1 分散データの統合とセマンティック Web

中尾 光輝
情報・システム研究機構
ライフサイエンス統合データベースセンター

片山 俊明
東京大学医科学研究所
ヒトゲノム解析センター

統合データベース

文部科学省委託研究開発事業である「統合データベースプロジェクト」（以下，統合 DB）では，生命科学研究を支える基盤整備として，ライフサイエンス分野の統合データベース構築を行っている．これまで，生命科学のデータベースは研究プロジェクトごとに散在しておりアクセス性が必ずしもよくない，時限制の研究プロジェクトでは完了後にそのデータが埋没しがち，統合的な利用のために必要な技術基盤の整備が立ち後れている，といった問題があった．このため統合 DB では，国内の完了した研究プロジェクトのデータの公開状況を調査し，データベースへの収録，共通の学術用語など辞書の整備，ライセンスや公開方法の整理などを行っている．統合的に利用可能なデータの種類と総量を拡大することにより，有用なデータの複合的な活用と再利用が促進され，データベースの組合せによる相乗効果が期待される．

統合 DB を推進するライフサイエンス統合データベースセンター（以下，DBCLS）では，独立に存在している多種多様なデータベースを統合し，これまでに蓄積された情報の利用価値を高めるために，２つのアプローチをとっている．1 つは先に挙げた既存データの収集業務であり，国内のリソースを統合したレポジトリを構築しさまざまなデータベースに対する横断検索が可能なポータルサイトを提供している．一方で，データベースや解析サービスを相互運用できる形で活用するためには，国内のリソースだけを統合しても十分ではなく，すでに広く使われている海外のリソースや技術的な規格を参考にしつつ，国際的にも相互運用性の高い統合のあり方を検討する必要があった．このため，もう 1 つのアプローチとして，分散するデータベースや解析サービスの Web サービスによる仮想的な統合を行うことにした．しかし，現状では海外においてもデータ型の標準化とサービスの規格化が依然として課題となっていることが分かってきた．

そのため，DBCLS では，平成 19 年度にデータベースと Web サービスの構築に携わっている実務レベルかつ最先端の研究者・技術者を国内外から招待し，Web サービスにおけるデータとプロトコルの標準化を目的とした国際的なソフトウェア開発会議として BioHackathon 2008 を開催した．この結果，国際的な協力体制が構築され，統合 DB においてもその成果が TogoWS などの形で結実し，統合のためのインフラとして提供されるようになった．引き続き，平成 20 年度には BioHackathon 2009 を開催し，ゲノム研究者など既存のインフラを利用した解析を日常的に行うユーザを交え，統合的な利用環境を構築するために必要な技術開発を行った．これら２回に及ぶ BioHackathon の議論から，データ量が爆発的に増大している現在の医学生命科学においては，単なる大規模データの統合と横断的な検索を超えて，統合されたデータから意味のある推論を行うためのセマンティック Web による知識化が必要，ということが明らかになってきた．なお，BioHackathon の経緯については本特集の序文である「編集にあたって」をご参照いただきたい．

本稿では，これまでの BioHackathon で議論されたテーマの中から，特に分散データの統合に関するトピックと今後のセマンティック Web への期待をこめて，現状の紹介と今後の課題についてレポートする．

既存のデータベースの現状

近年，ライフサイエンス分野においては次世代 DNA シーケンサの登場などからデータの大規模化への対応が

課題とされているが，物理学や天文学など他の大規模データを扱う諸分野と比較して，生命科学では扱うデータの量だけでなく種類の多様性が大きいことが特徴となっている．たとえば，Nucleic Acids Research 誌が毎年年頭に発行しているデータベース特集号では，毎年約 200 種の新規データベースが取り上げられており，今年までに 1,170 種類のデータベースが論文として発表されている．実際に利用可能なデータベースの数はこれだけにとどまらず，1,000 生物種を超えたゲノムのデータベース，細胞内の分子をさまざまな視点から整理した化合物や酵素のデータベース，細胞内プロセスに応じたタンパク質の修飾や輸送，遺伝子の発現量やタンパク質間相互作用のデータベース，細胞や個体の変異体ごとの表現型データベース，遺伝病など疾患にかかわるデータベース，進化系統解析のデータベースなど，実にさまざまなデータが世界中の研究機関から分散したリソースとして公開されている．

このうち，古くからデータが蓄積されており利用頻度の高い，DNA 塩基配列，タンパク質立体構造，遺伝子発現情報，文献，などのデータについては中央集権的なレポジトリが確立している．これらのデータベースは，「ライフサイエンスのデータは共有財産であり，科学の進歩にはデータの公開が不可欠」という理念から，研究者のコミュニティが論文出版時に公共データベースへのデータ登録を義務づけるよう出版社に働きかけた結果実現したものである．中央集権的なレポジトリが成立するためには，「税金による研究成果は国民に公開すべし」という行政的な指導とともに，データ登録を論文の出版とリンクするような研究者に対するインセンティブがうまく機能することが必要であった．ここではまず，これら主要なデータベースを多数収録してきた欧米の著名なレポジトリと国内の研究拠点で提供されているデータベースについて現状を簡単に紹介したい．

