UCRL-ID-146327

Challenges in Microbial Database Interoperability

Interagency Microbe Project Working Group

Terence Critchlow

November 21, 2001

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Challenges in Microbial Database Interoperability
Interagency Microbe Project Working Group

Terence Critchlow

Challenges

Currently, data of interest to microbial researchers is spread across hundreds of web-accessible data sources, each with a unique interface and data format. Researchers interact with a few of these sites when they analyze their data, but are not able to utilize the majority of them on a regular basis. There are two significant challenges that must be overcome to integrate this environment and allow researchers to efficiently perform data analysis across the entire set of relevant data, or at least a significant portion of it. The first is to provide consistent access to the large numbers of distributed, heterogeneous data sets that are currently distributed over the web. The second is to define the semantics of the data provided by the individual sites in such a way that semantic conflicts can be identified and, ideally, resolved.

The first step in establishing any integrated environment, from a data warehouse to a multi-database system, is provide consistent access to all of the relevant sources. While the type of access required will vary based on the integration strategy chosen – for example federated systems use query-based access while warehouses may prefer access to the underlying database - the essence of this challenge remains the same. Thus, without sacrificing generality, the remainder of this discussion focuses on query-based access. Each data source independently determines the queries that it supports, how it will answer them, and the interface that it will use to make them. Even when the same query capability is provided by different sources the details of the interface are usually different. For example, while many sequence data sources support blast searches, they differ in the parameter names, available options, script locations, etc. These differences are not restricted solely to input parameters; the query results returned by different sources also vary dramatically, with some sources returning XML, others preformatted text, and still others a variety of formats. This set of disparate interfaces makes developing an integrated environment extremely challenging because a specialized *wrapper* needs to be created for each data source.

Once consistent data access has been provided, the next challenge is to provide a semantically and syntactically consistent environment for the scientists to use. This would allow them to smoothly transfer data between different query interfaces and applications. Unfortunately, this is an even more daunting task than providing data access because resolving semantic differences between data sources first requires understanding the semantics being used by them. Currently, a source's semantic description of its data is usually buried in its documentation, if it is provided at all. As a result, scientists have become adept at simply looking at the data being provided and divining a first-order approximation of the semantics used by the source. Often, this approximation is sufficient for the types of queries that are being asked. However, when precise semantics are needed, a tedious and time-consuming search must be undertaken. Fortunately, some communities are becoming aware of this problem and are developing ontologies that overcome it by precisely defining the semantics of commonly used terms. While this simplifies data integration for those sources that adhere to a specific ontology, the definition of a single ontology for the entire domain of genomics remains a (probably unachievable) dream. Resolving syntactic differences is relatively straight-forward once the semantic ones have been resolved.

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48.

Suggested Approach

One approach to dealing with these challenges is to explicitly encode in meta-data the information that is currently implicit in both the interfaces and the data itself. Specifically, meta-data could be used to define both functional information, such as how to interact with an interface or how to parse the results of a specific query, and semantic information, such as what the parameters and results actually mean. There are several benefits to a meta-data based approach to data integration, if it is done properly, including

1. Sources maintain their autonomy since they do not have to conform to a specific schema, data format, interface, or ontology.
2. Meta-data describing a source could be created by someone other than the data source provider.
3. The same meta-data format could be used to describe all sites (if this is not the case, the challenges encountered integrating the sources in the first place would also be faced integrating the meta-data).
4. Automated tools could use the meta-data to determine how to access a site, and the types of information the site contains.
5. A source could reference existing ontologies from different sources when defining the semantics of the data it contains.
6. Translations between common ontological definitions for related concepts could be defined consistently.

The Role of Federal Agencies

Given the approach suggested above, there is a clear role for the federal agencies interested in addressing the microbial data integration problem. These agencies could dramatically improve the current situation if they directly or indirectly:

- Selected a meta-data standard for interface definition and ontology definition.
- Required / encouraged the source and tool providers they fund to provide meta-data definitions of their interfaces and semantics.
- Developed tools that used these meta-data descriptions.
- Supported a repository when interface descriptions and ontologies could be placed and accessed by the community at large

Conclusions

In order to provide an effective and efficient environment for microbial researchers to perform data analysis in, the data access and semantic integration challenges need to be successfully addressed. One promising approach to achieving this goal is to use meta-data to describe, at both a functional level and a semantic level, the interfaces to the data sources and the associated results. If this approach were supported by the appropriate federal agencies, it could foster the development of a variety of tools addressing the data integration problems currently being faced.

University of California
Lawrence Livermore National Laboratory
Technical Information Department
Livermore, CA 94551