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分科会
【立体構造・植物、人材育成関係機関】議事要旨

【日　時】　平成２２年2月2日（水）１３：３０～１６：２５
【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】　豊田哲郎(理研)、国島直樹(理研)、淺田征彦(理研)、皿井明倫(九工大)、Shaji Kumar(九工大)、中村保一(遺伝研)、藤澤貴智(かずさ)、岡本忍(DBCLS)、中尾光輝(DBCLS)、森下真一(東大)、中谷洋一郎(東大)、池村淑道(長浜バイオ大)、瀬々潤(お茶大)、田畑哲之、菅原秀明、堀田凱樹 (ROIS)、中村春木 (DBCLS 客員教授)、大久保公策、高木利久(作業部会主査)
永井啓一、川本祥子、坊農秀雅、畠中秀樹、吉羽洋周、河野信、髙祖歩美、坂東明日佳、平井信一、箕輪真理(以上、DBCLS)

（敬称略・順不同）

【議　事】議事に先立ち、高木主査から、本日の会議については、タンパク質関連の議論もあることから、DBCLS 客員教示で運営委員会の委員でもある中村春木先生がテレビ会議で参加すること、植物関係の議論もあるのでかずさ DNA 研究所のメンバーが参加することが報告された。

１．平成 22 年度予算配分について(文科省代理として高木）資料 1-1～4

作業部会において各参画機関の取り組みについて相互に評価を行い、外部の評価結果も参考にしながら、来年度の適正な予算配分を行いたいので、各機関に自分以外のところで利害関係の無い機関について 22 年度の計画の評価をお願いしたい。次センターへ引き継ぐべき成果であるかどうか、確実な成果が出るかどうかについて精査いただきたい。資料 1-3 にあるスケジュールのように進めたい。参考資料として、資料 2（昨年度目標と年末までの達成状況）を参照いただき、〆切が厳しいが明日までにご意見をお願いしたい。いくつもの取り組みのある機関について、それぞれの取り組みについて意見のある場合は、自由記述の欄に記載ください。

◆質疑応答◆

○資料 2 にまとめてある内容は、それぞれの機関から出てきた報告に基づくものか？それとも例えば中核がまとめて書いたものか？

→各機関から情報を集め、それを中核で多少編集して掲載している。内容については各機関が了承しているはずである。

○評価についてはこのまとめたものでなされるのか？

→評価者については、すでに計画の文書やもう少し詳細な達成状況についての報告があり、また、公開情報(ホームページなど)も見ながら実施することになる。

２．中核機関より(川本、畠中)

プロジェクトの最終年度の取りまとめを前にして、各機関がいろいろな DB を公開していくにあたり、中核機関の活動内容をご理解いただくために、これまでの達成状況や 22 年度の目標が川本特任准教授から、また、統合のひとつの例としてタンパク質関係の情報の統合システム TogoProt が畠中特任准教授から紹介された。

◇資料説明◇

これまでの進捗の紹介と 22 年度に統合に向けてどのような考えで進めようとしているかをご理解いただくためにご報告する。最後に、タンパクの課題も中心になっているので、その点について特にご紹介したい。

取り組みとしては、統合ホームページの概要（サービス数の増加)、アクセス統計、カタログ系サービス(情報更新、今後も継続)、横断検索(集めた情報を一括検索、統合検索にどのようにつなげるかが課題)、アーカイブサービス(“受け入れ”課題への対応)、統合 TV(数増加、英語対応)、共通基盤開発(取り組みの内容と情報流通における位置づけ)、これらの年次計画上での進捗を紹介した。次年度の計画としては、統合化の全体像とステップを提