■ 海外のデータベース

欧米の主要サイトとしては，NCBI と EBI が挙げられる．米国の National Center for Biotechnology Information（NCBI）は，National Institutes of Health（NIH）傘下で図書館業務を担う National Library of Medicine（NLM）の下部組織として 1988 年に設立されて以来，塩基配列データベース（GenBank），文献データベース（PubMed），遺伝子発現情報データベース（GEO），ヒトや動物の遺伝病データベース（OMIM/OMIA），生物種系統（Taxonomy），化合物データベース（PubChem）など，さまざまな基盤データベースを整備している．これに対応するものとして，欧州では共同研究機関 European Molecular Biology Laboratory（EMBL）内にバイオインフォマティクス部門である European Bioinformatics Institute（EBI）が設立され，塩基配列データベース（EMBL），アミノ酸配列データベース（UniProt），遺伝子発現情報データベース（ArrayExpress），ゲノムデータベース（Ensembl）などを整備している．

このうち，塩基配列データベースについては International Nucleotide Sequence Database Collaboration（INSDC）による国際的な管理が行われており，NCBI の GenBank，EBI の EMBL，日本の国立遺伝学研究所の DDBJ の 3 極で相互運用されている．研究者がいずれかのセンタに登録した塩基配列は INSDC 間で相互に流通され，どのデータベースからも利用できるようになっている．GenBank/EMBL/DDBJ データベースには現在 1 億 6 千万エントリ，2,758 億塩基，ファイル容量にして約 1 テラバイト弱の塩基配列が登録されており，そのデータ量はゲノムプロジェクトの進展，次世代 DNA シーケンサの登場などから指数関数的な増加を続けている．一方，タンパク質の立体構造データベースである Protein Data Bank（PDB）も，現在では米国 Research Collaboratory for Structural Bioinformatics（RCSB），EBI の PDBe，大阪大学の PDBj の 3 極で国際的に管理されており，現在約 6 万個のタンパク質立体構造が登録されている．

■ 国内のデータベース

国内では，先に挙げた国立遺伝学研究所の DDBJ，大阪大学の PDBj のほか，京都大学の KEGG，理化学研究所（RIKEN），産業技術総合研究所（AIST）や生命情報工学研究センター（CBRC），かずさ DNA 研究所などで提供されているデータベースがよく使われており，さらに，文科省・経産省・農水省などの傘下でもさまざまなデータベースが提供されている．それぞれの内容を網羅することは困難であるが概要を示すと，KEGG ではゲノムの決まった生物種の遺伝子情報と酵素や化合物などのデータベースを統合した代謝系を中心とする生体内パスウェイのデータベースを，RIKEN では FANTOM などヒトとマウスのゲノム解析情報を中心としたデータベースを，AIST ではヒトゲノムのアノテーション情報として H-invitational データベースなどを，CBRC ではタンパク質の解析ツールや機能性 RNA のデータベースなどを，かずさ DNA 研究所では植物や藍藻のゲノム情報を中心としたデータベースを，それぞれ提供している．このほかにもイネゲノム，糖鎖や脂質のデータベースなど有用なリソースは国内にも多数存在している．いずれにしても，欧米の NCBI や EBI と比べ国内のリソースは統合化が進んでいない状況で，この改善を目指した統合 DB 事業の設立経緯に繋がっている．

データベースの統合化の試み

すでに挙げただけでも，多様なデータベースが各所に分散して管理されている現状が垣間見えたことと思うが，ライフサイエンスにおいては扱うべきデータが量的にも質的にも増加し多様化してきたため，すでに個々の研究者レベルでは必要なデータを統合的に扱うことが非常に困難になっている．これに対するアプローチとして旧来とられてきたのが NCBI や EBI などのようなポータルサイト型の統合化である．一方で，ポータルに収録されない多数のデータベースを仮想的に統合するアプローチとしては，交換データフォーマットやプロトコルの標準化と Web サービスの利用が試みられてきている．

■ ポータルサイト

ポータルサイト型のアプローチをとる NCBI と EBI では，さまざまなデータベースを統合的に検索するためのインタフェースや Web サービスを提供しており，世界中の研究者はこれらのリソースを自由に使うことができるようになっている．このようなポータルサイトの利点としては，一度の検索でさまざまなリソースを横断検索できることや，内部のさまざまなデータベース間で相互にリンクが張られており，遺伝子から立体構造や文献など関連データを容易に辿れる点が挙げられる．問題点としては，ポータルに統合されている情報は限られており，ロングテール的に存在するさまざまなリソースの分散状況は改善できないこと，ポータルを跨いで関連する情報を参照したい場合に，同一の生物学的オブジェクト（たとえば遺伝子）に対して各ポータルごとに異なる ID が振られており，対応をとる作業はユーザの負担になっていることなどが挙げられる．

