統合データベースプロジェクト研究運営委員会作業部会（第５回）議事要旨

【日　時】　平成２０年１０月１０日（金）１０：００～１３：１５
【場　所】　ライフサイエンス統合データベースセンター大会議室
【出席者】　高木主査、浅井委員、黒田委員、五斗委員、徳永委員、森下委員、菅原委員
【欠　席】　田中委員、田畑委員、豊田委員、成松委員、松原委員、五條堀委員
【陪　席】

文部科学省ライフサイエンス課	川上調整官、田中調査員
産業技術総合研究所	野口副センター長、新間研究員、鹿内研究員
東京医科歯科大学	下川特任講師
理科学研究所	神沼研究員
情報・システム研究機構	堀田機構長
国立遺伝学研究所	大久保教授、池尾准教授
ライフサイエンス統合データベースセンター	永井特任教授、西川特任教授、川本特任准教授 箕輪特任研究員

【議　事】

１．統合データベースプロジェクト研究運営委員会作業部会について

ライフサイエンス統合データベースセンター（以下、DBCLS)高木センター長（主査）より、運営委員会および作業部会のリニューアルについて説明があり、以下の理由が挙げられた。

①中核機関、中核機関の参画機関、分担機関、補完課題と参加形態が多様で、各機関からのメンバーによっても運営委員会や作業部会への参加等の関係が複雑になったので、一度整理した。

②5 月に行われた中間評価において、中核機関と分担機関、補完課題、その他の機関とより連携を深めるようにというご意見があったので、作業部会の構成を実際にデータベース(以下 DB)構築や教育を実施している担当者中心に変更した。研究運営委員会は、わが国の DB の方向性やそのための体制について議論する場として位置づけた。

２．統合データベースプロジェクト中間評価結果について(資料２)

中間評価結果について、文科省ライフサイエンス課川上調整官より説明があった。以下、その概要。

全体評価としては、計画に対する進捗状況は順調以上に進んでいる。特に、中核機関においては、短期間に横断検索・文献検索を統合、目に見える成果を出しており、非常に評価できる。

事業の進捗・体制については、相互連携や教育体制が必ずしも円滑に行っていない、中核機関が主導的に PJ 全体の管理・運営できる体制となるように見直すべき、DBCLS により PJ 全体が統括・管理され、事業推進体制の整備されることが不可欠、プロジェクト（以下、PJ）終了後には整備された「統合 DB」の維持・発展が肝要、とのご指摘をいただいた。

昨年度の総合科学技術会議（CSTP）の方から、DBCLS と JST バイオインフォマティクスセンターについて一本化を含めた検討を行うことが必要との指摘もある。

分担機関については、

①京都大学…評価は高いが中核機関との連携の姿が見えない。KEGG の強固な基盤を活用について深い議論が中核機関との間に必要。

②東京医科歯科大学…取組の優先順位を考慮し、フォーカスを絞る必要あり。小規模なプロトタイピングに重点を置くべき。

③東京大学グループ…今後の DB の発展計画が重要。

補完機関については、

①理化学研究所…研究体制が的確に構築され、順調に準備がすすんでいる。あらゆるデータを積極的に公開す

る方向に進んでいただきたい。

②産業技術研究所…糖鎖関連の情報基盤の整理、有効利用が促進されることはDB統合のモデルケースとしても評価できる。

③国立遺伝学研究所…達成度は概ね妥当。「新しい種類の、あるいは新しい発想に基づくDBの開発支援」のような連携の在り方を今後期待。

④九州工業大学…小規模DBと統合DBの関係のモデルケースとして重要。

３．中間評価結果への対応・平成20年度PJ進捗状況について(資料３－１、３－２は各機関・課題とも共通)

（➤以下は参加者によるコメントや質問及びその回答など）

①中核機関（全体評価への対応(資料３－１)についても説明）

全体的な体制、管理運営、開発項目の見直し、個々のDBの必要性等について研究運営員会や今回の作業部会の構成を見直した。今後は、作業部会での議論を通じて、全体コンセンサスを得ながら進める。

人材育成の問題は効果を計りにくいが、本PJでは、育成と同時に中核機関や個別機関でDB開発の戦力になっているという特長があり、今後も継続予定。

臨床との連携については分担機関を中心に、医学、医療系メンバーが参加。医療系全般の対応は予算・体制的にも困難なため、モデルDBの構築に特化したい。

市民向けのコンテンツとしては、パンフレットの作成、バイオジャパン2008での展示といった専門家向けの広報の充実は実施中。やや一般向けでは、サイエンスアゴラでの展示を予定。今後は、ゲノム特定における市民向け活動などと連携しながら市民向けのコンテンツの企画も進めていきたい。

モニターによる定期報告制度の導入については、既にアカデミアや産業界から52名のユーザ評価のアンケートを実施、指摘に基づく改善を実施予定。

中期的な課題として、PJ終了後の体制については、文部科学省で開催予定のライフサイエンス情報基盤整備作業部会への情報提供、当PJの研究運営委員会での議論、関連分野のオピニオンリーダへの働きかけなどにより、コンセンサス作りを進めたい。

