

Application of RDF-based models and tools for enhancing interoperable use of biomedical resources

Toshiaki Katayama

- Database Center for Life Science

Project links

- <http://dbcls.rois.ac.jp/>
- <http://biohackathon.org/>
- <http://med2rdf.org/>

Background information

Background

Our group is mainly from two national bioinformatics centers in Japan

- National Bioscience Database Center (NBDC)
 - Governmental funding agency
- Database Center for Life Science (DBCLS)
 - National institute for database technologies






























Running NBDC/DBCLS BioHackathon meeting over the past 10 years

- Mission: Integration of databases in Life Sciences and Biomedical domains
 - Improvement of Standardization and Interoperability of DBs
 - Currently through the Semantic Web technologies
 - Standard ontologies and identifiers
 - Promoting to provide and use of RDF-based data resources
 - Development of applications on top of the RDF data



RDF resources currently available

- Nucleotide seq & annot
 - INSDC (DDBJ/DBCLS) 
- Genome
 - Ensembl (EBI) 
 - RefSeq (TogoGenome) 
- Protein seq & annot
 - UniProt (SIB) 
- Protein structure
 - PDB (PDBj) 
 - BMRB (PDBj) 
 - FAMSBASE (Chuo U) 
- Compounds
 - PubChem (NCBI) 
 - ChEMBL (EBI) 
 - Nikkaji (JST) 
- Gene expression
 - RefEx (DBCLS) 
 - ExpressionAtlas (EBI) 
- Samples
 - BioSamples (EBI/DDBJ) 
 - JCM (RIKEN) 

- Biomedical (Med2RDF)
 - ICGC, COSMIC, CIViC 
 - DGIdb, OpenTG-Gates 
 - ClinVar, dbSNP, dbVar 
 - ExAC, gnomAD 
 - HiNT, INstruct 
- Glycome
 - GlyTouCan, GlycoEpitope, WURCS, GGDonto, PAConto 
- Proteome
 - jPOST 
 - The Human Protein Atlas 
- Pathway
 - Reactome (EBI) 
- Others
 - MeSH (NCBI) 
 - BioModels (EBI) 
 - MBGD (NIBB/DBCLS) 
 - Quanto (DBCLS) 

Goals of the hacking project

TogoStanza (TogoGenome)

- **Reusable Web Components for Genome DB**


- <http://togostanza.org/>
- (e.g. <http://togogenome.org/gene/9606:APOE>)
- SPARQL back-ended visualization modules (cf BioJS)
- Already used in TogoGenome, TogoVar, MicrobeDB.jp etc.

- **General goal of the hacking project**

- Expose embedded SPARQL queries as REST APIs
- Potential collaboration with BioJS?

- **Expected results** at the end of the hackathon

- Data retrieval from REST API through SPARQLList
- Better design ideas for the future



The screenshot shows the TogoStanza website. At the top, there's a navigation bar with links: About, Showcase, Documents, and FAQ. Below this, there are several circular charts and a map. A prominent orange banner reads "Welcome to TogoStanza!" followed by a paragraph describing the framework. Below the banner, there's a section titled "Stanza and NanoStanza" which explains the framework's purpose and its components. At the bottom, there's a "Users" section displaying logos and links for various projects: TogoGenome, MicrobeDB, CyanoBase, MGBD, and TogoVar.

Welcome to TogoStanza!

TogoStanza is a generic Web framework which enables the development of reusable Web components that are embeddable into any Web applications. Although it can be used for any purposes, the main focus of TogoStanza is to assist the development of Semantic Web components such as querying SPARQL endpoints and visualizing the returned results. This portal site provides a showcase of existing TogoStanza components and a set of documents for users and developers.

Stanza and NanoStanza

TogoStanza is developed by Database Center for Life Science (DBCLS) and its service, framework and source code are freely available. Currently the TogoStanza framework supports two types of components, Stanza (standard version) and NanoStanza (iconic version), both are embeddable into any Web applications.

Our motivation to introduce the TogoStanza framework was originated from the fact that there were huge redundancies in the database development in life sciences. For example, many existing genome databases have independently developed mostly similar elements in their Web applications. With TogoStanza, service providers can focus on the development of database components unique to their data and save the development costs by reusing and/or customizing existing components.

