

RDFポータルへのアクセスを容易にするTogoMCPサーバーの開発

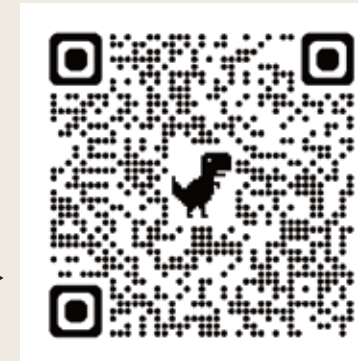
山本 泰智¹、藤澤 貴智²、金城 玲³

(1) 情報・システム研究機構 データサイエンス共同利用基盤施設 ライフサイエンス統合データベースセンター (DBCLS)

(2) 国立遺伝学研究所

(3) 合同会社アニマ・マキナ

データベースに自然言語で聞きたい質問を募集中 →



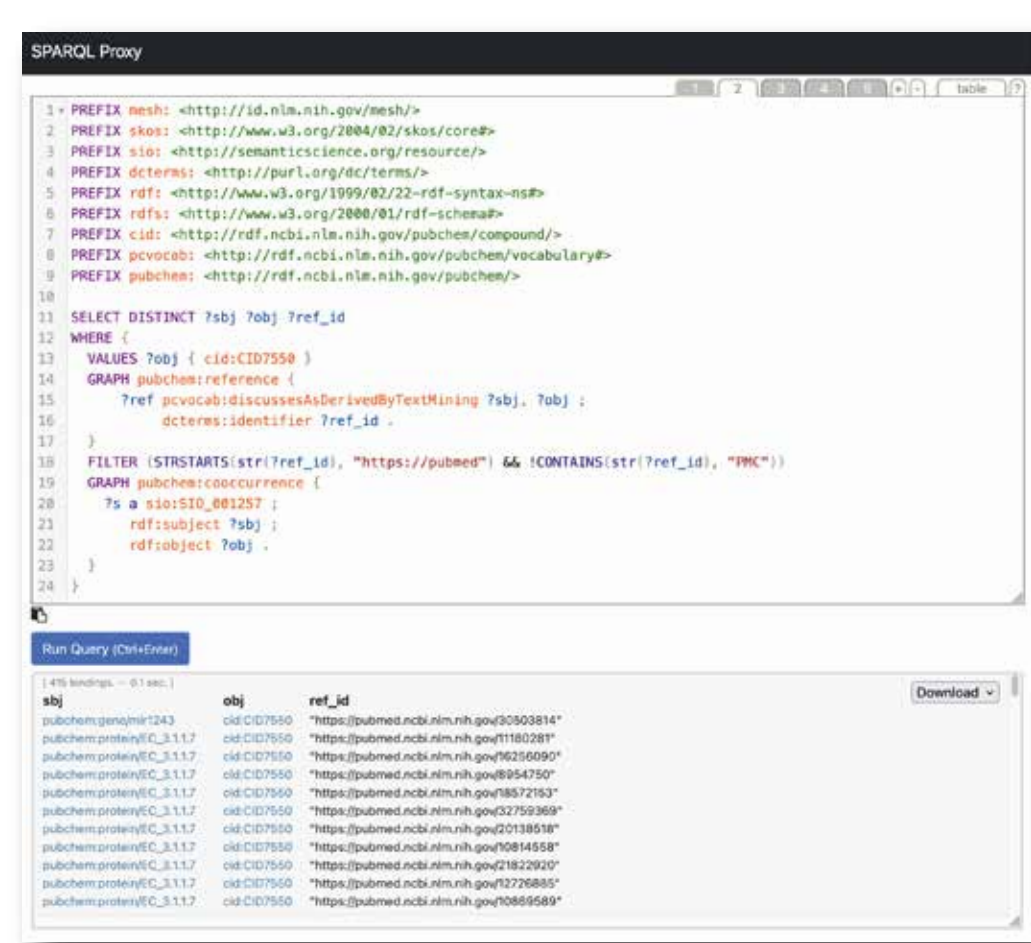
RDFポータルは、生命科学分野における遺伝子、タンパク質、パスウェイ、化合物、疾患など幅広いRDFデータセットを収載している。RDFデータセットから所望のデータを取り出すための問い合わせ言語SPARQLは多様な検索を可能にするが、言語そのものの複雑さや、予め正確なURIの把握が必要であることなどにより、欲しいデータを取り出すまでの道のりが長くなりがちである。さらに、同一の概念がデータセット間で異なるURIで表現されることも多く、その利用の困難さに拍車をかけている。そこで我々は技術的知識がなくても自然言語でRDFポータルへのアクセスを可能とするためのMCPサーバ「TogoMCP」を開発している。

