科学技術連携施策群の効果的・効率的な推進

補完的課題　事後評価

「生命科学データベース統合に関する調査研究」

責任機関名：大学共同利用法人情報・システム研究機構

研究代表者名：大久保　公策

（大学共同利用法人情報・システム研究機構

国立遺伝学研究所生命情報 DDBJ 研究センター）

研究期間：平成17年度～平成19年度

目次

III. 研究成果
1. 研究成果の概要

(1) 研究目標と目標に対する結果
1) ミッション① 関係府省におけるデータベース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。

試案の検討に当たっては、外部有識者を含む研究運営委員会において議論を行ったが、現状認識のずれから問題意識を共有することに困難があり、十分に合意形成を行うことができなかった。そのため、本調査研究においては研究代表者として以下の試案を提示する。
また試案の重大性に鑑み、これまでの検討を踏まえ、研究代表者として、今後の制度設計に資するものとして留意すべき事項も付記する。

【試案】

1. 「政府資金によるデータ産生型プロジェクト(注1)のデータを我が国の研究社会で早期に共有するためのルール」の作成を検討すべきと考える
2. 公共財としてのデータを保全・管理し、長期にわたるデータの育成と共有を行う公的機関の設置を検討すべきと考える

1. 「政府資金によるデータ産生型プロジェクトのデータを我が国の研究社会で早期に共有するためのルール」の作成を検討すべきと考える

我が国での DB 統合の要請については、「米国立バイオテクノロジー情報センター(NCBI、注2)相当の機関が我が国にも必要である」ということがしばしば言われており、 理想的なデータベース統合の例として NCBI でのデータ統合をあげ、我が国の現状と比較することが多い。例えば、文部科学省ライフサイエンスデータベース整備戦略作業部会報告 2007[i]や NPO 医学図書館協会による「国立ライフサイエンス情報センター(仮称)」推進準備委員会最終報告 2005[ii]の中では、予算、人員等の相違が議論されてきた。
本調査では、科学情報一般に関わる「日米の決定的な違い」について「政府資金によるデータの共有利用を保障するルールの差」を最大の原因として指摘する。すなわち「NCBI が必要である」というこれまでの要請は「政府資金によるデータ産生型プロジェクトのデータを早期に国民で共有する為のルールが必要である」と翻訳されるべき内容である。
これまでも研究成果の公開には積極的な取組がなされているが、今後、一定の国家プロジェクトについては、その研究により得られたデータが合理的な期間内に共有されるための仕組みが構築される必要がある。
具体的には、政府資金によるデータ産生型のプロジェクトについては、研究計画の申請時において産生を企図するデータの種別と量につき明確にし、データ共有の為の計画を提案採択時の重要な判断基準とすることなどが考えられる。また臨床医学研究などのプライバシーに関わる研究においても、データ

の匿名化などによる非プライバシーデータ化とその共有のための仕組みを採択条件の一つとしてその質を競わせる等の工夫をするべきである。

注1）データ産生型とはデータ生産に始まりその解析による発見を企図する研究開発である。詳しくはp193)-(B)「データベース統合の背景」を参照。

注2）National Center for Biotechnology Information：米国国立衛生研究所（NIH）に属する国立医学図書館（NLM）の一部門。医学生物関連の DB 構築・運用のほか、関連するソフトウェア、システムの開発などを実施している。

2. 公共財としてのデータを保全・管理し、長期にわたるデータの育成と共有を行う公的機関の設置を検討すべきと考える

契約時の計画に従って研究プロジェクトからデータが国民（もしくは合理的な理由により限定公開を受ける集団）に提供される場合、プロジェクトグループ自体は研究期間終了時に解散となるため、データを維持提供し続ける主体とはなりえない。

加えて科学データは一般に他のデータと組み合わさって価値が上がるために、データの生産者とは別に、安定に維持・運営できる機関（あるいはその役割を持つ仕組み）がデータの提供にあたることが合理的である。

この機関には公共財であるデータに対する優先的なアクセスが保障されるので、応用目的研究や企業活動と競合的になってはならない。すなわち集約されたデータの利便性において機関内外及び外部機関間の格差を作らない透明性の維持を義務付けられ、データの集約利用による発見研究や同目的の外部との共同研究を厳格に制限された中立的な性格の機関であるべきである。

一方で上記機関は、データの整理や説明に加えて、データの価値を上げるための組み合わせ等による育成を行うべきであり、その為の研究は認められるべきである。この集められ育てられたデータも、材料として提供されたデータと同様に公共財として利用者と共有するべきである。

また個別研究が独自に公開しているデータベース等を案内する書誌情報的な説明を行い、計算資源を利用して維持困難な情報資源やサービスの保存を代行し、わが国の情報資源やサービスに対する努力を能率よく国民に伝え、データを共有することの重要性を研究者、ひいては国民に説明する極めて重要な役目を果たすべきである。