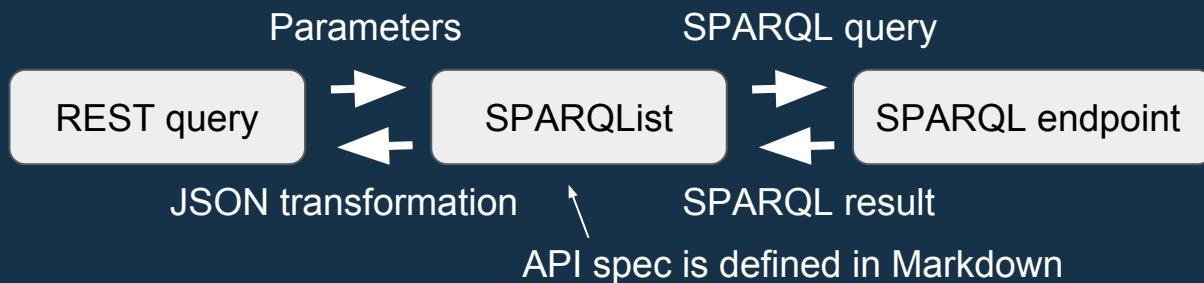
Users

- TogoGenome: <http://togogenome.org/>
- MicrobeDB: <http://microbedb.jp/>
- CyanoBase: <http://genome.microbedb.jp/cyanobase/>
- MGBD: <http://www.genome.ad.jp/>
- TogoVar: <http://togo-var.biocenterb.jp/>

SPARQList

- **REST API repository for SPARQL**

- <http://togostanza.org/sparqlist/>
- <https://github.com/dbcls/sparqlist>
- API definition can be written in Markdown
- Execute SPARQL on-the-fly
- Transform resulting JSON with JavaScript

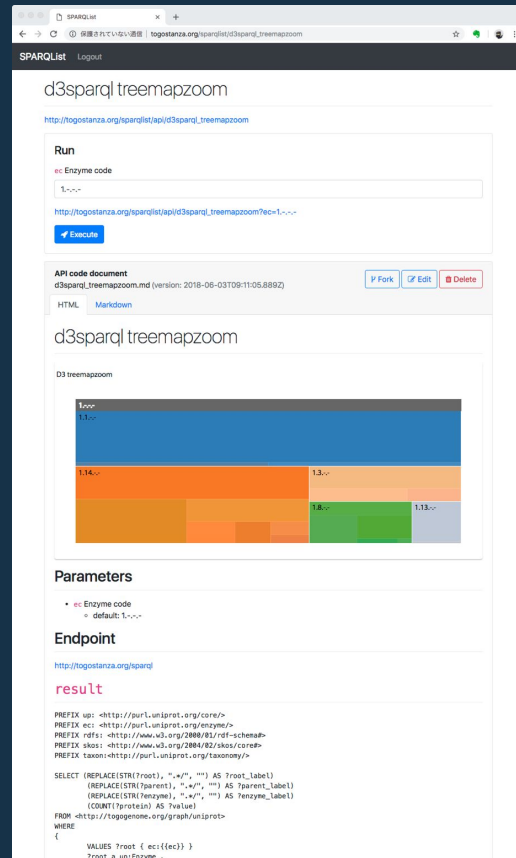


- **General goal of the hacking project**

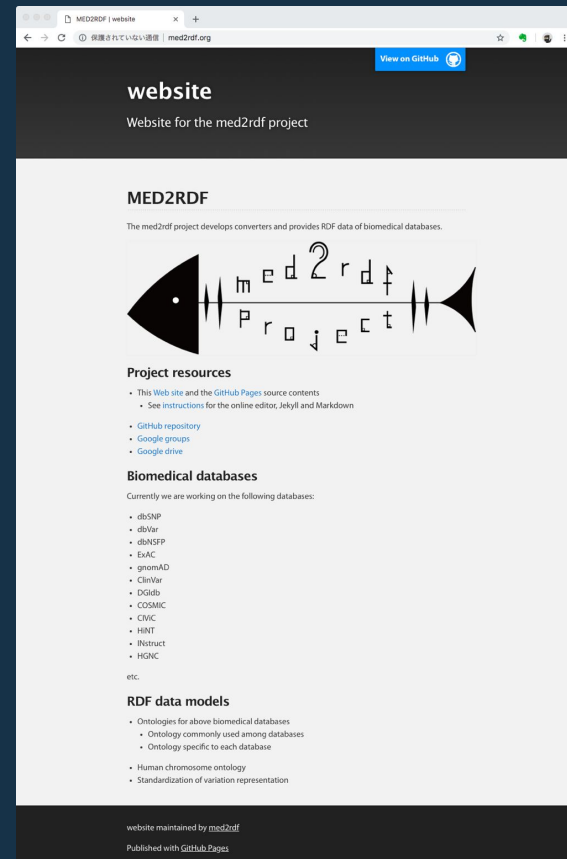
- Develop REST APIs for SPARQL embedded in TogoStanzas

