Integration of Life Science Databases through Knowledge Graphs

Database Center for Life Science (***DBCLS***) & ***BioData Science Initiative (BSI)***, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Japan

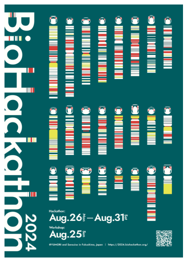
DBCLS Database Center for Life Science

APBJC 2024

At the Database Center for Life Science (DBCLS), we have been developing technologies for the integrated use of life science databases.

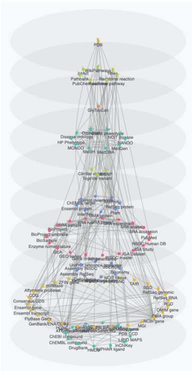
In the RDF Portal, numerous databases are integrated by converting them into knowledge graphs, and TogoID aggregates relationships between database IDs. Based on these developments, TogoDX/Human has realized data integration and visualization to explore diverse information related to humans.

For the integrated use of data, standardization and improved interoperability across databases are essential. To address this, DBCLS has been organizing the international developer conference, BioHackathon, since 2008, fostering international collaboration with database researchers and developers worldwide.



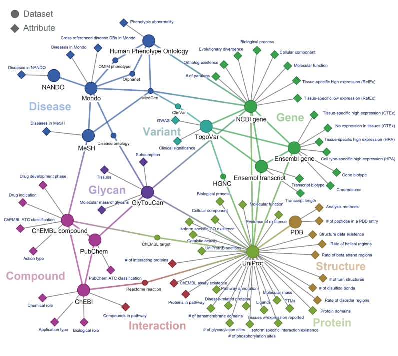
Compilation of Knowledge Graphs in the RDF Portal

Semantic-based integration of ID relations across DBs with TogoID



Structures
Pathways
Glycans
Diseases
Variants
Proteins
Literature
Taxonomy
Genes
Compounds

Consolidation of Human Data in TogoDX



Full-Stack Solutions for Knowledge Graph Development and Deployment

Data sources → (standardization) → RDF → (query) → SPARQL → (security + performance) → Proxy → (data retrieval) → GraphQL API / (transformation) → REST API

GraphQL API → (application) → Analysis → (reproducibility) → Workflow → (experiment)

REST API → (visualization) → Web Components → (end user) → Applications → (data science)

hypothesis; new data; new findings

JSON-LD; OpenLink Virtuoso; GraphDB; Amazon Neptune; jupyter; R Studio; Common Workflow Language

http://togows.org/

TogoWS can search, fetch, and parse major public databases and supports on-the-fly conversion to RDF.

http://togodb.org/

TogoDB publishes users' tabular data as a database and provides functionality to deliver the contents in RDF/SPARQL.

http://d2rq.dbcls.jp/

D2RQ Mapper can turn an RDB into a SPARQL endpoint by mapping an ER schema to an RDF data model.

http://med2rdf.org/

Med2RDF is a community research project that develops and provides RDF converters for major biomedical databases.

https://rdfportal.org/

The RDF Portal provides the world's largest collection of RDFs in life sciences and biomedical databases and serves SPARQL endpoints by utilizing SPARQL-proxy and RDF-config internally.

https://github.com/dbcls/sparql-proxy

SPARQL-proxy

SPARQL-proxy returns cached results for identical queries to improve response times. It also protects the SPARQL endpoint from DoS-like attacks by introducing a query queuing and job management mechanism, as well as rejecting UPDATE queries.

https://github.com/dbcls/sparql-ga

SPARQL-GA

SPARQL-GA optimizes SPARQL queries for Virtuoso by permuting graph patterns using a genetic algorithm.

https://github.com/dbcls/rdf-doctor

RDF-doctor

RDF-doctor scans RDF data, reports occurrences of classes and properties with their cardinalities, and generates ShEx for validation.

https://github.com/dbcls/rdf-config

RDF-config

RDF-config is a tool that can automatically generate SPARQL queries, RDF schema diagrams, ShEx, and GraphQL APIs by simplifying and describing the RDF data model with complex graph structures into tree structures.

https://github.com/dbcls/grasp

Grasp

Grasp turns a SPARQL endpoint into a GraphQL endpoint.

https://github.com/dbcls/sparqlist

SPARQList

SPARQList creates REST APIs instantly by combining SPARQL queries and JavaScript snippets for data retrieval and transformation in a simple Markdown document.

http://togostanza.org/

D3SPARQL

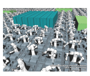
TogoStanza is a generic framework for developing web components, enabling the creation of web applications simply by combining visualization modules. Its reference implementation, MetaStanza, delivers no-code components such as tables and charts that accept SPARQL results, CSV/TSV, and JSON.