将来の継続的統合化のための予算の仕組み・構築については、低価格でDBを構築するような標準技術の開発を進めており、ファンディング機能についてもいろんな場で議論を深めていくべきと考える。構築済みDBの維持管理への民間資金の導入については、当面は国家PJデータを共有財として自由に活用できる環境づくりを最優先とし、受益者負担的な考え方はその後に議論したい。

省庁連携DBを促進できる体制の担保、医療情報との連携策の検討については、当PJ内に現在ある他省庁との連携を強化し、研究運営委員会等で省庁連携や医療情報との連携への議論していく。

計算機資源の融通制度について、サービスの大型化に伴い、計算機資源の不足は必至、豊富な計算資源を持つ機関の計算機システムの一部借用ということも交渉していく。

➤長浜の成果は市民に向けたコンテンツの充実という考えで実施されている。

➤計画は出しているつもりだが、もう一度、育てるべき人材の明確化とそれに向けて具体的な対策をもう一度詰めているところ。実際に効果は上がっていると思っている。

➤医学関係者との連携についての指摘は、医学臨床関係のデータを集めるには、データを持ってきてもらえるような臨床系や医学の先生に参画いただく必要があるのでは、という意味ではないか?

➤そういう関係者を入れれば済む問題ではなく、別の分野との連携を取ることこそ難しい問題。

➤市民向けのコンテンツの充実の必要という指摘に対する回答が、イベントやパンフレットとなっているが、中間評価の席で求められていたのはNCBIのような一般の人も利用できる形のサービス提供であると思われる。

➤そのようなサービスを実施するために必要なリソースが揃っていない。例えば、体系化された知識を提供する

ためには、テクニカルタームを整理して、一般的な表現で検索されてもきちんと検索がかかるようにすることが必要。NIH では、教科書など整理された知識の充実したコンテンツをあらゆるサービスに使うことができるが、本 PJ ではコンテンツ揃えからはじめている。

➤あくまで実施できている内容をまずは専門家向けに広報していることをご報告している。

➤遺伝研の WEB サイトでは、クイズやアニメーションのページから一般からの問い合わせが入ってくる。情報の出し方によっては、注目される可能性はあるのでは。また、米国の場合は、市民の情報を探そうという意識が高い。そういう意味では、市民教育としてのパンフレットやサイエンス・カフェ等のイベントも必要。

➤広報活動の対応部分では若干表現を変えたほういいかもしれない。

➤同じ質問を投げても、一般市民と研究者では違う答えが返るような環境を選べるようになっていれば、回答できているのではないか。

➤やさしい答えを新たに書き起こすのではなく、すでにきちんと編纂された教科書等を利用したい。そのような情報を提供してくれる協力者が協力したことへの見返りを感じてもらえるようにしたい。

➤教科書の利用に関しては、大学生向けのものを東大側で買い取った例もある。ユーザのレベルで検索結果を選ぶ方式は検討中。

➤現在使えるコンテンツは、非常に専門的か、あまりに頼りないかの極端。

➤専門用語をどうやって一般語に変えるかという問題は相手によって受け止め方が違うので、役所でもまだ難しい問題。

➤新たに文書を起こすのではなく、使わなくなった教科書の電子化して利用するというのは、一つの面白いアイディアだが、著作権等考慮すべき問題もある。

中核機関の進行状況について資料３－２と資料３－３（DBCLS 活動内容のみ）を説明。

DBCLS については、1)戦略立案・実行評価、2)統合 DB 開発、3)統合 DB 支援がある。

1)戦略…知財法律関係は、検討中にも問題が続出しており、早期対応が必要。調査に関しては、特に医用のリシークエンスに関する情報セキュリティの検討、次世代シーケンサの技術の調査を進めている。

2)統合 DB 開発…共通基盤技術では英文テキストと日本の文献のリンク技術についての調査、遺伝子名タンパク質名による full paper の対応付けを進行中。WEB サービスでは TogoWS の操作性を向上中。ヒト統合 DB の開発・運用では、文献の解読・処理技術の開発（論文正規化情報の取り出し、2 項関係エディターの開発、WiredMarker のキャッシュ機能)、遺伝子名辞書の高度化（不明瞭な専門用語定義するシステムの開発、種々の辞書の統合技術開発)、アナトモグラフィ高度化（医師との協力)、バンク目次（更新を継続）を実施中。モデル生物産業用統合 DB の開発では、微生物ゲノムのアノテーション用のパイプラインを構築計画中。

