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5月２５日（月）１４：００～１６：３０
【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】　五條堀孝(遺伝研)、菅原秀明(遺伝研)、黒田雅子(JST)、西村佑介(JST)、中村保一(かずさ DNA 研)、浅井潔(CBRC)、野口保(CBRC)、福井一彦(CBRC)、松本裕治(奈良先端大)、田中康博(文科省)、堀田凱樹 (ROIS)、高木利久(作業部会主査)
永井啓一、西川哲夫、川本祥子、山本泰智、箕輪真理(以上、DBCLS)

（敬称略・順不同）

【議　事】

平成 21 年度業務計画について

- 遺伝研

◇資料説明◇

今年度から、課題の位置づけが BIRD の中で実施することに変わった。TraceArchive の登録数は伸びている。NCBI とのデータ交換も実施し、DDBJ 受付分について公開。次世代シークエンサー対応として、ShortReadArchive にも対応。DDBJ でも今後 NCBI と同様な形で公開をしていく。現在は登録作業を代行。3 極でのアーカイブに関する会議に参加。これまでの国際塩基配列 DB と同様の運営を目指す。本 PJ の開発においては、登録受付システムの確立、DDBJ 独自のアクセッション番号発行、登録データを利用できるしくみづくりを目標とする。

◆質疑応答◆

○画像データの保存についての方針は？各研究機関が保管するものか？

→画像データもすべてアーカイブするのは無理だろうと考えられる。SRA を行う理由については、画像をとっておけば BaseCall がよくなってさらに解析できる可能性があるからだが、NCBI は当面安いストレージにとにかく取っておく、EBI は保存しても 1，2 年後には捨てる(案)。DDBJ は、その目的から考えてある程度は持っていようという方針。3 極で一致した見解ではない。

○完成品の配列情報はどこかに別途登録されるのか？

→DDBJ の中に通常のルートで登録。画像データについては３極で交換しない。

○第３者が生データにアクセスして解析しなおす、ということは現実にあるのか？

→SNP の発見の際に、元データを持っていなかったため発見できなかったという経験があるので、取っておくことになっている。ダウンロードするだけでもかなり大変ではあるが、基本精神としては全部取っておく。

○サービスの形態として現場で HDD に移すということがあるか？

→あり得る。所内 LAN でも転送が難しくなっているので。

○統合 PJ で担当する内容としては、昨年は TA サービス、今年度は SR サービス開発だと思ったが、資料に紹介されている 2008 年の実績は NCBI への取次のみとなっている。今後システムができ次第 DDBJ に NCBI からのデータが入ってくることになるのか。そうなると検索などもできるようになるのか？

→将来的には DDBJ でアクセッションを発行することができるようになる。今後の開発は SRA 中心。

○枠組みとしてはこれまでの INSDC の中でやっていくということか？

→2009 には「一緒にやる」という合意がとられたということである。

○予算的には BIRD プロジェクトに統合の予算を加えた形で実施していくのか？

→システム作りは BIRD 予算でできる範囲で行う。HDD は別の予算で手当てしている。

- JST

  ◇資料説明◇

  昨年度の実績ついては資料 1-3 で説明する。実績としては当初の目標について達成している。WingPro については、サーバトラブルなどもあり、内部で登録したものにとどまった。広報については継続的に実施。掲示板(情報交換サイト)の運用も実施し、DB 受け入れについても今後搭載のものも含めて実施した。21 年度については、今年度の作業項目について継続するが、更新内容の一覧表への自動追加機能をすでに実施(資料 3)済みである。

  （事務局から JST の 20 年度報告書を追加配布）

  ◆質疑応答◆

  ○21 年度の追加の機能について、ユーザーが指定しなくても追加してくれる機能なのか？

  →WindProList のどこかを指定する必要があるが、今まではページを追加した際にすべての関連ページを更新する必要があったのが、軽減されたという機能である。

  ○Mediawiki のカテゴリ機能と同じものか？

  →それとは異なり、開発したものである。

  ○JST の WingPro の意見集約システムと中核のカタログ DB との違いは？

  →（DBCLS から）DBCLS のほうは DB の初心者でも見て分かるカタログという位置づけ。意見集約の仕組みは持っていない。コミュニティの意見を拾うのであれば、JST のほうがいいのではないかと思う。

  ○JST のほうは意見集約がメインだとすればたとえばどんな意見が寄せられているか？

  →自分が持っている DB を登録することができる。こちらがリストを作るのではなく、DB を作っている人自身が登録できる。MediaWiki の仕組みを使っている。DB に対する意見も寄せられており、それを中核に転送したりして、DB の充実に一役買っていると思う。

  ○利用実績としては？

  →トータルアクセス数(約 24 か月)は 200 万くらいだが、アクセスの内容は見ないとわからない。DBCLS にもアクセス数情報は提供している。

  ○メタデータへ追加する作業については、DBCLS の受け入れとはどういう関係か？

  →受け入れの際にはメタデータを付ける必要があるが、新しいデータを作る際に、どういうデータを入れてもらったらいいかについて、検討する材料となる。

  →（DBCLS から）新しいデータを扱う DB について、今後メタデータの項目を検討する際に参照できる。

  ○今年度の計画はサイトの運用だが、これまでの実績ではデータの追加もされていたようだが。

  →今年度は運用のみとした。時間的にそれ以上の対応可能かどうか分からないので、現状（公開）維持。

  ○情報交換サイトがあまり使われていない。内部の情報交換のためのものだが、JST には使ってほしいという呼びかけも含めて運用してほしい。

  →掛け声はかけられるが、JST から発信する情報というのがないので、こちらから発信という機会がない。ROIS のほうでは、外部発表前の事前の情報発信としてよく使ってもらっている。また、年間のフェーズでよく使われるときと、そうでない時がある。意見交換という話があったが、議論する場にまで育てるのは結構大変。プロジェクトとして共有すべきものを事前に提出する場、あるいは何か声掛けをしたいときに使える場ではないか、というのが 1 年運用してみて思ったことである。

