

Generation and Analysis of Semantic Metadata Records Based on the Targeted Retrieval of Open-Web Resources for an Automated Article Search Engine Agent

In this project I describe CoVaSEA (Concept-Validating Search Engine Agent): an automated web crawler/query engine to act as a semantically-based scientific article search engine. Due to the machine-understandable nature of the semantic web, it is well placed to perform a variety of data analysis tasks. However, semantically formatted records are labor intensive to create and maintain due to the substantial and rapidly increasing quantities of scientific literature available on the open web. To remedy this, CoVaSEA was created with the intention of providing an automated method for users to navigate and expand the semantic records of scientific literature. To this end, CoVaSEA integrates multiple features including: (A) Targeted web-crawling for relevant articles from external literature databases to broaden semantic records, (B) Translation of free-form text into RDF triples to derive the graph-based semantic portrayal of lexical data for the NPDS database, and (C) An implementation of SPARQL query based search to allow for retrieval and manipulation of RDF semantic records. With the distinct advantage that the system is automated, CoVaSEA presents the capability to search “externally” to furnish large numbers of semantic literature descriptions on a regular basis and search “internally” to provide a method of retrieving those descriptions, thus laying the groundwork for a variety of future applications for which semantic metadata stores of scientific literature along with a method of analyzing these stores are a necessity.