参考資料（８）

ライフサイエンス分野の統合データベース整備事業

ライフサイエンス統合データベース基盤整備
１８年度　研究成果報告書

平成１９年３月

大学共同利用機関法人　情報・システム研究機構
独立行政法人　科学技術振興機構
国立大学法人　九州大学

本報告書は、文部科学省の委託業務として、大学共同利用機関法人情報・システム研究機構、独立行政法人科学技術振興機構、国立大学法人九州大学が共同で実施した、平成１８年度の「ライフサイエンス統合データベース基盤整備」を取りまとめたものです。従って、本報告書の複製、転載、引用等には文部科学省の承認手続きが必要です。

目　次

１．プロジェクトの目的 ····· 2
２．データベース統合戦略立案および評価 （情報・システム研究機構）····· 2
２．１　データベース統合戦略立案および評価の実施計画 ····· 2
２．２　データベース統合戦略立案および評価の実施内容 ····· 2
２．３　データベース統合戦略立案および評価のまとめ ····· 8
３．データベース統合化基盤技術開発（情報システム研究機構、九州大学）····· 14
３．１　データベース統合化基盤技術開発の実施計画 ····· 14
３．２　データベース統合化基盤技術開発の実施内容 ····· 14
（１）基盤知識表現技術開発（情報・システム研究機構） ····· 14
（２）癌研究知識表現技術開発（情報・システム研究機構） ····· 22
（３）多型知識表現技術開発（九州大学） ····· 23
（４）キュレーター支援技術開発（情報・システム研究機構） ····· 23
３．３　データベース統合化基盤技術開発のまとめ ····· 24
４．ポータルサイトの構築 （科学技術振興機構）····· 27
４．１　ポータルサイト構築の実施計画 ····· 27
４．２　ポータルサイト構築の実施内容 ····· 27
４．３　ポータルサイト構築のまとめ ····· 28
５．人材の育成 （情報・システム研究機構）····· 29
５．１　人材の育成の実施計画 ····· 29
５．２　人材の育成の実施内容 ····· 29
５．３　人材の育成のまとめ ····· 32
６．プロジェクトの総合的推進 （情報・システム研究機構）····· 33
６．１　研究運営委員会及び統合 DB 整備戦略作業部会 ····· 33
６．２　教育プロジェクトに関するミーティング ····· 34
７．プロジェクトの成果のまとめと評価 ····· 35
８．成果の外部への発表 ····· 35
９．実施体制 ····· 35

（注）フッターにある（ ）付き番号は、
参考資料内のページ番号です。

１．プロジェクトの目的

より多くのライフサイエンス研究者等がいわゆるゲノムプロジェクト・ポストゲノムプロジェクトの成果や多様なDB等をストレスなく利用でき、より高度な研究ができる環境の実現とその持続可能化のために，情報・システム研究機構、九州大学，科学技術振興機構が共同して，以下の４つの業務を行う。

（１）ライフサイエンスおよび知識情報処理の識者にライフサイエンスDBの専門家を加えた研究運営委員会を組織し､十分な情報の収集･分析に基づいて統合化戦略を立案する。

（２）統合化に不可欠な知識表現法、情報共有技術および文献からの知識抽出技術等の開発に着手し、順次戦略立案のための調査やポータルでの利用者誘導に適用する。

（３）DBのカタログおよび解析ツールなどのWEBリソースのカタログを作成し、利用者を目的にかなったDBやWEBリソースに誘導するポータルサイトを構築する。

（４）データベースの構築維持に不可欠な人材の育成に努める。

２．データベース統合化戦略の立案および評価

２．１　データベース統合化戦略の立案および評価の実施計画

DB 統合化の戦略は１）利用者であるライフサイエンス分野の状況、２）素材である個別データベースの状況、３）利用可能な情報技術、の３つの動向を常に考慮しながら継続的にかつ柔軟に立案されねばならない。ここでは3分野からの専門家を集めた研究運営委員会を組織し、同委員会管轄下に調査組織（統合DB整備戦略作業部会）を設ける。

調査組織は、

（１）ゲノム注釈とデータベース間の連携における課題

（２）国内外のDBの俯瞰と質的量的比較

（３）ライフサイエンス分野の研究の俯瞰調査

（４）検索アルゴリズムを含めた知識情報技術の動向調査

（５）臨床情報や医療統計の現状調査

を行い、上記委員会に適宜報告しながら戦略立案を支援する。

なお、本テーマは情報・システム研究機構で実施した。

２．２　データベース統合化戦略の立案および評価の実施内容

ライフサイエンス、知識情報処理、ライフサイエンスデータベース（DB）の 3 分野の専門家による研究運営委員会を組織し、各々の分野の動向に即したDB整備戦略について議論した。上記委員会に対し、3 分野の動向に関する俯瞰データを提示し、適宜与えられる調査課題に対し答えることのできる体制（統合DB整備戦略作業部会）を構築した（研

究運営委員会並びに戦略作業部会の議論の詳細は、６．プロジェクトの総括的推進の項に記載)。また、この体制を利用して戦略立案の基盤となる情報の収集・分析を、以下に示すように行った。

（1）ゲノム注釈とデータベース間の連携における課題

代表的モデル研究植物であり、全ゲノム塩基配列が決定済みである「イネ」ならびに「シロイヌナズナ」のゲノムアノテーション型公開データベースの基本項目を調査し、それぞれのデータベース間の連携と課題の整理を実施した。

「イネ」のデータベースについては、Nucleic Acids Research のデータベース特集、国際イネゲノム塩基配列プロジェクト(IRGSP)及び農業生物資源研究所(NIAS)のデータベース検索 Web サイト、PubMed 検索で rice 及び database キーワードにして、AND 検索した結果から重複を除いて得られた、46 種類を対象とした。「シロイヌナズナ」のデータベースについては、同じく Nucleic Acid Research のデータベース特集、国立遺伝学研究所、かずさ DNA 研究所、理化学研究所ゲノム総合科学研究センターのデータベース検索 Web サイト等から重複を除いて得られた、25 種類を対象とした。

調査項目としては、エントリー構成、エントリー数及び、一次データと二次データの区別、関連データベースや一般的公共データベースへのリンクの構成、画面レイアウトや表示ツール、用意されている解析プログラム群といった Web インターフェース構成、定期リリースの頻度やバージョンや過去のバージョン参照可能かどうかといったデータベースアップデートの状況、OS や管理プログラム、マシンスペックや開発言語といったデータベース管理システムに関する事柄をベースに、発表論文や開発者への連絡方法、見易さ、応答性、信頼性などの観点からの検索結果表示の問題といった点まで含めた。

同時に、主として実験生物系のデータベースの利用者を対象に、「これらのデータベースのなかでよく利用するサイトはどこか」、「複数サイトを利用する場合に困っている点はないか」など、聴き取り調査と郵送によるアンケート調査を（全 188 名を対象に）実施した。これによって、主として実験の現場でデータベースを活用している研究者が抱いているゲノムベースのデータベースの連係に関する現状の課題と、将来のデータベース統合にむけた要望を調べ上げた。

調査結果から、データベースのよりよい統合化は、以下のような比較的多数のユーザが抱く不満を解消する方向で行うべきであることを読み取ることができた。

1）DB 作成の時間差や異なる収集方針による遺伝子名や ID の相違が多く混乱の元になっている。これを吸収あるいは関連付ける基盤サービスが必要。

2）論文掲載情報や利用者からのフィードバックが直ちに反映されないことへの不満も大きい。

3）誤りの多さを不満とする声がある一方、仮想遺伝子にもなんらかのヒントが欲しいという要望も多い。提供するデータの分類や格付はできないか。

４）植物の分子の研究に於いては、頻繁に生物種横断的な検索や比較を行う。そのような情報が取得できるサイトがない。

反面、個別に現状のデータベースをみた場合の使いやすさや内容の充実度に関しては約半数が肯定的であり、将来の統合化データベースの作成にあたっては、現在、利用頻度の高い個々のデータベースが保持している有用な情報を活かしつつ、齟齬を解消し連係させる形での統合化を考えていくべきであるとの結論が得られた。

（２）国内外 DB の俯瞰と質的量的比較

科学技術連携施策群の生命科学データベース統合に関する調査研究と JST-DB（WING）の情報の調査を行い、主に分子に関する網羅性の高いデータベースカタログを生成した。データベースカタログの各エントリーには、データベース型分類情報を追加し、主要なデータベースに関する日本語解説を整備した。その結果、分子情報に基づくデータバンクには、索引はあるが目次が欠如していることが多いことが明らかとなった。そこで、一般の研究者でも、分子レベルの生物学研究の現状を容易に俯瞰できるようにすることを目的として、各種データバンクの内容を目次的に表現しようと試みた。これによりデータバンクの内容について、個々のバンクを区別することなく、自在に一次データを引き出し利用することも可能になると考えられる。今年度は INSDC(International Nucleotide Sequence Database Collaboration)が管理する DNA バンクと遺伝子発現バンク（GEO、Gene expression omnibus）について総合データ目次を作成した。これにより、目次項目ごとのデータのダウンロードが可能になった。

１）DNA バンク目次

DNA 配列読み取りをおこなった論文では、論文投稿時に INSDC への DNA 配列の登録が義務付けられている。従って数十塩基の配列から完全なヒト染色体の配列まで、学術論文で新規に報告された DNA 配列は全て INSDC に登録されているはずである。ここでは、INSDC に分類保管されている DNA 登録を登録の背景にある研究（プロジェクト）単位にグループ化し、研究対象や研究目的別に細分類した。これにより、生物種ごとに、さらにその研究対象ごとに、登録レコード数が多い代表的なプロジェクトを把握し、その配列データをまとめてダウンロードすることが可能になった。さらに、タイトルを日本語化することにより、プロジェクトの内容が一目分かりするようになった(図 2.2.1 参照)。

２）遺伝子発現バンク目次

NCBI が提供する GEO はマイクロアレイや SAGE などの遺伝子発現に関する実験結果を集積した、遺伝子、サンプル、値の三つ組みデータを対象としたデータバンクの代表である。進展の早い実験領域の１次データバンクは得てして利用者には難解である。そこで、少しでもデータが利用しやすくなるように、データの整理パイプラインを作成し、DNA

図 2.2.1 DNA バンク目次の表示例

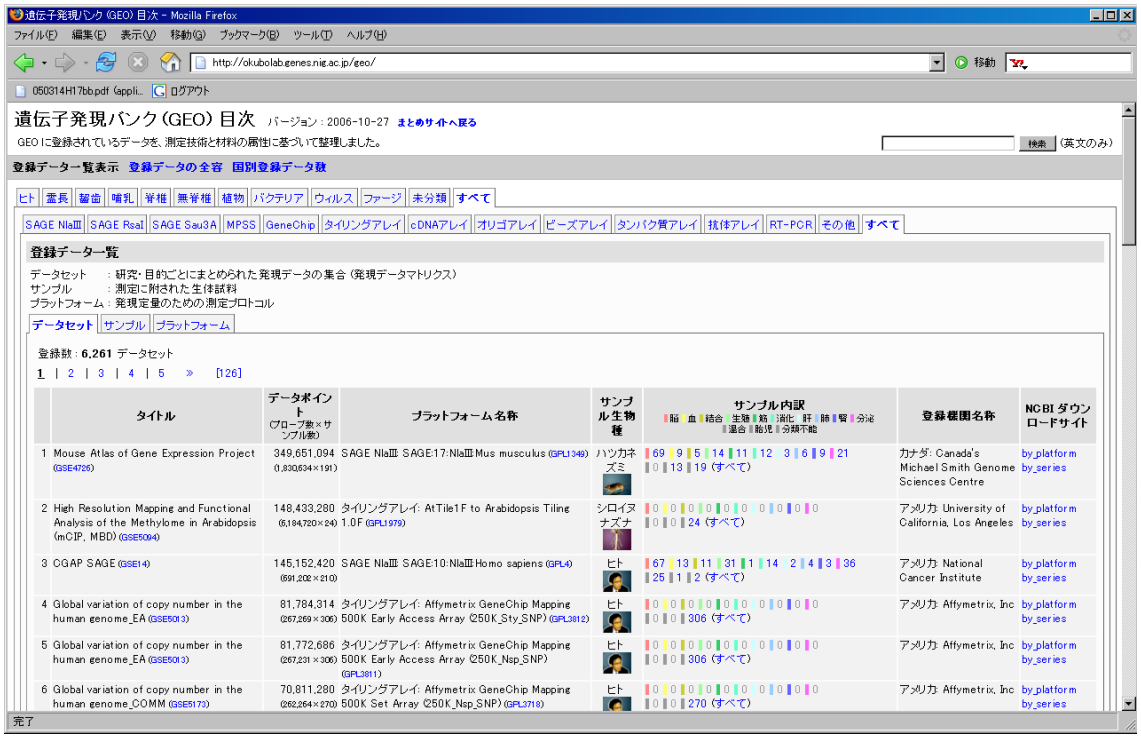

図 2.2.2 遺伝子発現バンク目次の表示例

バンク同様に研究対象や研究手法に基づく目次を作成した。これにより、どんな生物のどんな実験データが登録されて利用可能なのかが、閲覧可能になった(図 2.2.2 参照)。

