Analyzing and Expanding NPDS Biomedical Record Databases via a Semantic Search Engine

Shiladitya Dutta and Carl Taswell MD, PhD

2018-October-31

Abstract: In this paper we describe CoVaSEA (Concept-Validating Search Engine Agent): an automated semantic search engine that is interoperable with the Nexus-PORTAL-DOORS System. The Nexus-PORTAL-DOORS System (NPDS) is a data management system that organizes repositories of lexical metadata (in PORTAL servers) and semantic representations (in DOORS servers) of resources. Integrated with NPDS, CoVaSEA serves to apply SPARQL-based search to external biomedical literature databases by converting relevant article resources to a linked data format which can be searched by the query engine. This capability can prove to be valuable in a variety of tasks which benefit from a combination of a web-crawler's capability to search the open web with the versatility of a semantic query engine. CoVaSEA, mirroring its dual nature, has a SPARQL-based search engine facing inwards and a focused web-crawler facing outwards. The search engine acts as an "internal" search, allowing a user to explore semantic records via SPARQL-based search. For a more conducive user experience, the query engine employs a SPARQL builder, thus circumventing the need to know SPARQL syntax. The utility functions by having the user input a series of either filtering, mandatory, or optional conditions in either an independent or nested connection. Thus the user can form a search query via simple fill-in-the-blank statements. Though it cannot replicate the full power of SPARQL syntax, it is a potent resource for users who do not wish to use SPARQL. With the constructed or inputted query, CoVaSEA searches through the linked database with the query engine. Due to the dynamic and heterogeneous nature of the data that CoVaSEA handles, the SPARQL query engine divides up the queries into sub-queries such that only the necessary articles are compiled. This iteration based approach optimizes runtime and memory-load on computers by eliminating the redundancy of loading unneeded triples into a computationally expensive local graph. In contrast to the query engine, the web-crawler provides "external" search functionality to render the capability to search for outside resources. The web crawler retrieves articles along with their basic metadata from several of biomedical literature databases via REST API. In addition to basic metadata, key RDF triples which describe the abstract are constructed in order to record a full semantic description of the data in each article. To prevent redundant re-rendering of articles, CoVaSEA saves the triples which are built by the web-crawler in either a local quadstore or a DOORS directory. Each articles in the record has a corresponding named graph, allowing for the identification of the origin of triples. The graph contains two sections: metadata and the semantic representation. The metadata section stores key information about the article such as author, title, publication date, and database of origin. The semantic representation section stores a set of triples which outline the key ideas of the article's abstract. A secondary benefit of this storage schema is that CoVaSEA is not limited to converting and searching external lexical databases, but can also perform as a semantic search engine for the DOORS directory. In a broad sense, CoVaSEA consists of 3 components: the federated query engine searching, the web-crawler expanding, and the quadstore storing. The query engine builds and applies SPARQL queries for users to explore semantic triples. The webcrawler adds semantically formatted records to NPDS from the open-web. The quadstore/DOORS directory logs the formatted records for future searches. Acting in concert, these can provide a composite open-web SPARQL search utility. For researchers, this opens the possibility to search external biomedical databases via a semantic web approach and is a solution for situations in which SPARQL queries need to be applied to non-linked data resources.