Social Database for Biodiversity

Pannarale P[1], Scioscia G[1], Rubino F[1], Leo P[1], Pappadà G[2], D'Elia D[3], Grillo G[3], Vicario S[3], De Caro G[3], Gisel A[3], Mulè G[4], Susca A[4], Catalano D[5], Licciulli F[3]

Motivation

Biodiversity research concerns with data coming from many different domains (e.g., Biology, Geography, Evolutionary Studies, Genomics, Taxonomy, Environmental Sciences, etc.) which need to be integrated for leading to valuable Biodiversity knowledge. Collecting and integrating data from so many heterogeneous resources is not a trivial task. Data are extremely scattered, heterogeneous in format and purpose, and protected in repositories of several research institutes. Driven by the widely diffused trend of the web of sharing information through aggregation of people with the same interests (social networks), and by the new type of database architecture defined as dynamic distributed federated database, we are proposing a new paradigm of data integration in the Biodiversity domain. Here we present a new approach for the development of a Knowledge Base aiming to the collection, integration and analysis of biodiversity data implemented as a product of the MBLab project.

Methods

The implementation of the Biodiversity Knowledge Base is based on the integration of several components: a robust Database Management System (IBM DB2) managing the large volume of information from public databases like GenBank, a set of GaianDB nodes [1] to manage remote private collections of biodiversity data; the IBM Federator Server to implement the general conceptual schema integrating all biodiversity databases available across remote nodes of MBLab project partners.

Results

GaianDB is a Dynamic Distributed Federated Database of sources whose growth is regulated by biologically inspired principles and graph theoretic methods. By means of the GaianDB network architecture data remains on the remote research group servers, and each database owner is responsible for its integrity, availability and sharing. Each vertex of this network is a suitable entry point receiving the user query and responding with an output aggregating different pieces of information retrieved from the different data sources spanned all over the network. To integrate

[1] IBM Italia S.p.A, Sede di Bari, Via P.L. Laforgia 14, 70125 Bari, Italy [2] Exhicon I.C.T. S.r.l., Bari, Via avv. V. Malcangi 254, 70059 Trani, Italy [3] Istituto di Tecnologie Biomediche (ITB) - CNR, Via Amendola 122/D, Bari, Italy [4] Istituto di Scienze delle Produzioni Alimentari (ISPA), Via Amendola 122/D, Bari, Italy [5] Istituto di Genetica Vegetale (IGV), Via Amendola 165/A, Bari, Italy

GenBank molecular data in the MBLabDB we built an efficient and reliable ETL (Extraction, Transformation and Load) module, implemented with CLIPS Rule Based Programming Language. The ETL extracts information from the feature-based GenBank entries and fits them in the MBLabDB schema. Molecular data collections are structured following a Chado-like model [2], using Sequence Ontology entities and relations. This allows to retrieve data using the biological concepts expressed by the Sequence Ontology [3]. The main result of this work is the development of a standard conceptual schema and a knowledge base architecture tailored to biodiversity data collection, integration and analysis. The database is modeled on six main sections: Taxonomic, Individual, Collection, Supply chain, Experimental molecular data. Currently two biodiversity data collections have been integrated by using GaianDB: the ITEM Collection [4] located at the ISPA-CNR server, and the IGV Mediterranean Plant collection [5] located at the IGV-CNR server. As for Taxonomic area both the NCBI Taxonomy [6] and the Catalogue of Life [7] taxonomic classifications have been included. Using this new paradigm of data integration, biodiversity research groups can contribute to the information network by sharing their data sources with a reasonable effort. In this network, named Social Database for Biodiversity, information remains scattered, but knowledge becomes shared.

Contact e-mail

flavio.licciulli@ba.itb.cnr.it

Supplementary information

References

[1] G. Bent et al. A Dynamic Distributed Federated Database. Proceedings of the Second Annua Conference of the international Technology Alliance, London, UK, Sept. 2008, 238-244 [2] Christopher J. Mungall, David B. Emmert, The FlyBase Consortium (2007). "A Chado case study: an ontology-based modular schema for representing genome-associated biological information". Bioinformatics 23: i337-i346. [3] Eilbeck K., Lewis S.E., Mungall C.J., Yandell M., Stein L., Durbin R., Ashburner M. The Sequence Ontology: A tool for the unification of genome annotations. Genome Biology (2005) 6:R44 [4] http://www.ispa.cnr.it/Collection [5] http://www.igv.cnr.it [6] NCBI Taxonomy: www.ncbi.nlm.nih.gov/Taxonomy/ [7] COL: http://www.catalogueoflife.org/