（3）ライフサイエンス分野の研究の俯瞰調査

国内の研究を俯瞰するための情報源として各種学会の過去の抄録を統合し、検索や、施設別やテーマ別の再編成が可能なDB化を目的として、今年度は分子生物学会８年分の書誌事項に加え一部要旨を電子化し、施設名称など基本的な用語の統一を行い、要旨の検索、ソート、可視化が可能なシステムを開発した。本システムを用いて、要旨中に含まれるキーワードによる要旨の分類やその年次推移を観察することが可能になり、今後の分野の動向を逐次的に調査把握していくことが可能になった。また、同じ目的で日本語総説誌バックナンバー全文電子化作業の一環として、蛋白質核酸酵素のバックナンバー（約10年分）の電子化を行った。以上の検討により、日本語文献のオープン化、データベース化の重要性を認識できた。

さらに､ライフサイエンスの知識を俯瞰するためのデータや知識の整理法を開発する目的で、動物の脳に関する機能的、形態的、分子的、進化的なデータ・知識を集め、教科書的な知識と最新の知見をおりまぜて伝えるシステムの構築を行った。1,000個の脳細胞に対する40個の遺伝子情報のデータベース化と、細胞単位で発現を視覚的に把握することができるビューワー機能の構築により、脳細胞と遺伝子情報の相関性の3次元的な表現が可能になった。これにより、DB統合化の際の表示機能の重要性を把握できた。

（4）検索アルゴリズムを含めた知識情報技術の動向調査

生物情報を扱うデータベースは、いろいろな分野の異なる観点から作成され、また、それぞれ異なる形式で記述されている。また生物情報データベースに含まれるデータ量は計測技術の発展に伴い膨大な量となってきている。また、High Wire Press や PubMed Central をはじめとした文献の電子化・オープン化が進むことで、生物情報として利用可能なテキストや図表の量も増えつつある。さらに、これら生物情報の利用方法自体、ユーザによって様々である。このような背景をもつ生物情報のデータベースを統合するためには、高度な知識情報技術の利用が不可欠である。そこで、次世代の生物情報データベース統合に必要となる知識情報技術として、検索システム、データマイニング、Web 2.0 およびグリッドコンピューティングに焦点を絞り、聞き取り調査や文献調査によって動向を調べた。併せて、統合データベースに必要とされる計算機資源（CPU 数、ディスク容量など）を、知識情報技術の利用の観点から予測し、効率的な計算機環境を整えることを目的に、統合データベースに関する計算機資源の調査を行った。

検索システムについては、単純な項目検索やキーワード検索では、ライフサイエンス分野の、情報の膨大さ多様さゆえに対応が難しく、検索エンジン自体に高度な解析機能、可視化技術が必要になってきている。また、対象がグローバルなWWW上のデータへと広が

ったため、ローカルなデータベース内の構造化されたデータのみならず、WWW上の非構造化データをも扱える必要が出てきた。そこで、膨大かつ多様な情報へ対応する検索技術、および構造化データと非構造化データをともに扱う技術について調査した。

データマイニングとは、大規模なデータやデータベースから隠れた関係性や知識などの情報を帰納的に抽出する技術を指す言葉である。データマイニング手法は出力される情報の方向性と入力されるデータの種類から、おおまかに第一世代と第二世代のものに分けることができるが、第二世代のデータマイニング手法には、ベイジアンネットワーク、隠れマルコフモデルなどの確率モデルや、グラフマイニングなどの構造データからのマイニング手法、さらに、テキストマイニングやストリームマイニングなどの新しいタイプのマイニング手法が含まれる。この調査では、第二世代のデータマイニング、特に構造データからのマイニング手法について調査を行った。

Web 2.0 とは従来の WWW における静的なサービスに対し、次世代にあるべき新しいウェブのあり方に関する総称である。Web2.0 の特徴を持つタームとして、ここでは、web service、ロングテール、集合知、タグ付け、ブログについて調査し、さらに、これらとデータベースの関係について考察した。

グリッドコンピューティングは、元々は遊休計算機資源を有効に活用するために作られた仕組みだったが、現在は、大規模計算を効率よく行うための仕組みとして利用されている。このグリッドコンピューティングの目指す環境を実現するための様々な課題、例えば利用する計算機が別組織に属したり、そのプラットホームがばらばらであっても動的に連携できる仕組みの構築、を解決する必要がある。ここでは、こうした課題の解決策について調査した。

計算機資源の調査については、文献、インターネット、及び、公開されているプログラムの計算速度、及びメモリー使用量、ディスク使用量を計算し、また、今後、新たに利用されると思われる技術については、他分野での同技術を用いているプログラム速度を参考として見積もりを行った。Web クローリング、テキストマイニング、フェノタイプ情報の画像処理・画像検索、配列バンクの項目別検索および多型情報を含む配列解析などについて、必要とされる技術の調査を行い、それぞれで必要となるメモリー量、CPU 数、ディスク容量を概算した。

これらの調査によって、従来の検索技術には情報の膨大さと多様さに基づく限界がすでにきており、データマイニング技術を検索エンジンへうまく組み込む必要があることが分かった。また、グリッドコンピューティングを始めとした分散計算技術は、web service を前提としており、必要な web service をデータベース側で揃えていく事が今後より重要となることが分かった。また、知識情報技術を活用するために必要となる計算機資源の見積り根拠を得ることができた。

（５）臨床情報や医療統計の現状調査

臨床情報の調査に関しては、HL7等の標準規格の役割、データ抽出のためのカルテデータの電子化の状況やインセンティブ等につき、インタビューを含め調査を行った。また、生活習慣病を中心とした我が国のコホート研究の事例を分類・整理した。医療統計の調査に関しては、遺伝子多型解析に関わる遺伝統計学に焦点を絞り、遺伝統計学分野で用いられる解析技術に関して、インタビューを含め調査を行った。

我が国のコホート研究については、循環器疾患を対象とした大迫(おおはさま)研究、生活習慣病その他の種々の疾患を対象とした山形大学の地域特性を生かした分子疫学研究(21 世紀 COE プログラム)、地域住民、大都市検診を対象とした多目的コホートによるがん・循環器疾患の疫学研究、広範な疾患を対象とした久山町研究、癌、循環器疾患を対象とした放射線影響研究所コホート研究、虚血性心疾患を対象とした都市勤労者集団コホート、高血圧を対象とした端野・壮瞥町研究、一般住民の循環器疾患を対象とした NIPPON DATA 80、全国各地で行われている循環器コホート研究の個人データを統計的に統合し、リスク因子を定量的評価することを目的とした JALS (Japan Arteriosclerosis Longitudinal Study) について、その研究の背景と目的、対象地域、ターゲット疾患、特徴的な検査項目、対象人数、代表研究者、研究開始時期、資金源等について調査した。

遺伝統計学分野で用いられる連鎖解析、連鎖不平衡解析(ハプロタイプ解析）、QTL 解析等の解析手法と、それぞれの手法における代表的なアルゴリズム計 8 種類の調査を行ない、その特徴および長所・短所を評価した。併せて、各手法の代表的プログラム計 15 種類に関して、実装されているアルゴリズム、動作環境、入出力、利用形態、ダウンロード先などを調査し、その評価を行った。また、代表的な商用ソフトの 2 種類の機能、特徴などを調査した。

これらの調査から、医学データ活用における課題として、病名の標準化、前向きコホート研究の推進、人類遺伝学基盤の充実が重要との結果を得た。病名の標準化に関しては、現状は死因統計などの主として保険行政統計用のものか保険診療用のものしかなく、臨床研究向けには使いにくく、抜けもあり、また必ずしも真の病名が記載されない、といった問題がある。前向きコホート研究については、既存のカルテの活用（後向き研究）は、検査値などを除くと難しく、しっかりデザインされた一定規模の前向き研究によって初めて有効なデータが得られることが分かった。また、米国では家系を集めるプロジェクトにも多額の投資がなされているのに対して、日本では家系データが軽視される傾向にあり、これは日本における人類遺伝学基盤の不十分さに起因することが分かった。

２．３　データベース統合化戦略の立案および評価のまとめ

我が国のライフサイエンスＤＢは約２５０あると言われるが、統合化の目的はこれらバラバラに管理運営されている多種多様なＤＢを一つにすることであると一般に考えられている。もし、仮にこのような統合化が実現されるとすると、利用者にとっては、検索、仮説生成、解析が容易になり生産性が向上する。ＤＢ管理者にとっても管理運営がトータル

に効率的されるであろう。

しかしながら、上に述べたような「一つの統合 DB」は現実的に実現困難で有用性も低い。その理由は、分子データの統合化だけでは不十分であること、データの解釈や意味が研究の進展によって変化するため、ある時点で統合化を実現してもそれがずっと有意義かどうかは保障されないこと、ライフサイエンスの最先端はテキスト（論文）で表現されるため、それらとの連携が十分でないと最先端の知識がＤＢに反映されないこと、データをどう眺めたいかは研究者によって異なるため一つの視点で統合化を図っても多くの利用者の満足は得られないこと、などが挙げられる。

そこで、統合ＤＢ構築は完結しないプロセスであると認識し、研究の進展に応じて利用者の求めるものが変化することに柔軟に対応したＤＢを構築することが重要である。一つの統合ＤＢ実現は研究開発の生産性向上のためのあくまでも手段であるので、それを自己目的化することなく、いかにして研究開発の生産性向上を目指すかという原点に立ち返って方針を作成すべきである。

研究運営委員会、統合 DB 整備戦略作業部会での議論、及び上記の調査活動の結果に基づき得られた、統合 DB 構築の際の目指すべき課題、取り組みの際の基本的考えを以下に列挙する。

・ＤＢ構築者ではなく利用者の思考や意思決定を支援する DB を構築する
・利用者の興味、知識に応じて必要な情報、判断材料をもれなく提示する
・複数 DB をつなぎ異種データ・知識の関係が俯瞰（仮説生成）できるようにする
・いろいろなツールを簡便に組み合わせて解析（知識発見）できるようにする
・できれば、上記のことが日本語で行えるようにする
・ＤＢ化（構造化）されないもの（ 論文（テキスト、ポンチ絵、画像)、特許、教科書、報告書、解説記事、など）もうまく扱えるようにする
・ＤＢ構築に最先端の研究開発は必要だが、あくまでも利用者の利便性向上のためのサービス事業であることを認識する

以上の基本的な考えをベースに得られた、文部科学省が目指すべき統合データベースの概要は以下のとおりである。

まず、想定ユーザとしては、

・ライフサイエンスの研究者および医療、創薬などバイオ産業従事者
・ライフサイエンスプロジェクトの企画立案や評価に関わる人々
・ライフサイエンスデータベースの構築者

とする。

統合化の対象データとしては、ヒト・動植物・微生物の分子データ、文献データ、臨床などの表現型データを考える。

開発すべき提供機能としては、

・DB やツールの所在や利用法を網羅したポータルサイト
・分子データと文献知識（高次生命機能）を統合したデータベース
・分野の俯瞰や仮説生成が容易に行える検索機能
・DB 構築者へのインデックス、辞書、整理棚、書式、DB 構築ツールの提供
・データベース構築者（キュレータ・アノテータなど）の学習用教材
・上記の日本語による検索と表示

