科学技術振興調整費　「科学技術連携施策群の効果的・効率的な推進」
平成 17～19 年度実施「生命科学データベース統合に関する調査研究」成果の概要

研究代表者　国立遺伝学研究所　教授　大久保　公策

１）研究目的

大規模解析系を用いたプロジェクト型研究では、その成果はデータベースとして保存される。また多様化細分化を重ねたライフサイエンスの知的集積物も膨大化の一途を辿っている。しかし、それらは必ずしも実用化研究等に利用し易いものではないため、利用者が使いやすい、一元的に集約・統合したデータベースの構築が望まれている。この調査研究では、生命科学データベース統合に向けて、まず、国内のデータベースを調査し、統合化する際の技術的課題をフィージビリティ・スタディによる実現性検証結果を提示した上で、それらの成果を生かした制度設計やロードマップの作成に資する試案の提示を行うことを研究目的とする。

２）研究成果の概要　―ミッションステートメントの達成状況―

２．１）ミッション①　関係府省における DB 統合化に向けた取り組みの補完となる関係府省の制度設計やロードマップ作成に資する試案の提示

データの流通に関する制度等は、データベースの統合に関し大きな影響を与えていると考えられた。そのため、ライフサイエンスデータベースの整備が進んでいる国外の制度等の調査にも重点を置いたが、それらのルールのよき点は見習いながら、我が国の実情に合わせた柔軟なルール作りを今後検討する必要があると考えられる。また、学術情報の流通に関して広く、国民、研究者の理解を求めていくことも、今後、統合データベースの計画を推進するためには必要なことであるので、データベースの統合施策を進めていくにあたり、理解推進に向けた具体的な方策も盛り込んでいく必要がある。

２．２）ミッション②　複数の国内主要 DB を統合化する際の技術的課題並びに解決策の提示及びフィージビリティ・スタディを通じた実現性検証結果の提示

別々のライフサイエンスDBを一纏めにする場合には、それらの枠組みの不一致、同じ特徴を記載する場合の要素の呼び名（ID系列）や使われる知識表現（オントロジー）の不一致などが問題である。これらの問題に関して、i)解剖用語の自動分類機による全植物ESTの統合整理、ii)複数のDBにおける遺伝子ID系列

の自動更新等のフィージビリティ・スタディを行った。その結果、これらの問題に関して、情報工学分野で開発された技術が利用でき、解決可能であることを確認した。

解剖用語の自動分類機による全植物ESTの統合整理に関しては、植物ではいくつかのモデル植物以外公的なデータ整理が進んでいない、特にモデル生物情報の育種への応用が未着、という問題があった。また、ヒトをターゲットとした基礎生物学分野から見ると、植物は分類方法によってはカビ・シダなども一緒になっているので、種ばかり多くて整理に手をつけにくい、といった問題意識があった。そこで、植物データを非植物研究者や育種関係者が理解できるレベルにすることを目標に、比較的イネ麦などに近縁の植物（被子植物）と裸子、シダ、カビの分離、裸子植物に関して横断整理に必要な材料の提供、植物解剖にも基づいた植物材料自動分類機の設計を行った。それにより、全ての植物ＥＳＴプロジェクトを統合整理することが可能であった。なお、本フィージビリティ・スタディの成果は、文部科学省統合データベースプロジェクトで植物ボディーマップとして事業化すると同時に、農林水産省データベース統合プロジェクトのイネゲノムアノテーション更新時に発現データとして利用可能とした

また、複数の DB における遺伝子 ID 系列の自動更新に関しては、遺伝子や蛋白質に関する複数の ID 系列の参照表は DB の統合利用において欠かせないがこの更新は管理者にとって負担となる。比較的安定した欧米の ID 系列（NCBI の ID や SwissProt の ID）を選び、別の ID 系列を用いている DB(例えば H-inv）の外部参照情報をロボットで定期的にアップデートする仕組みと対応関係を管理するサーバを構築することで常にそれらの DB の ID の対照表を維持できることを示した。

具体的には、プロジェクト型分子網羅的スクリーニングによるデータの共有データベース GeMDBJ（国立がんセンター）と情報系研究者による知識集約型の二次的データベース H-InvDB（産業技術総合研究所）の試験的な統合を試みデータベースの統合の際の問題点を抽出することを目的とした。統合方法としては、H-InvDB の研究者と試験的統合・連携の基本コンセプトについて協議を行い、GeMDBJ からは、H-InvDB の強力なアノテーション機能、知識探索機能とを連携させるようにした。その結果、GeMDBJ が搭載しているゲノム・遺伝子多型データや、発現解析データの個別の検索結果から、H-InvDB が提供している該当遺伝子のゲノム上の位置、機能に関する生物学的知見や疾患関連情報等のアノテーション情報を参照できるように、データレベルでの連携を実現した。しかし、本統合の問題点として、どちらかのデータベースがアップデートした際、連携（リンク）が切れ、データ表示が出来なくなる問題が生じた。その対策として、“自動リンク管理システム”を開発しリンク切れの問題を解決した。