■ TogoDB

統合 DB では，既存のポータルで吸収できない雑多なデータを統合的に集積するため，TogoDB というシステムを開発している．これは，データを所有している研究者がデータを公開するためのシンプルなプラットフォームを提供するもので，ユーザは表形式のデータをアップロードするだけで，簡単にアクセス性のよい Web データベースとしてデータの管理・共有・公開を行うことができる．最近になって増えてきているデータの公開形態として，論文のサプリメントデータがある．PDF や Excel ファイルとして出版社のサイトを通して公開されているデータであるが，これらはユーザがダウンロードして利用するしかないので，データのアクセス性が低かった．TogoDB を利用してこれらのサプリメントデータを公開することで，生命科学におけるデータのアクセス性の最低ラインを，単なるファイルの公開というレベルから，レコード単位での閲覧と検索まで高めることができデータの流通性が向上する．また，TogoDB は Ruby on Rails のプラグインとしてオープンソースで公開されているため，統合 DB のサーバで運用されているレポジトリとしての利用だけでなく，ユーザが自分のサイトを構築してデータベース公開する場合にも活用できる．この場合も，TogoDB では Web サービスのインタフェースを自動生成するため，分散サーバとして相互運用が可能である．

■ BioDAS

分散型のデータ共有では，ゲノムの座標軸を基準として，遺伝子の位置や EST の発現情報などのアノテーション（付加情報）を相互に情報交換する BioDAS が古くから利用されている．ヒトゲノムだけでもカリフォルニア大サンタクルーズ校の UCSC Genome Browser，EBI の Ensembl Genome Browser，NCBI の Map Viewer と，インターネット上にいくつものゲノムブラウザが提供されている．これらのゲノムデータベース間や，研究者独自のデータを相互に流通するための仕組みとして考えられたのが Distributed Annotation System（DAS）である．BioDAS では，ゲノム配列上の「何塩基目から何塩基目までの区間に遺伝子 A の 2 番目の exon が載っている」といった情報を XML で表現し，ゲノム上の指定された区間で該当する情報を取得するためのデータ型と CGI 呼び出しのプロトコルが規格化されている．並行して，ゲノム上のさまざまな要素（遺伝子構造，リピートなどなど）を表現する用語の標準化も進められており，Sequence Ontology（SO）として定義されている．すでに，多くのゲノムブラウザが BioDAS 規格に対応しており，DAS プロトコルでデータを提供している研究グループも多数存在する．このため，ユーザは自分の指向に合ったゲノムブラウザを用いて，追加的に閲覧したい DAS のデータソースを指定するだけでさまざまなゲノム情報を統合して可視化することができる．BioDAS の大規模な利用例としては，欧州の BioSapiens プロジェクトが挙げられる．これは，地域的に分散している実験生物学者が遺伝子アノテーションの追加管理を行い，計算機生物学者が DAS サーバを構築し随時公開する分散アノテーションの取り組みである．

■ SOAP/WSDL と REST

Web サービスで使われる規格は SOAP/WSDL と REST が主流となっている．SOAP は XML を介したメッセージ交換プロトコルである．インタフェースを WSDL

（Web Service Description Language）で機械可読的に定義し，プログラム言語の中ではメッセージ交換を抽象化したまま扱えるのが特徴である．REST は HTTP の基本機能を利用したシンプルなリソース交換スタイルである．BioDAS は広義の REST サービスであるが，基本的には CGI を利用した独自規格のプロトコルである．同様に，NCBI E-Utils も古くから利用されているサービスで，データベース検索，エントリ取得，配列相同性解析などを行う CGI 群からなる．これらの古典的なサービス以降，ライフサイエンスでは SOAP/WSDL プロトコルによる Web サービスが普及し，近年再び REST にシフトしている傾向が見られる．SOAP サービスは，欧米の NCBI や EBI のほか，国内でも DDBJ WABI，KEGG API，PDBj など早くから Web サービスに対応してきた．しかし，実際にはいくつかの問題があり必ずしも十分に普及しているとはいえないのが現状である．SOAP/WSDL による Web サービスは本来プログラミング言語に非依存であり，さまざまな環境からこれらのサーバの機能（データベース検索，解析など）を最新の状態で利用できる．この前提は，実際問題として，サーバで使われる SOAP のバージョンや各言語の SOAP ライブラリの実装に依存しており，Java で構築されたサーバの特定の機能が Ruby のクライアントでは利用できない，といったことが起こっていた．また SOAP で規定されていないセッション管理の実装については，ID を返してユーザに任せる場合や Cookie を使う場合などサーバによってさまざまな方法がとられていた．さらに，Web サービスから返された結果がそのまま他の Web サービスの入力に使える例は少なく，結果のフォーマットも，テキストで返される場合，独自の XML で返される場合，Base64 エンコードされて返される場合などバラバラで，必要な後処理も利用するサーバごとに異なっている．

■ BioMOBY

上述のような問題はあったが，SOAP/WSDL による Web サービスの普及で，さまざまなサービスを連携しワークフローを構築することが可能となり，解析手順の自動化が進むことが期待された．ライフサイエンスにおいては，Web サービスでやり取りされるデータ型もさまざまであったため，BioMOBY プロジェクトでは UDDI（Universal Description, Discovery and Integration）に先がけて，すべてのサービスをディレクトリ（MOBY Central）に登録しデータ型に応じて連携可能なサービスを検索するサービスディスカバリの仕組みを構築していた．しかし，BioMOBY は SOAP メッセージの中に独自 XML としてデータを埋め込む方法をとったため，BioMOBY に対応していない SOAP サーバとの連携ができないほか，BioMOBY ライブラリが提供されている Java と Perl 以外の言語からは利用できないといった問題点があった．さらに，新しいサービスとそこで利用されるデータ型のディレクトリへの登録は利用者に任せられていたため，似たようなサービスやデータ型が整理されることなく雑多に登録されていく結果となった．

