生命科学系データベース統合に向けた取り組みの現状と課題

The current state and issues of life science database integration

畠中 秀樹¹

HATANAKA Hideki¹

.....

¹ 独立行政法人科学技術振興機構 バイオサイエンスデータベースセンター (〒102-8666 東京都千代田区四番町5-3)

Tel: 03-5214-8491 E-mail: h2hatanaka@jst.go.jp

¹ National Bioscience Database Center, Japan Science and Technology Agency (5-3 Yonbancho Chiyoda-ku, Tokyo 102-8666)

原稿受理 (2012-06-25)

情報管理 55(5), 347-353, doi: 10.1241/johokanri.55.347 (<http://dx.doi.org/10.1241/johokanri.55.347>)

著者抄録

多数乱立する生命科学系データベースの統合は、従来日本では関係する省ごとに取り組まれてきた。この取り組みをオールジャパンで進めるために、内閣府の主導のもと2011年度より、文部科学省・厚生労働省・農林水産省・経済産業省の間で協力体制が築かれ、カタログ・横断検索・アーカイブ・再構築の4つのステップに沿ってデータベースの統合が進められている。本稿ではこの4つのステップの現状を紹介するとともに、データ共有やRDF化など、各ステップにおける課題に触れる。

キーワード

生命科学系データベース, データベース統合, データベース横断検索, データベースアーカイブ, データ共有, RDF, セマンティックウェブ, オントロジー

1. はじめに

生命科学の諸分野には、多数のデータベースが散在する。なぜこんなに多いのか、なぜ複数のデータベースが1つにまとまらないのか？ この問いへの答えについては参考文献1) をぜひ参照されたいが、その中でまず取り上げるべきは文脈依存性の問題であろう。数理系データに比べ生命科学系データは、材料や実験方法などの細かい文脈により解釈や評価が変わるから、異なるデータベースの中に同じ名前がある場合に同一視してよいかを判断するには、それ

らの文脈を知る必要がある。つまり複数の生命科学系データベースを1つにまとめるには、まずデータに対する説明(メタデータ)をうまく整備する必要がある。これはデータの件数やサイズが大きいこととは別の問題である。

複数の生命科学系データベースを形式的・物理的に1つにまとめる必要は必ずしもない。しかし、ユーザーが必要なデータを見つけ、組み合わせで使い、正しく解釈し、新しい発見を行うためには、上記の文脈依存性などの問題が解決され、データ同士がうまくつながることが必要である。われわれはデータ

図1 4省データベース統合の段階的進展

ベースの統合を、そのような意味でとらえている。

本稿では国内の生命科学系データベースの統合について、特に筆者が関わっている省間連携の側面から現状と課題を説明する。

2. 4省のデータベース統合と4つのステップ

生命科学系データベースの統合は、従来日本では関係する省ごとに取り組まれてきたが、この取り組みをオールジャパンで進めるために、内閣府の主導のもと2011年度より、文部科学省・厚生労働省・農林水産省・経済産業省の間で協力体制が築かれることになった。4省のそれぞれから、(独)科学技術振興機構バイオサイエンスデータベースセンター(NBDC)、(独)医薬基盤研究所(NIBIO)、(独)農業生物資源研究所(NIAS)、(独)産業技術総合研究所バイオメディシナル情報研究センター(BIRC)の4機関が協力し、省間データベース統合の各種とりまとめ作業を担っている。なお、NBDCは2010年度以前に文部科学省のプロジェクトでデータベース統合を進めていたライフサイエンス統合データベースセンター

(DBCLS) から多くを引き継いでいる²⁾。

4省連携では、データベース統合を4つのステップ（カタログ連携、横断検索連携、アーカイブ構築連携、再構築連携）で推進している（図1）。カタログ連携では、各省のデータベースに関する情報を記載した、データベース単位のリンク集をまとめる。横断検索連携では、各省のデータベースサイトに含まれる無数のページに対し、キーワードで一括検索ができるようにする。アーカイブ構築連携では、各省のデータセットをダウンロード可能な形にして長期間維持する。データベース再構築連携は2014年度以降を想定しているが、各省からのデータベースを再編して収容し、統一化されたフォーマットで再構築された状態を実現する。

3章～6章ではこの4省のデータベース統合の4つのステップについて、現状と課題を述べる。なお各ステップの現時点での成果は、2011年12月に公開した4省の合同ポータルサイトintegbio.jp^{注1)}において利用可能であるのでぜひ参照されたい。

