

The National Center for Biomedical Ontology

Enabling the retrieval, integration, and analysis of “big data” in biomedicine

In June 2012, the Data and Informatics Working Group reporting to the Advisory Committee to the Director recommended that the NIH “promote data sharing through central and federated catalogs.” The foundation of such catalogs for data sharing rests in the ontologies and controlled terminologies that provide the standard definitions of the elements of biomedical data sets and of the contents of electronic health records. Without standard mechanisms to define the meaning of data, it is impossible to retrieve, integrate, and analyze the vast data sets that are now the norm in biomedicine and that form the basis of electronic health records.

The National Center for Biomedical Ontology (NCBO) has become the leading scientific organization for bringing this sort of semantic technology to biomedicine. The Center has its main activity at Stanford University (PI: M.A. Musen), with collaborators at the Mayo Clinic (co-PI: C.G. Chute), the University at Buffalo (co-Pi: B. Smith), and the University of Victoria (co-PI: M.-A. Storey).

Technology: The NCBO has created the world’s definitive online ontology repository, known as **BioPortal**, which stores some **350 biomedical ontologies and controlled terminologies** in a uniform, consistent manner. NCBO makes these disparate ontologies accessible in a standardized manner through a Web browser interface that attracts more than **65,000 visitors per month**—thousands of whom visit the site every working day.

The Center provides programmatic access to its ontology repository through Web services that are invoked by application programs in the laboratory and in the clinic. The programmatic interface to BioPortal receives a median of **1 million calls per month**. The NCBO offers additional Web services that put the BioPortal repository to use for recurring biomedical tasks that make use of our ontology content. The **NCBO Annotator**, for instance, is a high-throughput Web service that takes as input some text (such as a PubMed abstract or the textual metadata of a data set in an online repository such as GEO) and returns as output the terms from preselected ontologies that relate to that text. Thus, the ArrayExpress resource at the European Bioinformatics Institute makes extensive use of the NCBO Annotator to link the metadata associated with microarray studies to standard ontologies in BioPortal, enabling high-precision indexing, searching, and integration of gene-expression data for many thousands of users. At Stanford, investigators have applied the NCBO Annotator to electronic health records, analyzing the text of thousands of progress notes and discharge summaries *per minute* to identify adverse drug events and off-label drug usage. In a test to search for signals involving 9 drugs that caused known adverse events, the NCBO Annotator enabled identification of 7 of these—on average 1.9 years before the FDA called for the drugs to be withdrawn from the market.

User Community: The NCBO serves a vast and extremely vibrant user community that has developed more than **55 third-party biomedical applications** that rely on NCBO Web services for their infrastructure; our users created 15 of these applications in the last year alone. For example, the WHO relies on NCBO technology in its tools to develop the new revision of the International Classification of Diseases, aligning the terms in ICD-11 with those in SNOMED CT. The i2b2 data warehouse uses NCBO Web services to populate its “ontology hive.” Investigators at UCSD collaborated with Microsoft Research to create a plug-in for Microsoft Word that calls NCBO Web services as authors are working on a document to mark-up the text with appropriate ontology terms (for example, linking the name of an enzyme to a specific entry in the Protein Data Bank). caNanoLab is a resource that makes extensive use of NCBO technology to allow the NCI’s nanotechnology community to share precise information about the physical, chemical, and biological characteristics of therapeutic nanoparticles.

Training and Dissemination: The NCBO has a very active training and dissemination effort. We have attracted **16 visiting scholars**, and we have trained **18 graduate students** and **16 post-doctoral fellows** in the use of semantic technology in biomedicine. Our trainees have gone on to promote the use of NCBO technology in academia and in industry. We have organized **59 educational workshops and conferences** and have presented **17 tutorials** at national and international meetings. We regularly host webinars on the use of our technology. There have been **25 webinars** the past 2 years; some **3,500 people** have accessed our online recordings of these talks.