なお、医学研究等では、試料等の提供者個人毎の情報が、匿名化される等の措置が講じられたとしても、何らかの形で公開され共有される場合には、何よりも国民の理解と協力が不可欠であり、当該機関は社会に対しても事業の透明性を保ちながら、その説明責任を果たすべきである。

我が国の学術情報サービスの原型は1964年米国のWeinbergらがNSFとしてケネディに行った答申であるWeinberg Reportに存在する[iii]。これは我が国の情報基盤デザインの骨格として利用されている NIST(National Information system for Science and Technology)構想がWeinberg Reportの3年後に学術会議の答申として表された時に言及されている[iv]。

Weinberg Reportでは、科学技術情報基盤は研究情報源として、また科学政策立案の為の科学管理情

報(原文では科学インテリジェンス)として国家が科学を進めるにあたっての最重要課題であると明言している。加えてそれは図書館業務のような文書やデータの有形物のような保管と手渡し作業ではなく科学分野自体が行うべき情報の選別と濃縮作業を伴うものでなければならないと強調している。そして各専門分野に情報センターを設け、各大学などの図書館やデータセンターと結ぶ木構造の情報系統が提案されている。すなわち上記のような国立データセンターは既に1960年代から計画されていた。

しかし、我が国においては例えば医療・医学情報を中心に多数の情報サービスを使命とする情報サービスセンターは設立運営されているが、研究者を対象にした科学情報やデータの整理濃縮よりも一般市民への解説の著作・翻訳や広報を行うに留まっている。

【留意事項】

我が国では現代科学の直接的で確実なアウトプット(出力)である膨大なデータを共有するか分割所有するかなどの保存管理の方法が長期的に我が国の公的科学全体のアウトカム(成果) に与える影響について、あるいは過剰なデータ共有の強制が研究動機や知財取得機会に与えるマイナスの影響についての現実的な議論(の記録)はこれまでに見当たらない。 日本学術会議の 2001 年の声明に「公的資金(税金)を用いて実験や観測さらには調査から得られたデータは、一定期間、取得者の研究に使われた後は、原則として公開・共有されることが科学の発展に基本的に重要だ」[i]という「人類の共同事業としての科学」の立場からの記述があるが、「国のイノベーションの力を持続的に高める為」にという見地から、巨大なデータを研究材料とする科学分野において研究者の育成、動機付け、研究資源のすべてに関わるデータの保存管理の方法についての議論は尽くされていない。

したがって、本課題を契機に先ずは科学データの取り扱いについてより多角的な視点からの現実的な議論を透明に進め、一般社会や研究者社会での十分な理解に基づく合意形成と、その公正かつ効果的な履行に必要な制度の設計を行うことが望まれる。

公的科学の民営化(後述、p20 参照)後約 30 年を経て、次々と大型研究を進める米国では過去 10 年間に急速にデータや論文の公共財化を進めているが[vi]、情報インフラが充実した環境で価値創生の材料としての公共の電子コンテントが国家のイノベーション力の増進と発揮に欠かせないとの認識に基づき[vii]、学術誌のオープンアクセス化とあわせてデータ共有(Data Sharing) についての議論を公開してきた[vii]-[xii]。

その結果 2006 年の意見調査では 83%の米国民は「納税者は公的科学研究の成果に無料でアクセス可能であるべき」との考えを示している[xiii]。

現在我が国では、一般的に科学データ(ファクトデータ)や、それを整理したに過ぎない創作性のないデータベースに知的財産権が認められない[xiv]。 この為に生産者(事実上の保有権限者)の多くは当初の研究計画に従った解析と論文発表の後もある種の選別や加工を加えた後のデータを多様な利用規約を設けて個別に「公開」している。[xv] 統一的でない利用規約を持つデータの統合や利用は極めて煩雑であり、また不明の選別によって虫食い状態になった統合産物も十分な価値を持ち得ないであろう。従ってデータ全体の価値を損なわないために、加工や選別を受けていないデータ(注3)を十分な説明と

ともに保管し共有することが科学的に意味のあるデータベース統合の条件である[xvi]。英米では既に研究者の動機を損なわずデータの早期公開を競わせるためにデータの占有期限を決めたガイドラインに加えて、公募課題の選択基準への盛り込みや模範行為（Good Practice）の広報による教育、データ共有履行の調査による評価など科学社会全体での仕組みを作り始めている[xvii]－[xxi]。

さらに、今世紀に入り地球規模で広がる「知識に対するオープンアクセスへの潮流」という文脈の中でデータベース統合を考えることも重要である。 知識へのオープンアクセス運動は多国間の非政府運動として、ブダペストオープンアクセスイニシアティブ(2001-)やベルリン宣言(2003)によって知られているが[xxii] 、現在既に11カ国で政府系、民間あわせて 36 のファンディング機関が研究報告論文のオープンアクセス（36/36）、および研究データのオープンアクセス（16/36）に関するガイドラインを設けていることにも留意すべきである[xxiii]。