3)統合 DB 支援…ポータル整備・運用、広報普及では、６月２日にポータル関係の全体の成果として 23 のサービスを公開してプレスリリースをし、横断検索については、200 を目標に拡充中。生命科学系 DB については、昨年度分 350 に 150 を追加してカタログ化進行中。日本語文献としては、生物物理の年会要旨と論文誌を近日公開。それ以外は検討中だが、難航。また、クリエイティブ・コモンズやサイエンス・コモンズの活動との連携も検討中。広報としては、研究会共催やバイオジャパンの展示を実施。DBCLS と PJ を紹介するパンフレットを作成、個々の技術提供メニューに関するリーフレットの作成、ニュース配信を実行中。昨年度成果の評価を 100 名近くに依頼し、サービスのカテゴリーごとの評価やコメントを収集。特に多くのフリーコメントは非常に有用で、内部でそれぞれの対応策をまとめて公表していく。普及啓発については、統合 TV が予定どおり現在トータル 102 件の開発は終了。講習会もすでに 3 件開催(あと 4 件開催予定)。DB の受け入れと運用においては、基盤づくり(第 1 バージョン完成）と関連機関との連携(メタデータのやり取り)、タンパク DB の統合(理研・九工大の PJ 関連、仕様作成中)が進行中。本ポータルサービスのアクセス統計としては、7000 人／月のユニーク訪問者がある。現在、DBCLS のサービスについては、ログを収集、解析をしているが、この PJ 全体の評価の

ために参画各機関にも情報の収集をお願いする。

JST については 1)WingPro を継続して、新たに 7 つの DB を追加、2)事業の PJ の広報、サイトの運営は継続、3)DB の受け入れに関しては、Mouse Embryo の EST　DB を受け入れ予定。

産総研 CBRC については、ワークフロー(WF)を 3 段階で公開。１番目のアミノ酸の配列から立体構造予測を行う WF については、8 月に完成、現在限定公開中。２番目の予測とデータ取得による蛋白質の網羅的 WF については、今年末に公開予定。

かずさ DNA 研究所については、高度情報集積 DB（インターフェース改善、現在利用者 64 名）とゲノムアノテーション情報蓄積（7 万弱）進行中。

奈良先端大は専門用語辞書システムの開発(実装完了)、専門用語解析技術(進行中)、専門用語タグ付手法(今年度後半)を実施中。

九州大学に関しては、多型情報の第三者評価のためにシステムを組み込み、ゲノム特定領域からのデータで評価を進行中。

人材育成については、東京大学に関しては、DB 構築技術、DB 構築者の養成のため、DBCLS の２名を含む合計 12 名の受講・演習・実習が順調。

お茶の水女子大学に関しては、高度 DB 利用者の養成のため、養成プログラム実施中で、DBCLS の統合 TV 開発にそのうち２名の受講者が参加。

長浜バイオ大学は、5 段階のうち、初級アノテーション教育については、上期に終了。中級アノテーション教育については、下期の課題。自己組織化マップによる相同性によらない生物系統の推定については、２名の卒研生が OJT という形で、600 件の遺伝子の生物系統を推定した。シニア研究者と学生の共同作業である tRNA の DB が NAR の DB　Issue にアクセプト。長浜バイオ大学でも、４件の統合 TV を開発。

➤戦略立案・実行評価の項目に関して、調査以外に、この PJ の中の戦略検討や実行評価など、PJ 全体のマネージングやある程度のチェックする機能を期待。

➤実質的にはいろいろ行っていると思うが、明確化という点で少し弱いので、今後検討したい。

➤JST の意見集約システム WingPro の位置づけは不明確。素人から見てわかりづらいサービス。JST 広報で構築・運用している PJ のホームページを、対外的な統合 DBPJ の認知度アップや成果の公開サービスではなく、PJ のシナジーのために、進行状況をお互いが把握し、類似したツール開発等で強調的に機能拡充しようといった、コミュニティの一つの中心として機能するようにしてもらいたい。

➤WingPro は書き込み可能な Wiki で作られた DB 案内機能で、18 年度成果でもあるので研究者向けコンテンツの DBCLS とは重複しないように、米国政府系の公開データや EU 関係データを取り込んで維持している。DBCLS のカタログにも反映していただける内容。PJ ホームページについては、メーリングリストは DBCLS の方で検討中。掲示板については JST でテスト的に運用中で、近々メンバーにご案内する予定。

➤奈良先端大で成果目標にあげている機能実装はどのレベルか。まだ DBCLS のサービスに実装されていないが、いつごろ DBCLS の辞書等に使われるのかの見通しが知りたい。

➤実データを用いたチューニングが必要だが、そのための日本語のデータ収集ができない。日本語の辞書にするテクニカルタームの材料がない。文部省と学会が作ったものも著作権で使えない。今まで無料で創造的作業のために使える日本語の材料がなく、その収集から始めているため、それらを処理するテクニックはあるが、実作業ができない。

➤計画にも書かれているの、実行できるように努力する。

➤多型データを扱うグループの九州大学と東大の連携は？九大の目標設定や東大とのすりあわせが見えない。

➤日本の多型データの問題は、PJ 外の研究者にわかりにくく、データ自体のクオリティに対する不信があること。QC 自体にいろいろやり方があって難しい。

➤九大へのデータ提供も実施している。ただ、QC の考え方はいろんな観点の、いろんなレベルがある。例えば、タイピングのデータクオリティや、統計解析の段階など。それぞれの専門家が独自の基準を設けてやることは、結構なことだと思う。
➤初期段階のクオリティについては意見が分かれるものではないと思うので、頼れるデータを出して欲しい。
➤それぞれが勝手に目標を立てているのではなく、きちんと連携して進めていただけるのであれば問題は無い。そのような体制で進めていただきたい。

