報告書「我が国におけるライフサイエンス分野のデータベース整備戦略のあり方について」からの抜粋

文部科学省
科学技術・学術審議会
研究計画・評価分科会
ライフサイエンス委員会
データベース整備戦略作業部会

6-1 推進方策

前節「5. データベース整備戦略の基本的考え方」を踏まえて、「4-2 取り組むべき課題」を具体的に実現するためには以下の推進方策を遂行する必要がある。推進すべき方策とその留意点は下記の通りである。

(1)データベースの現状調査、評価、戦略立案機能の充実

現在、データベース整備の戦略立案機能はJSTバイオインフォマティクス推進センターBIRDや国立遺伝学研究所DDBJに設けられている委員会あるいは文部科学省のライフサイエンス委員会などによって一部担われているが、それらは非常勤の委員からなる委員会活動であり十分ではない。また各機関の委員会では、その守備範囲もその組織の活動に関するものに限られ、限定的なものとなっている。そこで、専門家による日常的活動（研究者の常勤）を基盤とし、データベースの現状や動向の定常的な調査および既存の戦略や活動の弛まぬ評価に立脚して、省庁の枠を超えて国家的視野に立って、ライフサイエンス研究全般やバイオ産業全般を見渡した戦略立案する機能が是非とも必要である。

なお、これらの調査、評価、立案に際しては、以下の点に十分な配慮・検討が必要である。

- データベースだけの問題と捉えるのではなく、ライフサイエンス研究の方向性も十分に踏まえた戦略を立案すること。
- データベース構築は、個々の研究者の創意工夫による研究とは異なる事業的な側面をもつことを十分に認識し、その推進および体制の整備に努めること。
- データベースは、ライフサイエンス研究全般、医療、バイオ産業全般の知的基盤、後方支援との明確な位置づけを行い、ニーズを的確かつ継続的に把握すること。
- データベースを構築する側の立場だけでなく、利用する側（例えば、医療や産業界）の意見が十分に取り入れられるように配慮すること。また、そのための仕組みを確立すること。
- 現在、ともすれば別々の戦略をもって収集・管理が行われている医学情報や薬学情報

との連携にも十分配慮すること。

・データベース間の連携強化のためのデータベースの形式や構造の標準化や知識の体系化に向けた用語の統一化（辞書作成・標準化）のための戦略もあわせて立案すること。
・また、用語の統一化やデータの記述形式の標準化などをデータベース構築の際に義務づけるための制度設計もあわせて行うこと。
・データベースの開発とそのための技術開発(研究）とを緊密に連携させる仕組みを考案すること。
・国として支援するデータベースや国として構築するポータルサイトの厳格な評価を行うための仕組みを検討すること。具体的にはモニター制度、利用者評価等を取り入れることを検討すること。
・文献データベースとの連携のための仕組みを検討すること。
・既存のデータベースだけでなく、ライフサイエンスの進展に対応した、新しい種類のデータベースあるいは従来にない発想に基づくデータベースの開発の振興にも十分配慮すること。
・データベース構築だけでなく、それを利用する技術開発の促進策も検討すること。
・長期的視点に立って、人材養成の促進を図る教育体制を構築すること。
・国家プロジェクトの成果活用の方向性を検討し、効果的な情報提供に向けた連携のための施策を考案すること。
・海外との連携をさらに進める方策を立案すること。特にアジア諸国のデータ生産者、バイオインフォマティクス研究者およびデータベース運営機関との連携について留意し、積極的な交流を図ること。

(2)基盤データベースの安定的な支援

我が国のライフサイエンス研究の基盤として欠かせないデータベース、世界的競争力の確保に向けて戦略的に重要なデータベースなどについては、安定的、永続的に支援することが必要である。現在この機能の一部は国立遺伝学研究所DDBJで実施されており、その他にもJSTバイオインフォマティクス推進センターによる支援が行われているが、データベースの数も限られており、また現在支援を受けているものについても、予算や期間の制約があり十分とはいえない。今後の更なる拡充が望まれる。なお、基盤データベースの安定的な支援に際しては、以下の点に十分な配慮・検討が必要である。