アジアでは中国も 2006 年に分野別データセンター新設を含めたデータ共有政策を進める旨を表明している[xxiv]。

各国の科学政策と歩調を合わせ、2004 年に OECD では我が国を始めとする加盟国および中国他を含めた 30 ヶ国政府による宣言 “Declaration on Access to Research Data from Public Funding”を採択している。この中では，経済の持続的な発展のためにも公的資金によるデータへの自由なアクセスと利用が望ましいことを宣言し[xxv]、科学データの扱いについての理想的な原則を端的によくまとめている。

(注３) 科学データは一般に利用目的によってデータの選別や加工などの前処理が異なる。従ってデータの選別や加工はデータの特定目的での利便性を上げる一方で利用目的 の多様性を減らすものである。また別々の基準で選別や加工を受けている場合、同種のデータであっても科学的に意味のある統合ができなくなる。米国の規約（OMB-A130）でも Raw content(生データ）の共有が奨励されている(p26) 。

（参考文献）

（ｉ） http://www.lifescience.mext.go.jp/download/34th/34-02.pdf

（ii） http://wwwsoc.nii.ac.jp/jmla/nlsic/index.html#houkoku

（iii） 科学と政府と情報－米国政府に対するワインバーグ報告－．日本ドクメンテーション協会，1966

（iv） http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu4/toushin/06041015/002.htm

（ｖ） http://www.scj.go.jp/ja/info/kohyo/ pdf/kohyo-18-k136.pdf

（vi） http://grants.nih.gov/grants/sharing.htm
その他制度は http://133.11.132.80/intelligence/report/law/index7.html 参照

(vii) Bits of Power: Issues in Global Access to Scientific Data. (1997) National Research Council. National Academy press (http://www.nap.edu/catalog/5504.html)

(viii) A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases (1999) National Research Council. National Academy press (http://www.nap.edu/catalog/9692.html)

（ix） The Digital Dilemma: Intellectual Property in the Information Age (2006) National Research Council. National Academy press （ http://www.nap.edu/catalog/9601.html)

（ｘ） Engaging Privacy and Information Technology in a Digital Age (2007)

(http://www.nap.edu/catalog/11896.html)

(ⅹⅰ)Privatizing the University--the New Tragedy of the Commons.(2000) Science. Dec 1;290(5497):1701-1702.

(ⅹⅱ)Can patents deter innovation? The anticommons in biomedical research(1998) Science. 1998 May 1;280(5364):698-701.

(ⅹⅲ)http://www.harrisinteractive.com/harris_poll/index.asp?PID=671

Wall street journal に紹介記事

http://commerce.wsj.com/auth/login?mg=wsj-users2&url=http%3A%2F%2Fonline.wsj.com%2Farticle%2FSB114893698047965609.html%3Femailf%3Dyes&mg=com-wsj

(ⅹiv) 中山信弘「著作権法」有斐閣(2007)p36-38, p119-120、p122-123

(ⅹv)http://lifesciencedb.jp/lsdb.cgi?gg=project

(ⅹvi) Seeking Consensus on Data Sharing (2002) ScienceNOW 26 February 2002: 3 http://sciencenow.sciencemag.org/cgi/content/full/2002/226/3

(ⅹvii)Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers(2008) PLoS Med. Sep 2;5(9):e183

(ⅹviii)Data sharing for computational neuroscience.(2008)Neuroinformatics. Spring;6(1):47-55.

(ⅹix)Sharing detailed research data is associated with increased citation rate.(2007) PLoS ONE. Mar 21;2(3)

(ⅹx)Towards effective and rewarding data sharing(2003) Neuroinformatics.;1(3):289-95.

(ⅹxi)Data sharing and intellectual property in a genomic epidemiology network: policies for large-scale research collaboration (2006) Bull World Health Organ vol.84 no.5 Genebra May 2006

(ⅹxii)ブダペスト宣言 http://www.soros.org/openaccess/help.shtml

ベルリン宣言 http://oa.mpg.de/openaccess-berlin/berlin_declaration.pdf

日本語での解説は「研究助成機関とオープンアクセス情報管理 48(3) 133-143」を参照

http://www.jstage.jst.go.jp/article/johokanri/48/3/133/_pdf/-char/ja/

(ⅹxiii) SHERPA で作成された世界の研究助成機関の成果に対するオープンアクセスガイドライン

http://www.sherpa.ac.uk/juliet/index.php

SHERAは Joint Information Systems Committee (JISC)と CURL (Consortium of University Research Libraries) がファンドした JISC FAIR (Focus on Access to Institutional Resources) programme の一部。UKの研究系大学や BritishLibrary などが受託して相互に連携してオープンアクセス e-print レポジトリを準備するための技術制度面の調査研究3ヵ年(2002-2005)プロジェクト.. 事務局はノッティンガム大。科学雑誌の著作権ポリシーと機関レポジトリに対する対応を網羅した調査 Romeo が有名。我が国でも筑波大学図書館など4図書館が情報学研究所最先端学術情報基盤事業の委託で日本版的 SCPJ(Society Copyright Policies in Japan)を作成している.