RDFポータルは大容量のデータを収載



<https://rdfportal.org/>

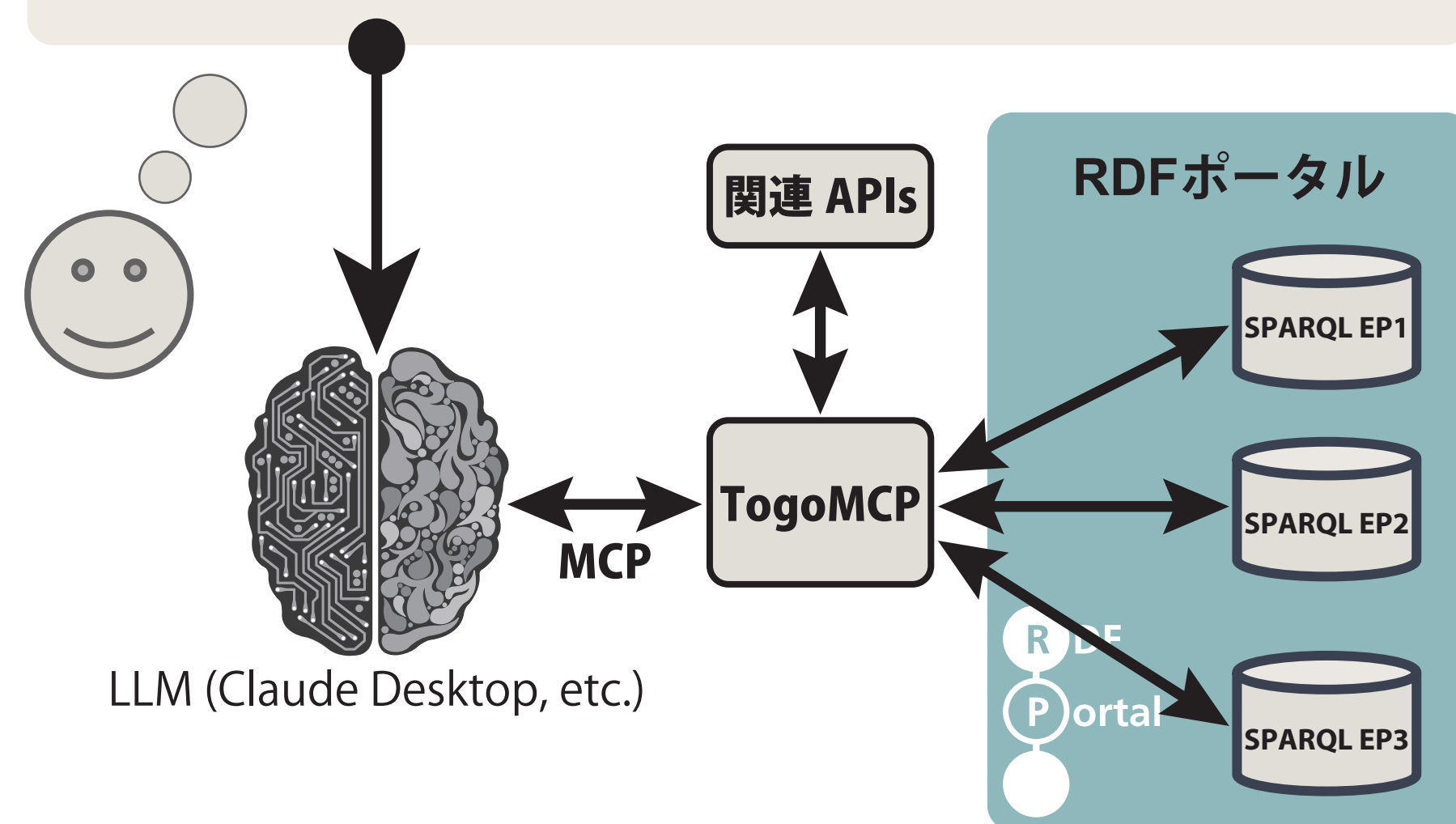
知りたいことをSPARQLで表現



Dukeys68

知りたいことを自然言語で表現

How do amyloid- β aggregation (molecular), neuroinflammation (pathway), synaptic dysfunction (cellular), and neurodegeneration (tissue) lead to memory loss and cognitive decline (clinical)? Also, explain the mechanism of drugs for treatment, if available. Integrate GO (molecular functions), Reactome (pathway dysregulation), UniProt (protein interactions), MONDO (disease phenotypes), ChEMBL (drugs), and MeSH (clinical manifestations) to build comprehensive models of how molecular defects lead to clinical symptoms. Use the RDF Portal to find evidence.