②分担機関・京都大学

中間評価への対応について五斗委員から説明があった。

中核機関との連携の姿が見えないということと、KEGG の強固な基盤をどう役立たせるかを中核と議論する必要があるという指摘について、中核との連携のところは確かにもう少し進めたい。今までは、キーワード検索エンジンは中核機関に、こちらでは化合物情報に特化した検索エンジンやツールを作るという方向で来ている。

医薬品化合物への DB の構築への配分の見直しについて、化合物 DB の統合には、構造検索や反応経路検索など様々な検索・解析のためのツールの開発費が必要になる。

医薬品 DB は JAPIC 以外のものをという点で、独自の情報をとりにいく必要があるということだが、JAPIC の導入に関しても初めての試みだったため、交渉等に時間がかかり難航したため、その他までは対応できなかった。現在はライセンス料を支払い、契約によりゲノムネットの医薬品 DB の形での公開について許諾を受けている。個別のデータは JAPIC のサイトに見に行く形。

➤厚労省の外郭団体が作成しているのに有償なのか?利用の仕方にまだまだ制限があるようで、それこそが「統合」に対する問題点。
➤添付文書については頻繁にアップデートが必要でその維持費等がかかるようだ。
➤もともと国からの資金以外に、企業の会費でやっている公益法人なので。
➤大学だったら無償提供してもらえるのか?
➤調べていないが、KEGG に対しては有償扱い。おそらく一般向けよりはかなり安価だったと思われるが。
➤もともと製薬会社が作成した文書集めているだけなら、個別にもらったら。
➤窓口を集約しているので、(文書収集のしくみが)回っている面もあると思う。

JAPIC の中でも構造が不明なもの(例:漢方)もあり、構造と薬効という形に特化して、化合物情報から天然物(漢方)の情報と医薬品をつなぐことと、ゲノムをつなぐことで、他の有料 DB との差別化は可能。そのためにゲノム情報に基づく合成経路検索で化合物を探索する方法を提案して実装していきたい。

ゲノムネット関係の開発費用は産総研の糖鎖関係の DB や中核機関の検索エンジン開発との役割を整理して予算配分見直しが必要という点については、糖鎖自体は産総研にまかせ、こちらでは配糖体や糖が結合した二次代謝産物を引き続き扱う。検索エンジンも、キーワード検索は中核機関に任せ、化合物に特化した検索機能の開発を継続したい。

KEGG の強固な基盤については、KEGG の3つの柱、パスウェイとゲノムと化合物の中でもパスウェイとゲノムを中心に合成経路との関係を検索可能にしたい。

20 年度の進捗について(資料3－2、資料3－4)、1．共通基盤開発は、今年度は、化学構造や反応ネットワークを中心に開発を実施している。昨年から作っていた構造検索のツールの効率が悪いので類似構造検索のアルゴリズムを高速化して、10 倍くらい速いものが完成。部分構造検索サービス(SUBCOMP)も昨年度来サポートされていなかったのを改良・バグ修正し、芳香環の認識をできるようにした。予備技術開発として構造情報を利用した方法を現在検討しており、今年度中にプロトタイプを作成予定。反応のネットワークの予測と酵素番号

自動割り当ては、化合物の構造の情報と反応の情報から、反応のパターンを DB 化している。昨年度までに作成した酵素番号の自動割り当てプログラムを改良し、DB 中の反応パターンをスコアリングする方法を改良、１２月中に公開予定。ネットワーク検索に関しては、反応のパターンを利用して、新しい化合物が得られたときに、それが合成される経路を調べるためのシステムを作っている。今年度中にはプロトタイプを公開予定。

２．統合 DB の開発について、ゲノムネットでは、欧州、米国、国内の DB をいくつかのカテゴリーに分け、全てミラーする DB と、キーワード検索だけできる DB に分けている。KEGG 以外に補足資料３－４の 4 ページに書かれている化合物 DB を対象としたキーワード検索を 7 月に公開した。また、手作業でリンク付けしたものとして、日本の脂質 DB（LipidBank)、アメリカ NCBI/NIH の DailyMed という薬の情報を、LinkDB で検索できるようにした。また、構造データもあった方がいいのではという以前のご指摘の対応として、PDB の LigandBox、農水の 3DMET、PDB-CCD と LinkDB で検索できるようにした。さらに、医薬品 DB については副作用や薬効のキーワードを抽出し、解析に使えるように開発中。

