

RDF Portalの刷新と機能拡張

多様なアクセス手段とAIによる生命科学RDF活用



トーゴの日
2026

川島 秀一

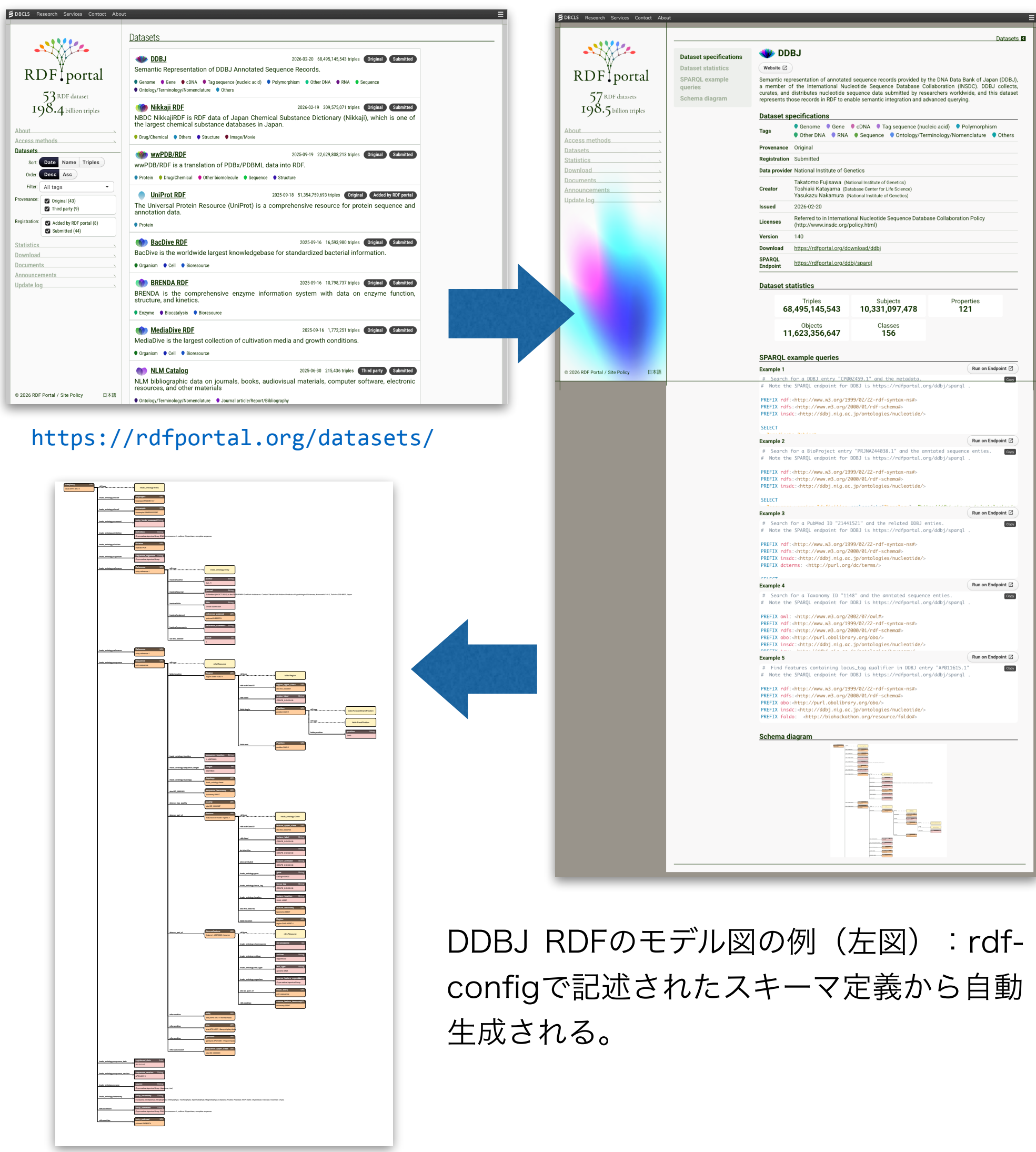
大学共同利用機関法人 情報・システム研究機構, 国立遺伝学研究所 バイオデータ研究拠点
ライフサイエンス統合データベース部門 (DBCLS)

