**Title of Work:**

KESeDa: Knowledge Extraction from Heterogeneous Semi-Structured Data Sources

**Conference:**

SEMANTiCS – 12th International Conference on System Semantics

**Rationale to ensure venue quality:**

The Conference featured keynote speakers from industry giants like Google and EA, among others and from several noteworthy research groups and academic institutions. Additionally, the conference has 25 sponsors and partners ranging across many areas of the technological industry and academia. Finally, the conference is partnered with, and releases papers under the umbrella of, the ACM, which is an established industry standard in computing research.

**Problem Statement:**

A great deal of information is present on the Web as heterogeneous, semi-structured data. However, this data is only weakly interlinked and does not include any semantic classification, which limits the usefulness of any extracted data. While some work has been done previously on extracting knowledge from structured and unstructured data, there has been little work done with semi-structured data. To that end, an examination of various extraction solutions and a tool capable of extracting knowledge from heterogenous, semi-structured data sources could drastically increase the amount of knowledge that can be retrieved from online data sources.

**Paper Synopsis:**

Seidel et al. begins by exploring the current landscape of the use of semi-structured data and motivates the need for an effective tool to capture that data. He explains that various semi-structured data formats such as HTML, XML, & JSON are being generated in increasingly vast amounts on the Web. He motivates the need for an effective tool by describing how just the unstructured, user-generated textual data is not enough to generate meaningful information. There also needs to be some amount of meta-data or structure to the values and types of the data to allow for programmatic knowledge extraction. He examines some of the existing technology for extracting & storing knowledge in various formats, and enumerates some of the many tools that implement data wrappers.

Data wrappers are a common approach to extracting data from semi-structured data sources, that extract content of a particular information source and translate (or “wrap”) it into a relational form. The biggest challenge to these tools on the web is that the format of data found on the web can vary greatly, and that relevant and irrelevant data, such as markup or inline code, are not separated and are frequently bound together. To that end, Seidel et al. propose KESeDA (Knowledge Extraction from heterogeneous, Semi-structured Data Sources).

The general approach for KESeDa is split into multiple processing steps. First, the data format of the inputted source data is identified, which allows the tool to implement the appropriate algorithm for data extraction (the article focuses on the procedure for JSON data). Next, the source file is prepared for later annotations by wrapping each JSON object value in an additional wrapper that contains an array structure which will later be populated with metadata. Each potential predicate has a relevance value and reference count attached that allows the authors to calculate the number of times the object was referenced correctly later on in the document. Next, values are analyzed by matching them against a set of dictionaries and examined with a focus on data type and format. After this, the keys of the JSON object are analyzed, attempting to match them to known predicates or to annotate the key with a possible relevance-based mapping. Finally, the annotated JSON source is transformed into a JSON-LD representation by selecting a suitable RDF predicate for each property. This resulting file is evaluated for syntactic correctness.

To evaluate the work, a prototype was built in node.js that implements the proposed JSON algorithm and processing. The example data contained 1,099 attributes and objects describing 140 different people, with a goal of extracting data to match as many possible attributes to RDF predicates for every listed person. A series of four predefined templates were used by a value-analysis component that attempted to parse a string as a date, email address, URL, and image file extension to appropriately classify data types. With this initial setup, KESeDa was able to distinguish all 140 people with 518 mapped predicates. Next, the algorithm was evaluated on its ability to associate between domain vocabulary and given JSON source properties. The algorithm was fed information about a person and it’s corresponding JSON object that had all properties correctly assigned to the according predicates. This allows the algorithm to evaluate if the automatically mapped properties extracted from the person match the given JSON object’s properties, allowing the algorithm to adjust the Relevancy value of properties as necessary. With this setup, all 140 persons were detected, but KESeDa was able to map 679 predicates.

**Future Work:**

One area of future work I would like to see explored is almost the opposite of this study. I would like to see an evaluation of if this information extracted from semi-structured data can be mapped back into unstructured natural language text data effectively. Within the scope of the study itself, I’d like to see these evaluations expanded to different unstructured data types and to different domains with more natural language components than a registry of people.