BioHackathon 2008

以上のような背景から，Web サービス間での相互運用性を向上させるためには，必要なサービスの整理とやり取りされる交換データフォーマットの標準化が求められていた．このため，BioHackathon 2008 では，EBI, DDBJ, KEGG, PDBj, CBRC の Web サービス開発者，BioMOBY の開発グループ，ライフサイエンスのリソースを利用しやすくする Open Bio* ライブラリの開発グループ，まだ Web サービスに載っていなかった糖鎖やタンパク質間相互作用のデータベース構築グループ，Web サービスの連携によってワークフローを構築する Taverna, MOWServ, soaplab, G-language, Cytoscape などのクライアントソフトウェアの開発者，の 5 グループを一堂に会し，データ型の標準化と相互運用性の向上について検討した．

■ データ型と ID の統一

生命科学の Web サービスはサービス提供者ごとに独自開発されてきたため，塩基配列など意味的には同じタイプのデータを扱っている場合もデータの表現形式はテキスト，FASTA 形式，XML 表記など異なるものが多数乱立し，ほとんどのサービスがそのままでは相互に接続不可能であった．さらに，指し示す対象が同一（たとえばまったく同じヒトの ALDH2 遺伝子）であっても，使用するデータベースやサービスプロバイダによって異なる ID が振られているという問題点がある．

データ型については，前述のように BioMOBY で公開ディレクトリが存在していたが，BioMOBY に対応していない Web サービスも多く，また，ディレクトリへのデータ型の追加はユーザによって自由に行われていたため整理された状況ではなかった．BioMOBY のデータ型をそのまま採用することに対しては BioMOBY 以外の大手サービスプロバイダから抵抗が強く，対案として WS-I 標準の新規格の提案や，BioPerl, BioRuby, BioPython, BioJava の主要 Open Bio* ライブラリで共通に利用できるフォーマットの検討などが行われた．結果として，非常に多様なライフサイエンスのデータ型を網羅的に標準化するには至らなかったが，C 言語による Open Bio* 共通ライブラリの設計と，それに対するバ

インディングを構築するプロジェクトが始まったことと，すでに Open Bio* 共通の O/R マッパー（プログラム内のオブジェクトとデータベース内のデータの関係を扱うライブラリ）として利用されてきた BioSQL の拡張を行い，主要な配列データと系統樹などツリー型のデータについて，格納と取得のラウンドトリップを保証するための開発が行われた．

ID の統一はこれまでもライフサイエンスにおいて何度も議論されてきた課題であり，DNS のような仕組みを持つ Life Science ID (LSID) の提案などが行われてきているが，現時点ではまだ広く使われるに至っていない．代わりに，後述するセマンティック Web では Persistent URL（PURL）を ID として利用し，OWL（後述の Web Object Language）で名前空間の対応をとることによって解決する方向が模索されている．

■ Web サービスの構築と API

Web サービスのサーバ構築では，既存のデータベースに対するインタフェースを公開したもの，既存の解析用コマンドラインツールをラッピングしたものなどがあり，その API と結果のフォーマットは Web サービス化されるシステムの影響を受けている場合が多い．たとえばデータベース検索では，単に search といったものから getEntry や find_XXX_by_keyword のような API が統一感なく利用されている．また，既存のツールを Web サービス化した場合も，実行のための API は exec, run, do などさまざまでツールごとに WSDL が分かれている場合やツール名も引数で渡す場合など多様である．さらに，実行結果も本来ターミナルで人間が閲覧するために 80 文字幅で整形されたツール独自のテキストがそのまま返される場合が多く，取得した結果から必要な情報を抽出し次の解析フローに投入するためには，パースなどの処理が必要となる．また，これらの解析ツールの実行には数分から数時間かかる場合も多く，タイムアウト処理のためにセッション管理が必要となるが，前述のようにその実装方法はサーバによって異なっている．

ライフサイエンスにおける Web サービスの別の問題点としては，Web サービスのサーバが必ずしも安定運用されていないことも挙げられる．通常 Web サービスは自動化のために利用されるので，利用するサービスの死活管理がクライアントに任されている現在の状況は使いやすいとは言いがたい．さらに，バイオインフォマティクスで必要とされるタスクのうち Web サービスで提供されていないものがまだ多いこと，実行のたびに比較的大きなデータをクライアントとサーバ間でやり取りする必要があり，一連のワークフローを実行する場合に効率がよくない，といった問題点も指摘された．これらは，サービスプロバイダにとって Web サービス公開のメリットを増やしていくことと，サーバ間での連携をとることで今後解決していく必要がある．

■ Web サービスの連携

BioHackathon 2008 会期中に，新規の Web サービスとして，糖鎖とタンパク質間相互作用のグループによるデータベースと解析ツールのサービスが開発された．糖鎖のグループは，糖鎖オブジェクトを LINUCS フォーマットで取得し，RINGS を用いて類似する糖鎖を検索，取得した KEGG の糖鎖データベース ID から，GLYDE-II を利用して最終的に SVG（Scalable Vector Graphics）形式の画像を生成するフローを Web サービス化した．タンパク質間相互作用のグループは標準データ型として PSI-MI 2.5 フォーマットを採用し，相互作用データベース IntAct でこれらのデータを呼び出せる PSIQUIC 検索サービスが構築された．さらに，ネットワーク解析アプリケーション Cytoscape の Web サービス呼び出し機能に PSIQUIC への対応が追加され，取得した相互作用ネットワークの可視化が行えるようになった．