RDF Portalは、生命科学分野のRDFデータセットを収集し、統一的に検索・利用できる環境を提供するデータ統合基盤である。2025年度末にウェブサイトを大幅に刷新し、データセットの概要、統計情報、来歴、更新履歴、サンプルSPARQL、RDF-configから自動生成したスキーマ図などを一つのページから参照できるようにした。さらに、従来のSPARQLエンドポイントとデータダウンロードに加え、SPARQL Composer、GraphQL API、LLMチャットインターフェース、TogoMCPを介したAIからのアクセスなど、利用者の技術レベルや目的に応じた複数のアクセス方法を整備した。TogoMCPについては関連発表で詳細が紹介されるため、本発表ではRDF Portal全体における位置付けを簡潔に示す。また、大規模RDFを高速に検索するためのQLeverトリプルストアによる運用試験と、新規データセットの継続的な追加について報告する。これらの機能拡張を通じて、RDFやSPARQLに詳しい研究者だけでなく、初学者やAIエージェントも生命科学データを発見し、理解し、横断的に活用できる基盤を目指している。

UIの刷新



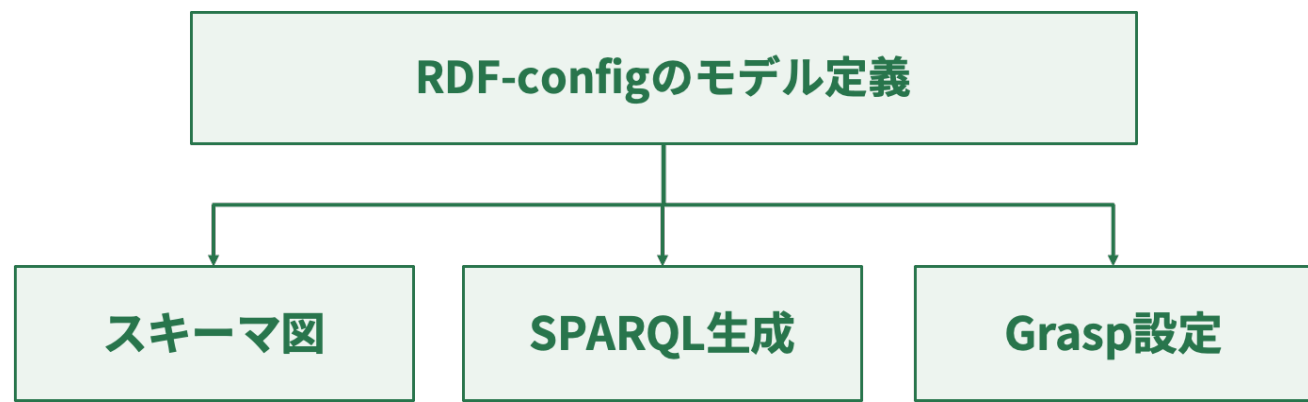
刷新した詳細ページでは、概要、提供機関、RDF作成者、来歴、ライセンス、バージョン、統計情報をまとめて確認できる。さらに、サンプルSPARQL、RDF-configから生成したスキーマ図、検索・ダウンロードへのリンクを配置した。データの内容を理解してから、具体的な利用へ進める構成である。