示した上で、中核（DBCLS）と各機関の連携状況、横断検索の機能向上から統合検索への展開、そのためのツールとしての辞書シソーラスと各機関で分野的に進めている統合 DB を連携させ、より大きな統合にしていくことが課題であることが示された。統合検索のプロトとしての TogoProt を紹介した。タンパク質関係のいくつかの DB 情報をつなげることにより、タンパク質に関するいろいろな種類の情報をつなげることを試みている。また、これに関連して既存 DB のミラー構築、文献からの情報抽出など、個別の DB についての試みも行っている。最後に、ユーザー評価について、各機関の協力に感謝するとともに、結果の公開について紹介した。

◆質疑応答◆

○産総研の構造予測のワークフロー(WF) 、理研のアノテーション(構造由来)、微生物アノテーションなどもできているようであれば、TogoProt につなげていくべきではないか。

→WF については作業部会での報告を伺ったので、これをきっかけに進めたい。理研の構造データについては、量が大量なので、データそのものを DBCLS で補完するのではなく、データを説明する情報を載せていきたい。

○理研についてはリンク情報のみ掲載するということか。

→個別データのリンクを想定している。

○表のデータを移すのは構わないが、理研で付与した、例えば、カルシウムイオンについての ID(同じ Ca でも意味が違うので異なるセマンティック情報を持っている) を取ってしまうと、再配布の際に問題になるとおもう。

→今後ご相談しながら進める。

○TogoProt について、再利用できるデータとしては各 DB 間の ID の関係情報がほしい。情報処理上必要。

→リンクのタイプについても記載することが必要ではないかと議論中。将来ダウンロード可能な形にしたときに、高度な利用ができるように考えていきたい。

○スライド 11，12(統合化の全体像とステップ、中核 - 各機関の連携状況)を見ていて感じたが、評価の観点としてこのようなことが期待されているのか？

→評価に関しては文科省の考えで進められるのであくまで推測だが、利便性が上がったとユーザーがどれほど感じるかを中心に考えてほしいとのことであった。私の理解では、この PJ の始まりにおいては必要な 10 項目があり、そのうちの 7 項目が PJ 側で実施すべきこととなっていたはず。それが大きな意味で評価基準。

３．プロジェクトの平成 21 年度の進捗状況および平成 22 年度業務計画について

- 理化学研究所

◇資料説明◇　資料 4、理研からの参考資料 1～7

資料 1－4 にある計画の 7000 万円については、シロイヌナズナ 3500 万円、タンパク質関係で 1000 万円＋1000 万円、アノテーションシステムに 1500 万円。シロイヌナズナについては、これをモデルとしてデータ統合の仕組みを作っていくことで、理研の中で他のデータについても統合できるように進める。アノテーションシステムは公開に結び付く部分。このプロジェクトで実施項目として挙げているもの以外に、マウスのデータも多いが、国際的な同意を取りながら進めている。また、医療に関するデータは、創薬プロジェクトや他機関との連携もあるため、すべてのデータを公開できていないが、データの構造に統一感を持たせている。理研全体の方向性を内外に示すためにもプレスリリースを行った。関連技術の紹介論文も NAR の WebServer issue などに掲載されており、評価の基準として利用度が高いことが求められている点にもある程度答えられていると思う。

22 年度の目標については、各種 DB については様々なオントロジーを利用したアノテーションを公開し、データ取得のための API 充実、DB を使ったプログラムの DB の公開も手掛ける。また、合成生物学に関するウェブ上のコンテストを実施、DB の充実を図るとともに有用な資源を生みだす DB を目指す。

