BioData Science Initiative
大学共同利用機関法人 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
バイオデータ研究拠点(BSI)

トーゴーの日シンポジウム2025
マルチモーダルデータ × AI

DBCLSでのデータ統合とAIに向けた取り組み

片山俊明 (ROIS・BSI/DBCLS)

2025-10-20

生命科学 × AIの根幹となるデータベース運用

生命科学の研究成果を集積し、適切にキュレーションすることで、研究者にもAIにも利用可能なデータを創出

生命科学研究

生命科学研究成果

- 論文発表
- データベース登録

生命科学的な問い

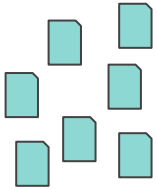
- 既存のファクト知識検索
- 元データにない知見
- データ統合による推論

データ利用申請

DB問合せ

論文・データ

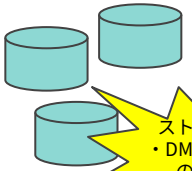
データ生産



データベース化

1次レポジトリ

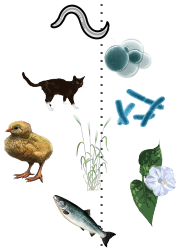
キュレーション



ストレージ・DMP・人材の欠乏

アノテーション

2次データベース



貴重な研究データの永続的な保全

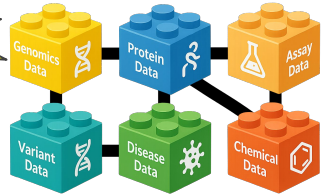
- 登録申請・利用申請業務
- 大規模ストレージ運用
- 解析環境の開発提供

データを分野別に規格化

- 専門家の整理で付加価値を創出
- 他の生物種での情報補完など

標準化・統合

2次利用に向けた専門家によるデータ整理



AIの性能は学習データの質に依存するため、「AIレディな高品質データベース」の継続的な整備

マルチモーダルなデータの標準化

- メタデータ
- データモデル・データ形式
- 標準語彙（オントロジー）
- リンク情報

基盤技術開発

- 高度なデータベース技術開発

研究利用・AI利用

データ利用の基盤を運用・開発中

科学データの取得・再利用

- 生データへのアクセス
- 統合データの機械学習
- 生物・AI由来データ選別

質問応答システム

- 自然言語問合せ
- QAタスクのデータ生成

AIエージェント

- MCPサーバ整備
- SPARQL自動生成

ナショナル DBに求められること

・国家予算で科学研究費を出す以上、得られた研究データを保全する恒久的なストレージと運用も国が責任を持つべき。

・膨大なデータは移動が困難なため計算環境がデータのそばに必要。

・生物のドメイン知識と情報技術をもつ専門家によるキュレーションの維持が重要。

マルチモーダルな生命科学データベースの種類

ナショナルセンターによるデータ統合の意義と、そもそも研究者によって「データベース」のイメージが異なる問題

ナショナル生命科学データベースセンター

データとスパコンは直結が必須 →

ゲノム・DNAデータ

NCBI・EBI

(>100PB)

全生命科学データの大部分を占める

欧米では100PB超のストレージを確保

日本の生命科学データを永続的に保全するDMPと提供申請・利用申請業務

ストレージと運用を維持するための国策が必要

DDBJ (25PB)

ヒトDB・JGA・AGD他は散在

ストレージ枯渇 中国等の台頭

マルチオミクスデータ

NBDC (1PB)

知識・文献データ

DBCLS (0.05PB)

この一部をAIが利用

マルチモーダル・画像データ

RIKEN・PDBj

AIの成果をDB化

医科学・医療データ

AlphaFold

実験由来データとAI由来データの識別

DBの管理機関や省庁が分散し、データが散在していることと、永続的な運用が課題。予算人員に比して対応分野は広い。

欧米はナショナルセンターで国家レベルのデータマネージメントを実施、主要なデータベースが集約され相乗効果で付加価値創出や、製薬などの産業応用にも繋がっている。
アジアでも、インドIBDC・中国CNCB・韓国KOBICなどナショナルセンターの設立が進む。