一方，国内の主要サービスである DDBJ, PDBj, KEGG の連携も検討された．BioHackathon 2008 はこれらのサービスの開発者が一堂に会する初めての機会であったため，まずは可能なワークフローの選定が行われた．結果として，タンパク質の機能をアミノ酸配列の類似性と立体構造から推定するために，(1) DDBJ のアミノ酸配列データベース DAD に対する BLAST プログラムによる相同性検索を行い，(2) 類似配列のアノテーションを取得，(3) アノテーションが得られなかった場合は BLAST 検索対象を PDB に拡張，(4) 類似構造を PDBj の Structure-Navigator で検索，(5) 類似構造のアノテーションを KEGG から取得する，という解析手順をワークフローエディタ Taverna を用いて設計した．この過程で相互運用性における課題に協調して取り組む必要性が認識され，今後も国内の Web サービス開発者間で継続的に連携していくこととなった．

■ クライアントアプリケーション

ワークフロー管理のための Web アプリケーション MOWServ では，ユーザに使いやすい形で BioMOBY や Web サービスの問題点を解決している．MOWServ はスペインのバイオインフォマティクス・グリッドで開発されており，BioMOBY のサービスとデータ型のオントロジーを専門家が見直すことによって整理し，統合的な解析環境を Web ブラウザ上に構築している．さらに，ワークフローの設計と状況把握，サーバ上での解析データの永続性，サービスの死活管理などの機能をサーバ側で

吸収することにより，相互運用性の高いサービスを構築している．

一方，EBI で開発されてきた Taverna は Java で作られたスタンドアローンのアプリケーションで，BioMOBY に限らずさまざまな Web サービスを連携したワークフローを GUI を用いて構築することができる．しかし，先の DDBJ, PDBj, KEGG におけるサービス連携の実証実験から，Taverna では分岐のあるワークフローが扱えないことや，出力結果のパースなどデータ型の不一致にはユーザが BeanShell スクリプトを記述して対応する必要があり，Java に不慣れなユーザには敷居が高いという問題があることが指摘された．

■ TogoWS

BioMOBY から Web サービスの連携までの議論で明らかになってきたように，現状ではライフサイエンスの Web サービスを統合的に利用する際に直面する問題点がいくつか存在する．DBCLS では BioHackathon 2008 の経験をふまえ，BioMOBY に準拠していない欧米の NCBI, EBI と国内の DDBJ, PDBj, KEGG の相互運用性を促進するため，統合 Web サービス TogoWS の開発を行っている．これらのサービスの内訳を見てみると，データベースの検索とエントリ取得を行うものがかなりを占め，残りは比較的計算時間のかかる配列比較や立体構造解析などのサービスであった．このうち，データベースのレコードは URL に容易にマッピングできるため，エントリの検索と取得には手軽に利用できる REST が向いていると考え，TogoWS では REST 型の Web サービスとして標準的な URL を提案しサービス提供を開始した．一方で，解析ツールの実行にはある程度計算時間がかかるほか，入力パラメータと出力フォーマットも複雑になりがちであるため，SOAP/WSDL の利用が適していると考えられる．しかし，既存の Web サービスには特定のプログラミング言語で使用できないといった問題があったため，TogoWS でプロキシサーバを運用し，主要プログラミング言語（Perl, Ruby, Python, Java）での動作確認を行うとともに，すべてのサービスのサンプルコードを作成し提供した．さらに，DDBJ, PDBj, KEGG の各 Web サービスの全メソッドについて稼働確認を毎日行い，稼働状況の記録を公開している．

TogoWS によるデータベース検索，エントリ取得，データ型変換は，それぞれ以下の形式で行えるよう統一されている．

http://togows.dbcls.jp/search/DB 名 / 検索文字列
http://togows.dbcls.jp/entry/DB 名 / エントリ ID
http://togows.dbcls.jp/convert/ 変換元 . 変換先

実際には TogoWS がデータベースごとに適切なサーバに問合せを行い，取得した結果をクライアントに返している．このため，ユーザは NCBI や KEGG などサービスプロバイダごとに異なるアクセス方法に悩まされることなく統一的にアクセスできるうえ，この URL を PURL として永続的な ID の代わりに利用できる．

しかし，このようにして得られるエントリのフォーマットはデータベースごとにバラバラであり，これまでは必要な情報を抽出するためには BioPerl や BioRuby など Open Bio* ライブラリを用いてプログラムを書く必要があった．TogoWS では BioPerl や BioRuby の機能をサーバ側に持たせることにより，エントリ中の特定フィールドの取得やデータ型変換を URL によって指定できるようになっている（図 -1）．これにより，Web サービスの連携で問題となっていた，BeanShell スクリプトの作成などプログラムによるパースや整形が必要な局面でも，それ自体を Web サービスで行うことができるようになった．