DDBJ RDFのモデル図の例(左図) : rdf-configで記述されたスキーマ定義から自動生成される。

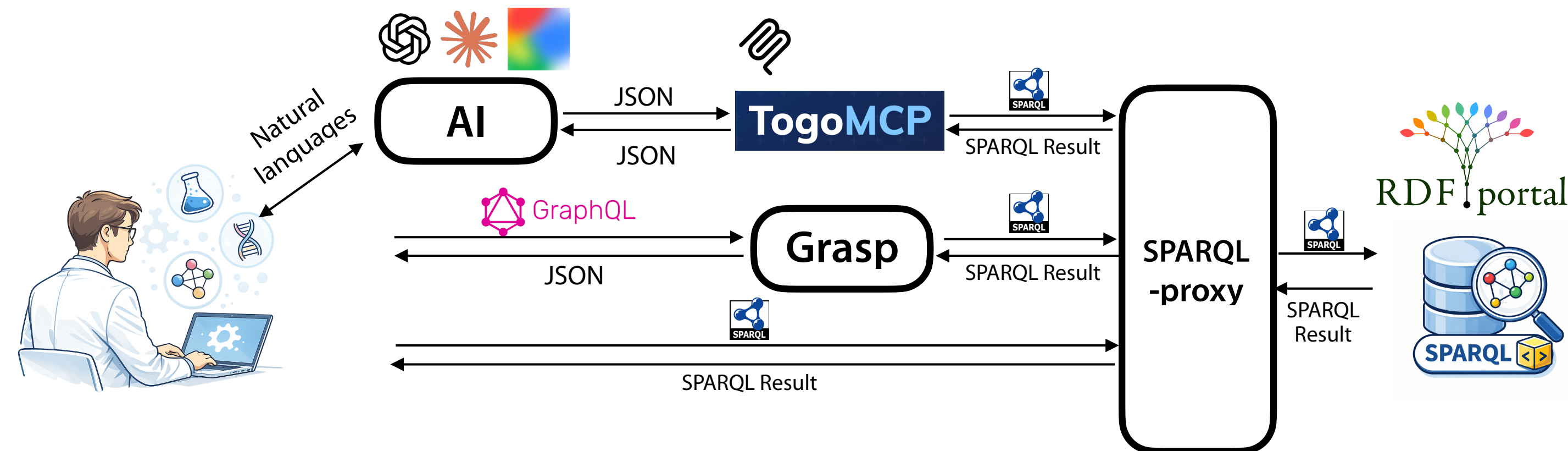
RDF-configを共通の設計情報に

RDF-configは、クラス、プロパティ、値の例など、RDFの構造をYAMLで記述する枠組みである。人が読むスキーマ図に加え、SPARQLの生成や、検証用のShEx、Graspの設定にも同じモデル定義を利用できる。表示・検索・APIごとに構造を別々に記述する負担を減らし、一貫した利用を支える。



同じRDFモデルの再利用。スキーマ図で構造を理解し、検索やアプリケーション連携へ進む。

アクセス方法



SPARQLエンドポイント

RDFデータに対してクエリを実行できるSPARQLエンドポイントを提供している。標準的なクエリ言語であるSPARQL 1.1を用いて、複数のデータセットを横断する複雑な検索やデータ抽出を行うことができる。以下に示すように、複数のVirtuosoインスタンスを設置している。

Endpoint name	URL
Primary	https://rdfportal.org/primary/sparql
DDBJ	https://rdfportal.org/ddbj/sparql
DBKERO	https://rdfportal.org/kero/sparql
PDB	https://rdfportal.org/pdb/sparql
SIB	https://rdfportal.org/dib/sparql
EBI	https://rdfportal.org/ebi/sparql
NCBI	https://rdfportal.org/ncbi/sparql
PubChem	https://rdfportal.org/pubchem/sparql