が必要との結論を得た。

なお、数年後に上記の統合データベースを実現させるための計画に関しては、図 2.2.3 に示すようなステップをとることが適当と考える。また、年次ごとに想定される具体的実施事項および成果を図 2.2.4 に示す。

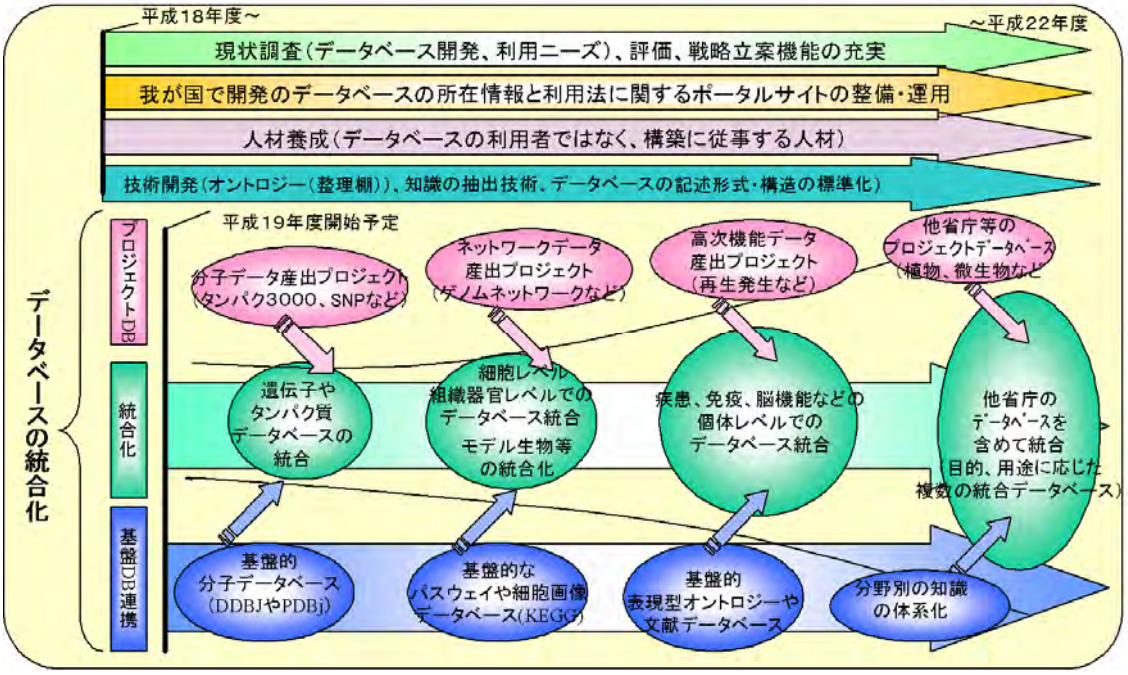

図 2.2.3 統合データベース事業展開の年次計画

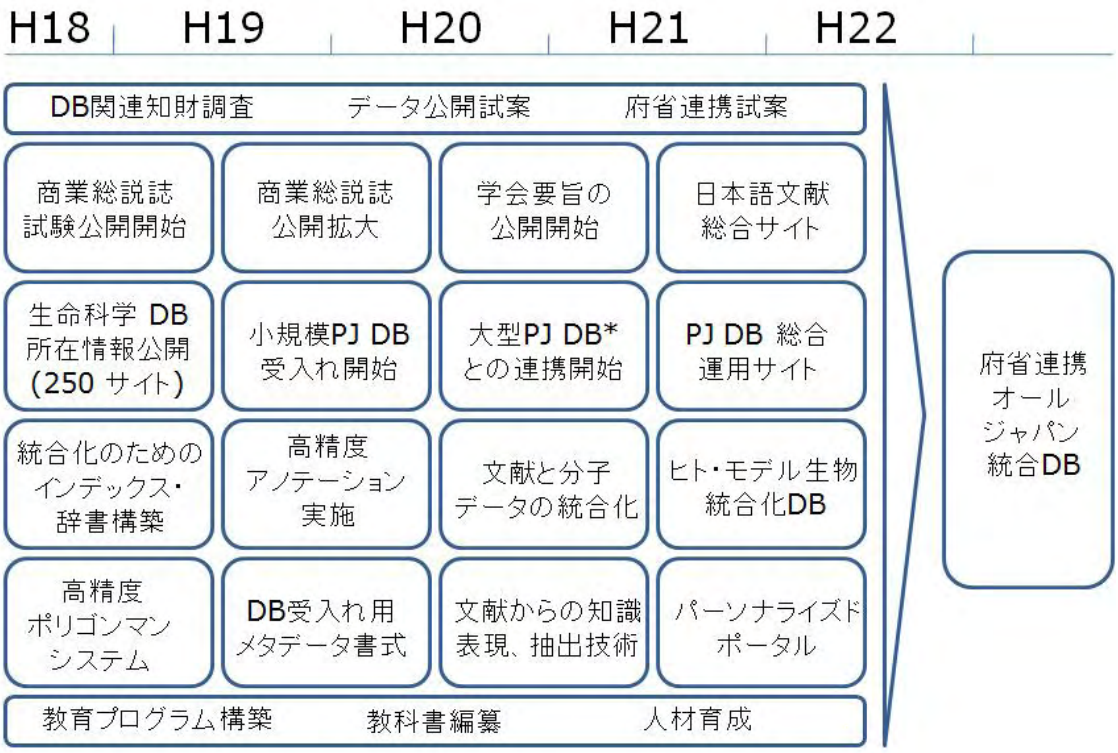

図 2.2.4 年次ごとに想定される具体的実施事項および成果

さらに、上記のあるべき姿の統合データベースを具現化する具体的な取り組みとして、以下の提言を行う。

（1）戦略立案・実行評価

統合 DB 構築はライフサイエンスやバイオ産業に従事する研究者や技術者に、より高度な研究開発ができる環境を提供することである。このような環境の構築には一般に長い時間を要することから、また、将来の研究の進展や変化を見越して計画を立てる必要があることから、長期的な視点に立って戦略を立案し、それにそって DB 開発を進めることが不可欠である。そのため、これまでのようにライフサイエンス DB 構築の専門家だけに頼って戦略を立てるのでは不十分である。そこで、引き続きライフサイエンス（基礎生物学および医・薬・農の応用生物学)、情報処理技術、ライフサイエンス DB の 3 分野の専門家による組織を構築し、各々の分野の動向に即した我が国の DB 整備戦略を立案する必要がある。また、この組織では、その戦略が着実に実行されているか、社会のニーズに適合しているか、などを定常的に評価し、必要に応じて戦略を機動的に変更する必要がある。また上記組織に対し、3 分野の動向に関する情報の網羅的収集・分析を日常的に行い、それら 3 分野の最新の俯瞰マップや動向マップを提示するなどして、戦略立案や実行評価を支援するチームも必要と考える。このような体制のもとで、戦略立案、実行評価の支援業務とそのための情報収集・分析だけでなく、関係府省、利用者、産業界、出版社・学会、国

内外の研究機関、等各種利害関係者との連絡調整、以下の各項目の統括、等の活動もあわせて行う必要がある。

統合 DB に関しては、本プロジェクトにより、そのあるべき姿のイメージ作りとそれを実現するための要素技術開発の立上げができたものと考える。しかしながら、統合 DB 構築は統合に必要な技術の開発だけで実現できるものではなく、著作権その他のデータベースにまつわる法律やデータの公開を遅らせる様々な要因といった、検討、解決すべき種々の社会的、制度的課題が存在することも分かった。それぞれが著作権やそのデータ産生の背景に存在する種々の権利関係を有する個別のデータベースを統合して公開する際には、著作権その他の法律をよく把握して、これを十分に遵守する必要がある。また、医療関係のデータベースの公開にあたっては倫理面への配慮が不可避である。一方、国家プロジェクトで産生されたデータに基づくデータベースに関しても、一般への公開が優先か、特許取得や産業振興のための一定期間のデータの非公開を優先するかといった問題が存在する。また、自分で産生したデータを大事にしたいという研究者が一般的にもつ気質もデータの公開を遅らせる大きな要因である。さらに、本年度の取組みでその重要性を認識することができた日本語文献情報に関わる、学会要旨や日本語総説誌のオープン化、データベース化に関しても、個別の学会や出版社との間で相互にメリットがある形の提携を行うための交渉が必要であり、そのためのビジネスモデル作りや著作権などに関する検討も重要な課題である。

以上の課題を解決し、真に役に立つ統合 DB を構築するためには、以下の取り組みが必要と考える。

- ニーズ面、シーズ面から見て今後重要となるデータベースコンテンツの調査
  解析技術の開発動向なども含めたデータ産出側の将来動向、併せてライフサイエンスおよびバイオ産業で今後必要となるデータニーズを調査することにより、ユーザニーズに合致した統合データベースのコンテンツの把握が必要。
- 統合対象の 250 の国内データベースの詳細調査と統合の優先付け
  プロジェクト DB の現状調査に基づく受入れ方法の検討および受入れの優先順位をつける方法を立案がすること必要であり、これに基づいた受入れと運用が望まれる。一方、レポジトリ制度の調査に基づいたレポジトリシステムの構築も重要。
- 国内基幹データベース、大量データ産出機関との連携策の検討
  データ量もユーザニーズも大きな基幹データベースや大量データ産出機関のデータベースとの連携策を検討し、統合データベースの提供する機能に合わせたデータの整形、リンク付けを実現。
- ライフサイエンスプロジェクトの現状調査と将来の統合化の阻害要因の洗い出しと解決策の策定
  各省の代表的なプロジェクトやそれを取り巻くコンソーシアムで産出されるデータの契約・権利関係、データ公開ルールを整理した上で、各プロジェクトや機関も納

得する形でのデータ提供を受けるための枠組みを検討。必要に応じてデータ公開を促進する法律立案も検討要。

・ データに関わる知的財産権や倫理的側面に関する調査と対応策検討

データベースレコード、文献など統合化に際して検討すべき知的財産権の調査と権利保有者との提携方針を策定。また、医療に関わるデータの個人情報の取り扱いなど、データベース公開に関わる倫理面の調査を行い、公開範囲に関する指針を策定。

（2）統合データベース開発

ライフサイエンス、バイオ産業にかかわる情報へのアクセスと利用に関する格段の利便性向上とそれによる研究開発の飛躍的な効率化と質的向上を目指すために、統合 DB とそれに必要な情報技術を開発する必要がある。具体的には、以下の取り組みが必要である。

1）共通基盤技術開発

・ インデックス、専門用語辞書の自動構築技術の開発
・ テキスト情報、画像・ポンチ絵情報ならびに異種 DB からの知識発見技術
・ 情報共有、情報交換のための WEB 技術開発
・ ワークフロー技術等との連携

2）ヒト統合 DB の開発・運用

・ 文献からの知識抽出システム開発とそれによるヒト知識の整理
・ 細胞、組織、器官、個体などの高次レベルの整理棚構築
・ 高精度アノテーションの実施と知識、機能を中心としたヒト統合 DB 構築、運用
・ 医療、医薬品に関するデータとの連携

3）モデル生物・産業応用生物統合 DB の開発・運用

・ 専門家集団からの意見集約による対象生物群の遺伝子名、機能、産物などの辞書構築
・ 環境ゲノム・メタゲノムを含む微生物ゲノム比較解析用 DB の構築と技術開発
・ 高精度アノテーションによるモデル生物・産業応用生物統合 DB の構築、運用
・ 解析ソフト充実と操作性の良いブラウザー開発による統合 DB の高度利用実現
・

（3）統合データベース支援

統合 DB の開発と運用に際しては、個々の DB としてどのようなものが現在開発・公開されているか、どのような DB 解析ツールが利用可能か、などを網羅的に調査し、その情報を一般の利用に供すること、我が国で開発された種々の DB を受け入れ、相互に連携して使えるようにすること、統合 DB の開発・運用やその利用技術開発に従事する人材を育成すること、などの支援業務が欠かせない。これを実現するために、以下の取り組みが必要である。