BioHackathon 2009

前回の BioHackathon 2008 や TogoWS の開発などを通じて，Web サービスの統合に一定の成果が得られたため，BioHackathon 2009 では，ゲノム研究者など既存のインフラを利用した解析を日常的に行うユーザを交え，統合的な利用環境を構築するために必要な技術開発を行うこととなった．このため，特にユーザが直接利用することになる解析インタフェースとして BioMart, Galaxy, Taverna, ANNOTATOR, FANTOM4 などの開発者を含め，大規模データへの取り組みや，セマンティック

NCBI GenBank に対してブタの p53 遺伝子を検索し最初の 10 件を取得
http://togows.dbcls.jp/search/ncbi-genbank/p53+pig/1,10
結果として得られた ID のリストから 1 エントリを取得
http://togows.dbcls.jp/entry/ncbi-genbank/6165622
このエントリを XML 形式で取得
http://togows.dbcls.jp/entry/ncbi-genbank/6165622.xml
このエントリを GFF 形式で取得
http://togows.dbcls.jp/entry/ncbi-genbank/6165622.gff
このエントリの配列の 3 〜 1163 塩基の領域を FASTA 形式で取得
http://togows.dbcls.jp/entry/ncbi-genbank/6165622:3-1163.fasta
このエントリの説明文を取得
http://togows.dbcls.jp/entry/ncbi-genbank/6165622/definition
このエントリの生物種系統情報を JSON 形式で取得
http://togows.dbcls.jp/entry/ncbi-genbank/6165622/source.json
BLAST プログラムの出力結果を GFF 形式に変換（データを POST する）
http://togows.dbcls.jp/convert/blast.gff

図 -1　TogoWS の REST URL による利用例

Web，テキストマイニング，可視化などのトピックについて，議論と開発が進められた．このうち，大規模データ，テキストマイニング，可視化については本特集でそれぞれ別個に取り上げられているのでご参照いただきたい．

ゲノム研究者などユーザに利用される環境の整備としては，BioMart と Galaxy の連携が進められたほか，国内の研究者に向けて DBCLS のグループにより Galaxy の多言語化（日本語対応）が行われた．さらに，TogoDB に登録された個々の研究者の小規模データベースを TogoWS の API を利用してアクセスするための仕組みを用意することで，TogoWS を経由することで Galaxy による解析が行えるようになってきた．

■ セマンティック Web への期待

一方で，さまざまなデータを横断的に利用するにあたり，それぞれのデータやサービスが持つ意味を明確にする必要性が出てきた．ユーザの手元にあるデータから，どのような関連データがあるかを提示したり，どのような解析サービスが利用可能であるかを示唆するような先進的な仕組みを構築するためには，セマンティック Web に対する期待が高く，BioHackathon でも RDF/OWL の利用や Web サービスのセマンティックアノテーションについて議論が行われた．

ここで，Resource Description Framework（RDF）は主語，述語，目的語のトリプルでメタデータを表現するものである．表現形式は XML, Turtle, N3 など複数あるが，Web での交換には通常 RDF/XML が利用される．しかし，RDF 自身では述語自体の説明や述語と他のリソースとの関連を表現できないため，RDF Schema（RDFS）が利用される．RDFS はクラスや属性について階層的な分類を行うことができる語彙である．さらにそれらの関係や推論には，オントロジー記述言語 Web Ontology Language（OWL）が利用される．OWL では新クラスを既存クラスの論理結合として表現するなどの機能が含まれており，表現力のレベルによって OWL-Lite, OWL-DL, OWL-Full の規格が制定されている．BioHackathon では，Web サービスの提供者がこれらに対する理解を深めるための解説も行われた．

■ セマンティックアノテーション

SOAP による Web サービスは入出力ともに XML によって行われている．この XML を RDF に置き換えると，データを意味論的に取り扱うことができ，推論や自動発見がたやすくなる．これを，Web サービスのセマンティックアノテーションという．最近になって，W3C は Web サービスのセマンティックアノテーション規格 SAWSDL をリリースした．これは，既存の Web サービスの入出力を外部データモデルとスキーママッピングルールで参照する方法を提供している．

Web サービスの相互運用性では，(1) シンタックス，(2) セマンティクス，(3) インタフェースの 3 つが重要となる．これらはセマンティック Web 技術における RDF, OWL, SAWSDL に対応する．BioMOBY のグループは「生命科学における Web サービスは入力と出力の生物学的関係を発見することである」と捉え，後継プロジェクトとして Semantic Automated Discovery and Integration（SADI）の開発を行っている．SADI は，入力を主語，出力を目的語，その関係を述語で表現するモデルであり，生命科学の Web サービスを再定義する意欲的な取り組みである．SADI の機能を利用している CardioSHARE は，RDF グラフ問合せ言語 SPARQL による問合せにサービスの自動発見を組み合わせることができる．いわば，問合せに応じて自動的に拡張する RDF グラフへの検索を実行することができる．

BioHackathon 2009 では，既存の Web サービスを SADI に載せる試みが行われた．SOAP サービスは SAWSDL を用いて容易に変換できたが，いくつかの問題点が明らかになった．DDBJ WABI サービスに対する適用例では，出力 XML の一部が非標準的な形式になっており，XSLT（XML Stylesheet Language Transformations）による変換ができなかった．この問題については，WABI を中継するサーバを立て，それに対して SAWSDL を用いることで対応した．TogoWS REST サービスは，WSDL に相当する機械可読なインタフェースの定義文章がないため SADI に対応することができなかった．この問題については，今後 REST サービスのための WSDL に該当する WADL を用いることで解決できるかもしれない．