List of SPARQL endpoints in RDF Portal

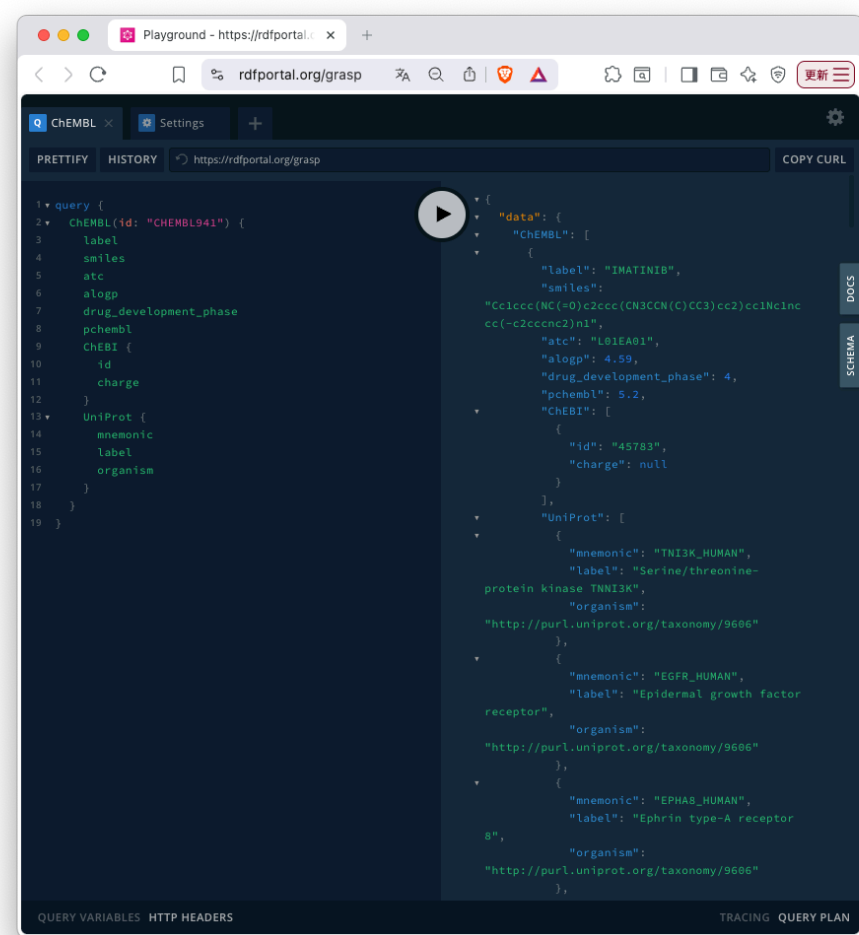
QLeverによる運用試験：データの増大に対応し、大規模RDFの高速検索を目指してQLeverトリプルストアで運用試験を進めている。アクセス方法の拡張と並行し、それを支える検索基盤の性能と運用性を検討する。

GraphQL API (Grasp)

GraphQLは、ユーザーが必要なデータを正確に指定することで、柔軟かつ効率的なデータ取得を可能にするクエリ言語およびAPIの設計方式。

GraphQLでは、ユーザーが取得したいフィールドを指定して問い合わせる。RDF Portalでは、DBCLSが開発したGraspがGraphQLをSPARQLへ変換し、検索結果をGraphQLの応答として返す。

アプリケーション開発者は、RDFのクエリを直接記述せずに、必要な情報を構造化データとして取り込める。ブラウザのGraphiQL画面では、スキーマや入力候補を確認しながらクエリを書ける。対応する関係を入れ子で指定することもできる。現在の対象は対応済みデータセットに限られ、各データセットで利用可能な型とフィールドに従う。

