CEDAR's Predictive Data Entry: Easier and Faster Creation of High-quality Metadata

Marcos Martinez-Romero, Martin J. O’ Connor, Ravi D. Shankar, Maryam Panahiazar, Debra Willrett, Attila L. Egyedi, Olivier Gevaert, John Graybeal, and Mark A. Musen
Stanford University, Stanford, CA 94305, U.S.A.

High-quality metadata are crucial to finding experimental datasets, to understand how experiments were performed, and to reuse data to conduct new analyses. Despite the growing set of standards and guidelines, quality of metadata in public repositories continues to be poor. A key difficulty is that the typical metadata acquisition process is time-consuming and error prone, with weak or nonexistent support for linking supplied metadata with semantic resources.

CEDAR is creating a computational ecosystem for development, evaluation, use, and refinement of biomedical metadata. Our approach centers on the use of metadata templates, which define data elements needed to describe biomedical experiments. The CEDAR Workbench is a set of Web-based tools for the acquisition, storage, search, and reuse of metadata templates. One of those tools is the Metadata Editor, which allows users to fill in metadata templates with metadata.

We have enhanced the Metadata Editor with value recommendation capabilities. CEDAR’s Value Recommender identifies common patterns in the metadata repository, and generates real-time suggestions for filling out metadata templates. The suggestions provided are context-sensitive, meaning that the values predicted for a particular field are generated and ranked based on previously entered values.

The Value Recommender supports both free-text values and controlled terms. In this particular example, the system suggests terms from the Human Disease Ontology (DOID).

BioSample is a database from the National Center for Biotechnology Information (NCBI) that captures descriptive information about the biological samples investigated in projects. The ‘BioSample Human’ template contains the fields used in BioSample for human samples.

CEDAR is developing novel methods and tools to annotate biomedical datasets with metadata. We have enhanced the Metadata Editor with value recommendation capabilities. CEDAR’s Value Recommender identifies common patterns in the metadata repository, and generates real-time suggestions for filling out metadata templates. The suggestions provided are context-sensitive, meaning that the values predicted for a particular field are generated and ranked based on previously entered values.

The Value Recommender supports both free-text values and controlled terms. In this particular example, the system suggests terms from the Human Disease Ontology (DOID).

BioSample is a database from the National Center for Biotechnology Information (NCBI) that captures descriptive information about the biological samples investigated in projects. The ‘BioSample Human’ template contains the fields used in BioSample for human samples.

CEDAR is developing novel methods and tools to annotate biomedical datasets with metadata. We have enhanced the Metadata Editor with value recommendation capabilities. CEDAR’s Value Recommender identifies common patterns in the metadata repository, and generates real-time suggestions for filling out metadata templates. The suggestions provided are context-sensitive, meaning that the values predicted for a particular field are generated and ranked based on previously entered values.

The Value Recommender supports both free-text values and controlled terms. In this particular example, the system suggests terms from the Human Disease Ontology (DOID).

BioSample is a database from the National Center for Biotechnology Information (NCBI) that captures descriptive information about the biological samples investigated in projects. The ‘BioSample Human’ template contains the fields used in BioSample for human samples.

CEDAR is developing novel methods and tools to annotate biomedical datasets with metadata. We have enhanced the Metadata Editor with value recommendation capabilities. CEDAR’s Value Recommender identifies common patterns in the metadata repository, and generates real-time suggestions for filling out metadata templates. The suggestions provided are context-sensitive, meaning that the values predicted for a particular field are generated and ranked based on previously entered values.

The Value Recommender supports both free-text values and controlled terms. In this particular example, the system suggests terms from the Human Disease Ontology (DOID).

BioSample is a database from the National Center for Biotechnology Information (NCBI) that captures descriptive information about the biological samples investigated in projects. The ‘BioSample Human’ template contains the fields used in BioSample for human samples.