セマンティック Web の現状と今後

ライフサイエンスにおけるセマンティック Web 技術の利用は徐々に浸透しつつある [1]．最初のステップは既存のデータリソースの RDF 化である．並行して，用語の標準化としてのオントロジーの整備が行われている．これらを元に，SPARQL による検索サーバの構築，ユーザインタフェースの向上，テキストマイニング技術の開発，可視化のためのソフトウェア開発などが今後の課題となってきている．

■ RDF によるデータの整備

W3C SWEO（Semantic Web Education and Outreach）Linking Open Datacommunity プロジェクトでは，自由に利用可能なさまざまなデータセットを RDF として収

集している．それらのリンクからデータ世界の全体像を描いた図を見ると，生命科学関連データを示すデータ群が 1/3 程度を占めているのが分かる☆1．このように，ライフサイエンスではすでに多種類かつ多数のデータが RDF で公開されていることが見てとれる．

先駆的な例として，タンパク質のデータベースとして最も参照されている UniProt は，2007 年から内部でのデータ管理に RDF を導入している．UniProt では，データ管理用のソフトウェアに合わせてデータの構造化をすすめたことによって，皮肉にもデータ構造が複雑化し，限られた資金や人員ではデータモデルや管理ソフトウェアの更新が困難になってしまった．この状況を解決するために，標準的な技術として RDF/RDFS と OWL を採用した．

また，Bio2RDF プロジェクトでは，NCBI, EBI, PDB, KEGG などさまざまな既存公共データベースの RDF 化に取り組んでいる．このプロジェクトのゴールは，インターネット上に分散したこれらのリソースをリンクしたグローバルなデータベースを構築し，現状では困難な SPARQL による問合せを実現するためのシステムを提供することであり，オブジェクト—関係データベース Virtuoso を利用した検索サーバが公開されている．

■ オントロジーの整備

一方で，多様なデータを分類整理するための共通用語であるオントロジーの整備も進められている．2000 年に論文発表された Gene Ontology（GO）プロジェクトでは，タンパク質などの遺伝子産物の機能を階層的に分類するための用語を管理している．当時，すでに真核生物のゲノムがいくつも決まっていたが，同じ機能を持つ遺伝子の説明記述が生物種間で統一されていなかったことから，biological process, molecular function, cellular component の 3 つの観点に分けて使用する用語を整理し共通化を図ってきた．

この動きは，ゲノム配列上の要素を分類する Sequence Ontology（SO）などに発展し，以後 Open Biomedical Ontologies（OBO）に引き継がれて，表現型，細胞種，分子修飾，発生，解剖学，疾病，環境などなど，さまざまな共通用語の蓄積と議論が進められている．

■ セマンティック Web の利用

このような背景のもと，セマンティック Web を実現した例がいくつか出てきている．NeuroCommons は ScienceCommons のプロジェクトの 1 つで，さまざまな公共データベースのデータとテキストマイニングした論文等の文献データを集積し，当初はアルツハイマー病など脳科学に関する知識の問合せを SPARQL クエリによって行えるシステムの構築を行っている．CellCycle Ontology プロジェクトでは，細胞周期制御のダイナミクスをセマンティック Web で表現することを目指しており，細胞分裂にかかわる遺伝子に関連するさまざまなデータを SPARQL によって検索できるシステムを構築している．BioGateway や RDFScape などのプロジェクトではセマンティック Web とシステムバイオロジーの融合を目指しており，将来的にはオントロジーをベースに生物学的な仮説を立て，実験を行って検証し，新たなデータからまた仮説を立てて検証する，といった生物学の流れが出てくることが期待されている．

日本でも，RIKEN SciNeS の PosMed でセマンティック Web に基づくシステムが構築されており，SPARQL を拡張して関連解析を行えるようにした GRASQL による問合せが可能となっている．これは，通常の SPARQL 検索で行われる RDF サブグラフの計算に加え，サブグラフ同士の関連の P 値を計算することにより関連するデータを効果的に提示するもので，ゲノム上の特定領域にある遺伝子群の中から病気などの表現型に関連する遺伝子のランキングに利用されている．

■ 今後の課題

セマンティック Web の先駆的な取り組みを概観したところであるが，現在のところ，データベース検索の多くはいまだ基本的にはキーワードによる横断検索である．多様なデータ型の混在やランキングに必要な情報量の不足により，検索結果の表示順を工夫するためのスコア付けが難しく，多数の検索結果の中から目的のデータに辿り着くための検索条件の指定は必ずしも容易ではない．また，検索対象が文献データベースの場合，検索結果として得られた論文リストから必要な情報を得るには，結局それぞれの論文を読んで取捨選択する必要があり大変時間がかかる．しかし，ゲノム解析の現場では，何千もの遺伝子ごとに類似の検索を繰り返すといった大規模処理が求められ，これらの問題点がボトルネックとなりやすい．この問題を解決するためには，複数のデータベースや検索条件を組み合わせた効率的な絞り込み検索，テキストマイニングによる文献データの事前処理，さらには，自然言語による問合せなどが可能になるとよいだろう．さらに，人手に代わる処理の効率化だけでなく，自動化の恩恵により思いもしなかった意外な関連データまでが提示されインスパイアを与えてくれるような思考支援システムが理想かもしれない．

