# 3. TogoID: データベース統合の基盤となるID変換サービス

池田秀也
ライフサイエンス統合データベースセンター




# ID変換の必要性

---

バイオインフォマティクスで様々なDBを活用するにはデータベースID間のリンクが重要

- 等価なものに付けられたID間の変換
  - 例: NCBI Gene ID ⇔ Ensembl ID
  - 使いたい解析ツールが、手元のIDを受け付けてくれない場合など
- 関連する情報の取得
  - バリアント→遺伝子
  - 遺伝子→トランスクリプト
  - トランスクリプト→タンパク質
  - タンパク質→立体構造
  - 立体構造→相互作用
  - 相互作用→化合物・医薬品
  - 化合物・医薬品→パスウェイ
  - パスウェイ→疾患

# 既存のID変換サービス

---

データベースID間のリンク情報を提供する既存のサービスの例

- 国内: LinkDB (ゲノムネット), Biodb.jp
- 海外: BioMart, UniProt ID mapping, Ensembl, Bio2RDF

既存のリンク情報の課題

- 対象としているデータベースのカバレッジが限られる
- 各データベースの毎年・毎月・毎日など更新への追従
- 対話的に操作するUIと、プログラムから自動化して利用するAPIの両方が欲しい

類似のものとしてIDの転送サービスがあるが

- ページを開いてみるまで転送先は不明: PURL, Identifiers.org
    - OK: 転送ルールだけ記述しておけばよいので維持管理は容易
    - NG: 事前に転送先のIDを知っておくには、データとして維持管理しておく必要がある

# TogoIDで実現したいこと

---

データベースのカバレッジを確保
- ライフサイエンス統合データベースセンターにおけるデータ統合のハブとして
- LINCなど生命医科学ドメインのニーズに応じて対応データベースを拡張する

ウェブ上で対話的に操作して変換し結果をダウンロード
- 始点となるIDから探索的に接続先のデータベースをたどる
- 始点および終点となるデータベースを指定して経路を探索する

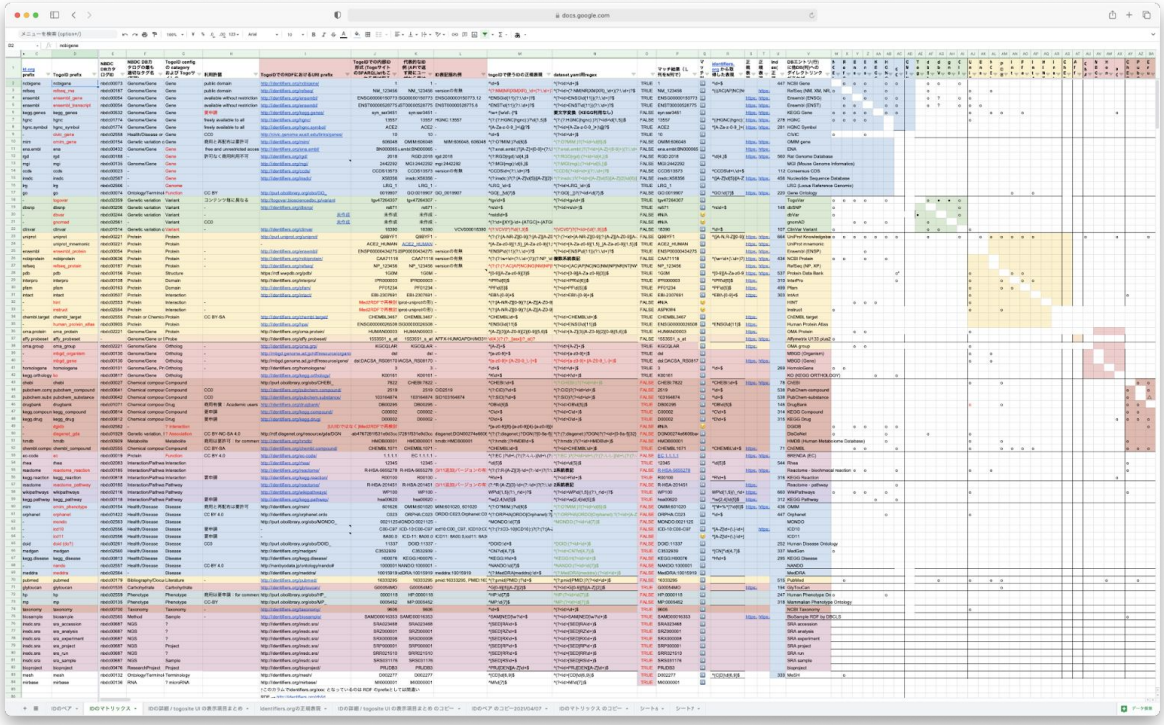
プログラムによる自動処理を実現
- 上記と同じ機能をウェブサービス(API)としても提供

