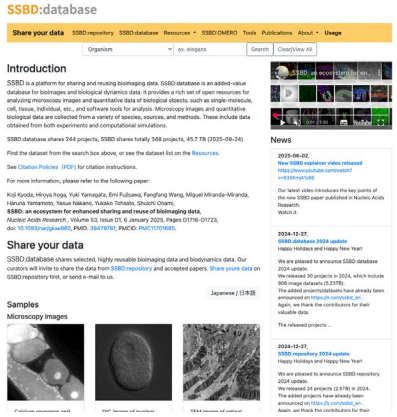
基礎・横断研究戦略 作業部会（第2回）	資料2-2
令和8年2月20日	

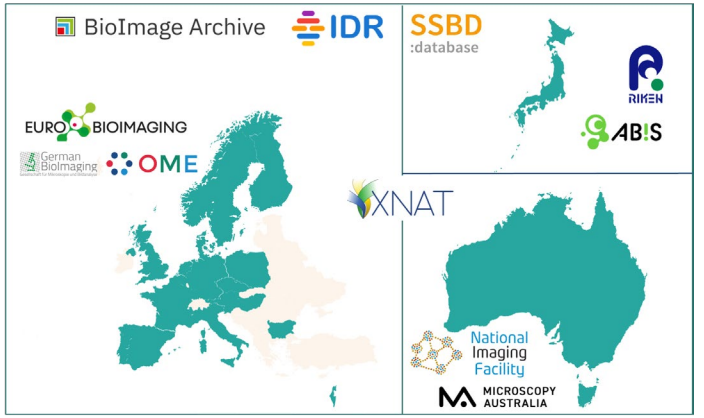
AI for Life Science基盤としての統合データベースの在り方について

大浪 修一
理化学研究所
生命機能科学研究センター チームディレクター
統合データ・計算科学プログラム プロジェクトディレクター
情報統合本部 部門長代理

SSBD

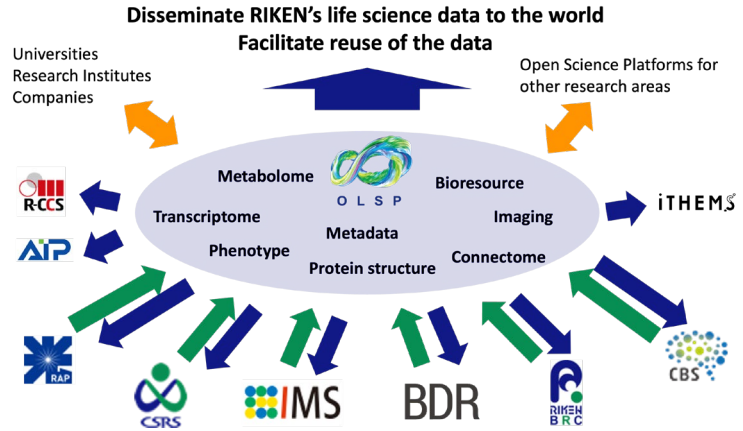
- バイオイメージングデータの共有のための統合データベース共レポジトリの構築（2013年～）
- JST統合化推進プログラム、CREST、科研費等から支援
- グローバルなバイオイメージデータ共有システムの構築（Horizon Europe）
  - 欧州の公共レポジトリと統合（BioImage Archive、Image Data Resource）





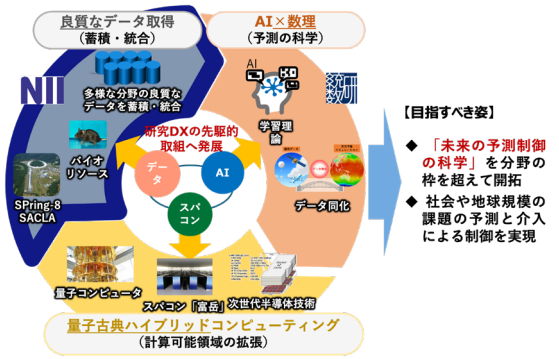
理研オープンライフサイエンスプラットフォーム (OLSP)