DDBJ+DBCLS
15億円・50人

DDBJ, DRA, JGA
BioProject, BioSample

AGD, CANNDs

GlyCosmos
Microbiome Datahub
JPOST
Mass
fanta.bio
RDF Portal
TOGO ID
PDBj
SSBD
TOGO VAR

NCBI
300億円・350人

GenBank, SRA, dbGaP
BioProject, BioSample

GEO
Taxonomy
PubChem
PubMed
ClinVar
MedGen

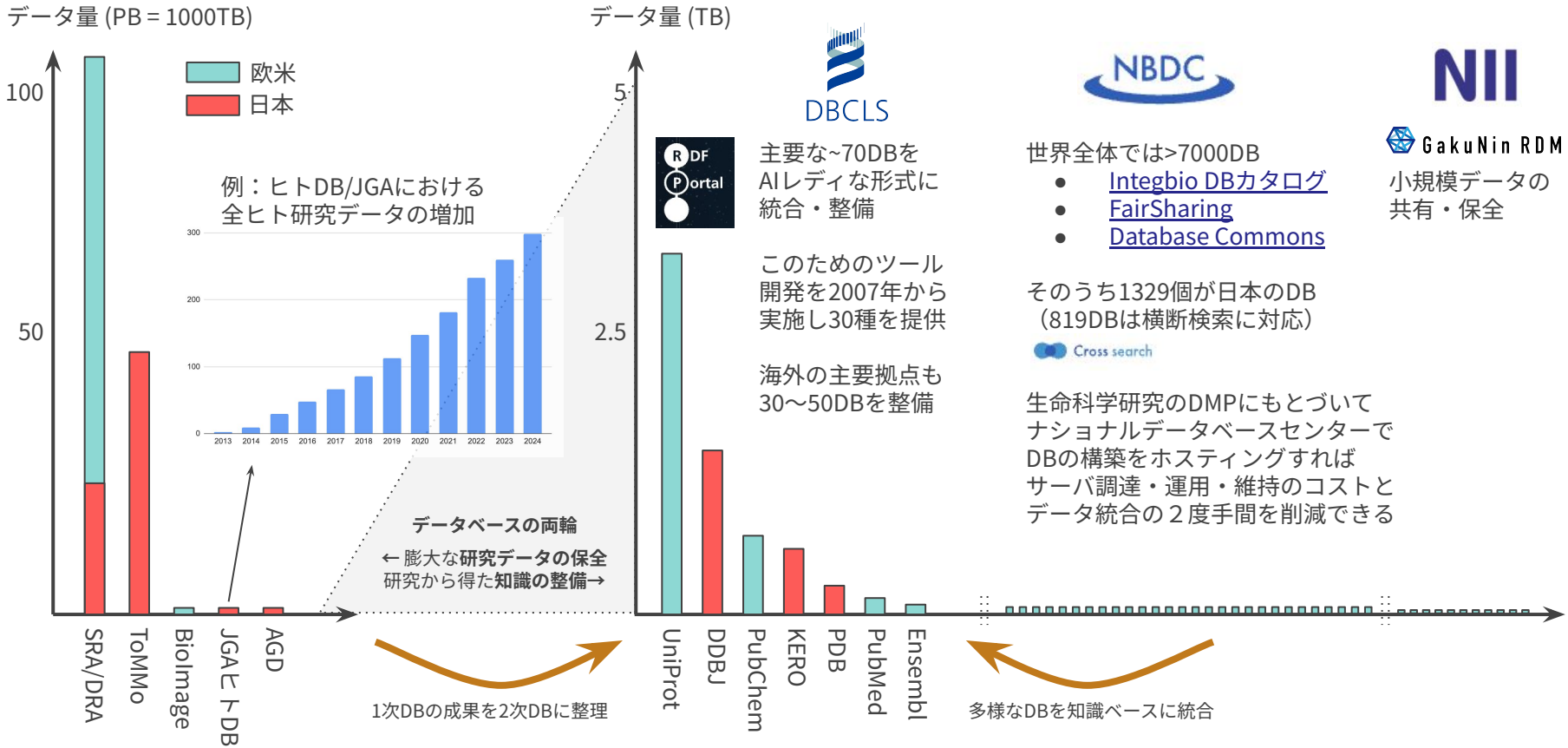
EMBL-EBI
180億円・650人

ENA, EGA
BioProject, BioSam

ArrayExpress
Ensembl
GWAS Catalog
MetaboLights
ChEMBL
ChEBI
Europe PMC
UniProt
Identifiers.org
reactome
PDBe
Protein Data Bank in Europe
BioImage
EVA