安定的なサービスの提供と定期的な更新
- サービスをクラウドで提供することでダウンタイムを解消
- データベース毎の更新頻度に合わせたアップデートを自動化 (TogoID-config)

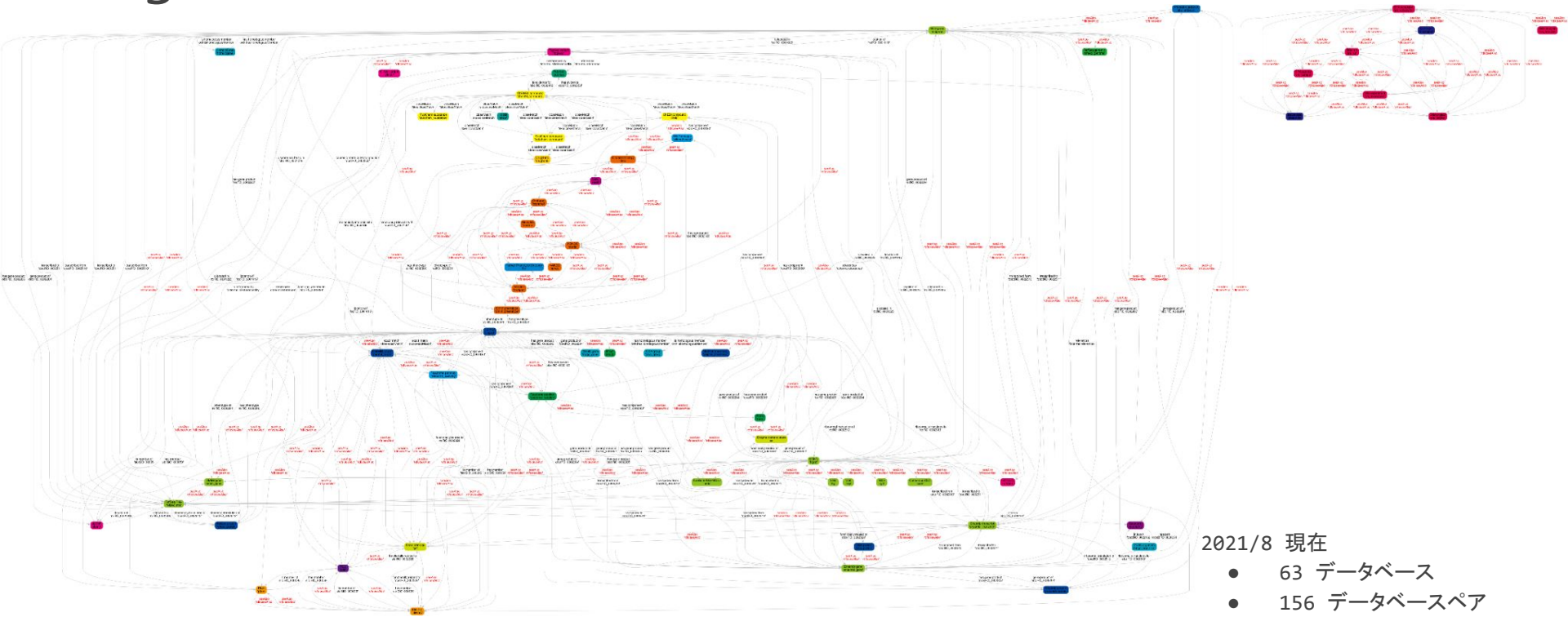
# TogoIDの対象データベース選定

---

- 遺伝子・タンパク質・化合物・パスウェイ・疾患など、対象とするDBをリストアップ
- NBDCデータベースカタログとの対応付け
- データ取得元・データ形式・更新頻度・ライセンスなどを調査
- 各DBからID変換可能なDBを調査
- ID体系（正規表現パターンなど）の調査



