Poster B-59

Design Principles of Ontology-based Query and Development of Ontology Wrapper

Authors:
Myung-Guen Chung (*Electronics and Telecommunications Research Institute*)
Myung-Eun Lim (*Electronics and Telecommunications Research Institute*)
Myung-Nam Bae (*Electronics and Telecommunications Research Institute*)
Yong-Ho Lee (*Electronics and Telecommunications Research Institute*)
Soo-Jun Park (*Electronics and Telecommunications Research Institute*)

Short Abstract: A typical mediator-wrapper system required practical linkage information among the database. Unfortunately, each database bestows a different. To solve this inconsistency problem in terminology, researchers have used several methods, one of which is ontology. But most system still remains ontology resource in the system. We propose ontology wrapper.

Long Abstract:
Joint efforts over the last several years for collecting and validating biological data have resulted in the construction of the largest elementary database, such as NCBI and EBI. It is difficult for biologists to use elementary databases directly in their experiments without the database integration. The purpose of integration is to decide on the information and to define the scope and direction of the experiment in terms of viable future work. This is the reason why we must develop the integration system. Several methods have been suggested for integrating the biological database.
A typical mediator-wrapper system required practical linkage information among the database. This is because each integration system calls for a common key for sharing identification (ID) between databases. Without the common key, the integration system will be rendered impossible. Unfortunately, a trait of the collection of bioinformatics sources is that similar data can be contained in several sources, but represented in a variety of ways depending on those sources. Each database bestows a different name for the same biological object, like a gene or a protein. So most existing databases do not have common keys with other databases. To solve this inconsistency problem in terminology, researchers have used several methods, one of which is ontology. But most system still remains ontology resource in the system. We propose ontology wrapper.
Most databases are used for the inner-ontology embedded within the system. Inner-ontology is designed for the specific database, as illustrated by the example of the TAMBIS system that only uses "TAMBIS ontology." There is another reason to use ontology wrapper. Public database of ontology in the biology research are evolved continuously. The choice of public ontology for integrations system is not a bad approach. To apply public ontology, we propose extracting remote information in real time with the "ontology wrapper" module. The ontology wrapper is suggested because unlike precedent systems, it is possible to use ontology resources from the outside in real time. The dynamic Ontology wrapper will meet the requirements of a large range of users.

Poster B-59

Full Text Search at SGD

Authors:
Qing Dong (*Department of Genetics, Stanford University*)

Short Abstract: To identify and provide access to relevant information within the full text of literature, SGD incorporated Textpresso, a vocabulary-based information extraction system developed by Wormbase. SGD built a semi-automated pipeline to collect full text documents. This resource searches 40,000 full text S. cerevisiae-related journal articles.

Long Abstract:
Full Text Search at SGD

Qing Dong1, Rama Balakrishnan1, Gail Binkley1, Karen R. Christie1, Maria C. Costanzo1, Kara Dolinski2, Selina Dwight1, Stacia R. Engel1, Dianna G. Fisk1, Jodi E. Hirschman1, Ben Hitz1, Eurie L. Hong1, Christopher Lane1, Stuart Miyasato1, Rob Nash1, Rose Oughtred2, Julie Park1, Anand Sethuraman1, Marek Skrzypek1, Chandra L. Theesfeld1, Shuai Weng1, Rey Andrada1, David Botstein2, and J. Michael Cherry1

1. Department of Genetics, Stanford University School of Medicine, Stanford CA, USA 94305
2. Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

The increased availability of full text for journal articles presents a unique challenge in identifying and providing access to relevant information within a growing body of literature. The Textpresso [1] software originally developed by WormBase was integrated into the Saccharomyces Genome Database (SGD, http://www.yeastgenome.org) so that users and curators can perform custom queries of full-text journal articles. This resource (http://www.yeastgenome.org/textpresso) searches the entire text of each journal article based on keywords (such as a gene or protein name) alone or in combination with optional categories of terms (such as cellular compartment) that comprise the Textpresso ontology. SGD programmers optimized the code for efficiency, and SGD curators made yeast-specific modifications to the underlying ontology used in the processing of papers. The resource currently searches 40,000 full-text S. cerevisiae-related journal articles.
One major hurdle to expanding the number of full-text papers included in the Full-Text Search has been automatically downloading the actual journal articles from the publishers' web sites, such as HighWire Press, and archival sites, such as PubMed Central. Although we have had success downloading recently published journal articles, many older papers are not available in electronic format. Manual download is performed when technical challenges prevent complete automation of the downloading process. All PDF documents are validated before archived and their status updated in database. Although this entire process takes extensive computational processing time, given the volume of literature, the limitations of Textpresso indexing method, our users find this resource extremely useful. Our goal is to maintain a complete and up-to-date set of full-text PDF files of literature relevant to S. cerevisiae. We are also actively developing and adopting tools to help identify the passages containing relevant information within the literature in order to increase the efficiency of data acquisition by curators and information retrieval by the general scientific community. SGD is

funded by the US National Human Genome Research Institute. Textpresso is open-source software and is part of the Generic Model Organism Database (GMOD; http://www.gmod.org) effort.

[1] Muller H, Kenny EE, Sternberg PW. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biol. 2004 November; 2(11): e309.