マルチモーダルな生命科学データベースの量

１次データ（実験の生データ）のレポジトリ運用と、２次データ（研究成果を整理した知識）の整備



１次データベース（配列や画像等）　**２次データベース**（研究者/AI用の知識）　**個別DB**（>7,000DB）　個別研究データ

マルチモーダルな生命科学データベースの統合

知識グラフによるデータ統合を技術開発と国際連携による標準化で実現

>2500 DBのカタログ化

- DBカタログ 2582DBs
- DB横断検索 819DBs
- DBアーカイブ 157DBs

DB間のリンク関係を整備

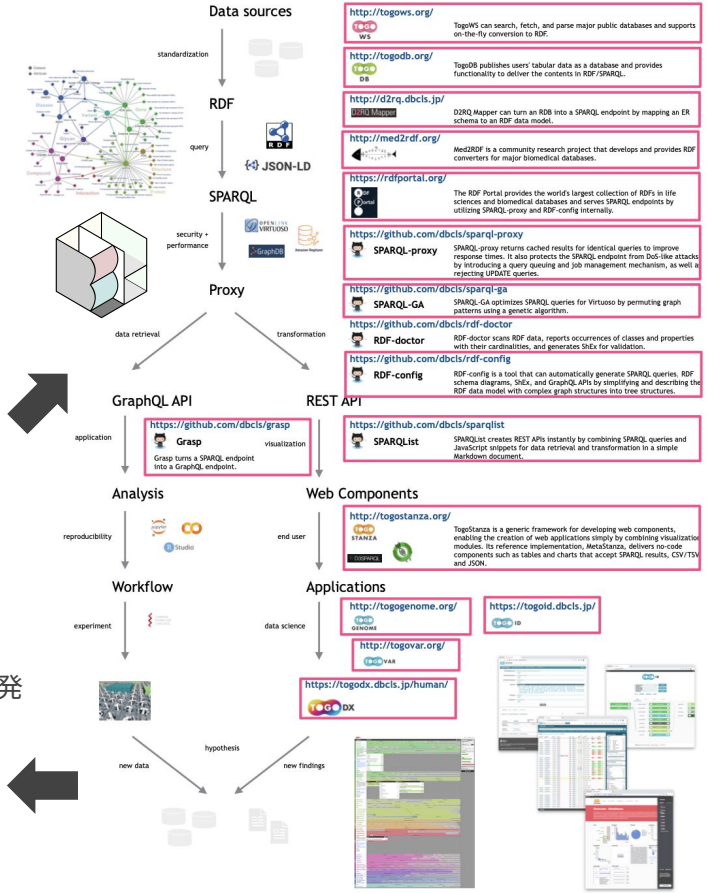
- TogoID 114DBs

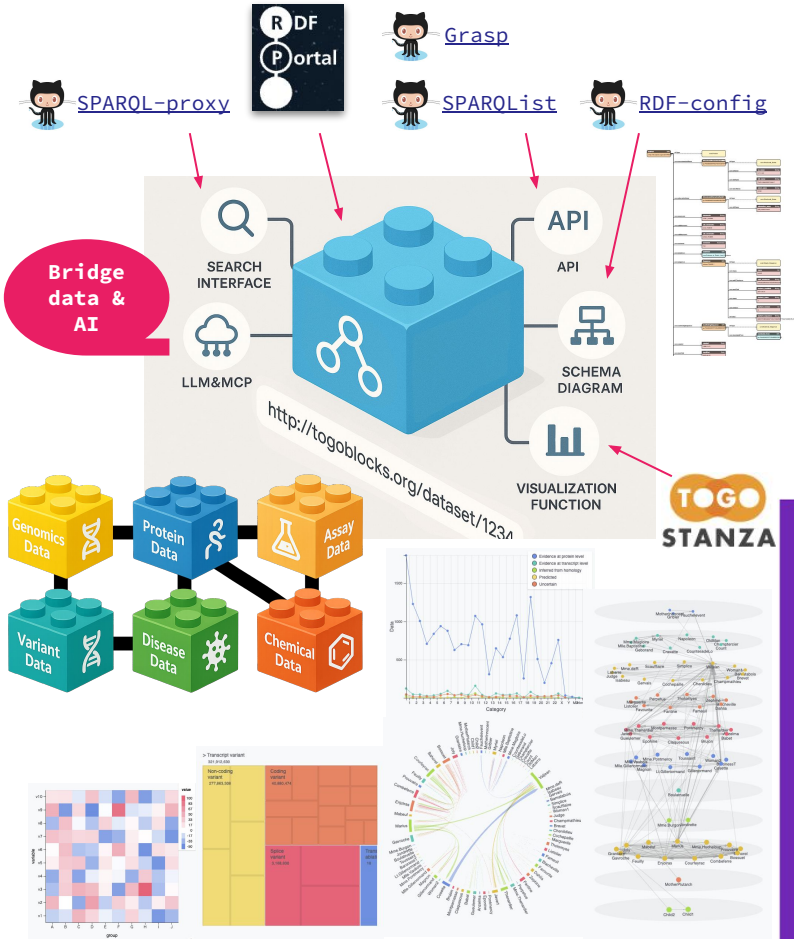
主要DBを知識グラフで標準化

- RDFポータル 70DBs
  + オントロジー開発

統合DBのアプリケーション開発

- TogoVar 約20DB統合
- TogoDX 約20DB統合
- Tabulae 統合DBの創薬応用





補足資料

マルチモーダルな生命科学データベースとAI

AIで用いる科学的知識データの提供と、データベースの品質向上や運用を効率化するAIの利用

1. 次世代AIを開発するための「学習・検証データ」となる情報基盤の　構築

- 正解となる典拠の提供と、良質な AIを育てるための構造化されたデータ整備
  → *AIは学習元の質に依存、高品質なデータがハルシネーション抑制の鍵*
- DBCLSの知識グラフは AI活用に最も適した形式で整備
- GBCやDICPなどの国内外の公的 DBを収載して拡張・カバレッジ強化

