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
【遺伝研・JST・中核の参画機関（基盤技術関係機関)】議事要旨

【日　時】　平成２１年1月２７日（火）１３：１０～１６：０５
【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】　菅原秀明(作業部会委員)、黒田雅子(作業部会委員)、西村佑介（JST)、中村保一(遺伝研・かずさ DNA 研)、岡本忍(かずさ DNA 研)、中尾光輝(かずさ DNA 研)、藤澤貴智(かずさ DNA 研)、浅井潔(作業部会委員)、野口保(産総研)、松本裕治(奈良先端大)、山中芳朗(文科省)、高木利久(作業部会主査)
大久保公策(遺伝研・DBCLS)、永井啓一、西川哲夫、川本祥子、箕輪真理(以上、DBCLS)
（敬称略・順不同）

【議　事】
平成 21 年度業務計画・平成 20 年度プロジェクト進捗状況について

議事内容に入る前に高木主査から作業部会構成の変更について、「昨年 5 月の中間評価からの助言に基づき、参画機関・分担機関・補完課題がもっと密に連携できるように作業部会をそのようなメンバーが集まって議論する場にするとともに、研究運営委員会については大所高所から日本の DB や PJ のあり方を議論する場とするために、メンバー構成を見直した。作業部会には他のメンバーも大勢いるが、話し合うべき内容が多すぎて時間が足りないなどの問題があったので、今回は関連テーマに合わせて 4 つの分科会に分けて開催することにした。」との説明があった。

高木主査から引き続き、「ライフサイエンスデータベースの統合・維持・運用の在り方」報告書について、「あと 2 年の本 PJ の終了後の体制について、文科省ライフ委の下にライフサイエンス情報基盤整備作業部会が設けられ、JST の BIRD と一体的な運営を図る体制を作るべきという報告がなされ、2 年後については存続が認められつつあるという状況。一方、CSTP の下に統合データベースタスクフォースがあり、そこでも在り方が議論されている」という報告がなされた。

- 予算配分額について

◇資料説明◇
資料 1-2 は中間評価において、助言として提出された費用配分資料。これを尊重して今年度と同じ額、総計 11 億円の予算を作成した（資料 1-1）→11 億円の内訳としては、プロジェクト自体の予算は 8.5 億円であり、当プロジェクトの継続性を考慮した将来的な JST との一体化に向けて、2.5 億円については JST からの予算執行分となる。そのうち、2.2 億円は情報・システム研究機構(→DBCLS)に、補完課題としての遺伝研分 0.3 億円は遺伝研へ直接配分される。予算に関する資料を提出いただいたので、資料 1-3 のスケジュールにしたがって、作業部会の議論を踏まえ今後微調整をし、松原委員長の確認を経て、確定していく手順。
＜DDBJ から＞
遺伝研 DDBJ の補完課題トレースアーカイブについては、JST BIRD の中の菅原 PJ に合体して実施することになった。

◆質疑応答◆

○遺伝研の名前は補完課題としては残る、ということでいいか？補完課題の代表は五條堀先生で、実施については菅原先生、という理解。

→補完課題 BIRD で評価の方向性がそろっていないとやりにくいので、その点については BIRD 課題と統合 DBPJ 分課題について評価するときに配慮いただきたい。

- 国立遺伝学研究所

◇資料説明◇

開発部分について、まだ本公開になっていない部分もあるがプロトタイプをもとに改良中。運用部分について、Trace Archive（以下 TA）については NCBI への代行登録を実施。計画書には載せていないが、運用の一部として Short Read Archive（以下、SRA）もスタート。代行登録については 2 月に NCBI 担当者と打ち合わせ予定。来年度予定については業務計画書にまとめられていないが、TA、SRA については実施すべき内容と把握。来年度新規取組みとして Barcode of Life(BoL)がある。これは国際的には PJ として実施、日本ではまだボランティアレベルで一部担当している状況。この PJ からは大量の SR が出てくる予定。NCBI の状況、日本での機器導入状況を把握し、Archive のための資源の見積もりを実施したい。JST 課題のためにはウェブサービス化が必要で、最低限の検索機能などは必要になる。TA、SRA のアクセッション番号については現在 NCBI が唯一の発行センター。塩基配列については、3 極でそれぞれ Prefix が決められており、自律的に発行できる。

◆質疑応答◆

○代行登録はずっとつづけるのか？DDBJ が自律的に番号を付与することにはならない?