・我が国が独自に保有することが不可欠のものや世界的に存在が認められる知識基盤に限定して支援すること。その際、存在意義が認められる期間、安定的に維持するための必要額を十分に精査し支援すること。そのための評価基準として、論文への引用件数、アクセス数、一次データ量などによる定量的評価、外部有識者や利用者による定性的評価、およびサービス体制の充実度等を用いること。
・データベースの存在価値を維持するためのデータの収集・精査、サービス向上に直接

関連する研究開発に限定して支援すること。新たな研究開発要素などは別予算（別途審査）（下記の(8)や(9)を参照のこと）で対応すること。

・ここで支援するデータベースについては、用語の統一化、データベースの記述形式や構造の標準化などの制約を課して、我が国のデータベースの統合化に寄与することを義務づけること。

・価値の高いデータベース、世界的に競争力のあるデータベースでありつづけるためには、それに関係した研究グループと密接な関係を常に維持していなければならない。そのための配慮を十分に行うこと。

(3)データベースの所在情報と利用法に関するポータルサイトの構築と運営

ライフサイエンス関係のデータベースに関する所在情報や利用法に関するポータルサイトを構築し運営することが必要である。これに関しても、いくつかの機関（JSTバイオインフォマティクス推進センター、国立情報学研究所など）でその試みはあるが、十分とは言えない。その理由は、常勤の専門家による運営が必ずしもなされていないこと、利用者からのフィードバックを常に活かしてサイトを最新のものに更新する仕組みが整っていないことによる。その背後には、このような仕事への評価の低さと予算面の手当てのなさの問題がある（国立情報学研究の活動は予算的な裏づけがあったが、平成17年度末で終了）。3-3節で紹介したように、我が国では数多くのデータベースが日々作られている。これらを十分に活用するためには、常に最新の情報を保持したポータルサイトが不可欠である。このサイトの構築・運用に際しては、以下の点に留意すべきである。

・何といってもポータルサイトにとって重要なことは、その網羅性である。日々、新しいデータベースが作られているような今日の状況では、個々の利用者が関連するデータベースすべての所在情報や利用法を把握するのは事実上不可能である。ポータルサイトにはデータベース作成者の意向も踏まえた上で、我が国のデータベースを漏れなく収載することが欠かせない。

・一方、ポータルサイトに掲載されるデータベースが玉石混淆ではかえって混乱を招く。これを避けるため、引用数、アクセス数、データ量等を調査し、利用者側から見て分かりやすいよう、掲載するデータベースの分類をすること。