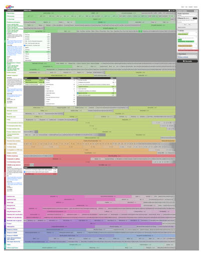
http://togogenome.org/

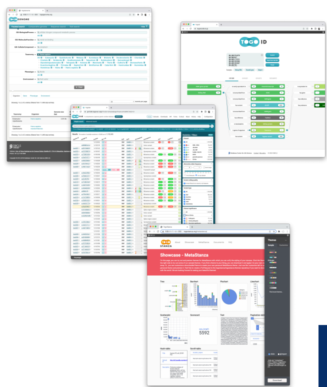
https://togoid.dbcls.jp/

http://togovar.org/

https://togodx.dbcls.jp/human/







DIVE: Data Integration Visual Exploration

APBJC 2024

Database Center for Life Science (***DBCLS***) & ***BioData Science Initiative (BSI)***, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Japan

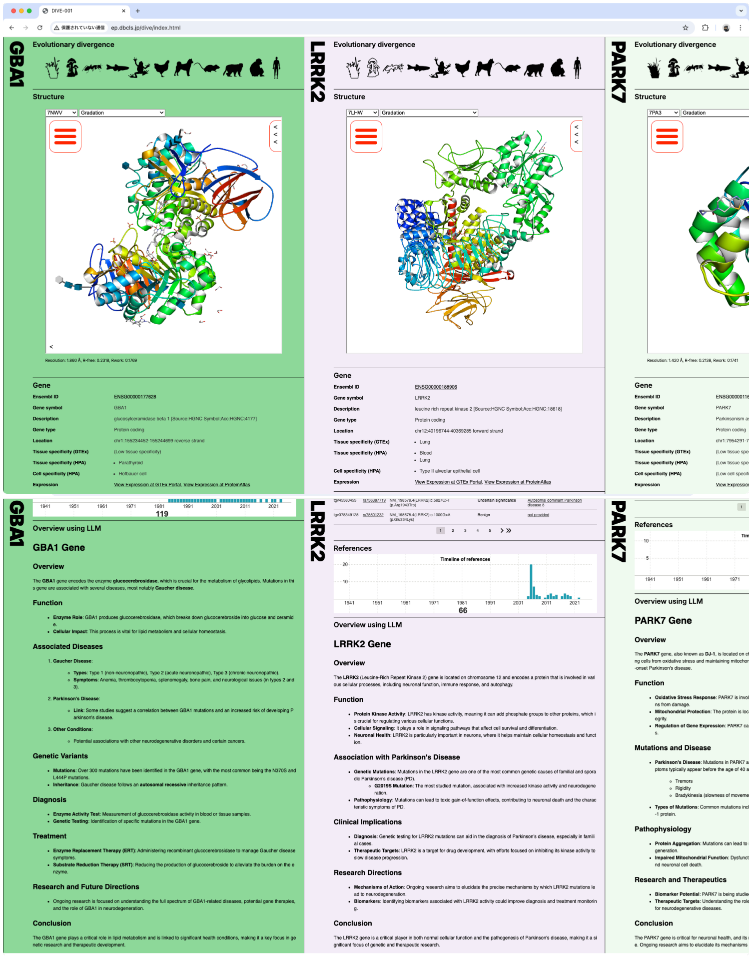
Pursuing an Integrated Database Designed for Everyone to Explore and Comprehend

Due to the complex graph structure of integrated data, it is not easy to browse and intuitively understand diverse information. Addressing this challenge is the project Data Integration Visual Exploration (DIVE), which aims to explore various exhibit applications that enable the interactive visualization of integrated data. The acronym "DIVE" also embodies the meaning of "exploration" into the depths of databases.

For our first project, we took on the challenge of visualizing the relationships between human genes, organs, and evolutionary connections. The human body model uses Bodyparts 3D, developed by DBCLS, while the data showing the relationships between human genes and organs, as well as the evolutionary novelty of each gene, is based on data organized through TogoDX.

Visual Showcase of Related Data with TogoStanza and LLM Summarization

Detailed information about each gene can also be displayed on the web using TogoStanza/MetaStanza. Expression data is retrieved from the RefEx, GTEx, and ProteinAtlas databases, protein function from the UniProt database, 3D structures from the PDB database, variant information and clinical significance from the TogoVar and ClinVar databases, and reference information is obtained from the PubMed database via TogoGenome. Additionally, gene descriptions are generated using Large Language Models.



Future Prospects

This time, we proposed an interactive interface that allows for exploratory analysis of gene set expression across various organs and their evolutionary medical significance, while also being highly intuitive for the general public with an interest in science. Although deriving meaningful hypotheses from the reference data contained in databases is not always easy, we will continue to explore how data can be represented in a way that enables intuitive and efficient data exploration.

About DIVE

DIVE started as a voluntary group at the BSI office in DMM.make AKIBA, with the goal of creating new scientific exhibits through hands-on craftsmanship.

For example, when considering the future of genome medicine, it is essential to improve the public's literacy regarding genomes, genes, and variants. However, we feel that there is still much room for improvement in genome exhibits at science museums both in Japan and abroad. At DIVE, we aim to create exhibits that allow people to fully experience cutting-edge scientific data.