Multi-Scale Model of Alzheimer's Disease Pathogenesis

From Molecular Defects to Clinical Symptoms: An Integrated RDF Portal Analysis

UniProt Gene Ontology (GO) Reactome ChEMBL MeSH

Integration Overview

This model integrates data from multiple RDF databases to show how molecular defects (amyloid- β aggregation) progress through cellular dysfunction (synaptic damage), pathway dysregulation (neuroinflammation), tissue degeneration (neuronal loss), ultimately manifesting as clinical symptoms (memory loss and cognitive decline). Each level is supported by evidence from UniProt, GO, Reactome, ChEMBL, and MeSH databases.

MOLECULAR LEVEL

Amyloid- β Protein & Aggregation

Core Protein

Name: Amyloid-beta precursor protein (APP)
UniProt ID: P05067
Function: Neural cell adhesion to extracellular matrix; regulates neurite outgrowth

GO Terms:

- GO:0005576 - Extracellular region
- GO:0005886 - Plasma membrane
- GO:0005737 - Cytoplasm
- GO:0003677 - DNA binding
- GO:0004867 - Serine-type endopeptidase inhibitor activity

APP Processing Pathways

BACE1 (Beta-secretase 1)
UniProt: P56817
Cleaves APP to produce β -amyloid peptides
Aspartyl protease activity

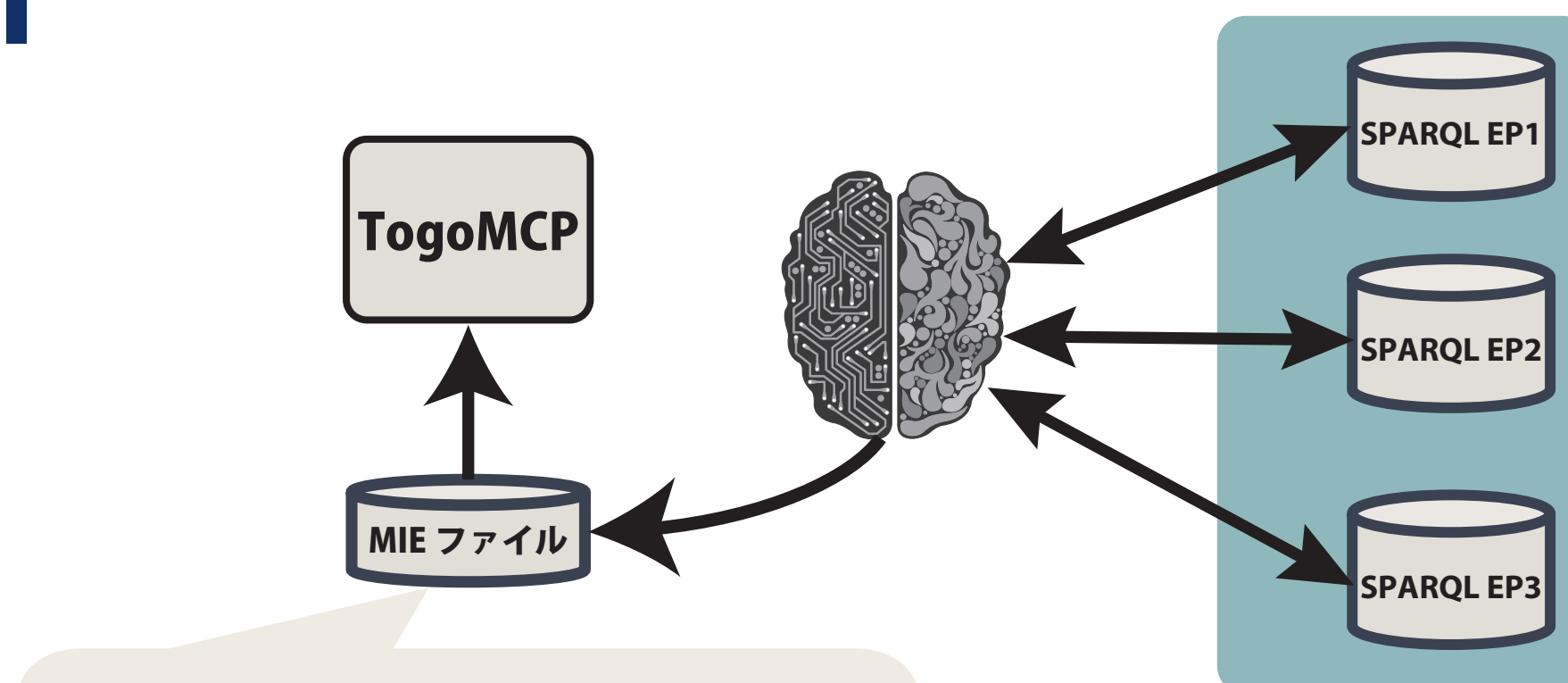
ADAM10

```
PREFIX up: <http://purl.uniprot.org/core/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX uniprot: <http://purl.uniprot.org/uniprot/>
```

```
SELECT ?fullName ?function ?goTerm
WHERE {
  uniprot:P05067 up:recommendedName/up:fullName ?fullName .
  OPTIONAL {
    uniprot:P05067 up:annotation ?annotation .
    ?annotation rdf:type up:Function_Annotation .
    ?annotation rdfs:comment ?function .
  }
  OPTIONAL {
    uniprot:P05067 up:classifiedWith ?goTerm .
    FILTER(REGEX(STR(?goTerm), "60_"))
  }
}
```

