Mapping Existing Data Sources into VIVO

Craig A. Knoblock, Ph.d., University of Southern California, Marina del Rey, CA, USA
Pedro Szekely, Ph.d., University of Southern California, Marina del Rey, CA, USA
Maria Muslea, M.S., University of Southern California, Marina del Rey, CA, USA
Shubham Gupta, M.S., University of Southern California, Marina del Rey, CA, USA

Mapping existing legacy data to the VIVO ontology and generating the corresponding RDF data that can then be loaded into VIVO can be very challenging. In general, the data users want to load into VIVO is stored in various formats using a variety of representations. In order to map the data, users must first understand how the data should be mapped to the VIVO ontology, they must perform various transformations on the data in order to get it into the needed format, and finally they need to translate the data into the RDF schema that VIVO expects. There are tools within VIVO for performing this mapping, but these tools require significant expertise to use.

We present Karma, an open-source information integration tool that allows a user to quickly map legacy data sources into RDF that can then be loaded into VIVO. Karma provides the ability to:

- Extract data from a variety of sources, including databases, spreadsheets, comma-delimited text files, and Web APIs.
- Semi-automatically map the data to the VIVO ontology by first producing an initial mapping and then allowing users to refine the mapping based on their own knowledge of the data.
- Transform the format of the data by example, which allows a user to quickly change the format, remove extraneous information, split fields, and so on.
- Integrate data across sources, which make it possible to combine data from different sources, decode foreign keys with the actual values, and merge data from different sources.
- Produce RDF data with respect to the VIVO ontology, which includes creating appropriate URIs, constructing the RDF, and generating inverse property links.

The capabilities provided by Karma allow users to quickly load a dataset, map each data field to the appropriate terms in the VIVO ontology, normalize and integrate data as required, and then generate the appropriate RDF. This RDF data can then be loaded directly into the VIVO application and immediately made available.

In this talk, we will describe the capabilities of Karma and then perform a live demonstration of the tool. We will show how one can rapidly go from an existing dataset with USC faculty members to the data loaded into VIVO. There is a demonstration version of the software and a video of the demonstration available from http://isi.edu/integration/karma. In order to have sufficient time to present the Karma system and demonstrate how it works, we would like to request a long presentation at the conference.