2. 登録が容易で、AIにも再利用しやすいメタデータ付きレポジトリの整備

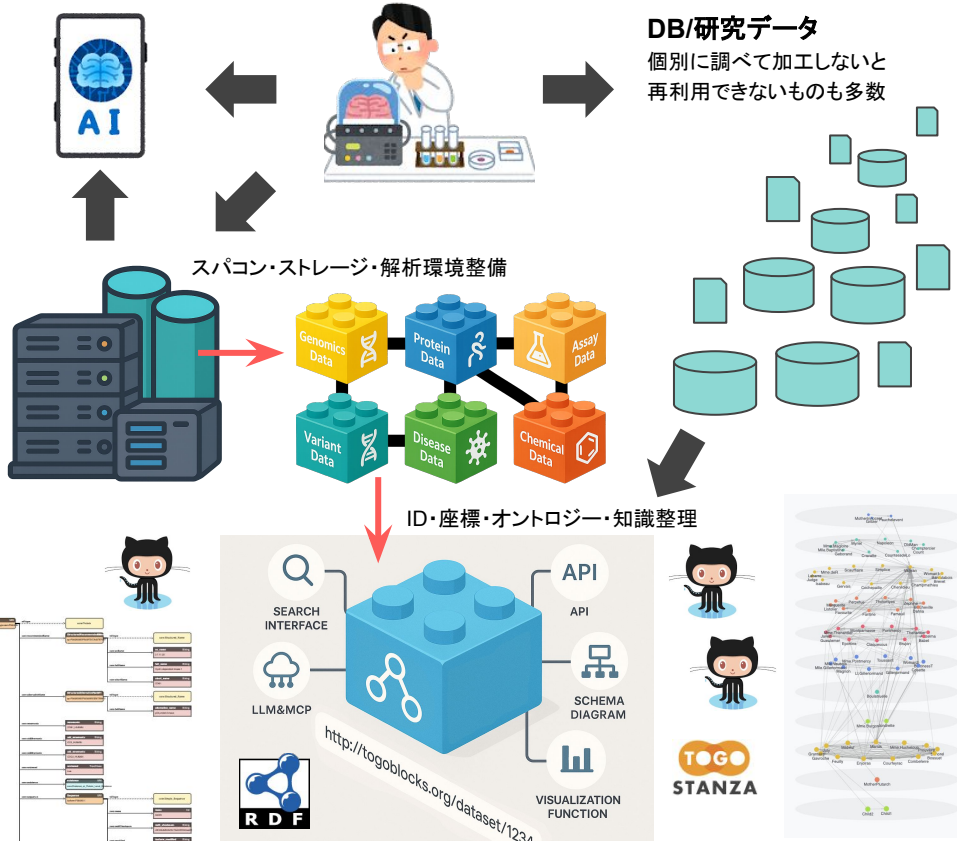
- 共通DMPの策定により、日本のすべての生命科学データを集約可能に
- データ登録をAI・UIで支援し入力負担を軽減、品質の高いメタデータを自動取得
- 集積データは永続的に管理され、他のデータと連結・検索・再利用可能に

3. 研究者が「使える」データベースの提供

- AIエージェントやLLM、LMM、MCPからのアクセスを想定した API・UI整備
- 情報科学の知識がなくても検索・再利用・発見ができる「ユニバーサルな設計」
- 産学連携に対応し、医療・創薬・基礎研究すべてのニーズに応える基盤

4. スパコンとストレージの拡充によるデータ利活用インフラの整備

- 大学共同利用機関法人で大学院生～研究者が公平に使える共用計算環境を提供
- ゲノムグラフによる構造多型解析や AI創薬など先端研究に直結
- 大規模データを散逸させずに利活用するための永続的なストレージ戦略を整備



知識グラフ(RDF)で統合された主要な生命医科学DB

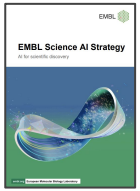
RDF: Resource Description Framework

- 塩基配列とアノテーション
  - INSDC (DDBJ/DBCLS)
- ゲノム情報
  - Ensembl (EBI)
  - RefSeq (TogoGenome)
- アミノ酸配列とアノテーション
  - UniProt (SIB)
- タンパク質立体構造
  - PDB (PDBj)
  - BMRB (PDBj)
  - FAMSBASE (Chuo U)
- 化合物
  - PubChem (NCBI)
  - ChEMBL (EBI)
  - Nikkaji (JST)
- 遺伝子発現
  - RefEx, GTEx (DBCLS)