・使いやすさによるデータベースの評価や利用法からみた分類などによるガイダンス機能の導入など、利用者の視点に立ったポータルサイトの運用に努めること。

・ポータルサイトの自動構築や評価のための技術開発もあわせて行うこと。

・ライフサイエンス分野の研究者、技術者を主たる対象とするが、一般の医療関係者あるいは育種家といった利用者も想定し、日本語での情報提供にも十分配慮すること。

(4)統合データベースの開発とそのための研究開発の促進

データベースの統合化に関しては、我が国においてもいくつかの機関でそれぞれの取組

みが行われている。それらには一長一短あり一概には評価することはできないが、いずれも我が国のデータベース全般を統合化するという視点は弱い。その理由は、そもそもそのような使命を負わされてわけでもないし、権限があるわけでもなく、そのための予算の裏づけがあるわけでもないからである。JSTバイオインフォマティクス推進センターにおいても、データベースの高度化・標準化が謳われているが、統合化は必ずしも視野には入っていない。しかしながら、上述のポータルサイトの構築・運営だけでは、我が国の様々なデータベースの価値を十分に引き出すことはできず、ライフサイエンス研究のみならず産業界からの要請にも応えることはできない。多種多様なデータが生物的医学的に整理された形で統合されなければ、膨大なデータの洪水に流されてしまうだけになってしまい、ライフサイエンスの発展が止まってしまう。逆に、バラバラだったデータベースを統合化することができれば、これまで別々のデータベースに収められていたデータ間の潜在的な関係（例えば、遺伝子と疾患と薬剤との間の新たな関係やゲノムの進化と表現型の進化の間の対応関係）を見出すことが可能になる。ポータルサイトだけでは、このような新たな知識の発見を直接的に支援することはできない。データベース構築の大きな目標の一つはそこから新たな発見をすることにあり、統合化はまさにそのためのものである。一朝一夕には無理でも、我が国のデータベースの統合化に向けた研究開発を強力に、かつ、地道に推し進める必要がある。

ただし、統合化と言っても生命階層のどのレベルの、どのような知識を発見したいのか、どのようなことに統合データベースを使いたいのかによって、その目指すところ、意味するところは異なってくる。仮に目指すところが同じでもいろいろなアプローチがありうる。そのため、我が国としてどのようなアプローチでどのような統合化を目指すべきか（これが一つとは限らない）に関しては、将来のライフサイエンスの動向や産業界からのニーズも十分踏まえた検討を行い、その議論に基づいて推進を図るべきである。幸い、現在、科学技術振興調費「科学技術連携施策群の効果的・効率的な推進」の一テーマとして調査研究が進められているところでもあり、その結果も踏まえて、前記(1)の戦略立案機能の中で推進策を練ることが望ましい。

ところで、どのような統合化を目指すにせよ、統合化にあたっては、そのための用語や概念の統一化、データベースの記述形式や構造の標準化が前提となる。これらの中はすでに欧米で開発が進んでいるものもあり、それらを採用することも考えられるが、5節の「データベース整備戦略の基本的な考え方」に述べたように、我が国の特徴や強みが十分に発揮できるように十分な配慮・検討が必要である。

この他にも、データベースの統合化とそのための技術開発に向けては、以下の点に十分な配慮・検討が必要である。

・国が支援するデータベースの構築者に対し、情報提供や技術指導を行うなど十分な連携をとり、用語の統一や記述形式の標準化を図ること。
・データベースの専門家（特にバイオインフォマティクス研究者）だけでなく、実験研

究者や医療やバイオ産業に従事する人でも簡単に使えるような検索ソフトの開発や日本語環境の整備にも努めること。

・欧米の後追いにならず、次世代の統合化を先取りするためにも、最先端の情報処理技術の活用や開発を行うこと。例えば、画像情報や新しい計測機器の出力結果等、新しい形式のデータに対応した情報処理技術や、新たな情報共有の枠組みのための情報処理技術を開発すること。

・概念や用語の統一が統合化の鍵を握ることから、また我が国独自の特徴を出す意味からも、分野毎に、実験系の研究者と情報系の研究者の双方からなる専門家集団を形成し、それらの専門家集団の知識の融合に基づく統合データベースを目指すこと。

・上のことと関連するが、データベース構築には実験研究者も深く関与できるような体制作りが必要である。

(5)維持が困難になったデータベースの受入れ

4-1節「データベースの問題点」に述べたように、各機関や各プロジェクトで開発されたせっかくのデータベースが、予算が切れると維持更新されなくなってしまうという問題がある。これに関しては、現在は研究者あるいは研究室の自発的な努力に頼るしかない状況であり、我が国のライフサイエンスにとって由々しき問題である。当然のことながら、すべてのデータベースを管理し続けるのは意味もないし不可能であるが、存続することが重要と判断されたものに関しては十分な支援が必要である。すなわちプロジェクトや科研費などの研究費が終了するなどして維持が困難になったデータベースの受け皿を、国として用意する必要がある。もちろん、闇雲に受け入れる必要はなく、存続価値を十分に厳正に評価して受入れや支援を判断すべきである。その際、以下の点に留意すべきである。