1）ポータル整備・運用、広報、普及啓発

・ DB サービス、解析サービスサイトに関する網羅的ポータルサイトの整備と運用

・ 検索に必要なインデックスや用語の収集とポータルサイト自動更新技術開発
・ ポータルサイト構築のための専門家ならびに利用者の意見集約システムの開発、運用
・ 日本語情報の収集、整理と日本語による研究情報の流通を促進する仕組みの整備
・ 海外情報日本語化のための技術開発と我が国の研究活動の海外への情報提供
・ ホームページの構築、講習会、シンポジウムの開催、ニュースレター等の発行

2）データベースの受け入れと運用

・ プロジェクト DB の受け入れと相互運用可能 DB への変換と運用、公開
・ 相互運用可能にするための標準化技術と用語の整理
・ 統合 DB 構築のための国内外主要 DB の更新、維持、管理

3）人材育成

・ 実データを用いたキュレータ・アノテータ教育の実践
・ 学部教育と連携した DB 構築者養成カリキュラムの実践
・ 大学院教育と連携した DB 高度利用者の養成と体系的なカリキュラムの作成
・ 教育プログラム・教材の実践・評価

３．データベース統合化基盤技術開発

３．１　データベース統合化基盤技術開発の実施計画

利用者が分子データや文献データを区別なくわかりやすい案内にもとづいて検索利用できる統合化を実現するための基盤技術の開発利用にむけ以下の４項目について開発に着手し利用を試みる。

（１）基盤知識表現技術開発

情報検索の利便性は索引付けの質に大きく依存するが解剖名称や細胞名称などの分野共通の基盤的な概念(用語)に関しても概念(用語)の整理と索引への利用は進んでいない。ここでは応用分野の区別無く索引利用可能な蛋白質名称、動物・植物の解剖用語、細胞名称および実験手法を対象にデータの内容を豊かに表現する索引系を開発しポータルサイト構築や戦略立案に提供する。

（２）癌研究知識表現技術開発

欧米では癌組織のヒトゲノム再配列解析プロジェクトが進行しており、そのサテライトとして統合癌 DB 化が検討されている。一方わが国では、臨床情報については、未だ DB 化に伴う倫理問題が議論されている段階である。そこで本プロジェクトでは大規模再配列解析プロジェクトの成果を念頭に置いて、すべての体細胞レベルでの遺伝情報と臨床情報を統合化した DB の作成において必要な表現法やデータ整理法の開発利用を試みる。

（３）多型知識表現技術開発

これまで多型情報の DB 化として、日本人ゲノムのプロモーター領域に存在する SNP を正確な頻度情報とともに記載したデータベース「dbQSNP」および日本人確定ハプロタイプ情報を記載した「D-HaploDB」の構築に取り組んできた。これらをさらに拡充するとともに、これを統合し日本人試料を用いた関連解析による疾患原因遺伝子の探索に不可欠な、高精度の情報基盤を確立することを目指す。また、XML 化等の標準化にも取り組む。これにより、ヒトゲノム多型データベースの標準化、統合データベースへの組み込みのフィジビリティースタディーを行い、多型情報統合 DB の医学分野におけるあり方を検討する。

（４）キュレーター支援技術開発

多数の文献に書かれた知識をまとめて表やサマリーの形で統合利用可能にする作業は、主に DB キュレーターと呼ばれる研究支援者によって手作業で行われている。論文などを読み解いて関心領域の抽出や記録を行う一連のキュレーターの作業を支援する技術を開発し支援環境の構築を行う。

なお、本テーマの（３）多型知識表現技術開発は九州大学が実施した。それ以外の項目は情報・システム研究機構が実施した。

３．２　データベース統合化基盤技術開発の実施内容

（１）基盤知識表現技術の開発

データベース構築の基盤となる知識表現技術開発の一環として、A．辞書シソーラス、及びB．オントロジー、分類機の開発を行った。さらに、開発したこれらの技術を用いたデータベース統合化の試みとして、C．分子データベース整理統合を行った。

A．辞書シソーラス

ライフサイエンスに関わる種々のカテゴリーのデータを統合し、情報を整理するときに必要となる辞書や、階層のついた同義語辞書であるシソーラスの作成を検討した。本年度は、１）遺伝子名称シソーラス、２）生物学名日本語一般名対応辞書、及び３）施設名称辞書の作成を行った。

1.）遺伝子名称シソーラス

検索やデータベース統合の混乱の主因の一つは、ライフサイエンス分野における同義語、特に遺伝子名称の同義語の氾濫である。そのため、遺伝子名称の同義語辞書を作成し、これを使用することで、種々のデータベースを統合的に利用する際の混乱を防止する必要がある。そこで、分子情報から医学文献までの遺伝子や蛋白質名称の正確な同定のための遺伝子固有名称（「遺伝子名」）と一般名称（「ファミリー名」）の辞書データの構築を目的として、ここでは、様々なデータベースで利用されている名称の収集と専門的キュレータによる編集を行い、遺伝子が持つ多様な名称の関係を明示した遺伝子名称シソーラス Ver1.0 を開発した。本シソーラスは、ヒトをはじめ９種類の生物をカバーしている。表 3.2.1 に、今回開発したシソーラスのファイルの構成を示す。また、表 3.2.2 に今回対象とした生物種、およびその遺伝子数ならびに遺伝子名称の数を示した。

表 3.2.1 遺伝子名称シソーラスのファイル構成

SWISS-PROTのID	EntrezGeneのID	その他DBのID	遺伝子名称				
SWISS-PROT:Q9Y6Y9	EntrezGene:23643	HGNC:17156	MD-2 protein	Lymphocyte antige	MD-2	MD2	ESOP-1
SWISS-PROT:Q9Y6Y8	EntrezGene:11196	HGNC:17018	p125	P125	SEC23-interacting	MSTP053	SEC23IP
SWISS-PROT:Q9Y6Y1	EntrezGene:23261	HGNC:18806	Calmodulin-binding	KIAA0833	calmodulin binding	CAMTA1	
SWISS-PROT:Q9Y6X9	EntrezGene:22880	HGNC:23573	MORC family CW-	Zinc finger CW-ty	KIAA0652	MORC family CW-	AC004542.C22.1
SWISS-PROT:Q9Y6X8	EntrezGene:22882	HGNC:18513	KIAA0854	Alpha-fetoprotein	ZHX2	zinc fingers and ho	AFR1
SWISS-PROT:Q9Y6X2	EntrezGene:10401	HGNC:16861	PIAS3	Protein inhibitor of	ZMIZ5	protein inhibitor of	FLJ14651
SWISS-PROT:Q9Y6X0	EntrezGene:26040	HGNC:15573	SETBP1	SET-binding prote	KIAA0437	SET binding protei	SEB
SWISS-PROT:Q9Y6W8	EntrezGene:29851	HGNC:5351	ICOS	CD278	inducible T-cell co	MGC39850	AILIM
SWISS-PROT:Q9Y6W6	EntrezGene:11221	HGNC:3065	MKP-5	Mitogen-activated	MKP5	DUSP10	MAP kinase phosp
SWISS-PROT:Q9Y6W5	EntrezGene:10163	HGNC:12733	WASP2	WASP-family prot	WAS protein famil	Wiskott-Aldrich sy	WAVE2
SWISS-PROT:Q9Y6W3	EntrezGene:23473	HGNC:1484	calpain 7	CAPN7	CALPAIN7	PalB homolog	Calpain-7

表 3.2.2 対象生物種と遺伝子数及び名称数

生物種等	遺伝子数	名称数
遺伝子ファミリ	12,110	27,923
ヒト	38,728	173,630
マウス	60,688	172,260
ラット	38,164	123,726
ゼブラフィッシュ	38,879	83,694
ショウジョウバエ	30,410	95,578
線虫	25,316	97,031
出芽酵母	6,190	33,030
分裂酵母	4,895	9,790
枯草菌	4,106	18,920
合計	259,486	835,582

2）生物学名日本語一般名対応辞書

データベースに記載されている生物の名称は、正確を期してラテン学名表記を基本としているが、利用者のほとんどは学名になじみがないのが実情である。そこで、研究分野でよく使われる生物種の基準として、学名に日本語一般名を対応させた生物学名日本語一般名対応辞書を開発した。対応付けは、塩基配列データベース(DDBJ)の登録エントリー数が多い生物種から順番に行った。また、標準和名が存在しない場合、その生物を説明する一般的な名称を用いた。登録データ数は合計 14028 種となっている。主要な 73 種類についてはさらに認識を容易にするアイコン画像を作成した。なお、ここでは農業環境技術研究所の日本野生植物寄生・共生菌類目録および日本産糸状菌類図鑑、日本爬虫両棲類学会の爬虫類のリスト、厚生労働省検疫所の届出対象動物種名リスト、哺乳類頭蓋の画像データベース（第 2 版）や The International Seed Federation (ISF)の植物病害関連生物リストなどを参照し、開発者が最も適切と思われるものを和名として採用した。その他、和名が不明な生物種については論文などから補完した。表 3.2.3 に分類ごとの登録した生物種数を示す。また、表 3.2.4 に辞書の一例を示す。図 3.2.1 には、アイコン画像の例を示す。

表 3.2.3 分類ごとの生物種数

分類	生物種数
霊長類	137
齧歯類	823
その他哺乳類	742
その他脊椎動物	4977
無脊椎動物	647
植物・真菌類など	6578
バクテリア	118
ウイルス	5
バクテリオファージ	0
未分類	1
合計	14028

表 3.2.4 生物学名日本語一般名対応辞書の一例

生物学名	日本語一般名称
Lemur catta	ワオキツネザル
Lepilemur mustelinus	イタチキツネザル
Varecia variegata	エリマキキツネザル
Cynocephalus variegatus	マレーヒヨケザル
Cheirogaleus medius	コビトキツネザル科
Otolemur crassicaudatus	オオガラゴ
Galago senegalensis	ショウガラゴ
Loris tardigradus	ホソロリス
Nycticebus coucang	スローロリス
Perodicticus potto	ポットー
Tarsius syrichta	フィリピンメガネザル

図 3.2.1 アイコン画像の例

3）施設名称辞書

日本のライフサイエンス研究を俯瞰するための重要な情報源として、各種関連学会の抄録などの報告文書があるが、これらをデータベース化する際に問題となるのが、例えば、大阪大、阪大、大阪大学大学院などといった、施設名称研究室名称の表記ゆれである。これに対応するために、施設名称辞書を開発した。これにより、同一の研究室の同一テーマを一塊として把握し国内の研究動向の把握を容易にすることが可能になった。表3.2.5に施設名称辞書の構成の一部を示す。

表 3.2.5 施設名称辞書の構成

標準化名称	標準化英語名称	科研費の機関コー	エリアス(エリアスが無いものは標準化名称を記載)
東大	Univ. Tokyo	172	東京大|東京大学
東大・医	Univ. Tokyo,Fac. Med.	172	東大・医学部
東大・医・一外	Univ. Tokyo,Fac. Med.,1st Dept. Surg.	172	東大・医・一外
東大・医・三内	Univ. Tokyo,Fac. Med.,Third Dept. Int. Med.	172	東大・医・3内|東大・医・第3内科
東大・医・整外	Univ. Tokyo,Fac. Med.,Dept. Orthop. Surg.	172	東大・医・整形外科|東大・整形

B．オントロジー、分類機

オントロジー、分類機として、１）動植物解剖学自動分類タガー、２）都市名国名自動検出タガー、３）解剖学3Dポリゴンマン辞書、４）3DアナトモグラフィーAPI、及び５）メソッドオントロジーとの連携システムを開発した。

１）動植物解剖学自動分類タガー

解剖学用語、すなわち臓器・器官・部位の名称を、専門家が作成したルール（振興調整費DB統合のための調査研究において作成）を用いて自動的にカテゴリー分類するプログラム、動植物解剖学自動分類タガーを開発した。