従って、技術的観点からライフサイエンスDBの物理的統合は十分遂行可能であると考えられる。しかし、その際には、統合しようとするDBの要素の内容を熟知し、統合過程の情報処理の両方を熟知した専門家が参加するか、もしくは双方の専門家同士の密接な議論が必要である。

２.３）ミッション③　国内外の医学分野・学術分野 DB、国内の産業分野 DB に関する技術的側面、制度的側面からの基礎調査結果の提示を行う

ＤＢの増加は1990年から急速に生じており我が国でも250を数える。 DBの増加は我が国においてもライフサイエンスがデータ依存を強めていることを物語っている。我が国で公開されているデータベース群は構築方により６群に分類できる。

<u>データベースの６分類</u>

①データバンク型（全ＤＢの5%）：
不特定多数の研究者の登録によるデータとドキュメンテーションからなるＤＢ。

②プロジェクト型(31%)：
特定の研究の方向性を持つプロジェクト型研究で作成された、大型のＤＢ。

③プログラム型（28%）：
バイオインフォマティストによる①②の再加工品。生データにバイオインフォマティックス研究者の開発した一定の処理をした結果を載せたＤＢ。　類似目的を研究者間で競うために専門家以外には区別の付かないＤＢも多い反面、特定の用途の標準となっているＤＢも見られる。

④キュレーション型（12%）：
典型的なドメインの視点を導入してキュレーターによって①-③を総合したデータを解釈レベルに高めたデータベース。

⑤知識モデル型（9%）：
測定データではなく特定分野の標準的知識（解釈）の要素の対応関係や要素同士の上下関係、木構造などによる形式的な表現データ。代謝ネットワークやシグナル伝達ネットワーク、遺伝子機能のオントロジーや解剖のオントロジーなどである。

⑥総説型（ 2%）：
特定分野の標準的知識（解釈）を読めば分かる程度の表などの形式にまとめ

たもの。

我が国では国際バンク事業への協力を欧米に次いではじめたことやプロジェクト型、知識モデル型のデータベースの先駆けを1990年台半ばまでに行うなどデータベース研究時代をリードして進めてきた歴史がある。これらのことから、我が国においても、データベース統合について技術的な課題は解決することが可能な状態であると考えられた。

1990年代から生命科学では理学工学の粋を集めてミクロな詳細さでマクロに対象を観察することを可能にする分析機器が登場し、生命現象を定量的かつ定性的に観察した巨大観察データが生まれた。これまでの実験データと異なり巨大観察データはデータ生産者以外の多種の研究に役立ち、企図しない研究を生み出す新しいカテゴリーのデータである。そのため、本報告書では「実験データ」と区別し「基盤データ」と呼ぶこととする。

我が国では2000年前後から科学技術基本計画に沿った大型の政府研究開発事業が一斉に開始され、生命科学の分野において、それらは、高精度大容量分析機器による「基盤データ」を産出するプロジェクトであった(**図1**)。