(ⅹxiii)http://www.scidev.net/en/news/china-unveils-plans-to-boost-scientific-data-shari.html

(ⅹxiv)OECD 2004 共同声明 http://www.oecd.org/dataoecd/42/ 12/35393145.pdf

翻訳は http://133.11.132.80/intelligence/finality/pdf/finality_01.pdf にあり 解説は

An International Framework to Promote Access to Data(2004) Science 19 March, 303,

２） ミッション② 複数の国内主要データベース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

DBとは世界の写しである。 計算機の中に作り上げた世界の枠組み（データモデル）に従って世界の中の要素の特徴を記録したレコード群（データ）である。多くの利用者によって世界全体の特徴を解析（データマイニング）したり世界の特定の要素を選び出（検索）したりするための道具である。 同じ対象世界について補完的な関係の対象や特徴を扱うDBがある場合にはそれらのDBを一纏めにしてより多くの要素につきより正確により多角的に記載したDBとすることがデータマイニングや検索利用に対して有利である。これが物理的なDB統合である。

同一世界を記述した別々のDBを一纏めにする場合には、世界の枠組みモデル（要素の特徴群や特徴群同士の関係）についての不一致、同じ特徴を記載する場合に要素の呼び名（ID系列）使われる知識表現（オントロジー）の不一致などがライフサイエンスでも例えば顧客DBでも共通の問題である。

これらの共通問題に関しては情報工学分野が解決方針を準備し様々な道具を開発し解決に牽引してくれている。異なるオントロジーの融合や機械巡回によるID対応テーブルの作成などすぐにライフサイエンス分野でも利用可能である。

本課題のフィジビリティスタディ(FS)でもこれらの技術的課題は情報工学領域で準備された方法が十分に有効であることを確認した。

用語統一の重要性とその解決例は二つのフィージビリティ・スタディによって示した。一つは解剖用語の自動分類機による全植物 EST の統合整理であり、いまひとつは複数の DB における遺伝子 ID 系列の自動更新機である。粒度や分類視点を統一することで全ての植物ESTプロジェクトを、同一プロジェクトのように統合整理することが可能であった。また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。

DBを質問に対して一定の処理をした答えを返すサービスととらえて別視点DBが供するサービスを組み合わせて新たなサービスを作り上げることは機能的なDB統合（マッシュアップ）と呼ばれる。機能的DB統合の為の技術としてはDBの提供しているサービス（利用者の質問に答えてデータを見せたり、データを加工したりする機能）自体をプログラムでも呼び出せるようにするWeb-API（application program interface）プログラムをサービス提供側に付加する技術が機能的DB統合に有効であると期待されている。例えばGoogle earthでは衛星写真で作った地図情報の部分的提供とユーザーによる座標へのコメントのマップをWeb-API経由で利用可能にしている。クライアントDB（例えばお店紹介情報）の検索結果をGoogleの衛星画像や地図画像の上にマップして表現することが可能になっている。Web-APIは上記FSで遺伝子発現DBやID対応機に既に実装しその有効性を確認することが出来た。

従ってオントロジーや辞書などによって物理統合は十分遂行可能であることが示された。その条件は統合しようとするDBが表現している領域の内容を熟知し、また、統合過程の情報処理の両方を熟知した研究者の参加、もしくは双方の研究者の密な議論による相互理解である。

【解説】 DB統合による価値の増加程度
「独占されている複数のDBが共有されるとその価値がDB数の組み合わせ数倍になる」Vinton Cerf氏の唱える法則が有名である。Vinton Cerfは通信プロトコル（TCP/IPプロトコル）をRobert Kahn氏とともに開発し現在のインターネットの基本構造を作ったことで「インターネットの父」といわれる。チューリング賞(ACM)を、コンピューターネットワーキングにおける功績に関して受賞。ICANNの会長職を務めている。いまひとつの技術基盤であるイーサーネット（Ethernet）技術を開発したBob Metcalfe(ボブ・メトカーフ)氏が1995年に提唱した「ネットワークの価値はノード数の二乗に比例する」という法則と似ている。

３） ミッション③ 国内外の医学分野・学術分野データベース、国内の産業分野データベース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

(A) 「データベース統合」の定義

「分野別推進戦略」（平成18年3月28日総合科学技術会議決定）において、ライフサイエンス分野の課題として、新規の医薬品や医療機器の産業化に向けた「実用化研究の基盤」が十分に整備されていないことが指摘されている。本課題にある「データベース統合」はこの「実用化研究の基盤整備」を目的とする事業である。