→番号の発行を DDBJ でできないとユーザニーズと合わない（登録タイミングの確定）が、EBI も同様のやり方。代行登録を継続する予定。ただ、（番号体系を見ると）EBI では独自に発行できるようになったらしい。何らかの基準をクリアすれば DDBJ で発行できるようになる可能性も。

○NCBI が発行する番号をもらえ、というのは論文投稿規程？

→NCBI の（1 事業である）TA が発行するものが必要という規定がある。DDBJ から番号を付与して管理することも手続き的にはできるが、Journal の Editor が受け入れてくれるかどうかが問題。

○NCBI では全部検索可能で、日本では日本にあるものだけ検索可能？

→メタデータだけは共通だが、DDBJ で扱う Trace のデータは日本での登録部分のみ。

○TA や SRA のデータ要件は 3 極間ではなく NCBI が決めたのか？

→Yes。

○日本では当初データが出てくる見込みがなかったのに、論文投稿の要件になったから、出てくるようになったのか?

→最初の 1 件はそのように始まった。

○データ保管の容量について、見えている範囲では、テラレベルで大丈夫か？ペタまではいかない？

→テラレベルで大丈夫そう。

○SR も論文で要求されるのか?

→むしろこちらの要求が強い。プロジェクト番号もつけろと言われる。

○アメリカのプロジェクト管理の番号ではないのか?

→もとはアメリカの PJ 番号で当初はゲノム関連のみだったが、今はゲノム以外の PJ まで拡張されて適用され、外国からのものでも登録の際にアサインされる。
○予算元が本来は何らかのルールを課すべきではないか？
→日本では JST の課題にはもともと PJ-ID が付いており、それを使えればいいが、(実際には)NCBI でつけたものを要求されている。
○科研費はついていないかもしれない。国内では DB の予算元を示すデータがないのが現状。
○NCBI に登録することのインセンティブは何か?
→論文化するときに登録が必要。ただし、最近では、論文と関係ない登録も増えている。
○TA と SRA では後れを取ったため、いろいろな混乱が起きているが、今後のヒト・リシークエンスなどでは早めに対応を考える必要がある。

- 科学技術振興機構

◇資料説明◇
今年度の内容については資料 1−5 にもまとめた。21 年度については、JST 内部で統合 DB 連携のための予算を取っており、引き続きウェブサイトの運用・アップデートを実施。予算の面では中核機関の予算のうち、JST で対応する部分について執行していく。補完課題の DDBJ 分についても、JST 課題に追加して執行予定。
◆質疑応答◆
○WINGpro のアップデート頻度
→搭載サーバを変えたため、編集時にフリーズしてしまうようになったので、予定していた一部作業はペンディング。年度末までに一覧表について WINGpro 収録の個別データベースのデータと連動できるようにする。
○（DBCLS より）Mouse EST DB はダウンロードサービスのために提供いただき、公開準備中。
○来年度 WINGpro への追加は予定されているか？
→マシンの不調によりサーバを移してから作業する必要がある。ただ、一覧を積極的に作るよりは、ユーザーが自分で登録できる場を提供するということで対応予定。
○登録を促す宣伝は？
→昨年は実施していたが、今年はマシンの不調もあり宣伝はしていない。ただ、MediaWiki の性質か、Google 検索ではかなり上のほうに出てくるようになっている。
○メタデータのサイトの更新は？
→国際標準もたくさんあるが、マンパワーをあまりかけられていないため、更新未定。
○掲示板は今後どのように運用？
→掲示板ごとに RSS を入れられるので、それは機能として来年度追加予定。