  - ExpressionAtlas (EBI)
- サンプル
  - BioSamples (EBI/DDBJ)
  - JCM (RIKEN)

- 医科学 (Med2RDF)
  - TCGA, ICGC, COSMIC, CIViC
  - DGIdb, OpenTG-Gates
  - ClinVar, dbSNP, dbVar
  - ExAC, gnomAD
  - HiNT, INstruct
- 糖鎖
  - GlyTouCan, GlycoEpitope, WURCS, GGDonto, PAConto
- プロテオーム
  - jPOST
  - The Human Protein Atlas
- パスウェイ
  - Reactome (EBI)
- その他
  - PubMed/MeSH (NCBI)
  - BioModels (EBI)
  - MBGD (NIBB/DBCLS)
  - Quanto (DBCLS)
  - :

https://rdfportal.org/

国際的なデータベースとAIの戦略および科学データのサステナビリティへの提言



1. **EMBL Science AI Strategy**
欧州EMBLは生命科学の総本山ともいえる研究組織で、欧州バイオインフォマティクス研究所EBIもここに含まれる。2025年に公開された本レポートではEMBLが生命科学とAIにどのように取り組んでいくかについて、分析とビジョンが表明されている。ウェットな研究機関とドライのデータベース機関を擁するEMBLでは、AI時代において、AIの理論的な基盤構築、AIレディなデータ整備、自動実験ワークフローによる研究の加速3つの主要な柱と位置づけている。
https://www.embl.org/news/connections/ai-in-the-life-sciences-philanthropy-fuels-embls-strategy/