一方「研究の基盤」とは「情報や材料などの研究成果が①蓄積され②体系化されたものであり、③広く共用されるものである」と3つの特徴で定義されている。（平成19年度4月 知的基盤整備計画 科学技術・学術審議会、技術・研究基盤部会） 従って本課題では研究基盤整備の一部としての「データベース統合」を「データやデータベースを保管し、体系化し、広く共用可能にする行為」と定義する。

(B) データベース統合の背景

a) 巨大データの登場とイノベーションプロセスの変化

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。そのために「実験データ」と区別して「基盤データ」や「基礎データ」と呼ばれる。多くの研究者が同じデータ群を共有し様々に解析やアイデアを競うプロセスはこれまの生命科学が経験していない研究プロセスである。本報告書では「巨大観察データ」を習慣に従い「基盤データ」と呼ぶ。

b) データ依存を強める生命科学における DB の役割

ゲノムに始まる「基盤データ」を中心にすえた研究として①既存のデータ解析法をあてはめて行う古典研究の推進と②データ中に未知の特徴や制約を求める基礎研究の二つの典型が出現し、さらに③多種類の巨大データを組み合わせて新たな多角観察データとし、①②を助ける統合データ研究も生まれた。このようなデータ中心の研究の登場にもまして重要なのは、生命を扱うあらゆる実験研究者が研究遂行時の参照情報として「基盤データ」を利用することで、研究にブレークスルーを見出したことである。すなわち「実験データ」が誰かの研究の最下流に位置し大多数には無縁のものであったのに対し、「基盤データ」

は多くの研究の遂行上欠かせないものとなった。そして分野全体が共通の「基盤データ」を利用しながら研究を進めるために出現したのが今世紀の生命科学のデータベースである。分子進化学や数理生物学などのごく一部の研究者の机上研究に使われた20世紀のデータベースとは分野に対する重要性が全く異なる。

c) トップダウン研究の登場

科学技術基本法(H7)の成立により科学に対する社会への貢献責務が明確にされた我が国では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生命科学でも一斉に開始された。特に生命科学ではヒトゲノムプロジェクト終了時期と重なり基盤データ」を作る「ポストゲノム」や「オーミック研究」と呼ばれるプロジェクトであった(**図1**)。分野全体が依存できる「基盤データ」を特定の政府事業が短期間に生産する時代の始まりである。

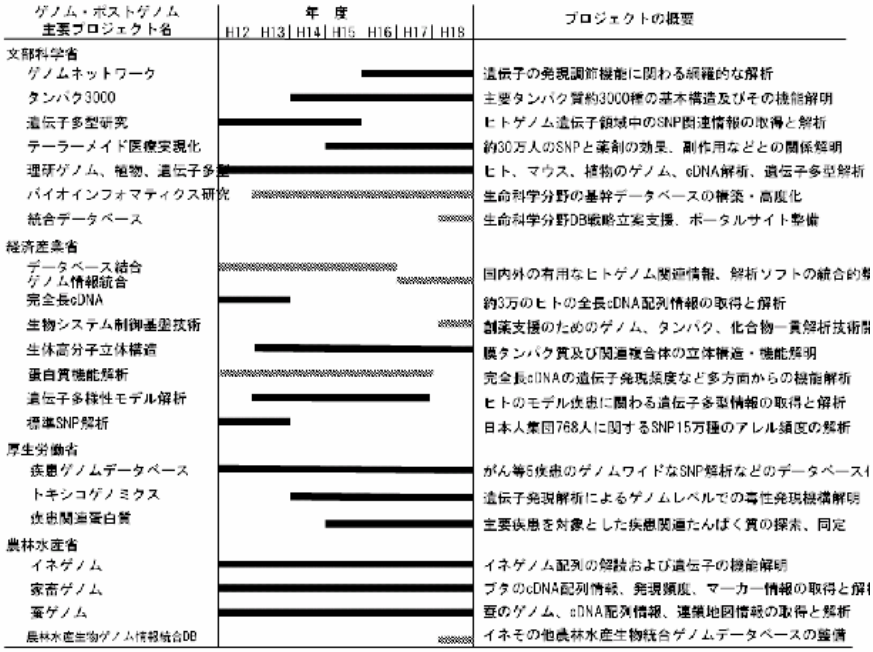

図 1　現行および終了間もないライフサイエンスの政府科学技術研究事業。

黒ぬりは「基盤データ」生産が目標に含まれる事業。

d) 知財立国政策と科学の民営化

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。また科学が経済的生産性を持つことは社会の持続的発展には欠かせない要件かも知れない。 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「政府資金による科学」の「民営化(privatization)」（以下　科学の民営化）が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2**）