動物解剖学分類タガーでは、表3.2.6に示すように、動物の臓器、組織を、大きく10のグループに分類(大分類)し、さらにそれぞれのグループを細かく分類、合計40の小分類グループに分類する。基本的には、与えられた解剖用語に対して、解剖用語辞書、病理関連語彙の分類辞書、形容詞の解剖用語辞書、一般的な臓器名称の分類辞書、の４種類の辞書を順番に検索し、上記のカテゴリーに分類する。

植物解剖学分類タガーは、表3.2.7に示すように、植物(維管束植物)の部位、組織を大きく6のグループに分類し、さらにそれぞれのグループを細かく分類、合計11の小分類グループに分類する。分類における検索対象としては、生物種に合わせて、種子を持たない維管束植物の解剖用語辞書、イネ科の解剖用語辞書、その他被子植物の解剖用語辞書、裸子植物の解剖用語辞書、トウモロコシ属の解剖用語辞書、フウチョウソウ目の解剖用語辞書を用いる。

表3.2.6動物解剖学分類タガーにおける解剖学用語の分類

大分類	小分類								
脳	大脳	小脳	脳幹	脳梁	松果体	末梢神経	脊柱	網膜	目
血	動脈	静脈	リンパ節	末梢血	脾臓	胸腺	骨髄		
結合	脂肪	骨	皮膚						
生殖	胎盤	子宮	前立腺	卵巣	精巣				
筋	心臓	骨格筋							
消化	食道	胃	腸	結腸					
肝	肝臓								
肺	肺								
腎	膀胱	腎臓							
分泌	下垂体	甲状腺	副腎	膵臓	乳腺	唾液腺			

表3.2.7 植物解剖学分類タガーにおける解剖学用語の分類

大分類	小分類		
地上構造	葉	茎	
若い地上構造	若い地上構造		
根	根		
成長点	成長点	カルス	
花・生殖	花粉	子房	花・生殖
種子・果実	胚	種子・果実	

２）都市名国名自動検出タガー

論文やデータベースレコードに見られる国の名称の未記載や国の名称にみられる表記ゆれを吸収することを目的に、国別に分類するための辞書、都市名国名自動検出タガーを開発した。ここでは、DNAデータバンク(INSDC)のDBレコードの国別分類を行う自動検出タガーのフローを説明する。まず、DBレコード(FlatFile)を構造分解して、国名分類に使用するフレーズを抽出し、抽出した文字列を国名辞書、国名シノニム辞書、国別コードトップレベルドメイン１(ccTLD)辞書、国別研究機関名辞書からなる分類辞書群で順番に検索し、国名分類を行う。

３）解剖学３Ｄポリゴンマン辞書

解剖学用語、すなわち臓器・器官・部位の名称やそれらをくくる概念をモデル人間中の３次元座標（３Ｄポリゴンマン）で定義した辞書である解剖学３Ｄポリゴンマン辞書を開発した。これは、科学技術振興調整費「生命科学データベース統合に関する調査研究」におけるボクセル人体モデルの検討結果を受けたものである。従来のツリー型表現（いわゆる解剖オントロジー）と違い、多角的に破綻しない表現が可能で、ボクセルデータに比べエディットが飛躍的に容易である。ポリゴンマンの各臓器・器官の空間座標は、数値人体モデルデータベース（独立行政法人情報通信研究機構が開発）を基盤に、人体解剖模型・図譜等を参考に詳細化を行った。PubMedアブストラクトで出現の多い用語を中心に定義を行い、約 130 語の定義が完了した。

ポリゴンマンの各臓器・器官の空間座標は、数値人体モデルデータベース（独立行政法人情報通信研究機構の長岡博士らが、北里大学、慶應義塾大学及び東京都立大学と共同開発した電磁波影響計算のための Voxel モデルファントム、分類組織数約 50)を基盤に、位置関係や形態を大きく損なうことなく人体解剖模型・図譜等を参考に詳細部分を書き加えることにより求めた。肉付きや顔貌は創作によるものであり、定量的な正確さはある程度犠牲にしているが、形態や相互の関係は概念的な正確さを期して構築した。今後、用語数を増やし詳細化を進める計画である。また、利用環境が限定されるデータということもあ

り、ダウンロードの形ではなく利用者が自前のデータを貼り付けたり、入力したり、共有したりできるようにアナトモグラフィー（次項）も開発中である。

４）３Ｄアナトモグラフィー ＡＰＩ

解剖学用語が付与された手持ちのデータ（例：器官別の発現解析データ、疾患別症状分布など）を解剖学3Dポリゴンマン辞書にマッピングして俯瞰可能なアナトモグラフィー（新造語）を開発した。マッピング結果は、静止画像もしくは動画で得ることができ、webサービスとして提供予定である。膨大かつ詳細な”体に関する情報”同士の関係を理解する際に有用であると期待される。図3.2.2に３Ｄアナトモグラフィーの出力結果の例を示す。

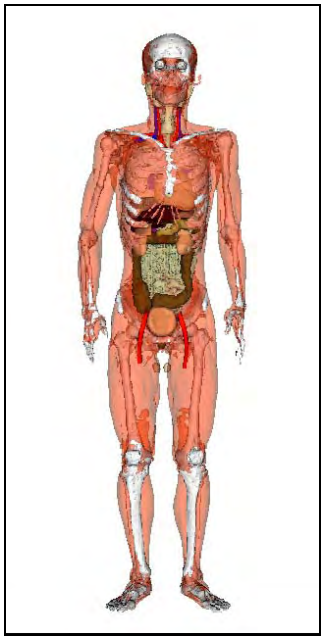

図 3.2.2　３Ｄアナトモグラフィーの出力結果例

５）メソッドオントロジーとの連携システム

日々増加するデータベースに登録される情報のほとんどはウェットの実験によって生み出される。実験結果そのものでもあるデータベースを理解するためには、実験の目的、材料、手法や条件を理解する必要がある。しかし最近では手法そのものが高度化、結果も膨大で解釈が難解になりつつある。そこで本プロジェクトではベンチとデータベースを結ぶために、実験手法の整理を辞書とオントロジーによって行い、データベースや論文を実験手法と目的から分類することに着手した。

今年度は、ドライ系メソッドオントロジーを利用した検索プログラムを作成し、Web リソースポータルに実装することにより、様々な名称がつけられているバイオインフォマテ

ィクスのメソッドを利用者の観点から分類し検索可能にした。表 3.2.8 に、構築したウェット系の実験手法名辞書の一部と図 3.2.3 に Web リソースポータルサイトの検索システムの一部を示す。

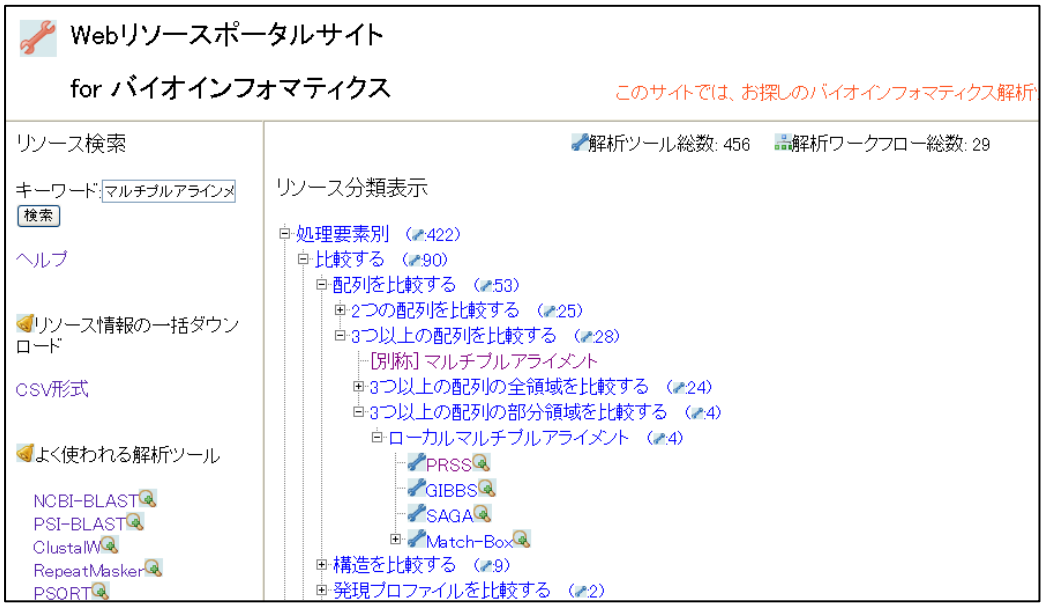

図 3.2.3 ドライ系メソッドオントロジーを利用した検索プログラム

C．分子データベース整理統合

本プロジェクトおよび関連プロジェクトで開発した統合技術およびリソースを用いて未整理のデータを統合した統合データベースとして、１）ヒト遺伝子発現統合の開発を行った。情報が安定した分野についてはこれからもこのような物理統合を進めてゆく考えである。

１）ヒト遺伝子発現統合

ヒト遺伝子の解剖学的な発現パターンデータの統合サイトを構築した。発現パターンは、測定法毎に異なる場合があることが知られており、ここではできるだけ客観的な発現パターンの解釈を可能にするために、5種類の発現データ、即ちiAFLP、GeneChip、EST、NCBIのSAGEmapによるタグマップ, SAGEデータの独自タグマップに基づく発現データを表示可能にした。また、遺伝子の発現パターンを、3種類の生物学的な分類（似た発現パターン、染色体上での隣接、同じ遺伝子ファミリー）に応じて表示することを可能にした。組織情報は、開発した動植物解剖学自動分類タガーで処理し整理分類しており、これにより異なるプロジェクトから得られた発現データ間の比較が可能になった。また、開発した遺伝子名称シソーラスと3Dアナトモグラフィーを、検索部分と表示部分にそれぞれ用いた。

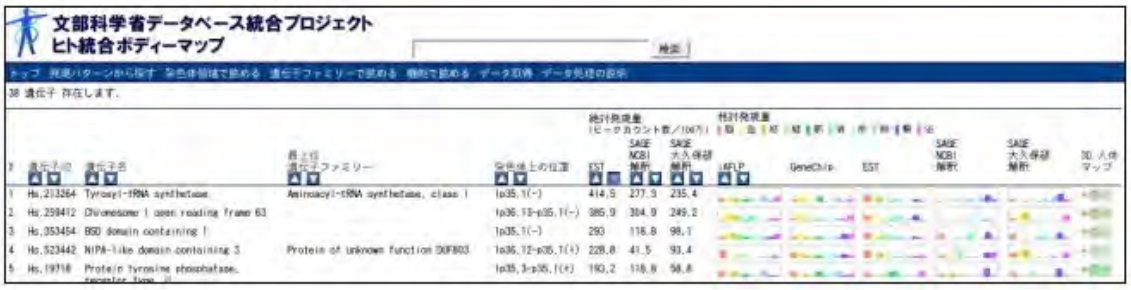

図 3.2.4 ヒト遺伝子発現統合の表示画面の例

（２）癌研究知識表現技術開発

実験的なデータベース作成等を通じて癌の分子データと臨床情報の統合、表現を想定ユーザーにわかりやすい形で実現することを目的として、癌遺伝子発現臨床情報データベースの機能拡張、及び大阪府立成人病センター乳腺内分泌外科の症例を対象に臨床情報の整理を行った。前者については、CGED (Cancer Gene Expression Database)の機能拡張を行い、従来の機能に加えて臨床情報から遺伝子を検索する機能を追加した。図3.2.5に「転移のある癌とない癌で発現の異なる遺伝子の検索」を例とした追加機能を表示する画面の例を示す。また、新規データとして乳癌(抗癌剤耐性研究)、胃癌、甲状腺癌に関するデータのアップロードを行った。臨床情報の整理については、成人病センター内の各種固形癌の臨床情報収集をターゲットとして想定し、乳癌については終了することができた。