Integbio データベースカタログ β 版 1.0 ~ データベースを探す

integbio.jp 「DBカタログ事業」へ
最終更新日: 2012/06/26

NBDC「生命科学系データベースカタログ」, 「WINGpro」, MEDALS「データベース便覧」, AgriTOGO「データベース便覧」, 医薬基盤研究所「各種データベース」の情報を集めています。

国内外の生命科学系データベースの情報を閲覧、検索、ダウンロードすることが可能です。

<ご利用にあたって> 【Show】レコードの詳細表示 【Show Advanced Search】詳細検索 【一覧の項目】項目別ノート 【Download】検索結果の一括ダウンロードが可能です。【本カタログについて】 【利用許諾とお問い合わせ】

<新着情報>
2012/6/25 以下のデータベース13件を追加しました
(GTPS, BSD, GIB-IS, BioDataBaseShowcase, バイオ基幹情報資源の高度化と共有化, xPSSS, PDB) Mine, Sequence Navigator, Structure Navigator, CiAID, 3DPL, ACED, Kyoto Human Embryo Visualization Project)
2012/6/15 経済産業省関連のデータベース6件を追加しました (Fulcrum, MCG CNV Database, PMP-Blast, MFSearcher, PoSSuM, DoBISCUIT)。

Show Advanced Search Search Reset Download

1139 Found Columns ... 15 1 2 3 ... 76 Next

データベース名称とURL	説明	運用機関名	稼働状況 (6/4)
Integbioデータベースカタログ 【名称:】 http://togodb.dbcls.jp/integbio_dbcatalog	国内外の生命科学系データベースのカタログです。各データベースの所在情報と、データベースについての説明や生物種などのさまざまな属性情報 (メタデータ) をまとめたリストを作成し、提供しています。2011年12月12日にベータ版1.0を公開しました。本カタログでは以下の既存のデータベースカタログのエンTRIESを統合し、記載内容の統合を行ったものをレコードとして収録しています。 *文部科学省 バイオサイエンスデータベースセンター「生命科学系データベースカタログ」, 「WINGpro」*経済産業省 MEDALS「データベース便覧」*農林水産省 AgriTOGO「データベース便覧」*厚生労働省 医薬基盤研究所のホームページから公開されているデータベースリンク集	独立行政法人科学技術振興機構	稼働中 Show
3DinSight 【名称:】 http://gibk26.bio.kyutech.ac.jp/jouhou/3dinsight/3DinSight.html	生体分子の構造、機能及び物性に関する統合型データベースです。蛋白質立体構造データベース (PDB) を中心に、PROSITE, Protein Mutant Database (PMD) などの機能情報, PIR, SWISSPROT, アミノ酸の物理化学的性質などのデータをリレーショナルデータベースに統合しました。構造、機能、物性を組み合わせたさまざまな検索を行い、これらの間の関係について調べることができます。構造と機能及び物性の間の関係性については可視化を行っている (例: 自動的に機能部位をマッピングした構造データのリアルタイム3D表示など)。	九州工業大学	稼働中 Show
5'-end serial analysis of gene expression database	ヒト細胞株11ラインについて、5' SAGE法を用いて網羅的に遺伝子発現解析した結果を収録したデータベースです。ライン間の発現の違いが分かりやすく表示されます。オリジナルSAGE法は3本の特定の断片を用いて遺伝子を特定するものであり、最大27bpの断片を用いる		

図2 Integbioデータベースカタログ (β版)

3. データベースカタログ

カタログ連携では、データベースの所在情報とデータベースについての説明やさまざまな属性 (メタデータ) をまとめたリストであるデータベースカタログを作成している。

従来は生命科学系データベースカタログとWINGproという2つの大きなカタログをNBDCが公開する一方で、BIRCのMEDALSデータベース便覧、NIASのAgriTOGOデータベース便覧、NIBIOのデータベースリストが各々のサイトで公開されていた。現在、これらをまとめて4省にまたがる新しいデータベースカタログ「Integbioデータベースカタログ」の作成を進めている。2011年12月、integbio.jpのサイト公開にあたり、現存する情報を1か所に集めて各リストの照合や記載項目の策定、記載内容の統合を行い、国内で約700件、国外も含めると1,000件を超えるデータベースを記載したIntegbioデータベースカタログβ版を公開した (図2)。

現在は正規公開版の準備を進めており、今後も各

省で情報収集を進め、4省のデータベースを網羅していきたいと考えている。カタログにとって網羅性は最大の課題であり、研究費単位、研究機関単位など、複数の方法を用いて調査を進めていく予定である。