・ライフサイエンスの進展とともに、支援しなくてもよくなるデータベースも出てくるが、その一方で新たに支援すべきデータベースも出てくる。このような変化に柔軟に対応できるような制度（例えば、データ産出プロジェクトの設置に際しては、そのプロジェクト経費の一部をプロジェクト終了後も一定期間データベースの維持更新が可能なように積んでおくことを義務付けるなど）を導入すべきである。

・文科省以外の省庁が整備したデータベースについても受け入れを検討すること。その際、内閣府の委員会、調査なども踏まえて検討すること。

・データベースの受け皿機関への移管に関しては、権利関係、事務手続きなどに配慮すること。

・ここで支援するデータベースについても、移管する際に、可能な限り、用語の統一化、データベースの記述形式や構造の標準化などの制約を課して、我が国のデータベースの統合化に寄与することを義務づけること。

(6)文献情報との連携

3-1節「データベース開発の世界的な動向」に述べたように、機能情報のデータベース化が重要な課題になりつつある。機能情報の多くは論文の中にテキストとして記述されていることから、文献中に記述されたデータや知識と、配列や立体構造などの実験データとの連携と統合に今後取り組まなければならない。米国NCBIでは、同じ組織で実験データも文献データも管理されていることから連携は比較的スムーズであるが、我が国ではこれまで別々に扱われてきたことから、今後連携を図っていく方策を講ずる必要がある。具体的には、遺伝子名や塩基配列のアクセッション（用語解説参照)などによる共通識別キーでの統合的検索を可能とするほか、ライフサイエンス分野の知識を計算可能な形へ変換し、概念対概念の関係を自動生成することにより、増大する論文データに対応できる知識の体系化を実現する必要がある。また、これにより提示される知識体系を活用しつつ、各データベースで利用されている各種用語の標準化を図る必要がある。また、このようなことを可能とする技術開発（概念・知識の収集の自動化やデータベースからの知識発見など）を並行して進める必要がある。なお、これらは上記(4)の「統合データベースの開発とそのための研究開発の促進」と重なる部分がいくつかあり、十分な連携のもとに進める必要がある。

(7)アノテーション（情報解読による実験データの注釈付け）の実施

基盤データベースの支援や維持が困難なデータベースの受入れ、さらにデータベース統合化および文献情報との連携といった活動と連動して、我が国で産出されたデータにもかかわらず未解析、未解釈のまま放置されている種々の実験データの意味付け（生物学的、医学的な解釈）を強力に推進すべきである。また、すでにアノテーションされているものでも正確さを欠くものもあり、それらについても再度アノテーションを実施すべきである。これについては、現時点で未解析・未解釈・不正確なデータのアノテーションを実行するだけでは不十分である。今後出てくるデータに対しても常に最新の技術、知識をもった専門家によるアノテーションを施す体制の確立が望まれる。アノテーションされていないデータは統合化しても意味がないし、ポータルサイトで所在が明らかになっても利用価値は低い。アノテーションを実施する際に留意すべき点は以下の通りである。

- アノテーションは独自の基準でバラバラに行うのではなく、上記(4)「統合データベースの開発とそのための研究開発の促進」や(6)「文献情報との連携」で開発された用語やガイドラインに基づいた注釈を行うこと。これによりデータベースの統一化が可能となる。
- 実験系と情報系の研究者が協力できる体制を構築して、より正確で意味のある情報解読・注釈付けを実施すること。
- cDNA、イネゲノム、微生物ゲノムなど日本の強みを発揮できるデータについては、統一基準でより信頼性の高い形での再アノテーションを実施し、それを公開データベースに反映することを検討すること。

