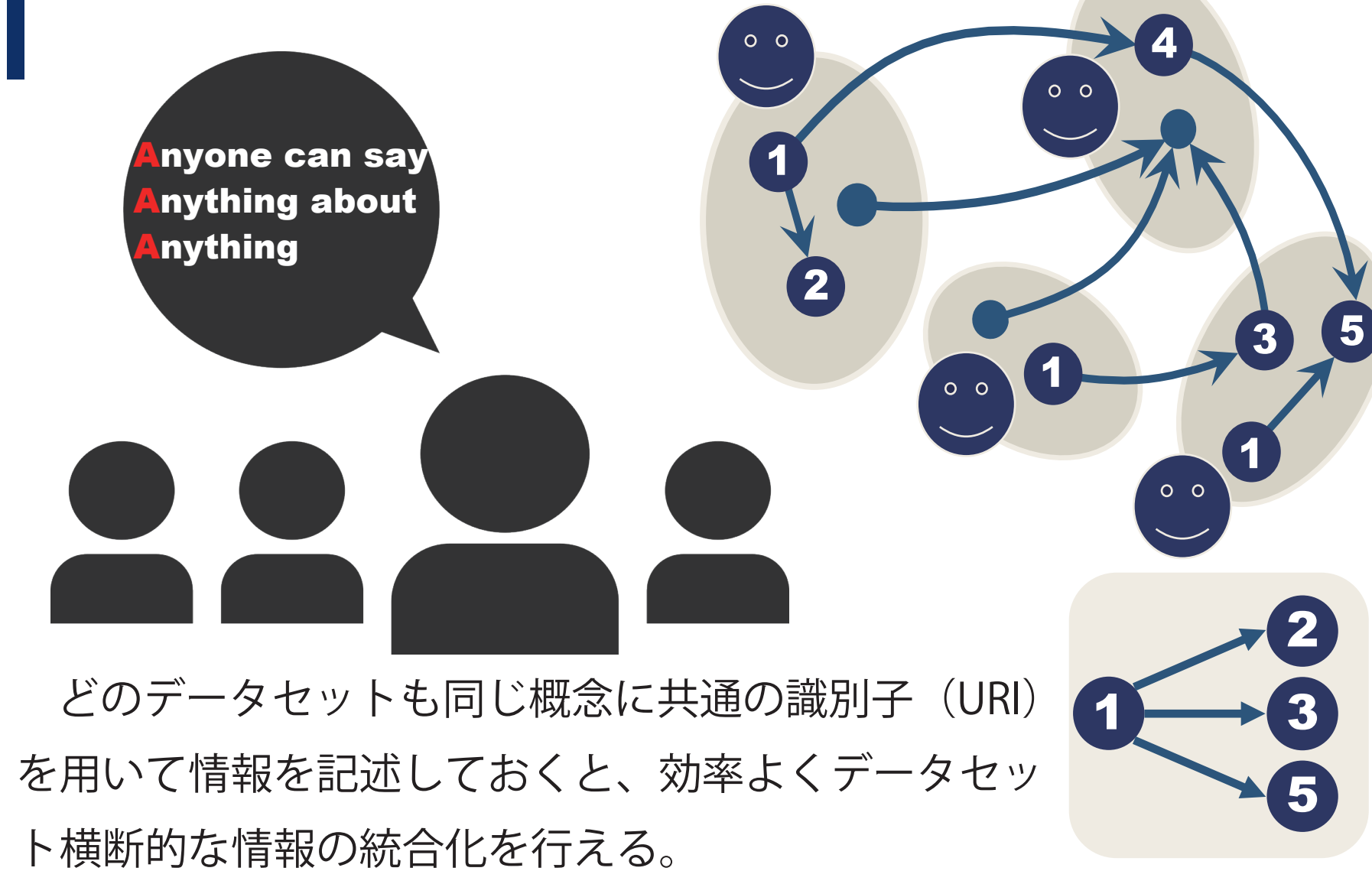


RDFポータルに含まれる同義URIデータベースの構築

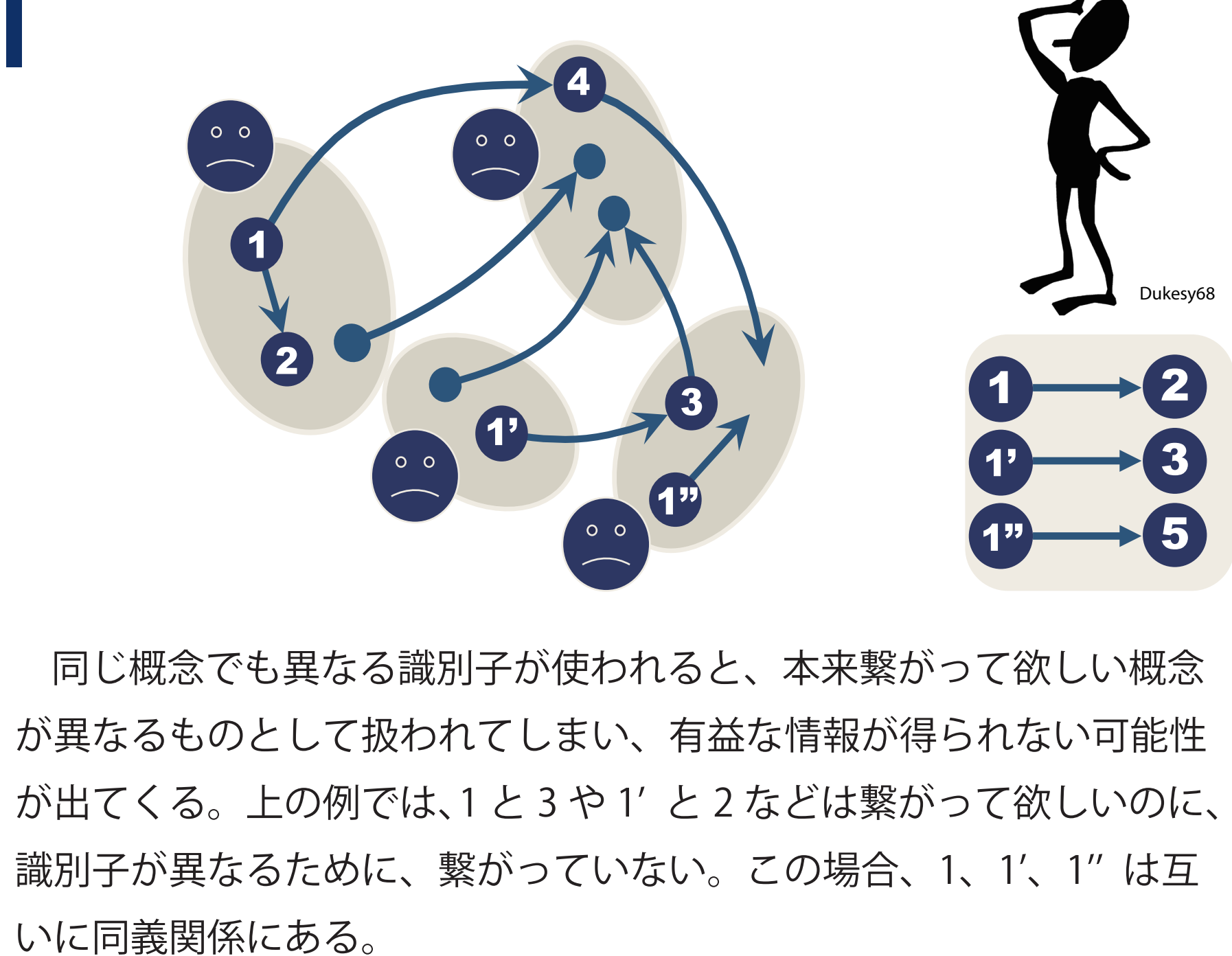
ライフサイエンス統合データベースセンター (DBCLS)
山本 泰智

RDFポータルには現在、生命科学分野のRDFデータセットをオントロジーを含めて61収載している。そのトリプル数は合計で1240億を超えている。そこには同じ概念を表現しながら異なるURIで識別されている事例も多く、従って、RDFデータの利点であるURIを介したデータ統合が妨げられている。この同義URI問題に対しては、Identifiers.orgやTogolD、Biomappingsといった取り組みがなされている一方で、RDFポータルに収載するRDFデータへの同義URIデータベースとしてはデータや機能が十分ではない点がある。そこで、これらのデータを活用しつつ、RDFポータルにおける同義URIデータベースの構築を進めている。

知識グラフ



実際には



実際の例

RDFポータルを調査すると、以下のような互いに同義関係にあるURIが見つかる。

- <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

繋げるための作業は大変

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。

FAIR原則*

Findable (見つけられる)、
Accessible (アクセスできる)、
Interoperable (相互運用できる)、
Reusable (再利用できる)

これらの原則は大切な一方で、上記の同義URIに関する課題については触れていない。

[*] <https://biosciencedbc.jp/about-us/report/fair-principle/>

Linked Data原則



- すべてのモノやコトをURIを用いて識別する
- httpで始まるURIでインターネット経由のアクセスを可能に
- RDFなどの標準仕様を用いて有益な情報を返すように
- 関連する外部データのURIを含めて情報を芋づる式に取得可能に