4. データベース横断検索

横断検索連携では、散在する多数のデータベースサイトに含まれる無数のページをキーワードで一括検索できる仕組みを作っている。既に、文献や特許、分子データなど多種多様な300以上のデータベースについて、検索用索引データ (転置インデックス) を作成し、複数のサーバーへ収納している。ユーザーがキーワードを入力すると、インデックスサーバーが連携してそのキーワードを含むページの抜書き (スニペット) の一覧をユーザーに示し、その一覧を見てユーザーは望むデータベースの望むページにアクセスできる。

現在、ユーザーが4省のデータベースを一括検索できるようにするために、上述のインデックスサーバー

図3 横断検索の省間連携の仕組み

に省間で相互乗り入れできるシステム（図3）の作成を進めている。文部科学省と経済産業省のデータベースに関しては2010年5月以来、P2P方式で容易に連携を実現できる全文検索エンジンHyper Estraier^{注2)}を用いて、両省のデータベースが一括検索できる状態にあった。厚生労働省と農林水産省のデータベースにもこの仕組みの導入を現在進めており、今後それぞれの省でカバーするデータベース数を増やす必要はあるが、近い将来4省の多くのデータベースが一括検索できるようになる予定である。

横断検索における大きな課題は、各データベースの更新に対応するためのインデックスの更新である。データがいつどのように更新されるかを把握できなければ、データベースとインデックスの間の対応が取れなくなり、検索漏れやリンク切れが生じることになる。特にデータベースのサイトの構造が変化すると、インデックス更新手順も再検討する必要があるため、データベース数が増えるとともに過大な負担となる。今後、インデックス更新や作成の効率を改善するために、データベース運用機関の協力を得ることが重要になると思われる。

5. データベースアーカイブ

アーカイブ構築連携では、国内の研究者が産生したさまざまなデータセットをダウンロード可能な形（アーカイブ）で長期間維持する活動を行っている。作成されたアーカイブは、「生命科学系データベースアーカイブ」³⁾で公開されている。データセットが広く再利用されるように、データベースやデータに対する統一形式での説明（メタデータ）や、クリエイティブ・コモンズ・ライセンスの仕組みを利用したわかりやすい利用許諾表示、データの一覧・検索・絞り込みなどができる簡易検索ページを併せて提供している。

アーカイブには大きく分けて2つの目的がある。1つはアーカイブという名前が示す通り、公的資金で生み出されたデータが失われないように保管するという側面である。生命科学の分野はゲノム・ポストゲノムの時代を迎え、時限性のプロジェクトにより作成されるデータベースが急増しているが、プロジェクト終了後にデータベースの維持が困難になりがちである。アーカイブはこのようなデータの散逸を防ぐため、データセットを維持費がかかりにくい形で受け取り、無料で維持・公開している。さらに、データを表形式でTogoDB^{注3)}に搭載し簡易検索ページを

設置すれば簡易データベースの形で公開を維持することもできるし、種々のカスタマイズを行って元のデータベースに近い形で公開を維持することも場合によっては可能である。

アーカイブのもう1つの大きな目的は、データの再利用の促進である。一般にデータには作成者が想像していなかった活用法や価値があるものだが、データベースの検索などのインターフェースは作成者が想定した利用法に利用を制限することになる。そこで、データをダウンロード可能にして、「自由」に使える状態にすることが重要になる。このためアーカイブでは、単にダウンロードを可能にするだけでなく、適切な利用許諾条件を明示するとともに、十分なメタデータを用意している。アーカイブの標準利用許諾には、データの改変や営利利用も可能なクリエイティブ・コモンズ・ライセンス「表示-継承」を採用し、誰もが自由に使えるように努めている。

アーカイブ事業では2011年度から、各省でアーカイブ作成を分担できるように、文部科学省側から「アーカイブ作成ガイドライン」を提供してデータ加工や利用許諾設定を支援するとともに、データベース寄託者や各省のとりまとめ機関がアーカイブサイトに直接ログインしてアーカイブ作成を行える仕組みを用意するなど、連携方法の改善に取り組んでいる。省間連携の成果として、170の化合物による動物個体や肝細胞の遺伝子発現の変化や毒性のデータベース「Open TG-GATEs」など、重要なデータベースがアーカイブ化され、ユーザーに多数ダウンロードされている。

アーカイブにおける最大の課題は、データの共有である。データの公開と共有は混同されがちだが、データベースを公開するだけでなく、他の研究者がデータを手元に複製して解析を加えその結果を公開することを、許すことが重要である。米国では「Taxpayers should not have to pay twice for the same information」⁴⁾とされ、科学データの公共財化が進んでいる。最近日本でも文部科学省や厚生労働省の