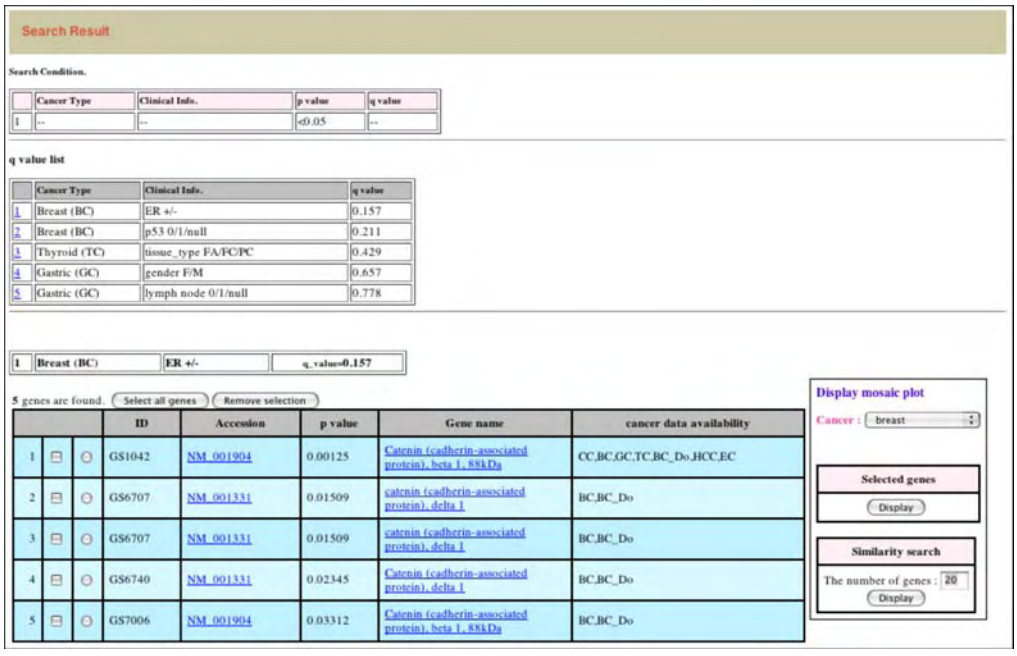

図3.2.5 臨床情報から遺伝子を検索する機能を示す表示例

（３）多型知識表現技術開発

ヒトゲノム多型情報は疾患・体質等の遺伝的背景を解明するための必須の情報基盤である。特にわが国にとって日本人のゲノム多型情報の整備は緊急の課題である。現在最も有効とされる情報源は国際 HapMap 計画での日本人のゲノム多型（SNP）情報とされているが、未だに調べられている SNP のゲノム上の密度が不十分である。また上記の遺伝的背景を明らかにするのに必要なハプロタイプ（SNP の並び方）の情報が必ずしも正確ではない。さらに同計画で調査された個体数が 45 人と少なく、日本人全体のゲノム情報を正確に反映したものとは言い難い状況 にある。そこで新たに日本人ゲノム多型情報に関する一次データを拡大・整備し、医科学等で利用可能な形態として提供することが、ゲノム多型情報を医学 情報と統合するために極めて重要な課題となる。

そこで、日本人ゲノム多型情報を高度化し、医療情報との統合のためのデータポータビリティーを図ることを目的に、疾患の主要な遺伝的要因である遺伝子発現調節領域多型の公開データベース「dbQSNP」の拡充、要因遺伝子探索に必須な全ゲノム確定ハプロタイプ構造の公開データベース「D-HaploDB」の拡充、及び他のデータベースとの多型データ相互利用促進のため上記２つのデータベースの標準言語化を行った。「dbQSNP」の拡充については、自己免疫疾患、がんへの関与が疑われる約 100 個の遺伝子のゲノム領域にある SNP 配列及びその正常日本人でのアレル頻度を直接配列決定及び定量 SSCP 解析により決定した。この結果、約 1.0 X 10^4 個の SNP 情報が記載されることになった。
「D-HaploDB」の拡充については、新たに 100 個の胞状奇胎について Affymetrix 社アレイチップを用いた、既存データの約 2 倍にあたる 5 x 10^5 個の SNP タイピングを行い、ゲノムワイド連鎖不平衡地図を飛躍的に高精度化できた。データベースの標準言語化については、データベース記述標準言語 XML のゲノム多型記述のための機能拡張版である PML を採用し、上記二個のデータベースの PML 版を構築した。これにより、他のデータベースとの多型データ相互利用促進手段を確立できたものと考える。

（４）キュレーター支援技術開発

本事業では、これからの研究に欠かせないのが各分野専門家の手による文献やデータベースからの事実の切り取りと再配置による知識の整理であると考えられる。一方で生命科学の専門家には情報技術を駆使することは容易ではない。そこで、論文から抽出したデータのデータベース構築作業を支援するソフトウェアの開発を目標に、論文情報解析・編集ソフトウェアのベースシステムの開発と論文情報解析・編集用各種解析モジュールの開発を行った。具体的には、論文や Web を渡り歩いて重要箇所だけドラッグドロップすると自動的に URL やページ座標情報が記録され、さらに記事を並べて後からメモ書きを行える環境を FireFox の PlugIn アプリケーション（ScrapParty）として開発した。収集記事は xml 形式で書き出すことができ、他人と共有することも可能である。また、”ScrapParty” の追加モジュールとして、辞書マッピング、ナビゲーション、論文構成認識、引用論文情

報収集を行う基本モジュールを作成した。

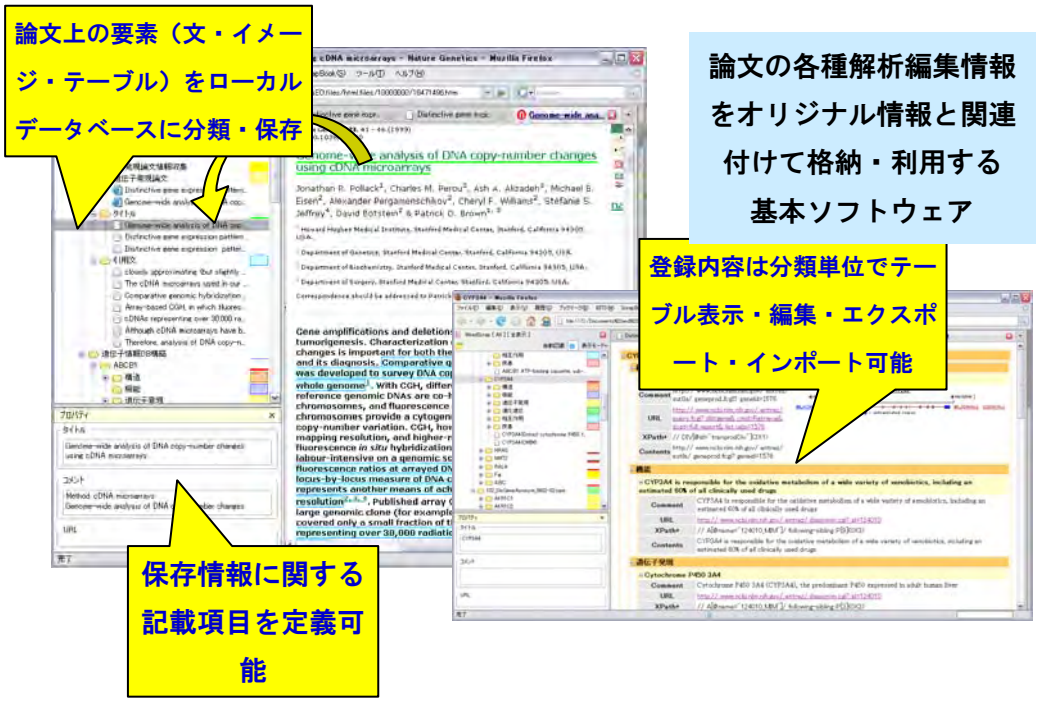

図 3.2.6　ScrapParty の概要

３．３　データベース統合化基盤技術開発のまとめ

① データベースレコードの統合（基盤知識表現技術開発）

検索だけでしか内容をうかがい知ることが出来ない膨大なレコードが与えられたとき、

「要素についてわかるとは全貌のなかでの位置を知ることである」、

「分類整理はＤＢレコード全貌についての最善の表現である」、

「分類整理はＤＢへの入り口であり検索結果の出口を提供する」、及び

「同様な分類整理は統合の一形態である」

という思想に基づいて開発を行った。

整理を行う「整理棚」として、配列や構造とは独立の、

１）研究場所、研究分野、

２）材料の生物種、解剖、

３）研究目的方法

を用いて、それぞれ辞書または分類機を開発し、実際に提供されたサービスに対して開発技術を適用した。

開発技術とサービスの対応は下の表のような関係である。今後も幾つかの対応が追加可能である。

	施設		研究		材料			表示		
1次データ	施設名称辞書	国名辞書	プロジェクトソータ	実験オントロジ	生物アイコン	生物名称辞書	解剖ソータ	ポリゴンマン辞書	アナトモグラフィ:ＡＰＩ (GoogleMap:API)	サービス名称
学会抄録	■									学会俯瞰
INSDC		■	■			■			■	INSDC 目次
GEO		■				■	■	■		GEO 目次
IAFLP		■				■	■	■		発現統合
EST		■				■	■	■		発現統合
SAGE		■				■	■	■		発現統合

（黒ますは今年度適用試験済み）

例えば、様式や表記が同一でない別年度の分子生物学会の年会抄録を材料に、施設名称辞書を用いて研究ループを同定し人名同定を進めることで、複数年会の要旨が統合されたわけである。複数年の要旨が統合されると、年会のメニューでしかなかった情報が研究室単位の研究内容や年次研究変遷の歴史の情報を持ち、試験利用者からも「研究室ＤＢ」としての価値が指摘されている。また同様な整理棚は学会によらず適用可能であるために、日本の研究全体を表現するＤＢを作ることも可能である。この例では一つの辞書によりＤＢは統合可能で、統合によって新たな価値が生まれることが多いことが示されている。

今後も上記の枡目を全て埋めるような対応をすることで、学会要旨から発現データまでが統合利用され統合表示が可能になると考えられる。

② 複雑応用領域の知識表現（癌研究知識表現技術開発、及び多型知識表現技術開発）

臨床医学では人の共通性ではなく個人の違いが研究される。種内の共通性を研究する生物学とは世界モデルが違う為に、臨床医学領域については領域内で利用可能な整理棚も自明ではないため、領域の専門ＤＢに協力を依頼した。

癌医学については、材料を分類するための癌の臨床分類が整理法としてある程度有効であること判ったが、癌遺伝子遺発現研究は患者背景や治療など研究ごとに多様な特徴を発現と比較する為のコントロール疫学研究であり、データを統合することには殆ど意義がないと思われる。むしろ比較したり和をとったりすることが出来る同種のコントロール疫学研究を集めるようなボトムアップ型の統合が有効であるとの結論に至った。今後は、論文を材料として治療効果やマーカー同定などを含めて、コントロール研究のクラスター化を

行うべきであると結論した。

多型研究知識については、ゲノムワイドな多型同定データ解析時に欠かすことの出来ない日本人のハプロタイプに関するモデルが不在であるために、まずはハプロタイプ情報を集積し信頼できる日本人のモデルつくりを第一歩とするべきであると結論付けた。

③ キュレーター支援技術開発

統合 DB 構築に際して、対象データベースのレコードに注釈づけをおこなっていく際に、いかに効率よく各種文献やデータベースの情報を収集、整理するかが課題である。今回開発した論文情報解析・編集ソフトウェアとその追加モジュールは、各種文献やデータベースに記載されている情報を自由に切り取り、分類、保存するシステムであり、今後のキュレーション作業の効率化の強力な武器になるものと期待している。

４．ポータルサイトの構築

４．１　ポータルサイト構築の実施計画

（1）データベース（DB）等ポータル構築

現存するDBのカタログは遺伝子や蛋白質等のデータ対象によるおおまかな分類が与えられるのみで利用者が目的にかなったDBを選択することは容易でない。ここでは中核機関の情報･システム研究機構が中心となってまとめる DB の俯瞰や戦略立案の目的で作成するDBディレクトリに、科学技術振興機構が維持してきたWINGDB案内の日本語解説を加え、利用者を最適なDBに案内する仕組みを構築する。同時に、DB中のデータを利用するために必要となる解析のためのツールや環境を案内するWEBリソースカタログ作りを増強する。利用者にとってわかりやすいインターフェース作りに配慮し、これらのポータルサイトを構築する。また、我が国に存在するライフサイエンスのいくつかのそれぞれ内容が異なるポータルサイトの機能を生かし、相互に利用すべき部分は利用し、全体としてより高い機能を果たすべく連携する仕組みを考案する。また、本課題を広く周知するためのウェブサイト公開用サーバを準備する。