➤資料にある DB はこちらにデータを取ってきて、こっちで自由に索引付したりできるのか。
➤カテゴリー１および２が全部ミラーして自由に使えるものだが、そういうものはあまりなく、COMPOUND というのが京大で作成した DB がメイン。カテゴリー３というのは（クローリングでは無く）キーワード情報をもらって検索できるようにしたもので、検索結果はオリジナルサイトを参照するというもの。
➤カテゴリー４はキーワードがもらえないので、KEGG の DB と対応関係がとれるものを手作業で全部取得し、その対応関係だけ検索できる。だから KEGG 側の DB に無いものは入れていない。
➤NIH のサービスでも、全部パブリックドメインではないのか？
➤PubChem や PDB-CCD は使える。ミラー化を検討する際に、検索対象の範囲を検討すると同時に、こちらで新たに見せるページを作るより最新のオリジナルページを参照したほうがいいこともあるので、完全にミラーできるけどやっていないものと、本当に完全にミラーしにくいものがある。
➤PubChem は、ダウンロードして第三者へのサービスという形では使えないのか？
➤PubChem には terms of agreement があり、使ってもいいが、データソース（製薬会社等） との契約等も必要になる。
➤米国ではデータポリシーやアグリーメントなどがわかりやすく置かれているが、日本ではわかりにくい。
➤しかし、そのようなアグリーメントもどこまで何をしていいか、実はよくわからない、というのがある。また、米国内であれば誰が使ってもいい、という条件であったりする。
➤書いてあるだけましで、日本の公開 DB の多くには書かれていないので、実際使おうとすると困る。権利関係等も不明。

③分担機関・東京医科歯科大学

中間評価への対応および 20 年度の進捗について田中委員の代理下川特任講師から説明があった（資料の３－５）。

中間評価では、「DB 開発ではなく統合への課題の提言や小規模プロトタイプ構築等のロールモデル提供」を要望されたので、医科学 DB 統合のロールモデルの提示を目標として業務を進行中。要素技術として、DB の統合に際して必要になる共通キーワードの検索エンジンや、オントロジー等について、プロトタイプとしての設計仕様をまとめている。ISO や WHO の ICD11 などの国際標準化などに盛り込むように計画を進行中。また、国際的で広範囲な倫理規定の基本調査を実施しており、HIPPA 法の観点（患者個人を特定できる情報を出してはならない）から、種々の医療関連情報についての制限が問題であるが、統合化にあたり、想定されるような法的、倫理的、社会的課題に関する調査を進めていく予定。

「２つのモデル DB をどのように表現化して国内に広めるか」について、医科歯科大学が保有する網羅的疾患

分離病態 DB のモデルとして、癌を扱うプロトタイプシステムを構築しているが、これに国立がんセンター研究所の GeMDBJ を加えて、今年度中の公開を目指している。現在、肝細胞癌や大腸癌に特化して、使用されている用語やシソーラスを手作業で収集・作成中。今後、自動作業化を試みる予定。検索 GUI の改良を実施中で、より直観的な操作を可能にするために、検索キーを入れる部分のインターフェース等を立案して試作中。また、他疾患への本モデルの適応を検証するため、大阪大学のパーキンソン病 DB の症例等を追加し、統合化を進行中。

「中核機関との連携が不明確」という指摘については、医科学 DB におけるプロトタイプ、倫理案の内容、論点を整理明確化し、中核機関と十分な協議を行い、連携方針を策定したい、と考えている。

➤国内に広めるというのは、技術的な問題等を解決すれば、本当に可能なものか?
➤海外でも類似の DB 等が公開されており、マスコミ等に取り上げられることで、意識としては浸透していくのではないか?我々がやるべきこととして、外から入りやすい、使い易い DB を作ることがまず重要と考えている。

④分担機関・東京大学

中間評価への対応および 20 年度の進捗について徳永委員から説明があった。

中間評価では明解なミッションであることを評価いただいたが、３点の指摘があった（資料３－１）。第１点に関しては、いろいろな関連学会、シンポジウム、ワークショップで、私あるいは関係者が説明・登録を呼びかけている。学術雑誌でも、日本あるいはアジア初の GWAS の結果を、論文が出たら登録してもらうよう呼びかけている。第 2 点については、いろいろな段階で様々な品質管理法が提案されており、国際的にもある程度使われているものが何種類もある。それらはなるべく搭載し、ユーザがある程度自分で基準を設けながら計算できる機能を付けていくことで、使いやすく結果もわかりやすいものにしていこうと検討中。第 3 点について、中核機関の計算機資源を既に利用していろいろな計算がスムーズにできるようになった。中核機関との定期的な相談を続け、いい環境で解析されたデータを DB に搭載する作業を円滑に進めたい。

成果目標と進行状況および資料の３－６について、標準 SNPDB に関しては、新しい解析技術で出てきたデータを順にとにかく搭載している。最近ではアフィメトリクス 6.0 で出したデータを新たに追加した。GWAS に関しては、何か月に１個位のペースで新しい疾患データ（関連解析結果）が搭載され、Copy Number Variation に関しても、今年度中に公開予定。資料に示したように GWAS の関連解析の新機能としては独立なスタディを比較できるようになった。CNVDB については統一基準の無い QC について私たちが考える基準で DB を作成した。リシークエンスについては、パーキンソン病について今年度公開準備中。原因遺伝子と臨床的な特徴を搭載した DB を準備した。