- かずさ DNA 研究所

◇資料説明◇
☆来年度の体制
中村保一（本務：遺伝研、兼務：DNA 研）PJ の代表としては引き続き PJ 終了まで担当予定

岡本忍（DBCLS 所属に変更）植物関連のほかの活動にも参加できれば。人件費を DBCLS に移動
藤沢貴智（かずさ統合 DB プロジェクト研究員）
中尾光輝（転出）
中村委員の本務が変更になっただけで、基本的には体制を変更せずに次年度も継続予定。

かずさは数多くの植物・植物関連微生物ゲノムを決めたが、これらの情報をもとにした機能解析の論文情報を自らアップデートしていくことは量的に不可能。ソーシャルブックマークを利用して、ユーザーの書き込みで情報を蓄積する仕組みを提案。実際に、いくつかの機能ごとのウェブツール（合わせて Kazusa Annotation Suite と呼ぶ）を作成し、情報蓄積の呼び水にするためにある程度の情報を論文から抽出して実際に入力。一部の生物種については、関連情報の載っている論文のカバー率が 99%に達するものもあり。アノテーター陣容（経費全部で 600 万円、博士 3 名、修士 5 名、他 1 名、女性多い、在宅多い→子育て中の女性の活用、遠隔モデル）。統合 PJ の講習会も含め、Suite ツールの使い方を中心に、発表を積極的に行った（含む海外）。
21 年度予定については、基本は継続で高度化が中心。そのためのツールや手法の改善を行う。理研とのデータ統合も視野に、シロイヌナズナについても文献情報の蓄積を実施する。
◆質疑応答◆
〇データの公開、ダウンロード（DL）可能かどうかなどについては？
→アノテーターが入力すれば個別データは即時公開。まとめて DL はできない。DL の仕組みについては技術的には対応可能だが、現時点でプランはない。どこかで DB をフリーズする必要あるため。
〇理研の植物 DB との関係は？
→理研は独自に持っているのでつなげることを考えている。データの取り込みは現実的ではない。
〇ある遺伝子の関係論文の 9 割が蓄積という意味は？機械的な検索での収集とはどう違うのか？
→この論文のどこに記述があるかという情報を蓄積。機械的に遺伝子名を拾おうとするとノイズが多くなる。生物種の区別（機械的には難しい）を実施して特定のものをアサインしていること、情報の在りかについて深いところにあるもの（たとえば Fig の中とか）も拾っていることが特徴。
〇何を 100%としているのか？
→PubMed で検索（生物種）できて、PDF が Available なもの。たとえばシアノ（バクテリア）関連は全部取ったとしてということ。シアノの中心的な種類に関する報文は年間 1000 件未満くらい。
→機械可読なものもやりたいと考えている。遺伝子名がそろってきたらできるので、アラビ（ドプシス）関連論文（25000 件くらい）については機械的な手法も検討している。
〇論文情報のサマリをまとめるということはしていないか？
→やってない。Yeast ではそのようなことをやっているサイトがあるが。報文数とやる内容から考えて、このくらいの作業量が適正規模。
〇まずは全部の報文について何らかの情報抽出をし、次の段階は同じ粒度のデータをほかの植物に広げるか、抽出するデータを詳細化するか、のどちらかか？
〇ほかに国際的に同じようなことを（シアノで）やっているところは？
→京大化研でジャンボリーみたいなのはあり、データが公開されている。米国でもプログラムでやろうとしているサイトがあったが、更新されていない。

○なぜシアノバクテリアなのか？かずさがゲノムを読んだからか？
→光合成の中心的なモデル。葉緑体の元であり、研究は盛んなため、本 PJ のモデルとして手ごろ。
○理研との連携の際に、理研のセマンティックとの相性は問題ならないか？
→遺伝子を中心にしているので、そこでつなげられるものの範囲で実施する予定。
○2011 年からの中期的な見通しとしては？この DB はかずさで継続的に維持更新するのか？
→かずさと離れるとしたら、中核で受け入れてもらう可能性もあるが、現在は未定。