- **Expected results** at the end of the hackathon

- Data retrieval from REST API through SPARQList



- **Reusable Biomedical RDF datasets**
 - <http://med2rdf.org/>
 - Provides RDF version of biomedical datasets for
 - ClinVar, dbSNP, dbVar, ExAC, gnomAD
 - ICGC, COSMIC, CIViC, HiNT, INstruct
 - Ontologies for Human Chromosome, Variation model
 - In collaboration with Kyoto University under the AMED project
- **General goal of the hacking project**
 - Develop Ontologies & TogoStanza for Med2RDF data
 - Pull more attention from international collaborators
- **Expected results** at the end of the hackathon
 - Completion and deployment of HCO and IDO at Id.org



SPARQ Builder / LOD Surfer

- **A search system based on class-class relationships**
 - Using metadata extracted from SPARQL endpoint, SPARQL Builder supports in writing SPARQL.
 - Based on the metadata for SPARQL Builder, LOD Surfer can extract data from multiple endpoints.
- **General goal of the hacking project**
 - Develop an efficient federated search system using the metadata and class-class relationships.
 - Find an application to these systems.
- **Expected results at the end of the hackathon**
 - Include more SPARQL endpoints especially for large datasets.

SPARQL Builder

SPARQL endpoint <https://www.ebi.ac.uk/rdf/services/sparql>

start class Protein end class Pathway

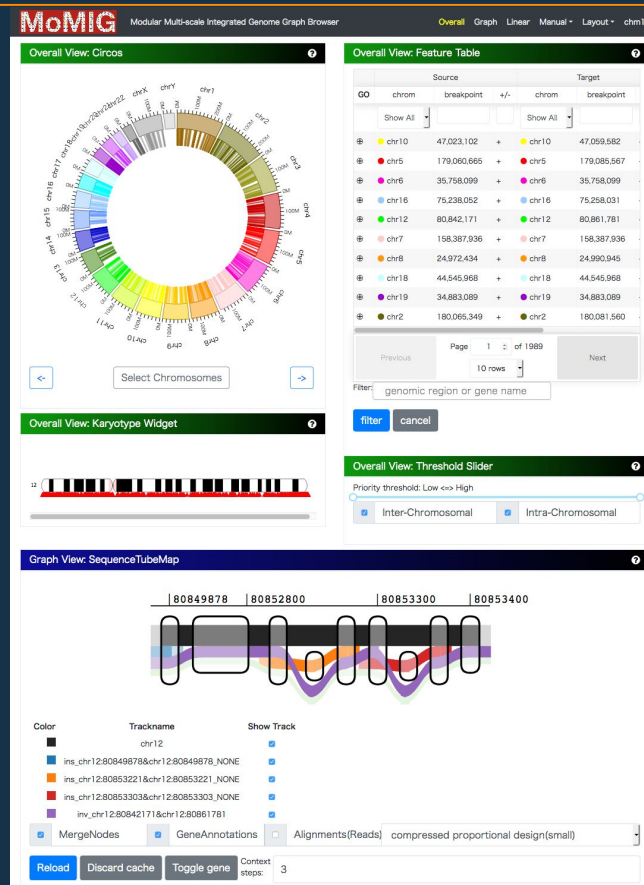
95 Paths found. [Permalink](#)

Select a path to generate a SPARQL query.

| | | | | |
|---------|-----------------|---------------------|----------------------|---------|
| Protein | - controller - | Control | - controlled - | Pathway |
| Protein | - left - | Degradation | - pathwayComponent - | Pathway |
| Protein | - product - | TemplateReaction | - pathwayComponent - | Pathway |
| Protein | - participant - | TemplateReaction | - pathwayComponent - | Pathway |
| Protein | - right - | BiochemicalReaction | - pathwayComponent - | Pathway |

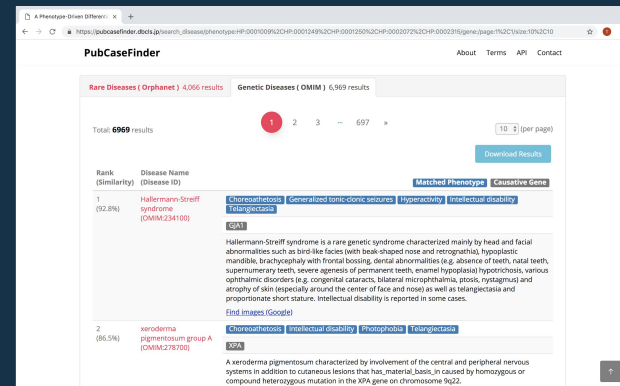
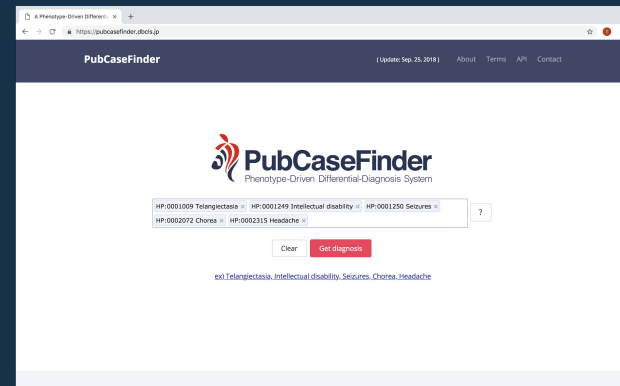
Graph-based Japanese Reference Genome