最後に、データの公開や共有について、第１案の最終版を作ったが、国際的な状況が変動しており、確定するまで議論が必要である。これまでは GWAS チーム内で検討してきたが、やはり統合 DB 全体の問題で、早急に中核機関と一緒に考える必要があると考えている。

加えて、最近 HGVBASE の genotype to phenotype の DB の担当者からミラー化の提案がきており、これについても統合 DB としての方針が欲しいが、今後相談していきたい。

➤向こうをミラーしないかという提案か。
➤相互にデータを完全に交換する提案である。
➤データは交換可能になった方がいい。
➤そのためにも基準の問題を要検討。公開と共有あたりの方針を決めなければいけない。
➤リシークエンス DB も個別のグループがそれぞれの持つ患者サンプルをリシークエンスしているのか。倫理の問題が障害になって個別のグループごとに独自のフォーマットで DB を作成するという現状では、疾患×グループの無数のリシークエンス DB ができてしまう。その解消のためには、特定の疾患については拠点を１箇所に決

めてやるという体制を作り、ガイドラインもそのグループ内で共通に納得できるものをつくることのほうが重要なのでは？東大でガイドラインを作っても、他の大学では、別の倫理委員会で別のガイドラインで、では問題は解決しない。
➤ガイドラインに関しては、統合 DB で作成することは、意義があると思う。検討している機関はそんなにないと思うので波及効果があると思う。
➤ガイドラインの議論については、来週の火曜日（10 月 14 日）にゲノム特定のフォーラムがあるので、ご意見を言っていただきたい。
➤試料の方の共有化については、（産総研からも）提案している。
➤うまく共有させるための仕掛けが何かあるのか？
➤厚生労働省の研究所などが国の研究費で使ったものは、そこに DB を整理して誰からも見られる形にすべき。今は所在情報もわからない。研究員が個別に交渉して試料をもらうような状況だと何も進まない。
➤さらに省を越えての仕組みはもっと大変。
➤それに対する提言も DBCLS のミッションの一つ。
➤データが出てこない話は、企業秘密のものまでが対象か?それとも、共有すべきものの話？
➤企業が絡む知財関係の話ではなく、医療機関が絡んでいる臨床サンプルの話。
➤出てこないのは、個人情報が関係するのか？もっと違う根拠で、そもそもやり方がうまくないのか？
➤それぞれの機関が責任をもって管理しなければならず、個別に倫理委員会を通して MTA を交わさなければ外に出せないことになっているので、個別に交渉して契約を結ばなくてはいけないという制度の問題だと思う。
➤もともと権利が保護されていないものは、中へ閉じこもる性質がある。一方、特許や著作権があるものは、権利が保護されているので流通する。データやサンプルにはそれがないので、所有していることでコントロールしている。
➤だから MTA を結ぶ必要があり、それ自体がかなり面倒になっている。
➤公金でできたものは全部皆のものにしてしまえば。
➤それを希望しているが、そのやり方として、医療機関が個別に実施ではなく、厚労省が音頭をとって、国の機関でやった試料の所在情報などが見える形にしてほしい。
➤データもサンプルも個人的に売買できないことは自明なのに、公開する部分を決めるといった扱いについてはあいまい。本質的には売買しているのと変わりないと思う。
➤本 PJ 内でも「自分で非公開と決めた」という例もあるが、これはデータの扱いに関する勝手なローカルルールでは? PJ 全体での公開に関する方針を共有して欲しい。
➤ローカルルールはよくないが、現在、いくつかの具体的なガイドラインづくりを進めているところで、あらかじめ全てのことにルールを決めてから実行することもできないので。

⑤補完課題・理化学研究所

中間評価への対応および 20 年度の進捗について豊田委員の代理神沼研究員から説明があった。

中間評価の指摘事項への対応について、１点目「国内外の他機関との協力を推進してください」については、現在開発中のアノテーションシステムを 11 月に公開予定で、他機関がシステムを利用できることで、情報交換が可能になり、連携が強化される、と認識。2 点目「理研内の全データの積極公開」について、研究者個人が公開を希望するかどうかは研究の発表状況により未発表分は公開できないので、個人ごとに同意をとり、公開を希望する場合は統合 DB に入れていく、という方針で運用。また、同時にライセンスについても公開時に問題になるが、理研内での話し合いが進行中で、その進捗によって対応する。

進捗状況（資料の３－７）について、植物のオミックス情報に関しては、6 件の DB の統合が完了し、公開準備中。アノテーションシステムの開発・運用については、database.riken.jp として、11 月に公開予定。横浜研

究所でのタンパク 3000 でのデータの搭載については、データの回収、ヒモ付が終わった段階。微生物由来のタンパク質構造（播磨研究所のデータ）については、今年度末まで公開予定で、データの整理中。