微生物由来タンパク質については、昨年 DB を公開し、さらに情報を追加した。来年度も、アノテーション情報の追加、後半は利用度を高めるような工夫をしていきたい。

◆質疑応答◆
広範な内容なので、タンパク質→植物→セマンティックウェブの順で質問をお願いする。
◆タンパク質
○コメントだが、国島先生の DB で、PDB への理解が足りない。重電子のデータのリンク先が、古い URL のままでリンク切れもある。ユーザーが困るのではないか。
→ありがとうございます。
○蛋白のアノテーションの内容はどのようなものか。
→データの回収のほうが先に必要だったので、まずはそのまま格納してアクセス可能にしただけ。どういうオントロジーで進めるかを、関連機関や先生方と協議しつつ進める必要がある。国際連携についても視野に入れるべきと思う。
→アノテーションについては、播磨の元データがいろいろな形式だったので、それらの整形やリンクの整理が主。それらの作業が終わり次第サイネスで公開しているという状況。
○ターゲットタンパク PJ との連携などはあるのか。共通のフォーマットとか、オントロジーとか。ターゲットタンパク PJ 関係者からコメントをいただけないか。
→ターゲットタンパク PJ の中に情報プラットフォーム(PF)があるが、PF 運用と公開がミッション。タンパク 3000PJ で決まった構造を検索・閲覧できる(実際に表示されるのは PDB の情報)。裏付けデータなどは公開していない。今回のデータ整備で実験情報などが整ってくれば、こちらの PF からも参照できるようにさせていただきたい。
○理研のアノテーションは理研での実験データ(回折などの生データ)が主であるとおもうので、プロトコールデータが主である PDB のほうとは重複が無いはず。相互参照すれば有用。PDB ではプロトコールに関するオントロジー開発がほぼ終了している。
→まずはデータの保全優先。オントロジーは変化していくものでもある。データが見えているからこその議論。来年度中に何らかの情報を出すためにも、関係者で協議しながら早めに進めていってほしい。
○アーカイブサービスの場合、入れにくいデータが存在する。サイネスの場合、実験の結果どのように役立ったかという情報を見つけるのは難しいようである。
→未公開のレベルのデータを管理して公開するところまでとりあえずこぎつけるのがサイネスの目的。アノテーションについては、ラボ内サービスに近いレベル。複雑な検索ができるようにするのは 2 次利用。むしろマシン上で利用できるようなデータ構造（RDF 構造）になっているが、すべてのレベルのユーザーに対応しているものではない。

◆植物関係
○表現型の統一について、必ずしも理研の基準とは合わない部分もあると思うが、国際会議への出席などにより、オントロジー設計への貢献予定があるのか。
→オントロジーデザインまで作れないが、シロイヌナズナの分野での国際協調のためにも出席予定。理研からリンクをはる利用方法を検討。BioMart への応用はかずさがやられるので理研では実施しないが、共通の ID については、DL するときに保存できるようにして、その後の統合に生かせるように残したい。
○PATO はすでに世界標準になっているのか、かずさでも採用しているのか。
→かずさでは使っていない。遺伝子ではなく、表現系のオントロジーである。
○どのくらいの規模のオントロジーか。あまり大きいと扱いにくいのではないか。
→資料にあるように 7000 はある。オントロジーとの対応付けは経験者の人手で行う必要があり、さらには実験をした本人に聞かなくてはならない情報もある。
○生物種ごとにオントロジーが既に存在していると思うが、それらが横につながったものか。

→OBO というオントロジーの標準化の一環。植物固有の現象などの用語の組み合わせで表現するもの。植物のオントロジーは、種ごとに成立しているものや、研究者独自に構築したものなどもあり複雑。

◆セマンティックウェブについて

とくになし。

- 九州工業大学(皿井)

◇資料説明◇　資料 5

熱力学データと構造データの統合により、生体分子の機能解析を支援する各種 DB を構築。DB の構築のための構造化の技術開発、および文献からのテキストマイニングの技術開発（DBCLS と連携）を実施。DB 構築プロセスは非常に多くのステップからなり、一番時間のかかるのは研究者が論文を読んで情報を抽出するところなので、この部分をなるべく自動化できることを目指している。21 年度のそれぞれの DB(熱力学／構造、蛋白質 - 核酸相互作用)のエントリ数増加に加え、新たな DB(蛋白質間相互作用)のプロトタイプ構築、オントロジー構築に向けた Vocabulary 整備、データ公開など行った。22 年度は、既存 DB のデータ追加、クロスリファレンスの作成に加え、オントロジーの整備、ツールの自動化(DBCLS と連携)を進めたい。