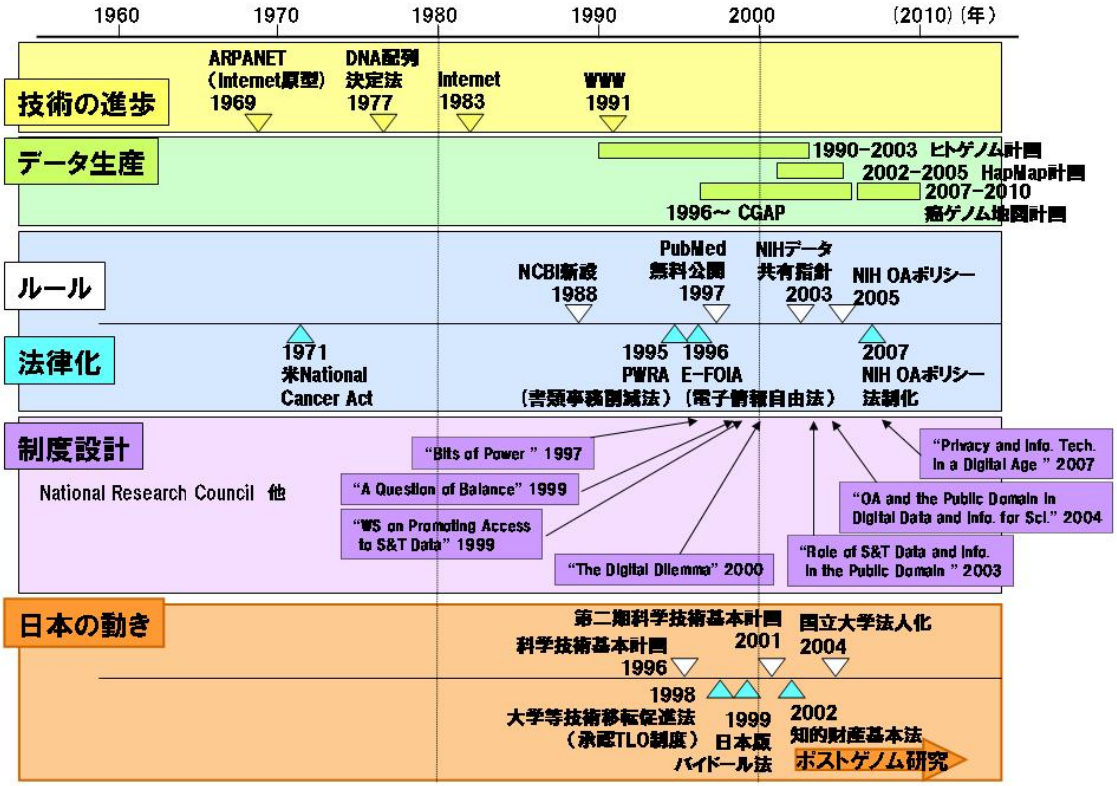

図2 統合を可能にした制度整備(米国)

(C) データベース統合の現状

a) 我が国の生命科学 DB の状況

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には、構築方法、及び所有権や公開統合の考え方で区別すべき6群に分類できる。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。しかしながら、現在我が国で、米国のNCBIやEUのEBIなどのようなデータベースセンター的な統合がなされてきていないことは、データベース統合は、我が国と米英との情報工学における技術の差ではないことを意味している。

<u>データベースの6分類</u>

①データバンク(全DBの5%):

不特定多数の著者(分野全体)からデータをドキュメンテーションとともに受け付ける。著者数が利用者数に近い「皆が作って皆で使う」ことで誰も私物化できない古典的な公共データバンクである。登録時に著作権の主張や利用制限を課さないと宣言している事業であり、知財権の対象にはならない。科学雑誌のような著作権の移譲契約を結んではいない。

②プロジェクトDB(31%):

特定の著者が大型予算でバンクに匹敵する量の分析や観察をした記録。 資金が公的な場合データとして共有化するべき部分が主体である。通常公共データ(主に米国NCBI)と組み合わせて出来るだけ

統合を行っているので、独自データと公共データを組み合わせてサービスを構築している。それぞれに論文公開にともない末端利用を許す公開をしている場合が多いが、競合サービスが出来ないようにデータの完全な共有はしていない場合が多い。 独自データには知財権が与えられず、統合DBも統合に使用した原材料のデータの利用条件によって商業利用は制限を受ける場合が多い。

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。オリジナル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。素材データとちがい個人のアイデアによる改変物であり著作物に準じた著作権が生じる。ただし原材料のデータの利用条件によって商業利用は制限を受ける場合が多い。プログラム自体は新規な部分については知財権が生じる。

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。人的な努力と判断の結果が主であり③同様著作物に準じた扱いが必要。

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。 情報源である辞書や図譜を一度理解して人的に表現しなおした場合には著作物である。ただし情報源として他の著者の編纂物である辞書や図譜を使いそれらを単近年は電子的情報源と公共用語集を材料に単にプログラム的に用語の共起などで単純に関係を作る場合も多く、その場合にはプログラムDBと同様の扱いを受けると思われる。

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。総説記事と同様の著作権により保護される.