➤ゲノムからフェノームまで多様なデータを加工しているが、横断的な検索は単なるキーワードサーチか？
➤今のところは、キーワードサーチのみだが、いろんな検索方法を実装中。例えば、理事長ファンドで開発しているセマンティックウェブのシステムを利用したものなど。
➤パターン認識などもフェノームに関して必要になってくる。セマンティックについては、機能情報や機能推定といった出口もあるのでは。
➤タンパク 3000PJ で理研が構造決定したタンパク質全てが対象か？
➤メンテナンスが難しくなったものを優先して搭載中。新規データについては、状況未確認。
➤中間評価の指摘事項では「理化学研究所内でのあらゆるデータの積極的公開」とあるが、それに対する回答が「研究者が公開化を希望する DB 化」となっている。指摘の趣旨は、研究者が公開を希望しなくても公開してください、ということではないか。
➤強制的にというわけにもなかなか。また、研究者に聞き取りをしたところ、提供者側にとってメリットのあるサービスを提供していないと出す気にならない、という率直な意見を聞いている。
➤国費を使っている独法組織からデータが出てこないという問題はこの PJ 全体の問題。良いデータを持っている理研が補完課題として加わるに当たっては、そういうものを出してくれることを評価委員も期待し、ライフ課も支援したはずだが、やっぱり出してこないというスタンスでも PJ 費を付けている責任は、付けている側にあって、それをコントロールできない DBCLS にあるのではないということははっきりしていただきたい。
➤公開していない機関から無理やりださせるということを DBCLS に期待しているのではない。むしろ重複して開発されているものを戦略的にコントロールして欲しい。すでにあるデータについては、公開可能なものについて、積極的に提供して広めていくということで、理研は植物関係のデータについては公開できるので、プロトタイプとして公開し、順次理研内のデータを公開していくという計画で参画していただいている。また、理研の組織は非常に大きく、考えや整備が決まっていないので、理事を頭に協議会をつくって進めてもらっている。
➤内部的にも提供者側にもメリットがあるように、と働きかけを進めており、少しずつ協力者は増えているので、前進していると認識している。
➤では、「理化学研究所内のあらゆるデータの積極的公開」という指摘事項は、ライフ課あるいは採択委員の考えとずれているのか？
➤ライフ課・文科省としては機関に対して強制的に出させることはできる話ではない。方向としては、やりながら一歩一歩進めていく話で、自発的にやっていただくように持って行くというのもライフ課・文科省の仕事だと思っており、そのためのこの事業。
➤外部から評価への対応はできないということか？ 個別研究者が公開したくないと言っていたらそれを打破するのが必要なのでは。
➤指摘事項に対しては、理研も積極的公開に動いていると認識している。シロイヌナズナを成功事例に理研内の他のデータも同じようにしていく。最初のプロトタイプがこの事業の中での約束。
➤シロイヌナズナができたのは、内部の人間が整備・加工するならいい、ということではないか?外部の人も自由にオリジナルデータが使えるようにはしてもらっていない。
➤加工の中身は整理しているに過ぎないと思うが。
➤セマンティック化はかなり込み入った加工である。
➤ゲノムや cDNA は定型的な処理があるので変わらないが、フェノームや画像データは、生データの処理の仕方によっては知識抽出の深さが全然違ってくるので、生データで公開してほしい。
➤フェノームのデータから、高次レベルの情報を抽出するということ自体が研究。その研究をこの PJ の中でや

ることがおかしい。
➤フェノームのデータに関しては、画像を収集されているデータ提供者の先生方と MTA を結び、オリジナルの画像も掲載。全て論文発表後はオープン。
➤ゲノムは発表前から論文化に関する条件付で公開する。画像データも同様にできないのか？
➤研究者のお互いにインセンティブがあったら出すというのは、統合 PJ の精神に反している。理研のトップがこの精神を共有していないとすると、この PJ に入っているのはおかしいということ。
➤理事長の方たちもそういう認識を持っているので、協議会を作って話を進めている。
➤個々の研究者を説得しているということか？
➤説得するためのルールづくりをしている段階。

⑥補完課題・産業技術総合研究所糖鎖医工学研究センター

中間評価への対応および 20 年度の進捗について成松委員の代理新間研究員から説明があった。

「自前データのダウンロードなど様々な利用者のニーズに備えてほしい」という指摘について、生データそのものはかなり難しいが、国の税金を使ったデータは基本的に出すものだということを話している。一方で、われわれの研究費は民間企業と共同研究して産業界に役立てるためなので、いろんな縛りがある。そこで委託研究契約書、共同研究契約書、特許許諾契約書を検討し、何が本当に外に出せないデータなのかを関係者と議論した。それ以外は基本的に出すことにした。

次に、「なるべく多くの関連 DB の統合を進めてほしい」については、順次進行中。GlycoEpitope、GALAXY、ノックアウトマウス、LipidBank、を今年度公開予定。来年度に向けても、複数の研究機関と交渉中。