--- Tim Berners-Lee

(<http://www.w3.org/DesignIssues/LinkedData.html>)

FAIR原則との関連性が高く、上記のLinked Data原則に加えて、データがオープンであるとする事項を加えた、Linked Open DataはFAIRとほぼ同等と言える。

より多くのデータベースが、より多くの関連データを参照することで、繋がりのあるデータベース全体の利用価値が高まる。

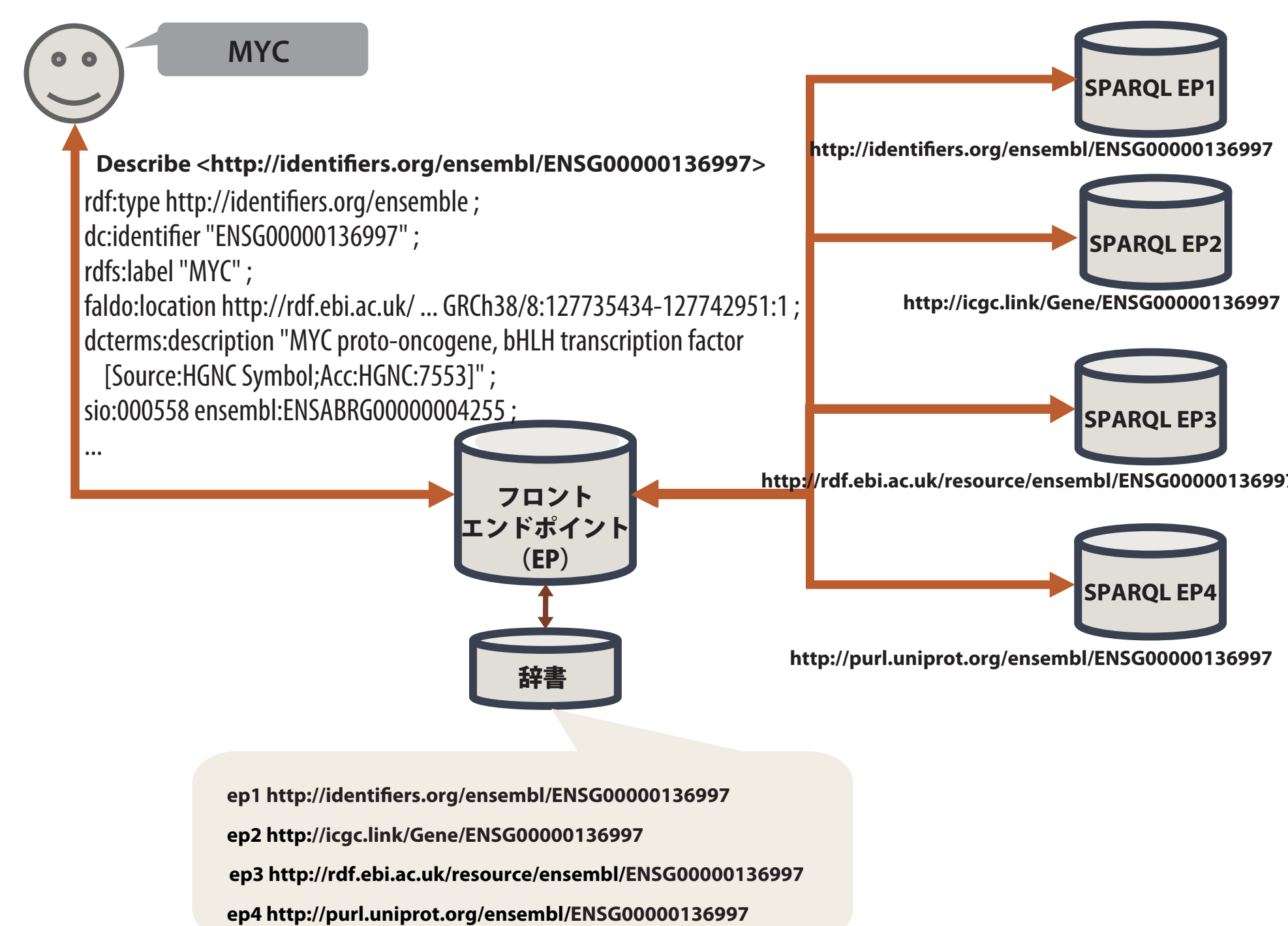
RDFの場合は、URIを使いまわすこと



同一概念であれば同じURIを利用すること

対処方針

同義URI辞書を用いて関連情報を漏らさず取得

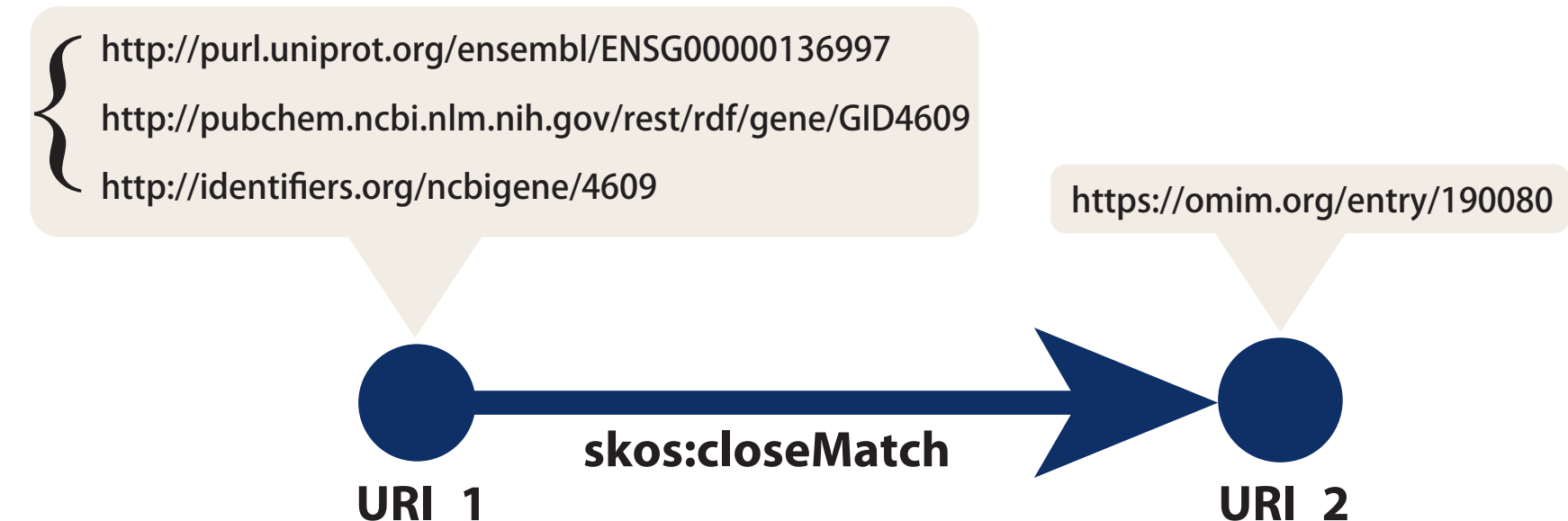


それぞれのデータセットに合わせて同義のURIで関連情報を取得できるように、同義関係にあるURIの辞書を用意する。

同義URIの収集

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、データセット毎に辞書を用意する。

- <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>



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な辞書とする予定。

結果

得られたトリプル数は以下の表の通り (RDFポータルのトリプルストア単位、頻度昇順、2024年8月末時点)。今後、内容を確認して、同義関係の有無やidentifiers.orgなど、上記の名前空間へのリンクを構築する。