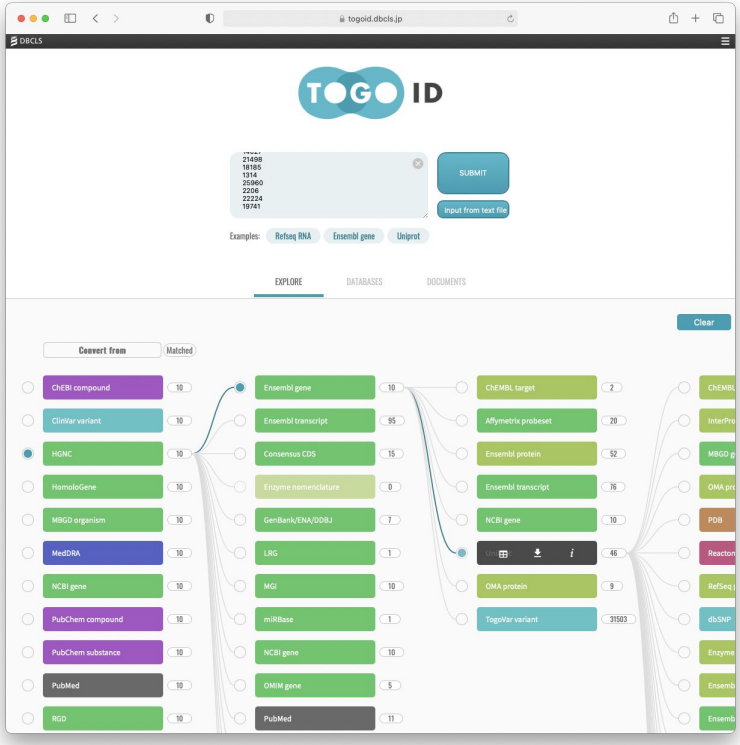
# TogoIDによるリンク情報の集約と管理



2021/8 現在

- 63 データベース
- 156 データベースペア

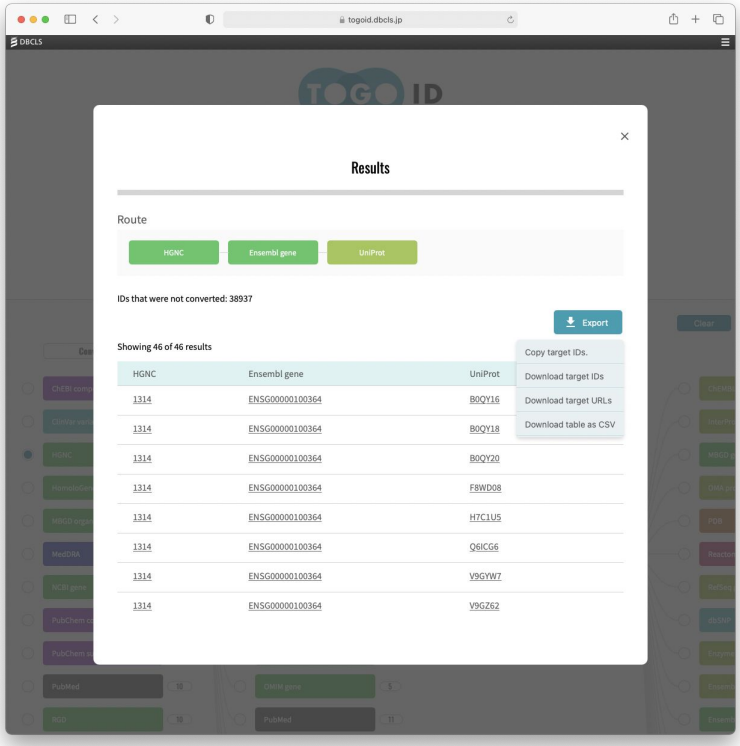
# TogoIDのウェブインターフェイス




https://togoid.dbcls.jp/

- IDリストを入力（もしくはファイルアップロード）
- 自動判定されるDBを確認して選択
- 変換先のDBを選択
- 必要なら数ステップ先の変換先DBまで選択

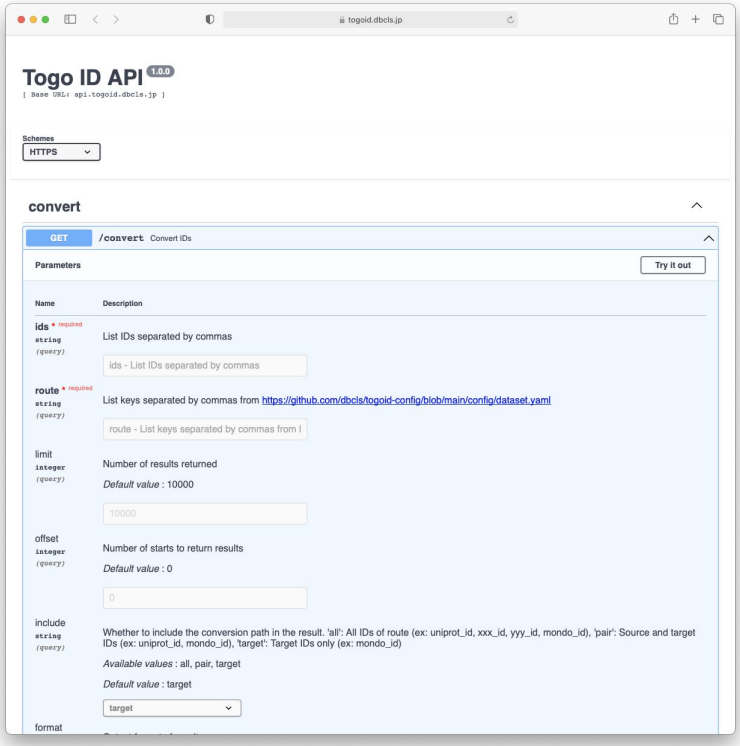
# TogoIDのウェブインターフェイス




https://togoid.dbcls.jp/

- IDリストを入力（もしくはファイルアップロード）
- 自動判定されるDBを確認して選択
- 変換先のDBを選択
- 必要なら数ステップ先の変換先DBまで選択
- プレビューして変換状況を確認
- 問題なければ変換表をダウンロード

# TogoIDのAPIによるプログラムからの自動変換処理

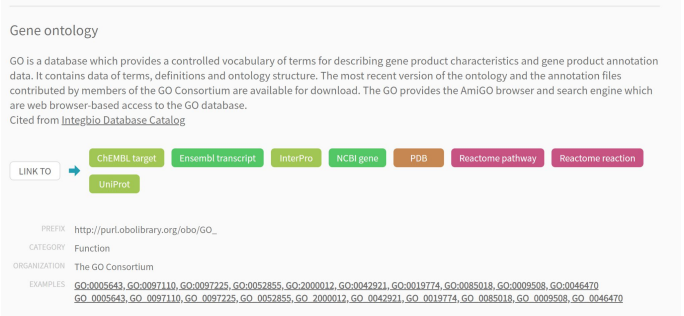



https://api.togoid.dbcls.jp/convert

- ?ids=5460,6657,9314,4609
  - 変換元のIDリストをカンマ区切りで渡す
- &route=ncbigene,ensembl_gene
  - 変換ルートをカンマ区切りのデータベース名で渡す
- &format=json
  - 取得するデータ形式を指定 (csv, tsv, json)
- &include=target
  - 変換先IDだけ (target)
  - 変換元IDと変換先ID (pair)
  - 中間の変換ルートすべてのID (all)
- &offset=0&limit=10000
  - 大量に取得する場合のオフセット・リミット値

# TogoIDを構築して分かった課題

---




元々のデータベースに内在する問題

- IDの表記ゆれが激しい
  - PDB: `1G0M`, `1g0m`
  - Gene ontology: `GO:0019907`, `GO_0019907`
  - Orphanet: `ORPHA:101078`, `ORDO:101078`, `Orphanet:101078`