なお、本テーマは科学技術振興機構が実施した。

（2）文献情報との連携調査

科学技術振興機構が開発運営を続けている文献情報提供事業では国内外の情報が日本語で幅広く提供されている。これを活用し、遺伝子機能に関連する情報について、文献情報との連携を検討調査する。

４．２　ポータルサイト構築の実施内容

（1）データベース（DB）等ポータル構築

中核機関の情報･システム研究機構が中心となってまとめる DB の俯瞰や戦略立案の目的で作成するDBディレクトリに、科学技術振興機構が維持してきたWINGDB案内の日本語解説を加え、利用者を最適なDBに案内する仕組みを構築した。同時に、WEBリソースカタログ作りを増強し、これらのポータルサイトを構築した。また、我が国に存在するライフサイエンスの異なるポータルサイトを調査し、連携について考察した。また、本課題を広く周知するためのウェブサイト公開用サーバを用意し､ウェブサイトを公開した。

１）ＤＢポータル

ライフサイエンス分野のデータベースのカタログサイトで、利用者を最適なDBに案内することを目指し、利用者からの書き込み追加が可能な「Wiki」を利用した。コンテンツの充実と即時性が見込めて、利用方法などの情報交換の場を提供できる仕組みである。公開時、詳細記事は371データベース分が収録されている。関係４省庁の協力による過去の国内調査資料にNucleic Acid Research誌Database Issue2006, Science誌Netwatch/Database等を加えリストアップした。また、データベース一覧（構築型分類）

は、科学技術連携施策群での調査研究の成果を提供した。

図 4.2.1　DB ポータル（WINGpro）のデータベース記事の例

2）WEB リソースポータル

実験データの解析や公的データの加工に使用する解析ツールや環境を案内する WEB リソースカタログサイトを構築した。 本統合技術開発の成果であるベンチメソッドオントロジーの一部をカタログの上に乗せてプロトタイプとして提供した。解析ツール 456 をリストしている。　図 4.2.1 に具体的な検索画面を例示している。

3）ポータルサイト連携のための調査

「Ｊａｂｉｏｎ］日本語バイオポータルサイト、「ライフサイエンスの広場」(文部科学省ライフサイエンスポータルサイト)、「ＵＭＩＮ］大学病院医療情報ネットワーク（一般公開）について調査し、連携の仕組みを考察した。

（2）文献情報との連携調査

独立行政法人科学技術振興機構が実施している文献情報提供事業で提供されている文献情報を活用し、遺伝子情報に遺伝子機能を付加する調査を実施した。100 程度の抄録から遺伝子関連の表現型や疾患名を抽出した。テストサイトを公開した。

４．３　ポータルサイト構築のまとめ

ポータルサイト構築にあたり、多数あると言われるデータベースから、我が国では研究者等の所属機関が把握している DB と海外の DB を加えた 371 について、案内サイトを構築した。今後の方針によるが、収録 DB 数の増加と利用者の意見を反映した内容の充実を図るための仕組みを構築し、利用者を最適な DB に案内し、さらに最適なデータを得ることを可能とするサービスを提供する研究開発を継続すべきである。ポータルサイト連携の仕組みも今後の課題でさらによりよいサービスを検討したい。

５．人材の育成

５．１　人材の育成の実施計画

DB 統合や維持管理のためには①キュレーター（論文の内容を理解して情報抽出整理し定型化した表現に変換して DB 構築を支援する学芸員）、②アノテーター（プログラム処理の結果に総合判断を加え、データに生物学的医学的な解釈を付与する学芸員）、③DB マネージャー（DB について理解しておりデータの参照情報などを自立的に更新できる技術者）の３者が必要である。しかしながらこれらの専門業務の存在は DB 構築運営を行ってきた組織でしか知られておらず、広くこれらの業務内容について衆知し、人材養成のための学習の材料を整備することは将来の DB 統合事業に参加可能な人材の裾野を広げることであり、これらの専門職のキャリアパス作成の第一段階である。ここでは３者の業務内容を区別し、業務に必要な基礎知識や技術について解説した教材の作成に着手し DB の本格的統合化事業に備える。なお、本テーマは情報・システム研究機構が実施した。

５．２　人材の育成の実施内容

キュレーター、アノテーター、DBマネージャーの業務内容を整理し、業務に必要な基礎知識や技術について解説した教材を作成してデータベースの本格的統合化に備える目的で、まずwikiと電子メールによって知識の蓄積が可能な教育資料編纂・閲覧システム「MotDB」を作成した。これを用いて、（１）キュレーターの育成のためのキューレーションDBの調査、（２）アノテーターの育成のための、アノテーターからのノウハウ抽出実施と教育用テキストの作成、（３）DBマネージャーの育成のための実習書の作成を行った。

図 5.2.1 教育資料編纂・閲覧システム「MotDB」

（1）キュレーターの育成

ハイスループット化されたゲノム研究手法により、大量のゲノム配列情報が作り出され、膨大な生物種のゲノム情報が刻々と蓄積し続けている。同時に、従来の小規模で精密な実験やゲノム規模の実験による、発現プロファイル解析、タンパク質相互作用解析や構造ゲノミクス解析、また、それに伴う計算機を用いた大規模予測がなされ、さまざまな角度からタンパク質機能の解析が行われている。それらのデータの統合活用には、自動アノテーションのチェックや文献からの情報抽出、コンピュータ解析によるアノテーションの修正と付加などをタスクとするキュレーション過程の整備と、その担い手であるキュレーターの育成が必要かつ急務であると考えられる。

このような背景から、キュレーターの育成のために、実際にキュレーションが行われている機関での作業内容や流れを調べることを目的として、文献等の資料、及びWeb公開情報をもとに、キュレーション作業の実際を調査した。その結果、著名な生物学データベースの領域でのキュレーションとキュレーターの作業の実際が明らかになった。また、著名なキュレーション型データベースの比較を行い、それぞれのデータベースの人員的な規模や、キュレーションがデータベース全体のなかでどのように関与しているのかを明らかにした(参照)。

表 5.2.1 代表的なキュレーション型 DB のスタッフ規模

DB	組織	キュレーター	コーディネーター	ソフトウエア技術者、プログラマー	参考
RefSeq	NCBI	30	3	38	http://www.ncbi.nlm.nih.gov/RefSeq/staffcredits.html
UniProtKB	EBI	26	5	36	http://www.ebi.ac.uk/Information/Staff/viewgallery_seqdb.php?cid=4
	PIR	12	2	3	http://pir.georgetown.edu/pirwww/about/staff.shtml
	SIB	52*1	5	13	http://au.expasy.org/people/swissprot.html
PATHWAY MAP BRITE KO	KEGG	21*2	-*3	4	http://kanehisa.kuicr.kyoto-u.ac.jp/people.html

*1. キュレーターという肩書きがなくアノテーターとなっていたが、その方々がいわゆるキュレーション作業もすると判断した
*2. KEGGは基本的な生物情報だけでなく、化学物質から疾患情報、薬剤情報など非常に広範囲にわたるDBを作成している。21名という数字はそれら多岐にわたるDBの作成にあたるスタッフの総数であり、実際に生物分野を担当しているスタッフは数名であると考えられる
*3. いわゆるコーディネーターという肩書きはない、前出の21名スタッフが自分の担当分野でおいて状況に応じてコーディネーター的役割もはたしてると考えられる

（2）アノテーターの育成

アノテーターの育成のために、ゲノムアノテーションの実務に携わるアノテーターを支援する教育用システムを作成し、同時に実践的なアノテーター教育テキストを作成することを目的として、①ゲノムアノテーションのノウハウの抽出、②アノテーターによる手動アノテーションの手順を模倣・再現するプログラムの作成、③アノテーションにかかわる知識や注意点を文書化しゲノム解析型統合DB構築に役立つ「アノテーション教育テキスト」のWiki上での作成を行った。

①については、かずさ DNA 研究所で実際に大量ゲノム解析に携わる高度専門技術を有するアノテーターからの聴き取りならびに実務調査を行った。その結果、実際に用いているツールや DB の利用手順、解析結果の解釈法などのノウハウを抽出できた。②については、抽出した情報の解析を元に、アノテーターによる手動アノテーションを模倣するプログラムを作成した。これにより、DB やツールの利用方法やその結果の解釈、さらに実行したアノテーションの根拠を明示することで初心者の学習を支援するプログラムが作成できた。③については、上記項目で抽出されたノウハウをふくめ、具体例をあわせた実践的「アノテーター教育テキスト」コンテンツを作成した。

（3）DB マネージャーの育成

ライフサイエンス関係のデータベースを構築・維持管理の実際を行う DB マネージャーの育成ための教育資料編纂・閲覧システムと実際のコンテンツ構築を行うことを目的に、上記の wiki と電子メールによって知識の蓄積が可能な教育資料編纂・閲覧システムの作成を行い、このシステムを用いて DB マネージャー養成のための実習書を作成した。DB を管理する DB マネージャーに必要不可欠なスキルとして大きく DB 構築とその維持があるので、教科書もそれに合わせて、構築編と維持管理編の二つのカテゴリーからなる構成とした。本年度は既存の DB の維持管理に必要なスキルを解説する維持管理編の構築を主に進めた。二つのカテゴリーには入らない事柄を、基礎編（DB マ

図 5.2.2 DB マネージャー教育用教科書の一部

ネージャーに必要な基礎的な知識を整理）、及びリファレンス編（各編に共通したリファレンスとなりうる事項を収集）としてまとめた。維持管理編は、日々の管理、アップデート、バックアップを中心に作成した。

５．３　人材の育成のまとめ

ライフサイエンス統合データベースを開発、維持していくために不可欠な専門職員であるキュレーター、アノテーター、DBマネージャーの育成に関して、そのベースとなるwikiと電子メールによって知識の蓄積が可能な教育資料編纂・閲覧システムを開発した。さらに、代表的キュレーション型データベース調査などに基づきキュレーションの実際を明らかにし、かずさDNA研究所の実績をベースにしたアノテーター教育システムの構築、およびDBの維持管理をターゲットにしたDBマネージャー教育用教科書の編纂を実施した。これらにより、上記専門職員の人材育成に関する基盤が構築できたものと考える。現在不足している、これら専門職員の育成を図っていくためには、今回開発したシステム、コンテンツをベースにその内容を充実していくことが重要と考える。

６．プロジェクトの総合的推進

６．１　研究運営委員会及び統合 DB 整備戦略作業部会

プロジェクト全体の連携を密としつつ円滑に推進・運営していくため，ライフサイエンス、知識情報処理、ライフサイエンス DB の 3 分野の専門家による研究運営委員会を組織し、統合化 DB の整備戦略を議論した。また、研究運営委員会の実働部隊（情報の収集・分析、動向調査、戦略立案支援、など）として統合 DB 整備戦略作業部会を設けた。

（1）運営委員会及び作業部会の構成と活動経過

研究運営委員会のメンバーは下記の通りである。

情報・システム研究機構	堀田凱樹
	小原雄治
	五條堀孝
	大久保公策
自然科学研究機構基礎生物学研究所／JST	勝木元也
東京大学／JST	高木利久
科学技術振興機構（JST）	大倉克美
東京大学	吉田光昭
東京大学	辻井潤一
京都大学	金久　實
大阪大学	中村春木
理化学研究所	榊　佳之
産業技術総合研究所	秋山　泰
東京理科大学	増保安彦
JT 生命誌研究館	中村桂子
DNA チップ研究所	松原謙一
かずさＤＮＡ研究所	田畑哲之
九州大学	久原　哲

