

社会を支える生命医科学統合データ基盤の確立

① ビジョンの概要

ゲノム解析技術が成熟し、あらゆる生体・環境からゲノムデータを安価に得られるようになった。ゲノム医学やバイオ・遺伝子バンクが普及し、ヒトゲノムのみならず地球生命全体の多様性を網羅・共有できる。この超情報時代において科学は知識探究から責任ある情報管理・共有へと重心を移す。人間とAIの協働を前提に、政府主導で、倫理性・再利用性・公共性を基軸とした「データ基盤」の設計と運用が求められる。

② ビジョンの内容

人新世とよばれる時代に入り、生命医科学を取り巻く環境は急速に進歩している。さらに近年は、生成AIによる自然言語処理・画像解析・統合分析の能力が加わり、これまで以上に多様で大量のデータを即時に解釈・再利用可能になっている。これらのデータは社会の共有財として保存され、倫理的・法的・社会的課題(ELSI: Ethical, Legal and Social Issues の略)を考慮しつつ再利用されるべきである。未来の科学者は、こうした高度データを駆使し、地球という有限資源を持続可能なかたちで活用するための科学的基盤を構築する必要に迫られる。

日本が科学先進国として持続的な発展を遂げるためには、(1) ゲノムや医療・環境情報を適切に保存・管理するデータベース、(2) それら膨大なデータをリアルタイムに解析・運用可能とする高速・大容量ネットワークとスーパーコンピュータ、そして(3) AIと協働しながらデータを活用し、政策や産業や教育に還元する多様な人材の育成の3点が不可欠である。これらデータベース・スパコン・人材を一体的に捉えた「統合データ基盤」の設計と実装こそが、日本が世界の知的基盤構築をリードする鍵となる。国を挙げてその制度設計に取り組み、国際的なガバナンス基準とも整合して展開することが求められる。このような基盤は市民の理解と信頼の上に成り立つため、データの利活用に対する国民的合意形成や教育活動も不可欠である。

③ 学術研究構想の名称

社会を支える生命医科学統合データ基盤の確立

④ 学術研究構想の概要

本提案が目指す「統合データ基盤」には、には、データベース、スパコン、人材に加え、生成AIインフラの4要素が不可欠である。これらは分離して存在するのではなく、有機的に結合される。そして国際的なガバナンスや倫理的・法的・社会的課題(ELSI)を考慮した制度設計と運用が求められる。

データベース: 中核組織として既存機関を統合・連携させた国立バイオデータベースセンター(DBセンター)を設置する。収集対象の選定や標準化、メタデータ整備、用語統一などに積極的に関与し、研究現場との協働で高品質なデータを長期的に確保する。特にAIと親和性の高い語彙整備やオントロジー構築が鍵となる。

スパコンおよびAIインフラ: DBセンターと不可分の関係にあるのが、データ処理・AI学習・統合解析のための高性能計算基盤である。クラウドと複数拠点の計算資源を連携させた分散型計算環境を構築する。生成AIの学習・推論には専用ノード(GPU/TPU等)が不可欠であり、今後の科学における中核的役割を果たす。

人材育成・運用体制: DBセンターと教育機関が連携し、キュレーター、AI技術者、コンサルタントなど多様な人材を育てる体系的な研修教育プログラムを整備する。AI活用に必要な検証・説明責任も教育内容に含む。

倫理的・法的・社会的課題(ELSI): 個人情報やセンシティブデータを多数扱う「統合データ基盤」においては、ELSIへの対応が制度・技術の両面で必須である。DBセンター内に専門部門を設置し、ゲノム・医療・環境データに伴うプライバシー・同意管理・法制度整合の諸課題に対応する体制を整備する。AIが生成した成果物の責任の所在や再利用条件など、著作権を超えた新たな法制度への準拠も検討すべきである。



図1 国立バイオ統合データ基盤の設置

⑤ 学術的な意義

科学の究極的な目標が人類のより良い未来の構築にあるならば、ヒトという生命体の動作原理を理解することは極めて重要である。もはや理論やモデルだけでは科学の前進は困難になり、極めて多様かつ高次元なデータを、時空間的に大量かつ高解像度で収集・蓄積し、それをもとに AI を駆使して解析・仮説生成・検証を繰り返す、データ駆動型科学への転換が求められている。

