

When the World Beats a Path to Your Door: Collaboration in the Era of Big Data

Mark Musen, Stanford University

For many years, my laboratory has led major projects that provide computational infrastructure for work in data science. Protégé is a software system for editing ontologies that, over the past 20 years, has acquired more than 300,000 registered users, of whom nearly 200 have contributed more than 130 plug-ins for use by the community. The National Center for Biomedical Ontology (NCBO) provides BioPortal, a repository of most of the world's publicly available biomedical controlled terminologies and ontologies, as well as services that use those ontologies for a variety of tasks. More than 45,000 users access the NCBO ontology repository each month.

The Center for Expanded Data Annotation and Retrieval (CEDAR), supported by the BD2K initiative, is developing technology to assist scientists in the creation of comprehensive metadata to describe experimental datasets that will be stored in online repositories. CEDAR has already attracted the interest of many initiatives supported by the NIH, including HIPC, LINCS, HeartBD2K, and CaDSR. If CEDAR proves successful, it may likewise see many users and collaborations.

Experience with the Protégé and NCBO projects has taught us the challenges of interacting with large user communities and of supporting collaborations with a wide range of users. It is essential to have open communication channels, to develop software with collaboration in mind, and to manage collaborations actively to achieve maximum benefit. Effective collaboration in academic projects remains difficult, however, due to resource limitations—in particular, the availability of technically knowledgeable personnel who have the time to manage such collaborations and to help map out new directions and projects. I will discuss how the lessons of our experience with Protégé and NCBO are helping us to develop and enrich our collaborations on the CEDAR project, and how appropriate communication with users can help to scale these interactions.