(8)新たなデータベース構築への投資

上述の基盤的データベースや評価の確立したデータベースの安定的な支援のほかに、ライフサイエンス研究の進展に対応した新たなデータベース、新たな発想に基づくデータベースの構築にも投資すべきである。これは、現在一部科研費特定領域研究「ゲノム４領域」やJSTバイオインフォマティクス推進センターで実施されているが予算や期限が限られており十分ではない。データベースには長期的視点が必要である。今後このような観点にたった支援制度を是非とも設けるべきである。長期的視野に立つといっても、新しく作られるデータベースは玉石混淆である。5年程度の時限を設けた競争的研究資金により実施することが望ましい。そこで評価が確立したものついては、例えば、上記の(2)「基盤データベースの安定的な支援」により支援することが考えられよう。なお、新たなデータベース構築への投資を行う際には、以下の点に十分な配慮・検討が必要である。

- ここで支援するデータベースについては、最初から用語の統一化、データベースの記述形式や構造の標準化などの制約を課して、我が国のデータベースの統合化に寄与することを義務づけること。
- 新たなデータベース構築は機関で行ってもよいし、数人から個人の研究者レベルで行ってもよい。個人で行う場合の支援策に関しては、特に若手研究者が行う場合には、データベース開発に対する研究者の理解が不足している現状を考慮して、任期付きあるいは終身雇用の職をどこかに用意するなどの点に配慮すること。
- データベースの構築そのものでなくても、それの基盤となる、分散処理、高速通信、データベースマネジメントシステム等の基盤的技術開発を支援することも必要。

(9)データベースを活用した研究（バイオインフォマティクス）の促進

当然のことながら、データベースはそこから有用な知識を発見してこそ意味がある。逆に言えば、そのことを見越してデータベース開発を進める必要がある。そこで、データベース構築への支援と並行して、それを活用する技術の研究開発、いわゆるバイオインフォマティクスの促進も図る必要がある。バイオインフォマティクスそのものは、科研費特定領域研究「ゲノム４領域」やJSTバイオインフォマティクス推進センターなどで振興が図られているが、データベース構築と一体となった研究開発は必ずしも活発には行われていない。今後この面での支援策を講ずる必要がある。また、従来の施策の更なる拡充を図る必要がある。なお、これに関しては、若手研究者の育成、そのための任期付きあるいは終身雇用の職の確保、競争的な資金制度におけるバイオインフォマティクス分野の研究と連携、などに十分配慮して遂行すべきである。

(10)データベース開発のための人材養成

いくつかの大学において、21世紀COEプログラムや科学技術振興調整費人材養成などの支援を受けながら、バイオインフォマティクス分野の研究者や技術者の養成が行われてい

る。しかしながら、質の高いデータベース構築を行う上で不可欠の人材である、アノテータ（データに生物学的医学的な解釈を加える専門職員）やキュレータ（データベースの編集作業に従事する専門職員）を目的としたものはほとんどない。経済産業省の産業技術総合研究所生命情報科学研究センターや国立遺伝学研究所で一部実施されてはいるものの十分ではない。問題は、教育する側にあるのではなく、受け手の少なさにある。それはアノテータやキュレータの技術を身につけても我が国にはその職がないからである。また、そのような仕事の重要さへの理解が不足しているからである。我が国で世界的に競争力のある、また、意味付けがきちんとされた、有用なデータベースを開発するには、まずはアノテータやキュレータの安定的な職を数多く確保するとともに、それに相応しい人を養成することが不可欠であり、そのための体制を早急に確立する必要がある。また、そのためにその後の将来の処遇（キャリアパス）につながるような学会の認定資格などの方策も検討する必要がある。高度に専門的な知識や技術をもったアノテータやキュレータを養成するには、振興調整費人材養成プログラムあるいは大学の専門教育との連携がなくてはならない。このことを十分に考慮した人材養成の仕組みを構築する必要がある。上記の観点は、データベースのシステム開発や運用を専門的に担ういわゆるシステムエンジニアやオペレータの育成に関しても言えることである。