OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

To Infinity, and Beyond: Uniting the Galaxy of Biological Data

PETER A. COVITZ

THE DIVERSITY AND COMPLEXITY of biological research data present a number of important challenges to those involved in data management and integration. Sheer volume of data is no longer the most important problem given the continuing drop in cost of computational processing power and storage, and advances in database and indexing technology. Rather, the lack of interoperability between systems and of comparability between data sources present the greatest barriers to knowledge and value extraction.

SYSTEM INTEROPERABILITY

System interoperability can be thought of as the ability to access a data or analytical resource programmatically without requiring substantial assistance from staff at the hosting site. In practice, the optimal way to achieve interoperability is for the hosting site to provide stable, well documented application programming interfaces (APIs) that conform to both software and bioinformatics standards. Such APIs are best developed by abstracting the primary data storage and retrieval schema from the published query and presentation interface. The API should be platform neutral and accessible from any common programming environment.

The Distributed Annotation System (DAS) developed by L. Stein and colleagues, used to serve the human genome sequence and annotations, is a major milestone on the road to generalized system interoperability. More recently a number of major bioinformatics centers are experimenting with or adopting Web Services standards for their public APIs. Web Services standards such as Simple Object Access Protocol (SOAP) and Web Services Description Layer (WSDL) offer a mechanism to achieve system interoperability. Sites that provide a SOAP interface to their data and resources and publish the WSDL description of those resources go a long way towards minimizing the effort needed by a remote programmer to integrate the hosted data directly into an application. In practice additional documentation of the API and background information on the data and resources is still necessary; but the need for special systems or developer support is minimal.

European Bioinformatics Institute systems now deliver EMBL records through a SOAP interface (www.ebi.ac.uk/xembl). The DNA Databank of Japan offers both data and analytical tool services through SOAP (xml.nig.ac.jp/soapp.html). The National Cancer Institute Center for Bioinformatics (NCICB) provides access to a wide variety of integrated genomic, drug and clinical trial data through its caCORE architecture (ncicb.nci.nih.gov/core). As more open source tools become available from these and other initiatives, members of bioinformatics community will hopefully be more willing and able to implement SOAP interfaces to their local databases. A major goal of the NCICB is to provide such tools and software to any site that wishes to set up its own public SOAP server.

A missing link in the forging of the system interoperability chain is a universal namespace and identifier system that allows unambiguous resolution of any biomedical entity. Without such a mechanism, we

Bioinformatics Core Infrastructure, National Cancer Institute Center for Bioinformatics, National Institutes of Health, U.S. Department of Health and Human Services, Bethesda, Maryland.

will be forever plagued with a ever-expanding list of database cross-reference IDs and links that provide only scattershot connectivity between sources. The "Life Science Identifier" or LSID project is attempting to build such a standard upon a Web Services technology foundation (www.i3c.org/workgroups/technical_architecture/resources/lsid/docs/index.htm).

With such a capability in place, one can begin to envision programmatically accessible registries of data and analytic services that automatically direct information requests from clients to appropriate hosts.

SEMANTIC CONSISTENCY

While system interoperability is necessary, it is not sufficient in and of itself to enable straightforward aggregation and comparison of data from disparate sources. In order to make widespread scientific analysis of disparate data possible, there must also be semantic and representational standards connected to the data. Indeed, the LSID effort could very well fail in part due to the lack of such consistency across biological data stores.

Vocabulary and nomenclature control is notoriously difficult to achieve in the life sciences. This is in part due to the incentives investigators face when they report their discoveries: a finding is often accompanied by a novel way to describe that finding, an approach that burnishes the perception that the reported work is unique and therefore exceptional. In some cases a finding is unique and exceptional and indeed warrants novel nomenclature and establishment of new semantic concept.

Despite these occupational hazards, semantic concept and vocabulary control is possible through ongoing, systematic development of ontologies, thesauri, metadata, and classification standards. A number of efforts are chipping away at the problem, notably the Unified Medical Language System (UMLS) of the National Library of Medicine (www.nlm.nih.gov/research/umls); the Gene Ontology (www.geneontology.org); the NCI Thesaurus (nciterms.nci.nih.gov); the NCI Cancer Data Standards Repository (caDSR, ncicb.nci.nih.gov/core/caDSR); the Veteran's Administration National Drug File (www.va.gov/vdl/Clinical.asp?appID=89); and the Logical Observation Identifiers Names and Codes (LOINC, www.loinc.org). These and related projects are bringing some order and consolidation to the universe of terminology in the life sciences.

SEMANTIC WEB AND BEYOND

Annotation and indexing of data with controlled terminologies is useful, and it brings us closer to the larger goal of semantic consistency across different sources. The Semantic Web project has been brewing a number of modeling languages that will ideally enable sites to more fully describe and advertise in a structured format the semantic underpinnings of the data they serve. The Web Ontology Language (OWL) is the most active standard under development (www.w3.org/TR/owl-ref). The optimal relationship between OWL and Web Services implementations is not clear for the present. The coming years should see some interesting attempts to implement these standards and improve our ability to aggregate and compare data. The success stories that emerge from these efforts will guide our hands as we build the bioinformatics star ships and space stations that propel us beyond the charted universe of biological knowledge, toward infinity.

Address reprint requests to:
Dr. Peter A. Covitz
Bioinformatics Core Infrastructure
National Cancer Institute Center for Bioinformatics
6116 Executive Boulevard
Suite 403
National Institutes of Health
U.S. Department of Health and Human Services
Rockville, MD 20852

E-mail: covitzp@mail.nih.gov