◆質疑応答◆

○XML 化やオントロジー整備について、数式との対応は。

→実験の定量的データなので、将来的には数式も含めた統合の可能性があるが、現在はなし。

○オントロジーについて、Vocabulary を整理しているということか。

→今のところは整理のみ。熱力学的な数値情報も含め、今後体系的にオントロジー化していきたい。

○DBCLS とのツール開発の連携の結果として収集されたデータの公開についての取り組みはどうなっているか。

→(DBCLS から)すでに TogoProt に一部のデータは組み込まれているが、ID の整理の過程で気がついたところなどあるので、議論中である。DBCLS としてはアーカイブでの公開を望んでいるが、アーカイブについては以前ライセンスの関係でちょっとむずかしいという話を伺ったと思う。

→(九工大)アーカイブでの公開も可能である。

○アーカイブに置くことも可能で、TogoDB を利用した公開も可能という理解でいいか。

→そうである。

- 東京大学

◇資料説明◇　資料 6

バイオ DB にかかる外注費の圧縮のために、内容がわかり、構築もできる人材を育成することが目標。演習計画については、①バイオ DB サーバー構築演習、②プログラミング演習、③独創的サーバー構築演習を予定したが、②をクリアすることがかなり難しいようだ。20 年度の修了生は 3 名が就職(DB 系企業も)し、21 年度にはプロジェクト関係機関からの参加もあった。DB サーバー構築演習については 10 段階の目標設定を行った。一部は受講者のレベルや専門性に合わせた内容を提供。4 名のほとんどが既に 10 段階を終了。独創的サーバー構築演習についても、それぞれの設定したものを構築。21 年度は 20 年度の受講者による講義・演習も行った。

◆質疑応答◆

○独創的サーバー構築演習において、活用された論文が出たとあるが、そのような論文が出た際に報告してもらえるとありがたい。活用情報だけでもいい。

→論文がまだアクセプトされていないので、まだ報告していないが、論文が出たらご報告する。それほど例が無いが、今後出てくると思うし、来年度は実績の一部として報告できると思う。企業に行かれた方の情報

についてはよくわからない。
〇ツールを作成した人についても報告もしてほしいが、そのツールを使った人の発表情報もあるといい。
→その通りだと思う。
〇テキストを作っていると思うが、アクセス可能になっているか。
→公開しているので、自習もできるはず。
〇独学で本当にできるか。つまずくのはどこか、テキストだけではダメなところなどの情報はあるか。
→バージョンの違いなどでつまずいていることが多い。パーミッションが正しくなくてつまずく場合もある。エラーが起きた時には対処も含めて演習ノートに書くようにしているので、ある程度自習でもできると思う。
〇プログラミング能力に依存するというよりは、分野の知識の量に依存するのか。
→コンピューターの知識に慣れていない人が受講しているので、基本的なところでつまずくことが多い。
〇誰でもプログラムできるようになるか。
→それは無理な人もいるかもしれない。アルゴリズムなどは難しいかも。構築演習はこなせるが、アルゴリズムには数学のトレーニングも必要。
〇アルゴリズムも知識か。それともセンスか。
→帰納法の考え方が必要で、高校生は苦手。それまでに積み上げた知識に依存する。計画の①は誰でもできる。②はむずかしかった。ちょっと目標が高すぎたので、適切な目標設定ができるようになったことも成果。

- お茶の水女子大学

◇資料説明◇　資料 7

DB を使いこなす人材育成を目標としている。21 年度の数値的な目標は達成している。講義・演習は基礎、専門と分かれており、基礎の部分ではバックグラウンドによって補足する内容。また全体として社会人向け、一度離職した人にも対応する内容や講義日程を設定。また、モチベーションが持てるように、また実感が持てるような内容を工夫した。募集した社会人(のべ 10 名)、学生(主として修士 30 名)の受講があった。統合 DB の活動にも貢献している学生がいる。22 年度の受講生については現在募集中。資料はすべて公開。