LLMが生成したSPARQLクエリ例

適切なクエリをLLMに生成させる方策



データセット毎にデータ構造やデータの内容などを構造化して記述するデータセット

* (yaml ファイル)

* 2025年9月に開催されたバイオハッカソンで策定

LLMは文脈を適切に与えることで望ましいSPARQLクエリを生成しやすくなるため、RDFポータルに収載されているデータセット毎に必要な情報を予め収集し、MIEファイルとして保存する。

MIEファイルの生成については、基本的なデータ構造を記載したテンプレートを基にLLMが行うが、より適切な手法を検討中である。

同義URI辞書を構築して繋がり易く

LLMに対してもデータセット横断的な問い合わせを生成しやすくし、問い合わせ先の網羅性を高めるため、RDFポータルに収載されている全てのRDFデータに対して、下記の述語で結ばれるトリプルの全てを取得し、同義URI辞書を構築している。

この情報と、TogoIDから得られる同義URI情報を合わせて、どのようにMCPサーバーからLLMに提供するのが有効であるか検討中である。

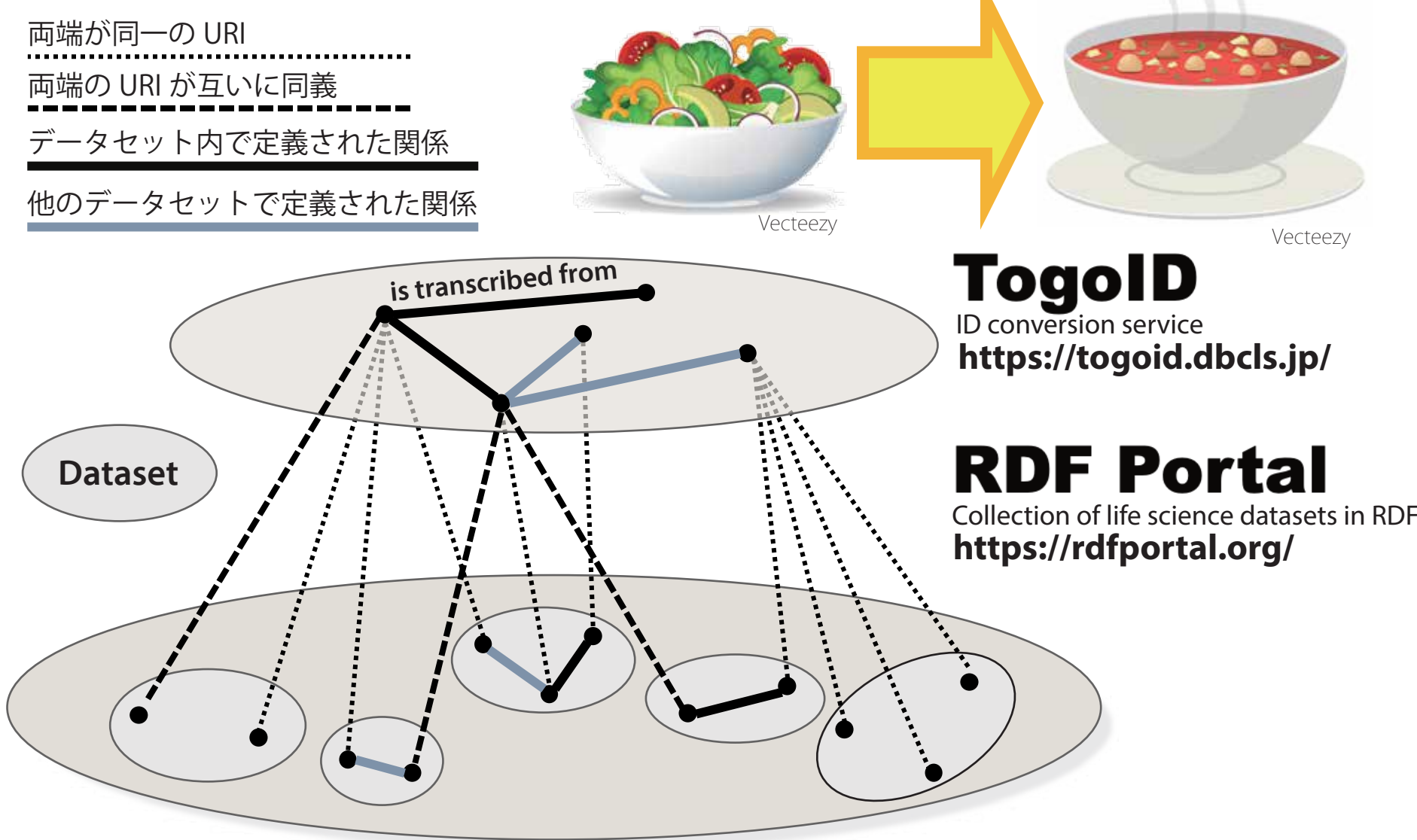
- <http://www.w3.org/2000/01/rdf-schema#seeAlso> (rdfs:seeAlso)
- <http://www.w3.org/2004/02/skos/core#exactMatch> (skos:exactMatch)
- <http://www.w3.org/2004/02/skos/core#closeMatch> (skos:closeMatch)
- <http://www.w3.org/2002/07/owl#sameAs> (owl:sameAs)
- <http://purl.obolibrary.org/obo/mondo#exactMatch>
- <http://www.geneontology.org/formats/oboInOwl#hasExactSynonym>