この統合 DB の契約の形態に一つ問題がある。データを公開・統合用に改良していく費用がかかるが、産総研以外の大学や民間企業等他の団体は統合 DB の契約に入っていないので、研究費を渡すことができないという点。「他分野とも連携したより上位の統合の具体策の検討」については、糖鎖の DB についてさまざまな場所で発表したところ、日本脂質学会からの申し出があり LipidBank を今年度公開統合した。宣伝は重要であると認識した。糖鎖研究者以外に向けても実際に DB に触れることのできるタイプの紹介をしていきたい。DB が統合されることにより、糖鎖に関係する情報がいつも見えるようになれば、ライフサイエンス分野の研究レベルが一段と上がると期待している。

進行状況として、資料３－２の１．については水面下で交渉を進行中。２．は提供可能。３．はキーワード検索の方は運用を開始。糖鎖構造による検索はシステムを開発中。糖鎖構造のひな形をクリックし、構造を編集して、検索するというもの。４．未着手。５．ノックマウスを題材にモデル生物の情報を集約するカテゴリーの構築を行うことだが、名古屋大学と連携。６．立命館大学、名古屋市立大学の DB を引き出す API については現在進行中。

➤産総研 GGDB のデータを引き出す部分の API を経産省側の PJ で作るので、それ以外を本 PJ で対応予定。
➤経産省とはカタログ作成、インデックス、横断検索的なことを共通してやる話は、別途進行中。ある程度連携はとれつつある。どこの省でやっているか区別する必要はない。

⑦補完課題・国立遺伝学研究所

中間評価への対応および 20 年度の進捗について五條堀委員の代理池尾准教授から説明があった。

中間評価への対応について、1 点目「新しい種類の、あるいは新しい発想に基づく DB の開発支援」についてはなかなかむずかしい問題なので、中核機関と今後いろいろ相談し方向性を出せるくらいまでは、頑張る。

➤『「新しい種類の、あるいは新しい発想に基づく DB の開発支援」のような新しい連携の在り方』の意味は?

➤トレースアーカイブ（以下、TA）にとどまらず、今後予想されるいろいろなデータ（新型シーケンサなど）まで視野に入れた検討をしてはどうか？というコメントだったと思う。
➤報告書では、その前に「中核機関との連携においては」とあるので、一緒にということではないか。

新型シーケンサの対応については、提案時点ではあえて述べていないが、データの性質は TA とオーバーラップするので、新型シーケンサを想定した DB のツールの設計を実施した。ただ、国際的にも新型シーケンサ仕様が固まっていないこともあり、今後開発や標準フォーマットも変動していく可能性も。そのため、TA はベースとしてしっかり作り、できるだけ新型シーケンサの対応も進めていく。

進捗について、TA 処理概要は一通りこなせるようになったが、今後、日常業務的に数をこなすためのチューニングと効率化が必要。また、大規模のゲノム配列に関しては、米 NCBI が一手に ID を発行するので、データを向うに送付するステップが必須。日米間を大量データのやりとりを FTP でやっているが、新規技術の導入を検討中。検索について、データ公開は FTP 経由であり、大規模にデータを処理する情報系処理には使えるが、個別研究者には使い勝手が悪いので WEB 経由の検索機能が必要と考え、プロトタイプ構築中。今年度中公開予定。これまで公開したデータは 2 種類あるが、受付から公開までの時間がかかりすぎているので、運用上のシステムを作り、２～数週間で公開できるようにするべきと検討中。

➤新型シーケンサデータの登録に際し、３つ問題点がある。１．多数の項目を埋める必要がある。２．ABI Solid のフォーマットに対応していない。ヌクレオソームの項目に対応したものがないなど、NCBI の方も混乱。３．データ転送の問題。これらを NCBI と解決して欲しい。
➤転送については、速いツールを導入し、遺伝研で代行登録することにしたい。フォーマットに関しては、議論されているところ。登録時の記入項目については、検討中だが、ぜひご意見が欲しい。
➤登録のための記入項目は、整理されておらず、かなり複雑。アクセッションが必要だから、やっているだけ。転送速度にしてもかかりすぎるのであれば、一般の研究者がダウンロードして使うのか。
➤計算機パワーが必要だが、興味あるデータであれば使うであろう。

４．総合討論　その他
➤「PJ 内の情報の共有」について、委員の方々に伺いたいが、それはそもそもできるだろうか？他の PJ でも、情報共有と言われるが、ポータルサイトの情報から、一緒にやろう、という流れになっていくのかどうか。
➤なかなか難しいのでは。
➤ライフ課あるいは評価する方が、状況をみるための共有といった観点が強いのではないか。
➤例えば、中核による展示会のブース出展があればその機会を PJ メンバーが活用するための情報を共有する、といった使い方ができるのではと思う。メンバーの中である程度お互いの取り組みを助け合うために。
➤宿題事項として検討する。次回から、時間を長めにとるのか、議題をまず半分にしてやるのか工夫したいが、今日は大変有意義だったと思う。また、３か月か４か月に１回開催というペースで考えている。（終）