とりわけ、ヒトゲノムや医療・環境データは個人と社会に深く関わるため、収集・解析・再利用には倫理的・法的・社会的配慮（ELSI）が不可欠である。「知識のフロンティア開拓」という従来科学の価値観では不十分であり、個々の研究者が自らの研究時間という制約の中で「何を知るべきか」「どこまで知るべきか」という哲学的課題にも答える必要がある（例えば生殖医療）。ELSI の適切な制度設計は、学術への信頼と受容を保証する基盤であり、AI の学習データ整備や出力の検証可能性にも影響する。

また、本構想が対象とする生命情報はヒトに限らず、あらゆる地球生命を含む。日本は島嶼国として豊かな生物多様性を有するうえ、自国の資源を体系的に記述し、管理できる技術と人材も備えている。しかし、制度設計やスケーラビリティ、分野間の連携不足により、オープンサイエンスや学術 AI の本格的活用は道半ばである。本構想は、分野融合と ELSI を統合した科学モデルの実装であり、生成 AI 時代における日本の学術基盤の再構築と世界貢献の柱となる。

⑥ 国内外の研究動向と当該構想の位置付け

国内ではプロジェクト毎にデータベース戦略やデータ共有政策が異なり、一貫した国の方針が見えづらい。国外には米国の NCBI、欧州連合の EBI、韓国 KOBIC のように国立と呼べる中核データセンターが政府直轄でヒトゲノムを含む生命科学データの集約を実施する。EU では、AI 統合解析や倫理対応を含む分散型情報利活用基盤である欧州健康医療データスペース（EHDS）の整備が進み、「次世代データエコシステム」へ移行する。

⑦ 社会的価値

本構想が目指す統合データ基盤は、単なる研究効率の向上を超え、社会全体に恩恵をもたらす持続的インフラである。本構想は UNESCO の「オープンサイエンス勧告」を具現化し国際的にも先進的である。ヒトゲノム情報のような個人識別符号まで対象に含む本構想では、ELSI 対応を組み込んだ組織的運用が不可欠である。同意取得・ライセンス整備・様式の統一などを統合的に進め経済的・医療的インパクトの最大化を図る。

⑧ 実施計画等について

<実施計画・スケジュール>

第1～2年度：準備組織の設置と中核拠点の公募・制度設計

準備組織は、国内におけるトップレベル拠点を含む研究機関と省庁や研究資金配分機関（AMED、JST 等）が参画し、データ共有政策、予算制度、評価制度、ELSI 基本方針、生成 AI の整備方針等について協議する。中核となるデータ資源とその維持管理機関を公募形式で選定する。

第3～4年度以降：データ基盤の本格構築・運用開始

準備組織を DB センター執行部に位置づけ、データ整備・解析・再利用までが統合された科学情報基盤の構築を開始する。この段階で、ユーザ目線での柔軟なデータ利活用環境を確立する。

<実施機関と実施体制>

ナショナルバイオリソースプロジェクト（NBRP）や日本生物多様性情報イニシアチブ（JBIF）と理研・情報システム研究機構が連携し「AI for Science」の基盤として機能する。中核となる DB センターは、政府が「ナショナル」の名を冠して設置し、その組織は公募により選定する。とりわけクロスアポイントメントによる人的エフォート確保、文理融合人材の交流・共同研究制度、AI 基盤モデル活用による、企業・市民向けの情報提供サービス、恒久財源に基づく運用と意思決定会議体の設置、ELSI 専任部門の設置を実現する。

<総経費> 328 億円（初期2年の準備に4億円、本格稼働の8年間に40億円／年）

主な内訳：中核センター運営・維持費用、クロスアポイントによる人件費、データベース整備・キュレーション業務、生成 AI モデルの構築・運用、分散スパコン間ネットワーク運用、トレーニング・ELSI 部門運営

⑨ 連絡先

有田 正規（情報・システム研究機構 国立遺伝学研究所）