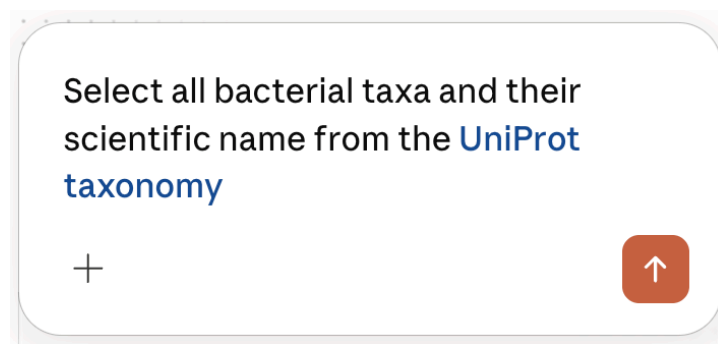


<https://rdfportal.org/grasp>

TogoMCP : AIエージェントの入口

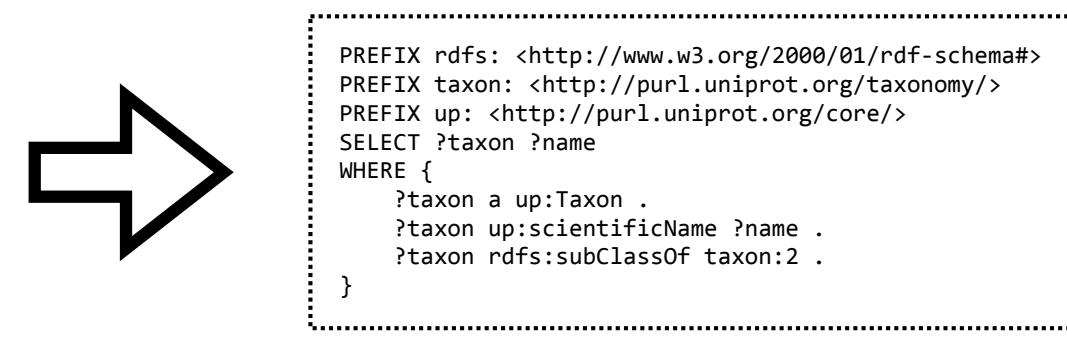
TogoMCPは、AIが外部の情報やツールにアクセスするためのModel Context Protocol (MCP) に対応したサーバーである。AIアシスタントは、データセットの説明、スキーマ、クエリ例を参照し、検索ツールを使ってRDF Portalのデータを取得する。利用者は自然言語で目的を伝え、AIを介してデータを探索できる。

RDF Portalでは、TogoMCPをAI向けのアクセス層として位置付ける。RDFの構造を説明する情報と検索機能をAIへ渡し、データに基づく回答を支援する。利用時には取得したデータ、出典、検索条件を確認する。詳細な機能と利用事例は関連発表で紹介する。



The Barrier: While highly expressive, SPARQL query construction remains a steep barrier for many biological researcher.

<https://togomcp.rdfportal.org/>



The Vision: Bridging the gap between 180 billion RDF triples and intuitive natural language inquiries. Empowering Large Language Models (LLMs) to natively navigate the knowledge graph, while grounding responses in structured knowledge to reduce hallucinations and improve reliability.

SPARQL Composer

データセット、取得する項目、検索条件を画面で指定し、SPARQLクエリを生成するインターフェースである。RDF-configのモデル定義を利用するため、利用者はまずモデル内の項目を選び、必要に応じて値を入力する。生成したクエリを対象エンドポイントで実行し、結果を確認できる。

SPARQLの構文を一から覚える前に、項目の組合せとクエリの対応を確かめられる。生成されたクエリを出発点として条件を追加する使い方もでき、初学者の学習と、日常的な検索の作成を支援する。対象となるモデルと項目の確認が出発点となる。