  実用される表記法をなるべく拾うような正規表現で対応、DATABASESタブに例示

- 1つのDBに複数のID体系が混在（ゲノム, 遺伝子, トランスクリプト, タンパク質…）
  - Ensembl: `ENSG00000186283`, `ENST00000638000`, `ENSP00000365411`
  - RefSeq: `NG_004671`, `NM_001199636`, `NP_001171968`

  TogoIDでは名前空間を分けて管理

# TogoIDを構築して分かった課題

---

変換後のIDが発散する問題

- １対多、多対多（変換後にその先の変換を続けると対応数が爆発する）
  - 遺伝子→Gene ontologyによる分類→タンパク質
  - タンパク質→Pfamなどの機能ドメイン→立体構造

  生物種で絞り込む、などが考えられるが未対応

どこまでID変換でやるのか問題

- ID変換の意味（セントラルドグマ、相互作用ネットワーク、関連文献）
  - ウェブページのクリックで辿れるもの全部OKというわけではない…

リンクの意味を標準化する必要性（逆向きの変換も含め）

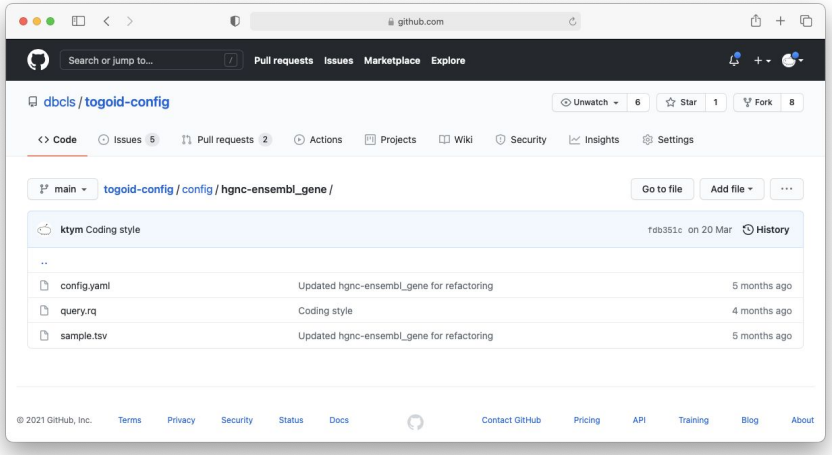
- セマンティック・ウェブ技術によるリンク関係のオントロジー整備 (TODO)
- 同じペアでも関係が同じとは限らない
  - 例) 糖鎖-タンパク質 (糖鎖を代謝する酵素？糖鎖で修飾されるタンパク質？)

# TogoIDデータ更新の自動化

変換元DBと変換先DBのペア毎に、IDの対応関係を抽出するプログラムを作成

TogoID-config
- https://github.com/dbcls/togoid-config

内容
- Rakefile 自動更新手順
- bin/　各種取得・変換スクリプト群
- config/ db1-db2ごとの変換規則群
  - dataset.yaml データベース一覧
  - db1-db2/config.yaml 更新手順
- input/　共通の前処理入力データ置き場
- output/ 生成される出力IDペア置き場
  - tsv/db1-db2.tsv タブ区切りファイル
  - ttl/db1-db2.ttl RDF版ファイル



# TogoIDの対象データベースの追加

---

変換元DBと変換先DBのペア毎に、IDの対応関係を抽出するプログラムを作成 ← TSVを生成

```
11099   ENSG00000071794
11114   ENSG00000126012
11115   ENSG00000012817
  :
```

TogoID-config

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

内容

- Rakefile 自動更新手順　　(← 前処理が必要なら追記する)
- bin/　各種取得・変換スクリプト群
- config/ db1-db2ごとの変換規則群
  - dataset.yaml データベース一覧　(← まだ載ってなければDBを追記する)
  - db1-db2/config.yaml 更新手順　← 上記プログラムの実行方法を記載する
- input/　共通の前処理入力データ置き場
- output/ 生成される出力IDペア置き場
  - tsv/db1-db2.tsv タブ区切りファイル
  - ttl/db1-db2.ttl RDF版ファイル

# ご意見お待ちしております！

---

- ウェブUIの機能面
  - ここが使いにくい/使いやすい
- 対象データベース
  - このIDを変換したい