- 産業技術総合研究所生命情報工学研究センター（CBRC）

◇資料説明◇
ワークフローの開発・公開予定については Phase 1~3 を今年度公開予定。基本的にユーザーが意識しなくても必要な処理が内部あるいは外部へのジョブ投入によりなされるようなものを考えており、必要な DB 等についても内部で持てる物は持つ。フローに組み込んだ、利用制限のあるソフトへの対応（例、Modeller[アカデミアフリー、民間有償]ついてはユーザー登録をしてもらい、その人だけは使えるようにするなどの対応）を検討。他の有償ソフトについても同様のスキームで対応予定。Phase 1（立体構造予測、PJ 内公開）については 8 月に、Phase 2（タンパク質アノテーション、一般公開）については 12 月に公開済み。Phase 3（タンパク質比較情報、一般公開）については 3 月公開予定。21 年度はプロジェクト内外（DBCLS、糖鎖 Gr、経産省統合 DB）との連携、CBRC 内グリッド環境連携・整備、タンパク質以外の分子についてのワークフローの検討などを進めたい。ワークフローの講習会参加者のニーズを掘り起こす予定。

◆質疑応答◆
○Phase 1~3 については包含関係にあるのか？
→それぞれの Phase ごとに完結。Phase 1 は内部公開のみでグリッド上での動作確認のためのもの。Phase2 は Phase1 を包含。Phase3 は機能的に異なるものである。
○ほかの予算（JST や大学）で開発しているツール（CBRC が関係しているもの）の組込みも可能？
→可能。一部の先生からは積極的に入れたいという話もいただいている。
○立体構造についてはこのサイトでやればいい、というものになっていくのか？
→そのようにしたい。
○OpenID の使い方は？
○（OpenID 担当者から）転用できないか検討中。Grid はたくさん情報が必要なので、OpenID で取っている情報が十分かどうかわからないので。
→経産省関連では様々な PJ があり、各 PJ で開発成果を単純に組み込んで用いることはできないが、それぞれの PJ 内部だけでやっていても限界があるので、なるべく統合したい気持ちはある。

- 奈良先端科学技術大学院大学

◇資料説明◇
専門用語解析技術と専門用語抽出ツールの設計と開発を通して、専門用語辞書の充実を図る。手法としては、ある程度の分量の用語についてそれぞれの処理を人手で行い、それを学習データとして

手法を機械処理できるように開発する。京大金子先生の辞書を元データとして採用。800 語程度について、専門用語の内部構造を付与。意味がわからないと単語の分割ができないので、作業は専門性を要する。分割の際には関係性情報を付与。文字（日本語）の重なり等により縮退減少、あるいは単語（英語）の並列句の省略などがあり、これが解析作業を複雑化している。解析の自動処理については、まだ精度の改善が必要ではあるが、既存の処理方法と比較して優れている部分もある。意味クラス推定は、新規専門用語の意味クラスを文章の構造などから照らし合わせ既存の用語との類似度から推定するものであるが、10 万語しかない LS 辞書に単語を追加する際の作業の補助ツールとなる。21 年度は今年度の継続として、手がけている項目をさらに拡張、高度化して開発する。

◆質疑応答◆

〇このような手法は何に役立つものか?

→構造を知ることで、専門用語の意味クラスを推定し、文章の中心（＝重要単語）を決めることができ、検索の精度を上げることができる。省略されている単語の理解に役立つ。

〇内部構造というのは一般的な辞書の中にも出てくるものか?