2. **How the National Library of Medicine should evolve in an era of artificial intelligence**
米国国立医学図書館NLMはアメリカ国立生物工学情報センターNCBIを擁する世界最大の生命科学データリソースの拠点で、本レポートでは科学知識の真正性保証、科学進展の観測、そして知識のわかりやすい社会還元の重要性を論点としてNLMがAI時代にどのように変化していくべきかの考察が与えられている。とくにAI時代において科学的データの整合性の保護者としてデータの真正性を保証する機関となること、多様な利用者向けに科学情報を翻訳、変換、要約するハブとなることなどが謳われている。
https://pmc.ncbi.nlm.nih.gov/articles/PMC12012362/

3. **Exploring Determinants of Longevity of Biomedical Databases**
Joseph Finkelstein et al., Joseph Stud Health Technol Inform, 6:290:135-139 (2022)
2009〜2016年に論文発表された518データベースについて調査したところ2020年にはその35%にあたるデータベースがアクセスできなくなっていた。
https://pubmed.ncbi.nlm.nih.gov/35672986/

4. **Longevity of Biological Databases**
Teresa K Attwood et al., EMBnet.journal, e803 (2015)
この論文では2015年までに発表された326データベースが18年間でどれくらい存続しているかを調査し、その60%が消失、14%はアーカイブされ更新がない状態であることを報告している。一方で研究機関からのサポートがあるデータベースは存続しているものが多いことを指摘し、現代の生命科学研究に必須なデータベースには長期的な戦略による予算とインフラの維持が求められることを訴えている。
https://journal.embnet.org/index.php/embnetjournal/article/view/803/1209

5. **On the lifetime of bioinformatics web services**
Fabian Kern et al., Nucleic Acids Research, 48(22):12523-12533 (2020)
2010年から2020年までに論文発表された生命情報科学のウェブツールは2010年のツールの50%から2019年と2020年のものの90%しか利用可能な状態を保っていなかった2週間以上にわたって利用不可能になっていた2019年と2020年のツール47個のうち、著者に連絡を取ることで約半数は再び利用可能になったが、長期的な維持のためには十分な支援が必要と考えられる。
https://academic.oup.com/nar/article/48/22/12523/6018434

6. **Biological databases in the age of generative artificial intelligence**
Mihai Pop et al., Bioinformatics Advances, 5(1):vbaf044 (2025)
生命科学研究は公共データベースに依存しており、誤りの導入や伝播が研究の妨げとなる。生成AIにより大量の誤情報が拡散するリスクも高まっている。これを防ぐため、研究者教育の強化、誤りやデータ来歴に関する研究推進、データベース維持への資金支援が重要とされる。
https://academic.oup.com/bioinformaticsadvances/article/5/1/vbaf044/8088229

7. **Global Biodata Coalition Highlights 2024**
Global Biodata Coalition（GBC）は、生命科学研究に不可欠な国際的データベースやバイオデータ資源の持続可能な運営を支援・調整するために設立された国際組織で、各国の研究資金機関が連携し、データ基盤の長期的な維持と国際協力を推進している。DDBJも含まれる52のGBCメンバーからなるGCBRフォーラムのレポートで下記のような記載がある（2ページ）:
Forum members highlighted key challenges, including ... concerns over AI-generated data potentially compromising trust in curated knowledgebase
フォーラム参加者は、主要な課題としてAI生成データがキュレーション済み知識ベースの信頼性を損なう可能性への懸念などを挙げた。
https://globalbiodata.org/wp-content/uploads/gbc-highlights-brochure-2024.pdf