- **Graph-based Japanese Reference Genome**
 - Improves mapping analysis of Japanese genome
 - Including SNPs and SVs observed in Japanese as branching of mathematical graph
 - Cooperates with related projects
 - Genome graph workshops: <http://genomegraph.jp>
 - MoMIG, a graph genome browser (right figure)
 - <http://demo.momig.tokyo/>
 - <https://github.com/MoMI-G/MoMI-G/>
- **General goal of the hacking project**
 - Publish graph-based Japanese reference genome
- **Expected results** at the end of the hackathon
 - Evaluation of graph-based Japanese reference genome



PubCaseFinder

- **A phenotype-driven differential-diagnosis system**
 - <https://pubcasefinder.dbcls.jp/>
 - Helps clinicians make a diagnosis for rare diseases and genetic diseases defined in Orphanet and OMIM
- **General goal of the hacking project**
 - Develop a REST API for providing a ranked list of genetic diseases based on phenotypic similarity
- **Expected results at the end of the hackathon**
 - Better design ideas for the future



Other tools we developed (to make use of RDF data)

Semantic Web tools developed in DBCLS

- NBDC RDF Portal: RDF data repository with summaries and SPARQL endpoints
- TogoWS: search, retrieval, parse, and convert entries into RDF
- TogoDB: deploy your datasets as a Web database and generate RDF
- D3SPARQL.js: D3.js based JS library to visualize SPARQL results
- SPARQL-proxy: safely deploy a SPARQL endpoint with cache and job management
- SPANG: easy-to-use command-line SPARQL client
- D2RQ Mapper: configure RDB as a SPARQL endpoint w/ intuitive GUI
- Umaka-Yummy: monitor and evaluate the quality of SPARQL endpoints
- :
- Also developing tools for NGS, Web apps, Text-mining, QA system, ML & AI

Hack organisation

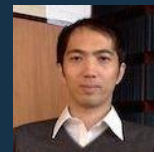
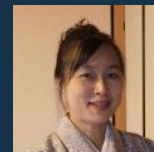
Organisation of the hacking project

- TogoGenome/TogoStanza & SPARQLList
 - Refactoring to extract embedded SPARQL queries as public REST APIs
 - Explore collaboration with BioJS
 - Toshiaki Katayama, Shuichi Kawashima
- Human Chromosome Ontology & Identifiers.org Ontology
 - Provide canonical URIs for human chromosomes
 - Provide semantics to the Identifiers.org databases
 - Update DBCLS RDFizing database guidelines according to the results
 - <https://github.com/med2rdf/hco>
 - <https://github.com/ktym/idorg-ontology>
 - <https://github.com/dbcls/rdfizing-db-guidelines>
 - Toshiaki Katayama, Shuichi Kawashima
- Japanese reference genome graph
 - Develop initial version of Japanese genome graph
 - Toshiyuki Yokoyama, Toshiaki Katayama



Organisation of the hacking project

- SPARQL Builder / LOD Surfer
 - Include more SPARQL endpoints especially for large datasets.
 - <https://github.com/sparqlbuilder>
 - <https://github.com/LODSurfer>
 - Atsuko Yamaguchi
- PubCaseFinder
 - Develop a REST API for providing a ranked list of genetic diseases defined in OMIM
 - <https://pubcasefinder.dbcls.jp/mme>
 - Toyofumi Fujiwara
- SPANG
 - Facilitate database integration through SPARQL
 - <https://github.com/dbcls/spang>
 - Hirokazu Chiba



Organisation of the hacking project

- Public data & workflows
 - Get data from public repo, then deploy CWL workflows on clouds and run
 - <https://github.com/pitagora-galaxy/cwl>
 - Manabu Ishii, Ryota Yamanaka, Tazro Ohta



Post-biohackathon perspectives

- More interoperable data
- More data science
- More applications
- More international collaborations
- Continued in the NBDC/DBCLS BioHackathon in Matsue this December :)