuracy is evaluated by comparing the results to the gold standard, hereon termed the answer set. The answer set consists of trace links that are deemed to be the complete, correct list of such links for a given data set. Most data sets do not have answer sets, or do not have answer sets that have been vetted or used by multiple experts or research groups. When data sets/answer sets are used by different research groups, there is currently no central mechanism for making/disseminating changes to the elements of the answer sets. MIDAS seeks to address these problems.

Classification tasks in requirements engineering also require researchers to label a given item as part of a class or not. MIDAS allows users to classify a trace link or classification item in an answer set as one of three types: black, white, or grey. A black link or item refers to a true trace link while a white link is a false trace link. A grey link is an ambiguous link: it may or may not be a true trace link or may or may not belong to the class. Given an answer set, identifying links and items that are "grey" via consensus will provide more insight on what links or items may be troublesome for humans to vet as well as identify answer set elements that can be included or excluded from an answer set for method evaluation. We believe that this is a key feature of our tool, which is described next.

III. MIDAS REQUIREMENTS/FEATURES

The University of Kentucky research group took the lead on eliciting requirements for a community answer set voting tool. The draft requirements document was developed and then shared with researchers at other universities who are also members of the Center of Excellence for Software and Systems Traceability (COEST), http://www.coest.org/. The resulting document can be found here: http://selab.netlab.uky.edu/homepage/publications/link-voting-requirements-version3.pdf.

Next, we identified the most important features for a minimum viable product: the ability to vote/comment on items in answer sets, and all features required to accomplish this. We list the features of MIDAS below.

Feature 1 –Manage User Data
Feature 2 – Upload Data Set
Feature 3 - List Data Set
Feature 3a – Vote for Links
Feature 3b - Add Comments on Links
Feature 4 - Average Votes on Answer Set
Feature 5 (future) - Browse, Comment, Vote on a Data Set
Feature 6 (future) – Clone a Data Set
Feature 7 (future) – Administer a Data Set
Feature 8 (future) – Administer Users

Ruby on Rails and sqlite3 were used to build the application. Sqlite3 databases store user information, data sets, comments, and voting results. The core features have been implemented, but a number of features have not. Within the existing features, we would like to enhance security, enable users to upload multiple data sets and switch between them, and enhance user interaction via an improved GUI.

IV. RELATED WORK

To our knowledge, MIDAS is the first online collaborative tool towards improving traceability answer sets. There is prior work on the topic of grey/ambiguous trace links. Niu et al. [1] studied grey links in the context of requirements change and reuse and found that grey links arose depending on the task. Niu et al. also studied consensus among study participants and found that grey links, or lack of consensus, tend to arise when participants are not involved in requirements change/reuse. Kong et al. [2] studied the impact of various factors such as environment and behavior on the accuracy of human analysts in forming and vetting trace links. In their experiments, grey links were those for which the participants could not easily assign a label of white or black link. Zogaan et al. published a large review on trace data sets [3].

V. CONCLUSION AND FUTURE WORK

We have a Trello board with our product backlog, our project hosted on GitHub, and are doing the following: continuing development of MIDAS for other requirements engineering research areas; encouraging COEST members and others to join in: developing the rest of the features, using the tool to develop and vote on answer sets, sharing their answer sets once voted on and use, and spreading the word to the broader software engineering community in hopes that an area outside traceability and/or requirements engineering will branch our project and use it for their data and answer sets.

Our project can be found in the following GitHub repository: [https://github.com/ladyskynet/MIDAS].

ACKNOWLEDGMENT

We thank NSF for partially funding this work under grant CCF-1511117 and CICI 1642134. We thank Eamonn Magner and Hussamuddin Nasir for their assistance.

REFERENCES

[1] Niu, N., Wang, W., & Gupta, A. (2016, November). Gray links in the use of requirements traceability. In Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (pp. 384-395). ACM

[2] Kong, W. K., Hayes, J. H., Dekhtyar, A., & Dekhtyar, O. (2012, September). Process improvement for traceability: A study of human fallibility. In Requirements Engineering Conference (RE), 2012 20th IEEE International (pp. 31-40). IEEE.

[3] Zogaan, W., Sharma, P., Mirahkorli, M., & Arnaoudova, V. (2017, September). Data sets from Fifteen Years of Automated Requirements Traceability Research: Current State, Characteristics, and Quality. In Requirements Engineering Conference (RE), 2017 IEEE 25th International (pp. 110-121). IEEE.