【解説】DBの公開度;末端使用と転用許可
DB利用には2種類のタイプを区別すべきである。ひとつは「末端使用（エンドユース）」であり利用者の質問への答えを求めるものである.Googleなどの検索はこれにあたる。いまひとつの利用は「転用（デリベーション）」である.ここではDBサービスの材料であるデータを別サービスやデータマイニングと呼ばれるデータからの発見に使うことを指す。たとえば「末端使用」は許すが「転用」を許さないDBには「電子辞書」がある。商用の電子辞書は購入すれば末端利用は可能であるがそこから用語データを抜き出して翻訳ソフトを作る派生利用は禁止されている。
DBの「転用」を許すことは、大学や企業において①新しい利用の創出②価値付加の競争③間違いの発見と訂正④新たなデータの組み合わせによる解析による発見　⑤第3者による自発的データアップデートとエンドユースの多様化など様々な実益がある。一方「公開」してエンドユースを許し「派生利用」を許さない理由は通常付加価値産物が生まれ第3者によって流通させられることで構築者の利益回収が妨げられる場合に限定される。しかしながら国家プロジェクトはもとより研究成果公開促進費などでつくられた公開用DBでも「エンドユース」を無償で認めるが派生利用を許さないものが多く見られる。分かりやすい例に学会と文部科学省が著作権を持つ「学術用語集」DBがある。研究成果公開促進費によって情報学研究所で電子化検索利用が許されているが「全データのダウンロード」は許されていない。

b) 純粋科学的視点からの巨大観察データの取り扱い方針

気象データ、衛星データ、などの地球科学データをはじめ今世紀に可能になった巨視的スケールの分析データは全て巨大な観察データであり「基盤データ」である。生命科学同様天体や地球環境など込み入った現象を扱う科学はこれらのデータへの依存を高めている。 人類の生活の向上や福祉への貢献という科学の純粋な役割を考えて世界中の政府資金によって作られる生命科学や地球科学などの基盤データを人類共有の財産とすることの重要性を国際科学会議・科学技術データ委員会(CODATA)、が繰り返し指摘している。（Bits of Power NRCpress 1997）　我が国の公的資金による科学研究事業の人類への貢献と国内への貢献がどのようなバランスにあろうとも巨大なデータが民間資金により競争的に市場に提供されない性質のものであることを考えると政府資金による観察データの保管や流通の重要性に関する主張には傾聴すべきものがある。

(注) CODATA(Committee on Data for Science and Technology) は国際科学会議（ICSU：各国アカデミーの協力を進めるNPO）によって1966年に設立された学際的な科学委員会である。我が国の窓口は学術会議

(注) ICSU (International Council for Science) は、人類の利益のために、科学とその応用分野における国際的活動を推進することを目的として、1931年に設立された国際的な非政府組織である。各国の科学アカデミーが参加主体となっている。ICSUの活動目標は次のとおりである。

* 科学や社会において重要な主要問題の識別とそれらへの対処
* あらゆる分野や国家の科学者間の関係の促進
* 国際的な科学の目標におけるあらゆる科学者の参加の促進
* 科学界や政府、市民社会、民間部門の間の建設的な対話を促進するための独立した権威あるアドバイスの提供

c) 経済的視点からのデータへの考え方

OECD加盟国が目指す持続的発展が可能な経済の為にも政府資金による科学データの扱いが重要であると2003年のOECD科学技術閣僚レベル会議の共同声明で述べられている。 我が国からは稲葉文部科学副大臣（当時）と野依理化学研究所理事長の出席したこの会議の声明では「政府資金に基づく科学技術データへのオープンアクセスは研究開発への投資以上の利益をもたらし、データの下流に豊かな商業的成果を生み出すのみならず、科学リーダーが複雑な将来の問題への正しい判断を行うにも欠かせない」と述べている。(Science 19 March 2004, 303, An International Framework to Promote Access to Data)

【解説】データ公開で国益を損なうか−

データの公共への納品（公共財化）に抑制的に働く「受益範囲を国内にとどめるための非公開」という思想は、慎重に検討されるべきである。仮に「我が国の科学技術力が国際競争力において保護すべき段階にある」としても非公開独占利用によって生じる不利益――たとえば自国の科学事業が相互にシナジー効果を生まない、科学技術データに価値付加（add value）したり実用化探索(mining)したりする企業や大学研究が育たない、事業の質（データの質）に対する批判機会が失われ向上しない等――の社会におよぼす不利益について十分検討されていない。非公開を続けることは幾つかの知財の獲得と引き換えに我が国のライフサイエンスから国際競争力を奪う可能性があることを意識すべきである。

d) 米国でのデータベース統合のケース研究

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。ゲノム研究以前には、DNA配列バンクはエネルギー省の国立研究所であるロスアラモスにあり、ゲノム地図は私立のエール大に、遺伝病の原因遺伝子地図は私立大学の教授のライフワークとして編纂を続けられていたが、これらが全てNIH管轄の政府機関であるNCBIに移管されすべて統合され転用可能な形で公開された。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3**)。