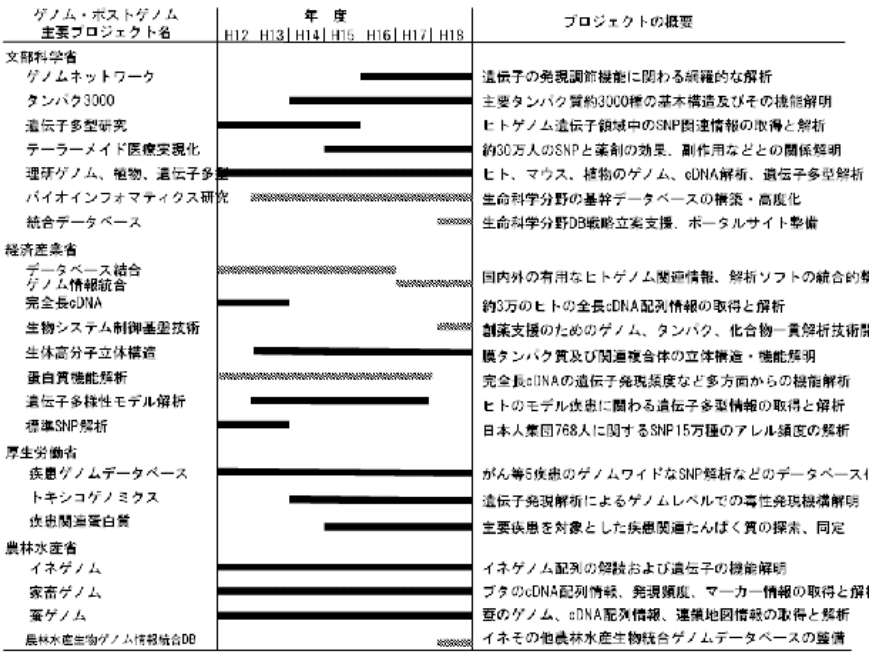

図１　現行および終了間もないライフサイエンスの政府科学技術研究事業。 黒ぬりは「基盤データ」生産が目標に含まれる事業

公的資金による科学的研究開発で研究者から生まれたアイデアに知的財産権（工業所有権や著作権）を認めることは研究の動機付けになると同時にアイデアの円滑な流通利用も促す。この科学社会と市場経済の連携は科学技術基本法の精神の一つの柱であり、産学連携、技術移転、知的財産保護をキーワードにTLO制度(H10(1998)) 日本版バイドール法(H11(1999)) 知財基本法(H14(2002)) 大学独法化(H16（2004）)と大学市場連携の為の制度改革が続いた。米国で20世紀の最後の４半期に起こったこのような流れは我が国では少し遅れる形で過去１０年間に急速に導入された。（**図3**）

米国のデータベース統合のケース研究

ライフサイエンスデータベースの統合が進んでいる米国を例にとり調査を行った。米国では各地に存在していたデータベースがヒトゲノムプロジェクトの開始直前にNIHに新設されたNCBI（国立医学図書館バイオテクノロジーセンター）に全てのヒト生命科学データが統合され、いわゆるポストゲノム研究はNCBIでのデータ公開と連携して行われてきた。我が国のプロジェクト型やプログラム型のデータベースは例外なくこれらのどれかを利用することで統合サービスを提供している。NCBIはデータ統合と共有の為に創設された国立データセンターでありまさにその機能を果たしたといえる(**図2**)。

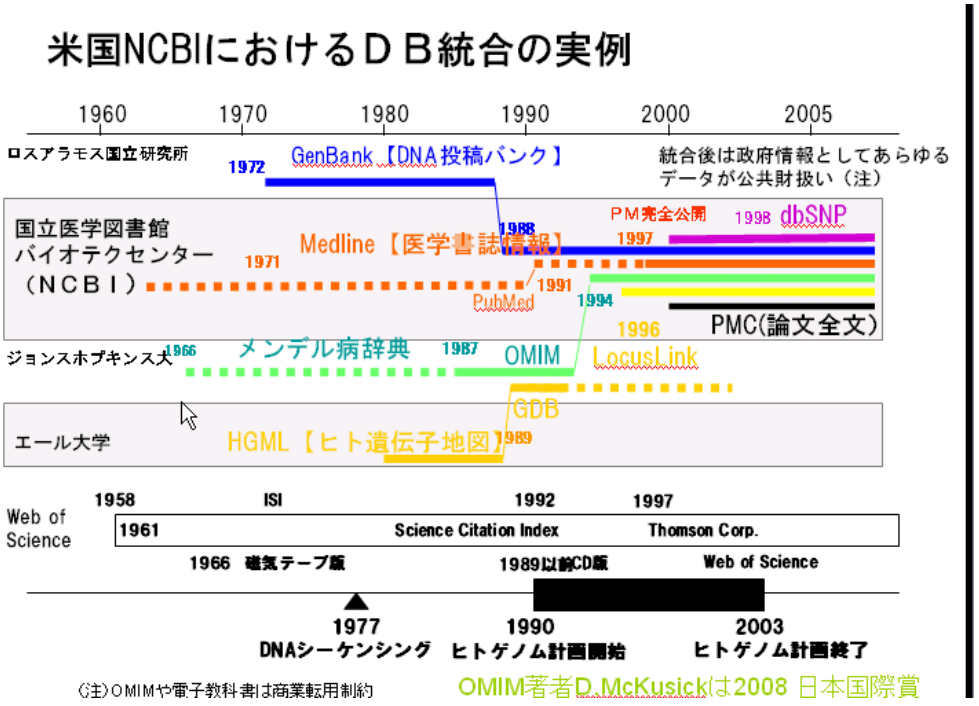

図２　米国 NCBI における DB 統合の実例

このような統合を可能にした背景には米国における政府情報およびその一部

である公的事業由来の科学技術データの厳格な公開制度があると考えられる。例えばMedlineもNIHの国立医学図書館(NLM)で100年以上続けられてきた医学文献の抄録作成事業であるが、電子化抄録が政府情報と看做されたためにインターネット経由で米国民にむけて無償で完全公開された。現在年間１億人以上に利用されているMedlineの公開事業をNCBIが担当したことがデータベース統合の成功を決定付けたとも言えるかもしれない。

