Center for Expanded Data Annotation and Retrieval (CEDAR)

Stanford University

PI: Mark A. Musen

Grant Number: 1-U54AI117925-01

The Big Data revolution requires that biomedical scientists be able to locate, analyze, and integrate the large datasets that now pervade biomedicine. Such work is possible only when experimental datasets are made available online and when they are annotated with metadata that explain how the data are organized, what the data represent, and how the data were collected. The Center for Expanded Data Annotation and Retrieval (CEDAR) will take advantage of the recent growth in community-driven metadata standards to develop innovative computational methods to ease the authoring and use of metadata annotations. Our specific aims focus on working with communities of investigators to standardize descriptions of the data generated through biomedical studies; creating a computational collective for development, evaluation, use, and refinement of metadata templates for describing laboratory studies; developing a comprehensive and open repository of metadata that will inform the learning algorithms that will drive much of our Center's technology; training the biomedical community in the use of metadata and in CEDAR's resources; and evaluating our work in the context of ImmPort, an NIAID-supported multi-assay data repository that will offer end-to-end opportunities to demonstrate and validate our ideas. We anticipate a growing community of users, starting with the Human Immunology Project Consortium, then the BD2K Center Consortium, then the Stanford Digital Repository, growing until we have developed a wide user base leading to measurable changes in the quality of the metadata used to annotate online datasets. The Overall description of our project provides a synopsis of CEDAR's activities and overall specific aims. PUBLIC HEALTH RELEVANCE: The ability to locate, analyze, and integrate Big Data depends on the metadata that describe data sets and the experiments that have been performed. This project will facilitate annotation of data with high quality metadata. The results of our work will lead to better data and, thus, to better science. Ultimately, such results will lead to better health.