

NBDL: A CIS Framework for NSDL

Joe Futrelle
NCSA, University of Illinois
Urbana-Champaign, Illinois
Futrelle@ncsa.uiuc.edu

Su-Shing Chen
University of Missouri
Columbia, Missouri
ChenS@missouri.edu

Chen-Chuan Kevin Chang
University of Illinois
Urbana-Champaign, Illinois
Kcchang@cs.uiuc.edu

ABSTRACT

In this paper, we describe the NBDL (National Biology Digital Library) project, one of the six CIS (Core Integration System) projects of the NSF NSDL (National SMETE Digital Library) Program.

Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries – standards, user issues, dissemination.

General Terms

Algorithms, Design, Standardization.

Keywords

Digital library, SMET education, Federated search.

1. INTRODUCTION

This NSDL project, NBDL, consists of participating institutions: University of Missouri-Columbia (MU), NCSA, University of Illinois-Urbana/Champaign (UIUC), and Missouri Botanical Garden (MOBOT) [6]. The project focuses on building an interoperable, reliable, and scalable Core Integration System (CIS) framework for coordinating, integrating, and supporting learning environments and resources provided by NSDL collections and services. We emphasize also integration issues of collections and services by building a testbed with a large biological collection, Tropicos, of MOBOT (Missouri Botanical Garden) and its educational services. This testbed can be extended to other disciplines and environments.

2. BIOLOGICAL INFORMATION

Biology has one of the most complex information-structures (concepts, data types and algorithms) among scientific disciplines. Its richness in organisms, species, cells, genes and their pathways provides many challenging issues for biological sciences, computational sciences, and information technology. The advances in biological science and technology urgently need development of very large biological digital libraries for analyzing and managing biological information: sequences, structures, and functions, arising from DNAs, RNAs, genes and proteins, and taxonomies. At present, biological databases are among the best archived, managed, and preserved. The earliest

phase of bioinformatics is perhaps the naming and classification of organisms invented by the Swedish biologist, Carolus Linnaeus (1707-1778). His "binomial system" provides a taxonomic hierarchy of types, species, families, orders, classes, and phyla (divisions) for biology, still in use today. Unfortunately, the biology community has never developed an integrated data resource of genomic, morphological, and taxonomic information. That is, users can not search and explore all such information objects serendipitously. The NBDL project will address this important scientific issue.

3. THE EMERGE ARCHITECTURE

The technical infrastructure supporting NBDL is the Emerge distributed IR toolkit [3]. Emerge consists of a set of components to enable federation of distributed, heterogeneous data collection by means of query and data translation. Data sources are proxied with a protocol-translation component called Gazelle and integrated with a broker component called Gazebo which translates client queries into the variety of query formats required by the distributed data sources. Thus, data can be retrieved and integrated independently of its location, access protocol, and query syntax. The following depicts the Emerge architecture:

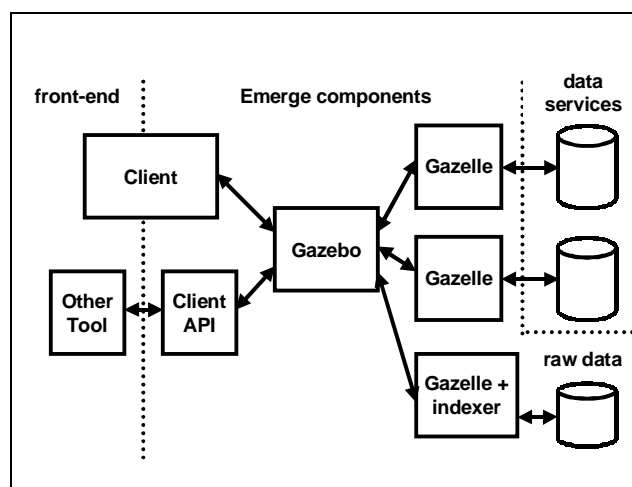


Figure 1. Emerge Architecture

Emerge components access data through a simple interface called the Gazelle Target API, which must be implemented for each data source. In some cases, classes of data sources can be

supported by a single implementation. In the case of NBDL, we have developed an implementation of the Gazelle Target API for TROPICOS database, which allows arbitrary queries to be executed against it [7].