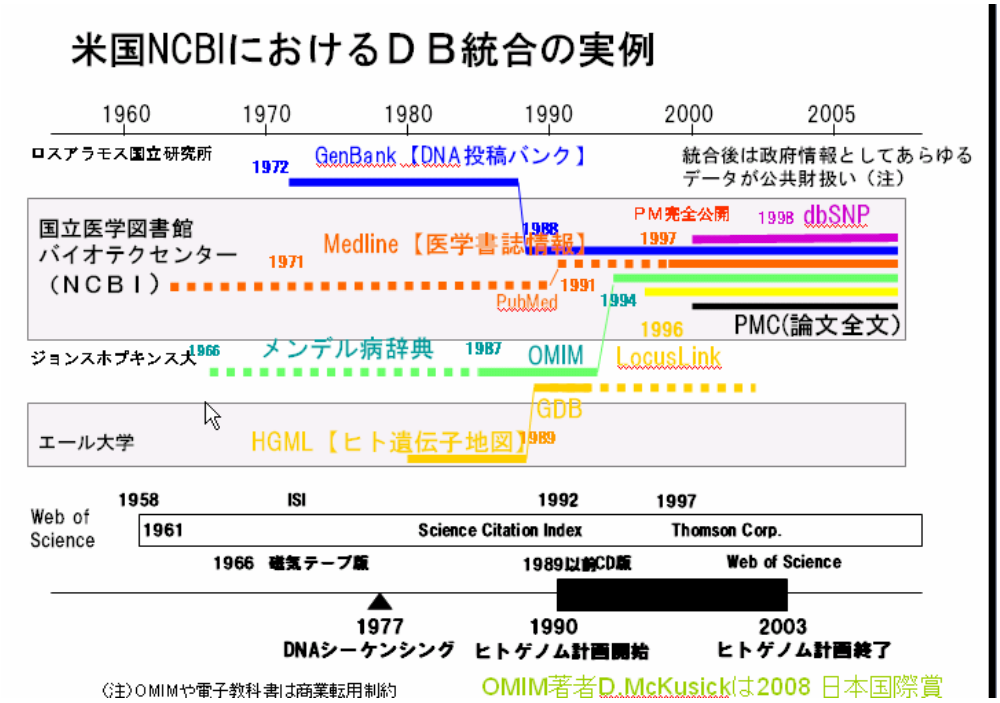

図3　米国 NCBI における DB 統合の実例

e) 米国でDB統合を可能にした制度

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。当時米国では1995年の書類事務削減法と1996年の電子情報自由法の成立により情報スーパーハイウェイを介しての政府情報の公開義務付けが厳格化された。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のである。 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。NCBIのスタッフは連邦職員であり彼らは最も厳格に市民の利益に相反する活動を禁じられておりデータは完全に市民と共有されると同時に十分に説明され最低限の整理統合が施されている。他の分野でも気象情報におけるNCDC (National Climatic Data Center) 国土地理情報においてのUSGS (US Geological Survey) や化学情報でのNIST (National center for Standard and Technology)など国立データセンターはNCBIと同列の法的制約を受けて運用されているために政府機関として大学などの非営利団体にくらべ厳格な法制度下でその事業内容に制約をうけていることで大学や企業での付加価値サービスと競合せず、かれらへの高品質なデータの提供者として機能している。このためにエンドユーザーは第3者が競争して生み出す多彩なサービスから最適なものを自由に選択することが出来る。

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年には国内プロジェクトに対しても同様に50万ドル(5000万円)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。

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するという投資価値の最大化を約束する二つの合理がここで成立している。

【解説】米国における科学技術データに関わる制度群 ①精神

米国は世界で最大の科学技術情報 ---連邦記録(federal record)と科学技術データベースを含む-----の creator, user, disseminator であり、それは価値の高い国家の資産(assets)であると考えられている。

米国の情報関連法で貫かれている基盤精神は「情報の創造と伝播と利用の多様さを育むことで民主主義は繁栄し情報の社会的恩恵が最大化される」というものである。もっといえば、政府の情報資産から、最大の経済的社会的な恩恵を引き出すために、政府情報資産は市民全員に最も効率的でタイムリーで公平な方法で入手可能にされるべきである. という考えである。

米国の法や政策には原則としてこの考えを盛り込んである。プライベートセクターの財産権を守ることを強調するのと対照的に連邦レベルの米国内の情報政策は ①strong freedom of information law, ②no government copyright, ③fees limited to recouping the cost of dissemination, and ④no restriction in reuse とまとめることができる。