統合 DB 整備戦略作業部会のメンバーは下記の通りである。

五條堀孝
田畑哲之
大久保公策
菅原秀明
高木利久
高野明彦
黒田雅子
藤山秋佐夫
久原哲
中村桂子
増保安彦

以下に示す研究運営委員会，統合 DB 整備戦略作業部会の開催を行った。

①第一回研究運営委員会・統合 DB 整備戦略作業部会（合同会議、2006 年 11 月 8 日））

②第二回統合 DB 整備戦略作業部会（2006 年 12 月 18 日）

③第二回研究運営委員会（2006 年 12 月 25 日）
④第三回研究運営委員会・統合 DB 整備戦略作業部会（合同会議、2007 年 2 月 1 日））
⑤第四回研究運営委員会・統合 DB 整備戦略作業部会（合同会議、2007 年 3 月 19 日））

（2）運営委員会及び作業部会の結論

第一回研究運営委員会・統合 DB 整備戦略作業部会、及び第二回統合 DB 整備戦略作業部会と第二回研究運営委員会においては、統合データベース整備事業の全体構成、及び平成 18 年度の実施状況について討議が行われた。第三回研究運営委員会・統合 DB 整備戦略作業部会では、平成 18 年度の実施状況についての討議と共に、今後の具体的進め方についての討議が行われた。第四回研究運営委員会・統合 DB 整備戦略作業部会では、18 年度成果の報告を行った。以上の運営委員会及び作業部会の実施によって「ライフサイエンス分野の統合データベース整備事業」のプロジェクトの総合的推進が達成された。

６．２　教育プロジェクトに関するミーティング

統合 DB の人材の育成プロジェクト（教育プロジェクト）に関して、外部有識者を招聘して意見を聞くための以下の会を開催した。

①統合 DB の教育プロジェクトに関する打合わせ（2006 年 10 月 17 日）
②「シニア世代と学部教育」第一回ミーティング（2007 年 1 月 29 日）

「統合 DB の教育プロジェクトに関する打合わせ」においては、外部有識者として、実際のゲノムのアノテーションを通じて教育を実践されている長浜バイオ大学の池村教授を招いて、ゲノムアノテーションを通じたアノテーション教育のあり方と今後の方針についての討論を実施した。「シニア世代と学部教育」第一回ミーティングにおいては、池村教授が提唱されている「シニア研究者の高度な知識の活用・継承によるアノテーション教育」について、シニア世代 6 名と現役世代 3 名の研究者を招き討論を行った。これらの討論の結果は、「ライフサイエンス分野の統合データベース整備事業」の全体計画策定において反映された。

７．プロジェクトの成果のまとめと評価

あるユーザにとっての情報の Utility（利便性）とは、
Utility = Relevance(質問への関連度)　x Validity(有効度)／Work to Access（入手労力）
として定義される。

１８年度の実施項目のなかでは、分子データに関してあらゆるユーザにとって Relevant な情報への Work to Access を減らすために分類整理を行った。Relevance を表現する手法としては、地理や施設、生物種、解剖など最も多くのユーザに共有されている実世界の整理軸を用いることで、多様なデータ群が統合的に分類可能で Workt to Access を低下させる効果があることが証明された。またこれらの分類は検索後の結果提示にも有効であり、キーワード検索と組み合わせることでより Relevant な情報に簡単に導くことが可能であると期待できる。分類に使用した辞書や分類機、オントロジーは今後も開発を続けることでより感度と精度を増すことが期待され、全ての開発物は次年度からの統合に利用可能である。

一方生命科学領域のデータは配列以外非常に Validity や Reliability, Accuracy が低いと考えられており、Valid なデータを作り出す努力も統合目的にかなう仕事であると考える。発現データについて Reference データを作ることを計画していたが、統合整理し比較表示までの開発は行ったものの相互に矛盾するデータセットが予想外に多く、Reference 作成の方針を立てることが出来なかった。

配列や構造などの説明不要な分子データ以外は多くのデータがコンテキストに大きく依存する観測であり、研究単位で詳細に検討をしない統合は意義が希薄である。すなわちデータは論文のサプリメント情報と扱うべきであると結論した。あわせて大型のデータを伴わない論文も基礎研究、臨床研究論文ともに個々の報告の解釈は説得力を十分に持つものでなく、治療法や診断法を変更させる信頼度は存在しない。従って個別論文は「作業報告」として「操作と結果」にあたる FACT 部分を切り出し相互に強く関連するものをまとめることにより「信頼できる判断」と「証拠」の対を形成してゆくことでおおきな「知識」をくみ上げてゆかねばならない。これは臨床研究における「Evidence Based Medicine」における「Clinical Evidence」と同様である。癌研究や多型研究についてはまさにクリニカルエビデンスを作成することが統合であると結論した。今後のプロジェクトでは「BioMedical Evidence」の生成とＤＢ化によって論文情報を統合してゆく方針である。

８．成果の外部への発表

次ページの添付様式を参照されたい。

９．実施体制

別表１を参照されたい。

添付様式

論文寄稿

業務コード	実施年度	和誌/洋誌	論文タイトル	発表者名	発表誌名	巻	号	ページ	掲載年月	メモ
06026018	18	和誌	生命科学データベースの現状と課題	大久保 公策	科学	77	4	364-369	2007 年 4 月	
06026020	18	洋誌	D-HaploDB: a database of definitive haplotypes determined by genotyping complete hydatidiform mole samples.	Higasa K, Miyatake K, Kukita Y, Tahira T, Hayashi K	Nucleic Acids Res.	35		D685-689	Jan. 2007	
06026020	18	洋誌	QSNPlite, a software system for quantitative analysis of SNPs based on capillary array SSCP analysis.	Tahira T, Okazaki Y, Miura K, Yoshinaga A, Masumoto K, Higasa K, Kukita Y, Hayashi K	Electrophoresis	27		3869-3878	Oct. 2006	
06026020	18	洋誌	Novel Mutations in Norrie Disease Gene in Japanese Patients with Norrie Disease and Familial Exudative Vitreoretinopathy.	Kondo H, Qin M, Kusaka S, Tahira T, Hasebe H, Hayashi H, Uchino E, Hayashi K	Invest. Ophthalmol. Vis. Sci.	48		1276-1282	Mar. 2007	

講演

業務コード	実施年度	国内/国際	講演タイトル	発表者名	講演会名	発表年月日	メモ
06026018	18	国内	データベースから見たライフサイエンスプロジェクト	高木 利久	日本分子生物学会 2006 フォーラムシンポジウム[プロジェクト型研究時代の生命科学の課題]	2006 年 12 月 8 日	
06026018	18	国内	知的生産性向上のための情報処理	大久保 公策	日本分子生物学会 2006 フォーラムシンポジウム[プロジェクト型研究時代の生命科学の課題]	2006 年 12 月 8 日	
06026018	18	国内	ライフサイエンスDB その歴史とわが国の現状と課題	高木 利久	日本分子生物学会 2006 フォーラム バイオテクノロジーセミナー	2006 年 12 月 8 日	
06026018	18	国内	オントロジーや辞書は役に立つのか	大久保 公策	日本分子生物学会 2006 フォーラムﾊﾞｲｵﾃｸﾉﾛｼﾞｰｾﾐﾅｰ	2006 年 12 月 8 日	
06026018	18	国内	使い倒し系バイオインフォマティクスによる知のめぐりのよい生物学研究のすすめ	坊農秀雅	お茶の水女子大学「魅力ある大学院教育」第 6 回バイオインフォマティクスへの招待	2007 年 3 月 16 日	
06026018	18	国内	統合 DB の構築に必要な情報技術	高木 利久	情報とシステム 2007	2007 年 3 月 1 日	

業務コード	実施年度	国内/国際	講演タイトル	発表者名	講演会名	発表年月日	メモ
06026018	18	国内	ライフサイエンスのデータベースの現状と課題	大久保　公策	情報とシステム 2007	2007 年 3 月 1 日	
06026018	18	国内	ライフサイエンスにおけるゲノム情報の高度利用に向けた生命知識の構造化	大久保　公策	知の構造化ワークショップ－知の構造化ツールは、新しいサイエンスを開くのか－	2006 年 12 月 4 日	
06026018	18	国内	知識発見のための癌臨床情報のデータベース化	加藤　菊也	第2回　大阪大学臨床医工学融合研究教育センターシンポジウム	2006 年 10 月 8 日	
06026020	18	国際	Analysis of Genes Affecting Susceptibility to Systemic Lupus Erythematosus (SLE).	OTahira T, Horiuchi T, Sakaguchi D, Yamai M, Miyagawa H, Tsukamoto H, Hayashi K	Annual Meeting of American Society of Human Genetics,	Oct. 9-13,2006	New Orleans, U.S.A
06026020	18	国際	Capillary array SSCP analysis of pooled DNA for association testing.	OHayashi K, Masumoto K, Y. Okazaki, A. Yoshinaga, K. Higasa, Y. Kukita, T. Tahira	Annual Meeting of American Society of Human Genetics,	Oct. 9-13, 2006	New Orleans, U.S.A
06026020	18	国際	D-Haplo: A genome-wide definitive haplotypes determined using complete hydatidiform moles.	OHayashi K	Human Genome Variation 2006	Sep. 14-16, 2006	Hong Kong, China. (Invited)
06026020	18	国際	The power of Definitive Haplotypes in association studies.	OMiyatake K, Kukita Y, Higasa K, Wake N, Hirakawa T, Kato H, Matsuda T, Tahira T, Hayashi K	Human Genome Variation 2006	Sep. 14-16, 2006	Hong Kong, China
06026020	18	国際	Periodicity of SNP distribution around transcription start sites.	OHigasa K, Hayashi K	Human Genome Variation 2006	Sep. 14-16, 2006	Hong Kong, China

プレス発表

業務コード	実施年度	発表タイトル	掲載新聞名	掲載日
06026018	18	「ライフサイエンス分野の統合データベース整備事業」の成果公開について－ライフサイエンス研究の発展に向けて－	―	―

別表１　平成１８年度に於ける実施体制

研　究　項　目	担当機関等	研究担当者
1.　データベース統合戦略立案および評価	情報・システム研究機構	◎堀田　凱樹
	東京大学大学院新領域創成科学研究科	○高木　利久
	情報・システム研究機構 国立遺伝学研究所	大久保　公策
	かずさDNA研究所植物ゲノム基盤研究部	中村　保一
	JT生命誌研究館	中村　桂子
	情報・システム研究機構 国立情報学研究所	高野　明彦
	情報・システム研究機構 国立情報学研究所	藤山　秋佐夫
	情報・システム研究機構 国立遺伝学研究所	五條堀　孝
	情報・システム研究機構 国立遺伝学研究所	菅原　秀明
2.　データベース統合化基盤技術開発	情報・システム研究機構 国立遺伝学研究所	○大久保　公策
	東京大学大学院新領域創成科学研究科	高木　利久
	情報・システム研究機構 新領域融合研究センター	川本　祥子
	情報・システム研究機構 国立情報学研究所	武田　英明
	情報・システム研究機構 国立遺伝学研究所	西川　　建
	情報・システム研究機構	三橋　信孝
	情報・システム研究機構	水田　洋子
	大阪府立成人病センター	加藤　菊也
	九州大学生体防御医学研究所	林　　健志
3. ポータルサイトの構築	自然科学研究機構基礎生物学研究所	○勝木　元也
	科学技術振興機構研究基盤情報部	大倉　克美
	科学技術振興機構研究基盤情報部	黒田　雅子
	科学技術振興機構研究基盤情報部	小池　俊行
	埼玉医科大学ゲノム医科学研究センター	坊農　秀雅
	東京大学医科学研究所ヒトゲノム解析センター	川島　秀一
	東京大学医科学研究所ヒトゲノム解析センター	山　俊明
	情報・システム研究機構 新領域融合研究センター	川本　祥子
	情報・システム研究機構 国立遺伝学研究所	大久保　公策
4.　人材の育成	情報・システム研究機構 国立遺伝学研究所	○大久保　公策
	情報・システム研究機構 新領域融合研究センター	川本　祥子
	科学技術振興機構研究基盤情報部	黒田　雅子
	かずさDNA研究所植物ゲノム基盤研究部	中村　保一
	埼玉医科大学ゲノム医科学研究センター	坊農　秀雅
	情報・システム研究機構	岡本　　忍

注1.　◎:課題代表者、○:サブテーマ代表者

注2.　本業務に携わっている方は、全て記入。