研究費の公募要領で、NBDCへのデータ提供を明文化するケースが増えてきており、今後日本でも科学データの公共財化・共有・流通が進むことが期待される。

6. データベース再構築

統合のさらなるステップとして、特に2014年度以降、セマンティックウェブなどの技術を利用してデータベースの再構築を進める予定である。各省から譲り受けたデータベースをRDF (Resource Description Framework) と呼ばれる統一フォーマットに変換し、オントロジーを適切に付加することにより、他のデータベースと容易に統合できる状態にする。これらを収容したシステムを公開し、適切なインターフェースを用意することで、複数のデータベースをまたぐ条件での検索や、複数のデータベースから情報を抽出した出力など、統合化された検索が容易にできるようになるはずである。

現在はこのステップの前段階として、4省の間で各種の技術交流を行っている。2008年から毎年行われている国際開発者会議バイオハッカソン (BioHackathon) や、2010年から行われている統合データベース技術情報交換ワークショップでは、1週間ほど実際の開発を行いながら技術交換を行う。毎年10月5日に行われるトーゴの日シンポジウムでは、この分野の研究者が講演やポスター発表を行うとともに、他分野の研究者や一般参加者を交えてテーマに沿った議論を行っている。その他、学会での共同ブース展示・各種実務連絡会・技術供与・共同研究を通じて、技術的な検討を進めており、NBDCの実務者連絡会サイト (wg.biosciencedbc.jp) などで情報共有を行っている。

7. おわりに

4省の生命科学系データベースの統合は本稿で述べてきた通り、省間の連携体制は整ってきている

が、それぞれの省の中でカタログ・横断検索・アーカイブの適用範囲を広げていくことについては、まだまだやるべきことが多く残っている。これを達成するには、どのステップにおいても個々のデータベース作成者の自発的な協力を得られるか否かが鍵となるだろう。カタログではデータベースの所在情報が自発的に提供されればカタログ拡充は容易になる。横断検索ではサイトマップや使用IDの一覧がデータベースサイトで整備されていればインデックス作成

の負荷は減るし、データベース作成者側でインデックス更新作業ができれば検索結果を最新の状態に保つことも容易となる。アーカイブでもデータベース作成者側でダウンロードサイトが作られ、適切なメタデータや利用許諾が明示されれば目的のかなりの部分が達成される。データベース作成者の協力を得るためには、単に協力を要請するのではなく、統合のための規格を整備していくことが必要であり、われわれも一層の努力をしていきたいと考えている。

本文の注

- 注1) 文部科学省, 厚生労働省, 農林水産省, 経済産業省による生命科学系データベース統合のための合同ポータルサイト。省間連携の取り組みの紹介と, 成果となるデータベースカタログ, 横断検索, アーカイブなどのサービスを掲載。 <http://integbio.jp/>
- 注2) フリーソフトウェアの全文検索エンジン。N-gram方式により日本語文書に対しても漏れのない検索が可能な転置インデックスを作成できる。また, P2P型の分散処理によるインデックスサーバー間の連携機能をサポートしている。 <http://fallabs.com/hyperestraier/>
- 注3) DBCLSで開発された簡易データベース作成システム。表形式データのCSVファイルをアップロードするだけでデータベースを構築・公開できる。 <http://togodb.dbcls.jp/>

参考文献

- 1) 高木利久. 我が国のデータベース構築・統合戦略 第1回 データベースの現状と未来. 細胞工学. 2011, vol. 30, no. 10, p. 1096-1100. (原稿を改変した記事が<http://events.biosciencedbc.jp/article/01>から閲覧可能)
- 2) 白木澤佳子, 高木利久. ライフサイエンス分野のデータベース統合を目指して バイオサイエンスデータベースセンター (NBDC) の発足. 情報管理. 2011, vol. 54, no. 3, p. 144-151.
- 3) 生命科学系データベースアーカイブ. <http://dbarchive.biosciencedbc.jp/>, (accessed 2012-06-29).
- 4) National Research Council. A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases. The National Academies Press, 1999, p. 53. http://www.nap.edu/openbook.php?record_id=9692&page=53, (accessed 2012-07-05).

Author Abstract

Integration of life science databases scattered around Japan used to be tackled by individual related ministries. In order to make this an all-nation task, four ministries in Japan, MEXT, MHLW, MAFF and METI, started systematic inter-ministry collaboration in FY2011, and have been promoting database integration through the four steps: database catalog, cross-search, archiving and restructuring. This article introduces the

current state of these four steps, and mentions issues in each step, such as data sharing and RDFizing.

Key words

life science databases, database integration, database cross-search, database archive, data sharing, RDF, semantic web, ontology