Client queries are translated by Gazebo by a general-purpose query translation engine capable of generating queries in a wide variety of syntaxes. We have developed a configuration for the Gazebo query translator, which generates TROPICOS queries. Semantic equivalences between query syntaxes are expressed in Gazebo using meta-attributes, which are similar to the Alexandria Digital Library's search buckets [4]. For TROPICOS, we have developed a set of meta-attributes which can be translated into TROPICOS as well as ZBIG's Darwin Core query syntaxes, allowing a single client query to be targeted at either TROPICOS or any ZBIG service. This will allow tools to be built that can transparently retrieve biological information regardless of whether it originates in TROPICOS or a ZBIG resource. The NBDLuser interface is given the following figure:

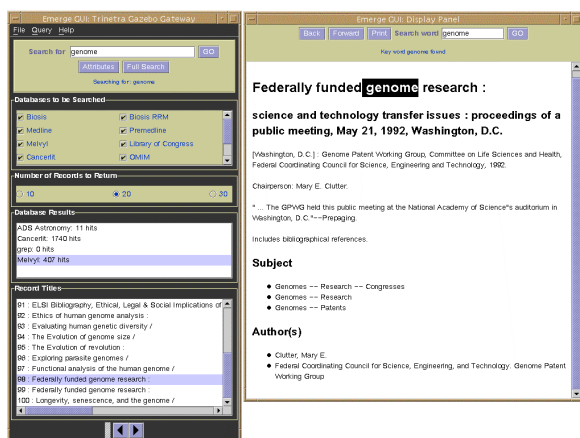


Figure 2. User Interface of NBDL

Semantic Mapping

A key requirement of the CIS architecture is the development of semantics (or ontology) of information, which are then captured in semantic mappings for the intelligent search engines, such as Emerge. Semantic mapping extends the existing full-text, hypertext, and database indexing and mapping schemes to include semantics or meanings of information content. Furthermore, semantic mappings will reduce the complexity of biological taxonomy and classification, and correlate semantically the genotypes and phenotypes at various levels of biological information in NBDL.

An example of semantic mappings is common/scientific name mapping. A significant barrier for the educational use of resources, such as TROPICOS, is the lack of common-name indexing: species are referred to by scientific names, which are unfamiliar to most educational users. To address this problem, we are extending Gazebo's query translation engine using an approximate query translation algorithm [1]. Common names will be translated into scientific names by querying a name directory service such as ITIS during the query translation process [5]. Although the resulting query will not have the exact

semantics of the original query, it will be a close approximation and will allow users to discover relevant information in one integrated step rather than requiring them to use two separate applications. This is an important innovation towards disseminating biological information into the educational community.

Under various current standards (e.g., IEEE LOM and IMS Standards), metadata structures are defined and collected for some specific educational domains. Since metadata structures will always be changing, we are developing "evolving metadata structures", which have adaptive semantic properties [2]. Metadata will evolve and grow through out time and utilization. Semantic mappings are extended to schema mappings of these evolving metadata structures about learning objects and applets in our resources and services, correlating different nomenclatures, syntaxes and semantics. In the LOVE (Learning Object Virtual Exchange) of NBDL [6], we are developing this novel feature so that interoperability, reliability, reusability, and scalability will be maintained even under changes of networked resources and services.

Rich NBDL Collections

The TROPICOS botanical database at the Missouri Botanical Garden (MOBOT) contains 851,000 name records for plants and associated information on bibliography, types, nomenclature, usage, distribution, and morphology. The data base currently contains over 1.5 million specimen records - mostly new collections gathered over the last 20 years with full locality data, coordinates, and elevation information. A literature file of over 80,000 publications used for vouchering distribution and usage and authority files of people, books and journals, and geographical place names. Most recently images from a variety of sources are included to add a visual impact to the wealth of textual data. At this time thousands of images of plant habitats, structure, type specimens and prologues are available for selected taxa. The web site provides a full range of on-demand and interactive html pages providing a scientific overview of the information accumulated around any of the scientific names in the production database TROPICOS.

3. ACKNOWLEDGMENTS

Our thanks to other NBDL project members: Chip Bruce, Ann Bishop, and Bryan Heidorn. Their contributions to biology and education materials are essential to this project.

4. REFERENCES

- [1] C. K. Chang and H. Garcia-Molina. [Approximate Query Translation Across Heterogeneous Information Sources \(Extended Version\)](#). *Proc. of the 26th VLDB Conference*, Sep. 2000.
- [2] S. Chen, *Digital Libraries: The Life Cycle of Information*, Better Earth Publisher, 1998, <http://www.amazon.com>.
- [3] "Emerge home page". <http://emerge.ncsa.uiuc.edu>.
- [4] J. Frew, M. Freeston, L. Hill, G. Janée, M. Larsgaard, Q. Zheng. *Generic Query Metadata for Geospatial Digital Libraries*. IEEE Metadata 99.

[5]"Integrated Taxonomic Information System". <http://www.itis.usda.gov/plantproj/itis/>.

[6] NBDL. <http://cecssrv1.cecs.missouri.edu/NSDLProject/>.

[7]W³TROPICOS. <http://mobot.mobot.org/Pick/Search/pick.html>