pubchem	5,918,318
pdb	63,410,584
ncbi	125,005,528
ebi	167,695,465
primary	179,221,126
kero	600,406,435
ddbj	1,319,657,122
sib	3,092,157,419
Total	5,553,471,997

第一列に書かれている項目名は各トリプルストアに収載している主なデータセットの名称。RDFポータル全体のデータのサブセットで、同義関係の候補トリプルだけに絞ることで、一つのトリプルストアに全トリプルが収載可能であり検索も高効率。このデータを対象として同義URI関係の調査や辞書構築に用いる。必要に応じてクラス情報などは連合検索を用いて取得する。

今後の予定

得られた関係から実際の同義関係を確認した後に、ワークフロー化など、辞書更新作業の簡略化の可能性を確認する。例えば、遺伝子IDである場合は、同義URI候補のクラスURIを取得し、当該URIやラベルにgeneという文字が含まれていることを確認するなどである。

また、identifiers.orgやBiomappings*などの外部資源も参考にして妥当性を判断するほか、上記の6述語以外の述語を用いて同義関係を示すトリプルの有無を調査する。

更には下記の図のように、DBCLSで提供している、様々な生命科学で用いられるID間の関係を収載するTogolDサービスを利用して、RDFポータルのデータが生命学的に意味のある繋がりをより多く持つようにしていく。

[*] <https://github.com/biopragnatics/biomappings>