それ以外にNCBIでのデータ統合と公開を可能にしたのは個々のプロジェクトのデータ公開ポリシーである。例えばヒトゲノムのバーミューダ会議やHapMap計画でのデータ公開ポリシーは極めて厳格なものであり、独占的利用が許されるのはデータ取得後１年のみであった。 これらのポリシーは国際協力プログラムで他国からのデータを速やかに集めることが解析力に自信のある米国を利するための政策的側面もあったと思われるが2003年には米国内プロジェクトに対しても同様に50万ドルを越えるグラント請求ではデータ共有の計画を明示することと堅持することがNIHポリシーで義務付けされた。

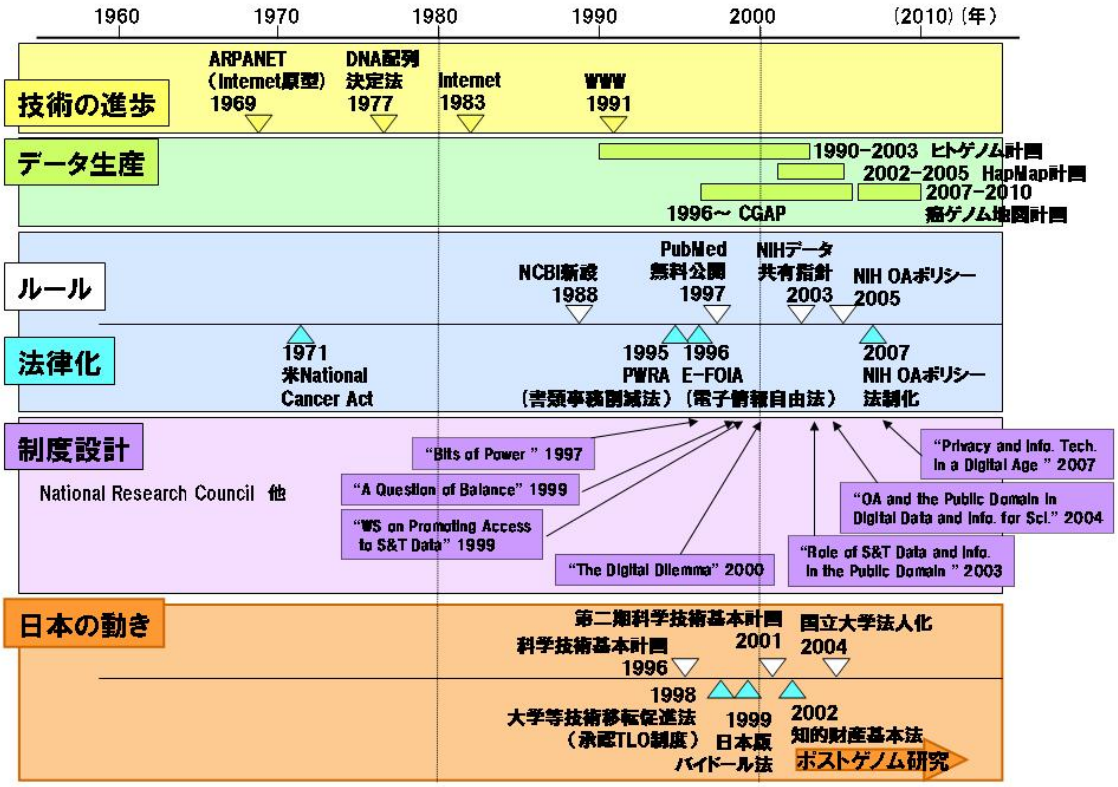

図３ 統合を可能にした制度整備(米国)

データの流れを作る米国の規則体系は明快である。政府科学事業ではデータ

は連邦調達規則(Federal Acquisition Regulation) に従い残らず一旦政府に納品され政府情報となる。一部の例外を除いて内部で作られたもしくは納品された政府情報は情報公開法 (Freedom of Information Access)により速やかに無制限に市民に提供される。しかも OMB-circular A-130(Office of Management and Budget) によって特に付加価値生産物が作成可能な原材料(raw content)の配布を強く奨励されている。情報企業や大学はユニークなアイデアと少ないコストでこの生データに対して様々な発見努力や付加価値競争（個人プロジェクト）を行い、結果として豊富な知財が生み出されエンドユーザーは多様な商品、情報、サービスから好みのものを適正な価格で手にいれることができる。国家情報由来の知財化は全く認められない一方公共データを用いた「個人プロジェクト」では政府研究助成金（NFS や NIH からの grant）を利用していてもバイドール法によって知財が保護されるために一層発見開発は活気を帯び科学の進歩とともに産業化が保障されている。すなわち、材料は政府が提供し、価値付加競争を知財化して保護することで活発化させている。

２　水素利用／燃料電池

・施策一覧

・俯瞰図

・本文

・ 補完的課題