Dataset	Major topic	Description	#triples	Dataset	Major topic	Description	#triples	Dataset	Major topic	Description	#triples
DDBJ	Sequence	Annotated nucleotide sequence records	684951455	LigandBox	Compound	3D small-molecule structures for docking studies	2987628	KNAPSAcK RDF	Compound	Data from the KNAPSAcK family of databases	348961
UniProt RDF	Protein	Protein sequences and functional annotations	388736045	MicrobeDB.jp	Microbiology	Archived genome-centered microbial information	2612535	Rhea	Biochemical reaction	Expert-curated biochemical reactions	204690
PubChem RDF	Compound	An open database of chemical compounds	290849054	BioSample RDF		BioSample RDF developed by DBCLS and distributed	1755948	MediaDive RDF	Culture medium	Microbial cultivation media and growth conditions	177225
wwPDB/RDF	Protein structure	Structural data from PDBx/PDBML	243788674	GlyCosmos	Glycan	Integrated glycans and related biological information	1563955	GlyYouCan	Glycan	An international glycan structure repository	174964
Ensembl RDF	Genome	Genome annotations for vertebrates and other eukaryotes	185236572	MeSH RDF	Vocabulary	The MeSH biomedical vocabulary	1547047	HomoloGene RDF	Homolog	Homologous gene information from HomoloGene	155312
MoGo+ RDF	Variation	Genomic variation across laboratory mouse strains	123768626	MedGen	Disease	Human medical genetics and genetic conditions	1257036	HGNC	Gene	Approved human gene nomenclature	134595
DBKERO RDF	Omics	Multi-omics data including SNVs and gene expression	111989132	RefEx RDF	Gene expression	Reference gene expression from CAGE and Genes	1234475	JCM	Bioresource	Microbial strain metadata from RIKEN JCM	111374
TogoID RDF	Identifier mapping	Links and conversions between database identifiers	77707321	GTDB	Taxonomy	Genome-based taxonomy of bacteria and archaea	1096451	NBRC	Bioresource	Microbial resources maintained and provided by	982005
Open TG-GATES	Toxicogenomics	Public toxicogenomics data	680038460	Quanto RDF	Sequence quality	Sequence quality statistics from SRA data	1077826	NLM Catalog	Literature	NLM bibliographic records for journals, books, and	215436
Bgee	Gene expression	Gene expression comparisons across animal species	675588575	AMR portal RDF	AMR	Antimicrobial resistance phenotypes and genotypes	7160299	NARO Genebank	Bioresource	Microbial resources from NARO Genebank	132210
PubMed	Literature	Biomedical literature citations	282590315	BioModels RDF	Biological modeling	Computational models of biological processes	4761056	Integbio Database Catalog	Metadata	The Integbio Database Catalog in RDF	117526
ClinVar RDF	Variation	Clinically relevant variants and associated phenotypes	199192365	GWAS catalog	Variation	Genome-wide association study results	4708257	PAConto	Glycan / Disease	Pathogen adhesion to glycans and related infections	81785
MBGD RDF	Genome	Microbial genomes for comparative analysis	160901814	MassBank	Metabolome	Mass spectra for small-molecule identification	2988909	TISMeD	Expression / Methylation	Tissue-specific methylation, gene and protein expression	55727
PubTator Central	Literature	Automated biomedical concept annotations in literature	118798774	Ensembl RDF (GRCh38)	Genome	Genome annotations for human assembly GRCh38	2615204	SSBD	Images	Metadata for quantitative data and microscopy	40300
JoGo	Haplotype	Human gene haplotypes at multiple sequence levels	630117311	Reactome RDF	Pathway	Expert-curated biological pathways	2161077	NANDO	Disease	Vocabulary of designated intractable and pediatric diseases	39977
ChEMBL RDF	Compound	Bioactive molecules with drug-like properties	740029339	FAMSBASE GPCR	Protein structure	Protein structure	2129778	GGDonto	Disease	Glycan metabolism disorders and their causative factors	39518
BMRB/RDF	Protein structure	Biomolecular NMR data from NMR-STAR	658230916	Cellosaurus	Cell line	Annotated cell lines used in biomedical research	1946425	GlycoEpitope	Glycan	Carbohydrate antigens and antibodies	30150
fanta.bio	Gene regulation	Promoters, enhancers, and transcriptional activities	643231015	BacDive RDF	Microbiology	Standardized bacterial information	1659398	HCO	Ontology	Assembly-aware identifiers for human chromosomes	25646
OMA RDF	Ortholog	Ortholog inference across complete genomes	630117311	PubCaseFinder RDF	Disease	HPO-based diagnostic support for rare and genetic diseases	1636738	PolyInfo BDG S, ID definition	Polymer	Biodegradation identifiers and links for synthetic polymers	5446
jPOST database	Proteome	Proteomics data reanalyzed using standardized criteria	584249484	PGDBj Ortholog RDF	Ortholog	Ortholog	1365217	Nucleic Acid Drug Data	Drug	Approved and clinical-stage nucleic acid drugs	948
Linked ICGC	Cancer genomics	Cancer donor information and somatic mutations	577714133	BRENDA RDF	Enzyme	Enzyme functions, structures, and kinetics	1079873	MCO	Ontology	Assembly-aware identifiers for mouse chromosomes	466
NCBI Gene RDF	Gene	Gene information from NCBI Gene	577082774	ChEBI RDF	Compound	Small-molecule structures, roles, and relationships	9521789				
Nikkaji RDF	Compound	Chemical compounds in the Japan Chemical Substances Database	550262723	Dataset of WURCS	Glycan	Glycan structures described in WURCS	6213789				

67 RDF datasets 231 billion triples

データセットの一覧