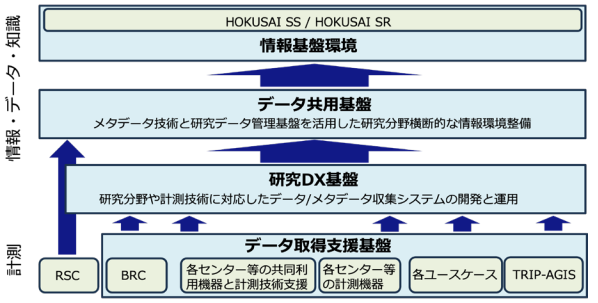
- 2017～準備研究、2019～本格研究
- 理研の最先端の生命科学分野のデータを統合し、世界と共有
- 理研の全ての生命科学系研究センターの参加の下で、生命科学分野のオープンサイエンスを推進



理研TRIP研究DXプロジェクト

- 理研の世界トップレベル研究から創出される最先端かつ多様な研究データを「良質なデータ」として最大量取得、蓄積（2023～）
- 理研の全ての研究センター等と連携し、研究DXに寄与する技術開発や技術導入、技術運用、技術支援等を実施し、予測科学を開拓





AI for Life Scienceの発展ためにデータベースは必要なのか？

- AIはデータを食べることにより賢くなる
- AIを賢くする/活用するためにはデータベースは必要

AI for Life Scienceの発展のために必要なデータベースとは?

- 現在のAIはよく整備された公開データをほぼ食べつくしている
  - よく整備されていない公開データはあまり食べていない
- 今後必要となるデータは
  - 既存の公開データをAIが食べやすいように整備したデータ
  - AIが食べやすいように整備されて提供された新しいデータ
- 今後必要となるデータベースは
  - 整備されたデータをAIが食べやすいように提供するデータベース

よく整備されたデータとは?

- 正確なメタデータが豊富に付与されたデータ
- メタデータをAIが正確に理解できるデータ

メタデータ：データがどのようなデータであるかを説明するデータ

メタデータが無ければデータは電子的な情報の塊にすぎない

このようなデータをAI-readyデータと呼んだりする

正確なメタデータが豊富に付与されたデータ

- 二つの立場
  - 標準化・構造化されたメタデータ
  - 自然言語で自由に記述
- データベースでメタデータをつけてきた経験
  - 論文等の自然言語から正確なメタデータを抽出することは困難
    - データ産生者の確認を要する
    - 自然言語で書いておけばOKという考えには単純には賛成できない
  - データ産生者に標準化・構造化されたメタデータを豊富に付与することを求めるのは困難
    - 作業が膨大
    - 標準化・構造化されたメタデータを人力で集めるという考えには単純には賛成できない
- メタデータの収集方法を根本的に変える必要がある
  - メタデータはデータ生産の場で取得する
    - 電子ラボノートブック
    - ラボデータ管理システム
    - 実験の録画・録音
    - 実験自動化

メタデータをAIが正確に理解できるデータ

- AIにメタデータの保存場所や保存形式を理解させる必要
  - 標準化されたインターフェース
    - MCP
  - 既存のAPI
- データベースを運用してきた経験
  - MCPの標準の成熟度が低い
- 国際的な議論・開発を続ける必要がある

AI for Life Science基盤として重要なこと

スピード感

- AIをとりまく状況は従来の科学のスピード感とは段違いに速い
- 理化学研究所科学技術基盤モデル開発プログラム（AGIS）
  - 2024年4月開始
  - 2023年の構想時とはAI for Scienceの中軸となる技術コンセプトが変わってきている。
    - 基盤モデル→AIエージェント→世界モデル
  - 技術コンセプトの変化のスピードに対応した柔軟なプロジェクト管理
- Bioimage indexプロジェクト（Biohub）
  - 日米欧のバイオイメージングデータのレポジトリを統合するプロジェクト
    - Biohub、理研、EMBL-EBI、Allen Institute、OpenRxiv
  - １ヶ月半で日米欧のDBが同意
    - 2026年1月8日　メールでコンタクト
    - 2026年1月13日 オンラインミーティング
    - 2026年1月27日 オンラインミーティング
    - 2026年2月4日　オンラインミーティング
    - 2026年2月10-12日　ミーティング@Biohub in SF
  - 2年程度のプロジェクト期間
  - MVP (Minimum Viable Product)
- AI分野のスピード感に対応する必要がある

