# BHAVI 2018 Fall Quarter Update*

Shiladitya Dutta

created December 14, 2018, revised December 28, 2018

## Primary Research Project

- Description of progress on your primary research project;
- goals expected at beginning of quarter;
- tasks completed by end of quarter;
- citations for any relevant literature found during the quarter not previously listed in your prior report references for the project;
- STEMM competitions, conferences or journals considered for possible publication of project;
- report title for each submission planned or completed;
- reports already submitted and status – rejected, accepted or published;
- request to continue primary research project versus change to different primary research project.

Most of the development on the primary research project has been in the area of developing the software. The software development has been focused on improving the usability of CoVaSEA. This includes optimizing the code and making it easier for a user to download CoVaSEA on their system. Most of the runtime optimization has been in the lexical to semantic converter. This includes creating an easier to use user interface, allowing the user to change the location of meta-data storage, and creating an instruction set. It also includes the development of a SPARQL Builder form to allow users to build a search query if they don't know SPARQL. The starting point of the software in the beginning of the Fall Quarter is detailed in the end of summer technical report: BHA201809ShiladityaDutta0930 [1].

The goals in the beginning of the quarter were to make CoVaSEA more reliable and improve the usability of CoVaSEA. This includes improving the web crawler, allowing for more search options for the user, and reconfiguring the code so that it could more easily be transferred onto another system. I believe that I have achieved many of these goals for CoVaSEA. The web crawler was optimized so that it could access a larger amount of references. The inclusion of the SPARQL Builder form allows for users to build SPARQL queries without SPARQL syntax. Finally, the code has been improved such that it is smoother for a user to download CoVaSEA.

Relevant Citations that have not been previously listed:

- Paper describing the Scribe API of the NPDS system [2]
- Paper describing the co-referencing method used by the StanfordCoreNLP toolkit [3]
- Paper describing the dependency parsing method that is used by the StanfordCoreNLP toolkit. [4]
- Example of paper that utilizes semantic triple extraction for summarization of texts that are relevant to the biomedical field. [5]

*Document created December 14, 2018, revised December 28, 2018; typeset December 30, 2018.

- Another example of a paper describing a method to extract RDF triples from a paper and discern the key triples via a trained machine learning classifier.[6]
- Paper describing the part of speech tagger that is used by NLTK. [7]
- Paper describing the DublinCore ontology which is used in order to describe citation meta-data in the RDF files.[8]

One competition that I wish to submit to is the Golden Gate Science Fair (GGSF). This science fair is the science fair for the San Francisco area and provides a pathway to participation at the Intel International Science and Engineering Fair (ISEF). One note that should be made is that GGSF doesn't preside over my region. Rather students in Pleasanton would normally participate in the Alameda County Science and Engineering Fair (ACSEF). However, ACSEF has went on a 1-year hiatus, thus all science projects are being routed to GGSF.

Titles:

- AMIA: A Focused Web Crawler and SPARQL Query Search Agent to Expand and Navigate NPDS Semantic Metadata Records
- Brain Informatics: SPARQL-Based Search Engine and Agent for Finding Brain Literature and Converting References to NPDS Metadata Records
- IEEE BIBM: SPARQL-Based Search Engine and Agent for Finding Biomedical Literature and Converting References to NPDS Metadata Records

The first submission was planned to be to the AMIA conference. However, the submission was decided not to be sent in at the last minute. The second submission was to the Brain Informatics 2018 Conference as an abstract/ oral slide presentation. The submission was accepted and presented by Dr. Taswell at the Brain Informatics 2018 Conference on December 9, 2018.[9] The third submission was a 3-page poster paper to the International Conference on Bioinformatics and Biomedicine (BIBM). The submission was rejected from the conference.

I would like to continue the development of CoVaSEA into the Winter Quarter. I have a few objectives that I will pursue during Winter Quarter. The highest priority objective is to create an efficient CLI and GUI for CoVaSEA. There should be two version of the GUI: the one for expert semantic web engineers who are proficient with SPARQL and the one for users who aren't familiar with the technical nuances of SPARQL. The approach for this is to review the search GUIs of popular literature databases and take note of features which are beneficial. Also purchase various textbooks on good GUI design. A GUI is absolutely necessary for the completion of the software so that users can more easily use CoVaSEA. The second objective is the expansion of the web crawler. This mainly consists of expanding the capabilities of the search system of the web crawler such that it can handle more search parameters. The third objective is the adding of more supported sentence formats to the lexical-to-semantic converter. This would allow the converter to parse more advanced sentences accurately.

Number of Records entered into the Nexus Diristry: **16**. All of the records were entered into the Davinci Diristry.

## Secondary Research Project

- Description of progress on your secondary research project;
- goals expected at beginning of quarter;
- tasks completed by end of quarter;
- citations for any relevant literature found during the quarter not previously listed in your prior report references for the project;
- STEMM competitions, conferences or journals considered for possible publication of project;

- report title for each submission planned or completed;
- reports already submitted and status – rejected, accepted or published;
- request to continue secondary research project versus change to different secondary research project.

No Secondary Research Project. Originally was to be second author on Arnav Bansal's project. However, Arnav Bansal left the BHAVI program thus I have not had a secondary research project. For a secondary research project, if possible, I wish to work on a research project in the semantic web field. The most appropriate project is the FAIR metric project since my RDF triple extraction has significance in that project area.

## Education Project

- CSSE = Computational Sciences & Software Engineering
- CSSE course work completed at school None
- CSSE course work completed independently online Linked Data Engineering by Open HPI: https://open.hpi.de/courses/semanticweb2016?locale=en
- CSSE textbooks purchased and read The Algorithm Design Manual by: Steven S. Skiena [10]

## Discussion

Any general comments, suggestions, requests regarding the BHAVI program?

One comment that I have is that it would be beneficial to have a list of mantras of Brain Health Alliance made available to the BHAVI students. There are many mantras in BHAVI including "RTFM" and "Real software that really works". These are general guiding principals for participants in the program, thus is would be beneficial to have a list of mantras for students, especially new students, to read and understand.

## References

[1] C. T. Shiladitya Dutta Adam Craig, "A sparql query search agent for finding web resources and creating npds semantic metadata records describing them," Brain Health Alliance Virtual Institute, Technical Report, 2018.

[2] A. Craig, S. H. Bae, T. Veeramacheneni, *et al.*, "Web service apis for scribe registrars, nexus diristries, portal registries and doors directories in the npd system.," in *SWAT4LS*, 2016.

[3] K. Clark and C. D. Manning, "Entity-centric coreference resolution with model stacking," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, 2015, pp. 1405–1415.

[4] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2005, pp. 363–370.

[5] Y. Shang, Y. Li, H. Lin, *et al.*, "Enhancing biomedical text summarization using semantic relation extraction," *PLoS one*, vol. 6, no. 8, e23862, 2011.

[6] J. Leskovec, M. Grobelnik, and N. Milic-Frayling, "Learning sub-structures of document semantic graphs for document summarization," 2004.

[7] K. Toutanova and C. D. Manning, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," in *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, Association for Computational Linguistics, 2000, pp. 63–70.

[8] S. Weibel, "The dublin core: A simple content description model for electronic resources," *Bulletin of the American Society for Information Science and Technology*, vol. 24, no. 1, pp. 9–11, 1997.

[9] S. Dutta and C. Taswell, “Sparql-based search engine and agent for finding brain literature and converting references to npds metadata records,” in *11th International Conference on Brain Informatics*, 2018.

[10] S. S. Skiena, *The algorithm design manual: Text*. Springer Science & Business Media, 1998, vol. 1.