☆1 W3C SWEO Linking Open Data community, http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/ http://www4.wiwiss.fu-berlin.de/bizer/pub/lod-datasets_2009-03-27_colored.png

ライフサイエンスの分野では，すでにデータの種類と量が研究者個人で把握できる範囲を超えてきており，多種多様なデータを効率的に組み合わせたマイニングが新しい知識発見につながると期待されている．このような知識管理システムの構築には，データへの意味付け，そのための用語体系の利用，意味論的な問合せ機能が必要となるが，これらはセマンティック Web 技術の進展で解決されていく可能性がある．しかし，現状では生物学者が SPARQL クエリを書く必要があるなど問合せのためのインタフェースに問題があるほか，データ量に対してシステムがスケールしないという大きな問題もあって，ごく限られた局面でしか利用されていない．測定データと既存の知識から推論する場合，生物学の複雑かつ不完全なデータ，たとえば部分的に欠けている時系列データの取り扱いなどに注意する必要があるほか，測定条件の違いや測定誤差によって異なる推論結果が導かれることもあるため，それぞれのデータの由来やバージョン情報などのメタデータをきちんと扱えるような仕組みを考える必要がある．さらに，問合せ結果の可視化やシミュレーションのためのソフトウェア開発も課題である．セマンティックズームやデータセレクタのような機能を持ち，大量の情報の中からコンテクストに依存して関連データを効果的に提示する，いわばセマンティックナビゲーションが必要となるだろう．いずれにしても，人工知能の黎明期と違って，大規模な計算資源と大量の電子化されたデータが利用可能な現在，このような目的にセマンティック Web がいくつか解決策をもたらしてくれるのではないだろうか．

次の BioHackathon が開催できるなら，生命科学におけるセマンティック Web の実現による高度なデータベース統合化を目指したい．具体的には，(1) 国内外のデータベースリソースをセマンティック Web の枠組みで扱うための基盤技術整備，(2) 統合と推論を促進するためのオントロジーと関連データの整備，(3) コンテクスト依存の可視化およびユーザインタフェースの整備，の３点について重点的に議論を行い，その結果をそれぞれのソフトウェア開発プロジェクトに反映させることができれば，と考えている．

参考 URL

1) 統合 DB/DBCLS : http://dbcls.jp/
2) BioHackathon 2008, http://hackathon.dbcls.jp/
3) BioHackathon 2009, http://hackathon2.dbcls.jp/
4) NCBI, http://www.ncbi.nlm.nih.gov/
5) EBI, http://www.ebi.ac.uk/
6) DDBJ, http://www.ddbj.nig.ac.jp/
7) PDBj, http://www.pdbj.org/
8) KEGG, http://www.genome.jp/kegg/
9) RIKEN, http://www.riken.go.jp/
10) CBRC, http://www.cbrc.jp/
11) TogoDB, http://togodb.dbcls.jp/
12) UCSC Genome Browser, http://genome.ucsc.edu/
13) Ensembl Genome Browser, http://www.ensembl.org/
14) BioDAS, http://biodas.org/
15) BioSapiens, http://www.biosapiens.info/
16) BioMOBY, http://biomoby.org/
17) BioPerl, http://bioperl.org/
18) BioRuby, http://bioruby.org/
19) BioPython, http://biopython.org/
20) BioJava, http://biojava.org/
21) RINGS, http://rings.t.soka.ac.jp/
22) GLYDE-II, http://glycomics.ccrc.uga.edu/GLYDE-II/
23) IntAct, http://www.ebi.ac.uk/intact/
24) Cytoscape, http://www.cytoscape.org/
25) MOWServ, http://www.inab.org/MOWServ/
26) Taverna, http://taverna.sourceforge.net/
27) TogoWS, http://togows.dbcls.jp/
28) BioMart, http://www.biomart.org/
29) Galaxy, http://main.g2.bx.psu.edu/
30) SADI/CardioSHARE, http://sadiframework.org/
31) UniProt RDF, http://dev.isb-sib.ch/projects/uniprot-rdf/
32) Bio2RDF, http://bio2rdf.org/
33) OBO/GO/SO, http://www.obofoundry.org/
34) NeuroCommons, http://neurocommons.org/
35) Cell Cycle Ontology, http://www.cellcycleontology.org/
36) PosMed, http://omicspace.riken.jp/PosMed/

参考文献

1) Antezana, E., Kuiper, M. and Mironov, V. : Biological Knowledge Management : the Emerging Role of the Semantic Web Technologies, Brief Bioinform, Vol.10, No.4, pp.392-407 (2009).

（平成 21 年 7 月 22 日受付）

中尾 光輝
mn@dbcls.jp

ライフサイエンス統合データベースセンター特任研究員．博士（理学）．日本分子生物学会，ISCB，JSBi 各会員．オープンバイオ研究会実行委員．

片山 俊明
ktym@hgc.jp

東京大学医科学研究所ヒトゲノム解析センター助教．中尾氏らと BioRuby プロジェクトを設立 (http://bioruby.org/)．オープンバイオ研究会主宰 (http://open-bio.jp/)．最近はクマムシのゲノムを解析中 (http://kumamushi.org/)．