→日本語の一般語にはあまり複雑な構造がないので、学問としてはあるが一般的ではない。文章になったとき、日本語では助詞が入るので構造がつかみやすいが、助詞のない言語（漢語など）ではこのような解析が役立つ。日本語でも専門用語になると文字の並びが複雑になり、このような解析手法が必要になってくる。

〇文脈により（用語の）区切りが変わることはないのか？

→可能性としてはあるが、内容を確認しながら作業できる。今までそのような例はないが、切り方を間違えて訂正することはある。

〇内部構造づけでどこが困ったのか？

→専門用語の意味を取るのに、考え方はボトムアップであるのが通常だが、この解析手法の場合は作業の繰り返しを防ぐためにトップダウンの手法を取っていることもあって、分割作業自体が専門用語を理解する人もとってもわかりにくいようだ。

〇たとえば角結膜炎であれば、角膜炎と結膜炎の両方に登録しておけばいいのでは?

→手法としてはいろいろ考えられるが、どの方法が一番効率的か、という問題だと思う。また、この分野の専門用語としては、体の部位名や病名はある程度システマティックだが、薬品名になるとかなりアドホックなつけ方をしているので、複雑になると予想される。

〇誰でも使えるのか？

→Web で公開しているので、誰でも検索は可能。編集にはログイン必要。

〇インデックスの切り出しなどに役立ちそうか？

〇（DBCLS より）こちらからどのようなデータを提供したらいいのかがわからなかったが、だんだん判ってきた。こちらの手作業で作っている辞書などにこの仕組みを使えるのではないか。PNE でも手作業で情報を抽出しているので、結果を使ってもらえるなら、データの出し方を調整したい。

〇解析済みの 800 例のデータは単なる練習？

→データは蓄積できるようになっている。作業者が学生一人なので、エラーがある可能性はあるが。

〇松本先生の開発されている技術はこちらのどのようなところにいつごろ反映されていくのか？たとえば辞書の作成などに使わせてもらうということか？

→一つのやり方としては、金子先生の所にこのツールをインストールするということも可能ではないか？あくまでツール提供なので、どこで運用するかは相談しないといけない。

- まとめ（総合討論）　他

＜国立遺伝研＞

○TA や SRA などの動向について知っていることがあればお知らせいただきたい。台数把握までできていないので。

→シークエンサーの情報のことか？（東大）柏には 3 台入っていると思う。

→遺伝研はもっと多いと思う。

→沖縄は 3 台。CBRC にも入るような話がある。

→かずさでは一時期話はあったようだが、その後不明。

＜かずさ DNA 研＞

○シアノが扱うにはちょうどいいサイズだったとのことだが、ヒトまでスケールアップできるか？

→この方法では難しいのでは、どこか自動化するなどが必要。

○イネはどうか？

→イネはぎりぎりのサイズ。イネは独自にアノテーションジャンボリーをやっている。

○医薬品関係ではインドで安い人件費で文献 DB を作って売っているが。

→そういう形ではなく、アカデミアである程度まとめて、ユーザーが世界中にいるという状況ならヒトやイネでも実施される可能性はあるが、現在はやられていない。

＜奈良先端大＞

○LS 辞書エントリの内どのくらいを人手で処理すれば辞書の構築が自動化できるようになるのか?

→現在の 800 では（学習データとしても）足りないので、2,000 位を実施予定。実験的に増やしながら見ていくことはできる。800 をやるのに、ひとりで 1 ヶ月くらいだった。

○見せていただくと、データ作成がかなり難しいケースもありそう。すでに入っているものについて、確認する作業をすれば助けになるかも。

○対象として、LS 辞書はちょっと粗いのでは？

→確かに一般的な用語もかなり入っているので、何かほかに適当な学習データのお勧めがあれば。

○辞書材料と共に提案したい。10%くらい手作業でやればあとは機械的にできるか？

→精度の問題があるが、試してみることはできる。

・2009 年 6 月 12 日(金) に成果発表を兼ねたプロジェクトのシンポジウムを予定しているので、ご協力願いたい。

(16：05 終了)