戦略性

- 新しい様式のデータを取得する方法が加速的に開発されている
  - 新規DBの開発、DBの改廃を戦略的に行う必要がある
  - BioImage Archive
    - EMBL-EBIがバイオイメージングデータの将来的な重要性を予測して戦略的に開設（2019年）
    - 開設後に現在のプロジェクトリーダーのMathew Hartleyをリクルート
  - Biohub
    - 動画とオミクスの同時計測データの収集を開始
      - Bioimage Indexの会議の後で問い合わせを受ける
      - 具体的なファンディングの提案
  - 日本として戦略的にDBを開発していく必要
- データ量は増加、資金は有限
  - 研究資金の効率的な運用が求められる
  - 理化学研究所情報統合本部
    - 計算資源の経済的実務的に最適な配置
      - 中央計算機とローカル計算機の役割分担
  - EMBL-EBI
    - Federation modelを推進（PRIDE、BioImage Archive）
    - 全てのデータを欧州のDBでホストするのは経済的に不可能という考え
  - 効率化により得られた資金を開発に投資できる

データ生産現場との連携

- データ生産現場とDBの間に行動原理のギャップがある
  - データ生産現場はDBのために良質なメタデータを付与する動機が希少
  - DBはデータ生産現場に良質なメタデータの付与を要求
- 欧米では本問題はいまのところ放置されている
  - 欧州
    - EMBL-EBI　DBの研究施設、データ生産は行わない
  - 米国
    - Allen Institute　データ生産の研究施設、公共レポジトリは運営していない
- 理化学研究所統合データ・計算科学プログラム（CoRe）
  - データ生産～研究データ管理～公共リポジトリでのデータ公開のデータの流れをDXで一体化
    - 電子実験ノート、データ管理システム、試料情報GUI
    - プロテオミクスから実装開始、イメージング、ゲノム解析に展開中
- 日本が強みを発揮できる領域
  - データ生産、DBの主要ファンディング元が一つ

データ活用現場との連携

- データ活用現場
  - AI、データ科学者、計算生物学者
  - 実験生物学者
- AI、データ科学者、計算生物学者をDBの第一のターゲットとする傾向
  - 実験生物学者はAIを通じて情報を取得するようになるという未来像が主流
  - 1次データベースの重要性
- DBとデータ活用現場との間でのデータの活用の議論は十分でない
  - DBの開発会議にデータ活用現場の人材が入っていない
- 日本が強みを発揮できる領域
  - DBとデータ活用現場が近い
  - データ生産、DBの主要ファンディング元が一つ

AI for Life Science基盤として重要なこと

- スピード感
- 戦略性
- データ生産現場との連携
- データ活用現場との連携

これらを実現できる体制がのぞましい

AI for Science基盤としての統合データベース

- National Databaseとして基軸となる1次データベースを運営
  - DDBJ, PDBj, jPOST, SSBD, DBCLS等を統合的・戦略的に運営する組織を構築
    - 現場の遺伝研、大阪大、京都大、理研の体制はそのままで、全体をとりまとめるバーチャルな組織を構築
    - 大規模計算資源の統合などによる運営の効率化を個々のDB運用に支障をきたさない速度で行う
  - データ生産プロジェクト、バイオリソースプロジェクトと密接な情報連携
  - AI for Life Scienceプロジェクト、生命科学研究コミュニティと密接な連携
  - 我が国を代表して外国のデータベースとの交渉に臨む
  - 期待される効果
    - 各データモダリティ、AI for Scienceの動向変化にスピード感を持って対応
    - 計測技術開発、分野の動向を捉えて新規データベース/リポジトリを戦略的に開発
    - マルチモダル計測データへのスピード感を持った戦略的な対応
    - マルチモダルデータ活用技術の効率的な開発
    - 各データベースモダリティの技術、知識の共有
    - 国際的発信力・影響力の強化
    - 研究資金の効率的運用とそれによる開発力の強化
    - 人材育成、人材活用の強化
    - AI for Life Scienceのためのデータ産出・管理・共有・活用を戦略的に実施
    - 実験研究者との連携を強化