<http://purl.uniprot.org/ensembl/ENSG00000136997>
<http://pubchem.ncbi.nlm.nih.gov/rest/rdf/gene/GID4609>
<http://identifiers.org/ncbigene/4609>

<https://omim.org/entry/190080>



原則として、同義関係にあるURIは、identifiers.org、purl.uniprot.org、pubchem.ncbi.nlm.nih.gov/rest/rdf/gene/のいずれかに繋がるようにし、それらの間の同義関係を基本的辞書とする。



データベースに聞きたい質問募集中！

自然言語でRDFポータルに問い合わせられるようになりつつある中で、実際にデータベースから取り出したい生命科学的な質問例をなるべく多く集めたいと思っています。

普段GoogleやChatGPTなどに投げている質問など、なんでも構いません。こんな質問に答えてくれると嬉しい、という事例をお寄せください。



<https://tinyurl.com/nlq4rdfportal>



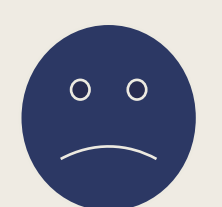
概念はURIで表現

RDFポータルを調査すると、遺伝子MYCを表現するURIが複数見つかる。これらを用いてSPARQLクエリを作る必要がある。

- <http://identifiers.org/ensembl/ENSG00000136997>
- <http://icgc.link/Gene/ENSG00000136997>
- <http://rdf.ebi.ac.uk/resource/ensembl/ENSG00000136997>
- <http://bio2rdf.org/ensembl:ENSG00000136997>
- <http://rdf.ebi.ac.uk/resource/ensembl/ArrayExpress/ENSG00000136997>
- <http://purl.uniprot.org/ensembl/ENSG00000136997>
- http://omabrowser.org/ontology/oma#GENE_ENSG00000136997
- https://bgee.org/bgee15_0/gene/ENSG00000136997

同義URIを繋げる作業は大変

```
BIND(URI( REPLACE( str( ?_uri ),
  "http://omabrowser.org/ontology/oma#GENE_",
  "http://identifiers.org/ensembl/" ) ) as ?_ya _uri )
```



更に、プレフィックスの違いだけでなく、識別子の表記体系が異なる場合もあり、単純な文字列置換だけでは適切に処理できない場合もある。

解決案：MCPとTogoMCP

MCP (Model Context Protocol) は、LLMなどのプログラムが外部知識にアクセスする方法を定めた通信規格で、この規格に準拠したリクエストに受け答えできるプログラムをMCPサーバーと呼ぶ。

具体的なSPARQLクエリなどの生成はLLMが行うが、何をどのように問い合わせるのが望ましいかという文脈(コンテキスト)をMCPサーバーが提供する。

LLMが自然言語で受けた質問を解析し、RDFポータルに適切なクエリを発行できるようにTogoMCPは必要な情報を提供する。