◆質疑応答◆

〇自習用テキストは利用されているか。
→閲覧数はある。書き換えが頻繁にあるので、現在はダウンロード（DL）できないが、最終的には DL 可能にしたい。
〇どのようなところでつまずくかのノウハウは貯まっているか。
→情報系の人はモチベーションが足りないので、ブラウザの導入などでわかりやすくなる。生物系の人は、プログラミングは難しいが、パラメーターの設定の違いでとれるデータが異なるなどを見せると、効果あり。
〇教育を受けた人の受け入れとして、DDBJ や理研からの意見は。
〇DDBJ では次世代シークエンサーデータの処理ができる人(既存ツールなどかき集めて、パイプラインを作れる人）がほしい。
→わかりました。パイプラインの理解まではなかなか到達できないが、それに向けて一歩進められる内容を検討する。
〇理研ではこのような人材はそろっているのか。DB 構築の人材を募集しているのか。
→(理研)万能の人はいないので、分業の際にそれぞれの人の長所を見ながら組み合わせている。人が足りているわけではないが、チームプレーでできる人がいれば。
→現在は、それぞれの人がそれぞれの能力に合わせて進めているので、そのようにはなっていないが、複数人で一緒に作業するという考え方もあると理解した。一方で必要だといわれるとやる、という面もある。具体的な人のニーズについても教えていただければ、ありがたい。

- 長浜バイオ大学

  ◇資料説明◇　資料 8(実際には、資料 8 に情報が追加された資料が別途配布された)

  今年度に追加された内容を中心に説明。長浜バイオ大では、実験と情報解析の両方を習得した人材を育成するために、今年度コンピュータバイオサイエンス学科を開設した。コンピュータバイオサイエンス学科 1 回生が DB 実習 I として、バイオ技術やバイオ装置に関する収集を実施した。今年度から、次世代シークエンサーデータに対応したアノテータ、キュレータの教育を、3 回生 50 名を対象に実施した。昨年度から継続した活動としては、2 回生全体 250 名は「バイオ DB の統合利用」、3 回生全体 250 名は「健康への貢献遺伝 DB 構築」を行った。他大学を含む外部からのテキストの利用、出前実習の依頼の例もある。高校生への啓蒙活動(含むテキスト作成)も実施。4 回生についてはシニア世代との共同作業としての tRNA 遺伝子の DB 作成を実施。来年度は、大学院生も参加させて高機能画像も取り込んだ DB への活用を試みたい。

  ◆質疑応答◆

  ○次世代シークエンサーデータを学んだ学生さんの進路(就職)は。

  →学習した学生はまだ学部 3 回生なので、就職活動を開始したばかりだが、次世代シークエンサーデータを取り扱う分野へ進みたいとの学生の希望を聞いている。卒業生の中には、次世代シークエンサーデータを扱っているものがいる。

  ○求職情報として、そのような人材が欲しいという声はあるか。

  →新卒の学生は Permanent 志望が多いが、研究機関だとどうしても任期付きのポジションが多いので、マッチングが難しい問題がある。但し、ライフサイエンス分野の DB に興味を持ち、研究機関の任期付きのポジションに進んだ卒業生もいる。

- 総合討論

  ○人材養成 3 機関の話に関連して、何か抜けているところはないか。3 機関の実施内容の連携のようなことは考えているか。

  →この 3 つの機関の活動は今後必要と思われる 3 種類の人材（データをコンピューターサイエンス的に作る人、データのコンテンツ[アノテーション]を作る人、データを活用する人）について対応している。作業部会は年に 2 回くらい開催しているので、テキストの共有など連携・調整が取れていると思う。

高木主査より、あと 1 年ということもありきちんと成果を出していくための協力依頼があり、閉会した。

（16：25 終了）