- かずさ

  ◇資料説明◇

  かずさにおいてはすでにいくつかの DB の構築を行っているが、スタッフ不足により DB の陳腐化がおこることを防ぐ仕組みづくり(外部アノテーターの活用、専門家ユーザーによる情報入力システムの構築)が目標。

理研との連携による植物 DB の統合についても検討中。かずさアノテーションの利用実績は昨年度末で昨年度当初より 50%増。21 年度の計画としては新しいものを足すのではなく、現在の課題をさらに推進し、質・量ともに向上させる。

◆質疑応答◆

○ソーシャルブックマークを使ったアノテーション技法は他にあるのか？

→バイオに限るという意味では例がない。

○外国からもアノテーションを入れてもらっているのか？

→今のところ雇用したアノテーターが入力したものがほとんど。まずある程度の情報蓄積が必要。

○有効性の検証はこれからか？たとえばイネプロとかでも使えないか？

→使えるとは思う。外部のプロジェクトに使ってもらうということも必要かもしれない。

○ユニークユーザー数が伸びているが、このユーザーはアノテーションまでしてくれる人か？

→現在は見て使っている人がほとんど。伸びている部分としては海外からのアクセス部分が大きい。

○かずさで決めたゲノム以外にも、大きなプロジェクトでこういう仕組みが動くといい。

○遺伝子名や ID についてはこのアノテーションで決めているのか？

→開始当初イネについての ID の問題が大きかった。日米の違いもあった。植物界(たとえばトマト)では大きな問題であって、この仕組みを使えばある程度解消できるが、まだ知られていない。宣伝不足。

○ID で対応関係を付けているのは自動でやられているのか？

→人手は介していない。閾値の設定によってはつかない場合もあるが、基本は自動。

○植物の統合ということでは理研のシロイヌナズナ DB との関係は？今後の公開予定は？

→BioMart を用いた連携を開始。まだ DB を合わせたレベル。ベータ版としてすでに DL は可能となっているが、統合的なレベルではない。今後も連携を進める。

- 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◇

まず、本年度の体制変更について、課題内容がアクティブ・ワークフロー（＝WF）に移っていくので、取りまとめ者が変わる。（主担当：福井、開発担当：田代（継続））

昨年度は、蛋白質にかかわる WF をプロトから公開版まで 3 種類構築し、一部は一般にも公開した。WF としては、既存情報があるものは外部 DB を参照し、情報のないものは新規に計算フローを走らせ、情報を提示する仕組み。今年度は（１）（使いやすさ重視の）固定の WF については JST BIRD で既開発の名古屋大学太田教授が開発した SAHG DB の構築 WF のウェブ版を開発。（２）（ユーザー自身が WF を組める）アクティブ WF については、環境開発。WF アプリを比較し、KNIME の採用を決定。モデルとして CBRC 開発の ASIAN システムを WF 化し、重い処理だけ CBRC のサーバで行うフローを開発中。実験のために DBCLS 内のサーバを１台使用する。

◆質疑応答◆

○どのソフトウェアが CBRC で作られたプログラムか？

→（資料５の３ページ下では）POODLE と TIMBETA, WoLF、（５ページ上では）PRIME が CBRC で開発したもの。

○（６ページ下の）WF としてはこれを用いた DB も作られているが、DB の提供とツールの提供の意味の違いは？例えば(PSI-)BLAST の際に、SAHG を対象とすることを想定しているか？

→自分が持つ配列についてアノテーションが付けられる。サーブ DB については別途公開される予定。ソフトウェアについては BIRD との権利関係は未整理なので、検討する。また BLAST の対象 DB としてヒトゲノムを元情報としている SAHG を用いるということではなく、この WF でヒトゲノム以外のものが解析で

きるようにする。
○ユーザー数については？想定ユーザーは？
→はじめて余り間もないので、あまり伸びていない。想定は立体構造にあまり技術的に対応できない人。ニーズはあると思う。
○産業界でのニーズは？
→産業界としてはもっと精密な構造等が必要となるが、入口の情報としては使えるのではないかと思う。実際には興味のあるものを直接サーバに投げるということはないと思われる。モックでツールの評価をしている可能性はある。（産業界対応という点では）投入された情報についてセキュリティを確保することも今後要検討である。
○機能性 RNA の WF についての提供予定は？
→予定はあるが、今年度までは NEDO PJ のほうで同様な開発を行う予定があるようなので、静観。予定としては 2 番目に作る AWF で対応。
○（WF 開発の）Step と GRID 環境の関係は？
→Step2 以降は GRID で対応しているあるいはする予定。重すぎる計算部分を対応させるとか、そのような部分は有償のサーバを使うと効率化できるとか、自由な選択ができるようにしたい。
○宣伝が十分ではないということだが、今日のような資料を使って、説明を詳しくしてほしい。
→了承。
○CBRC で開発したソフトウェアの部品（API のあるもの）を DDBJ で使うことは可能か？
→可能だと思う。なるべく多くのツールについて、ソース公開や API での利用公開を進めたい。SOAP 化もできているので、使えるのではないか？
○アクティブ WF の目的は、ユーザーフレンドリーなものを作るということだが、自分で好きな関数を組み込むことが可能になるのか？
→Input/Output が問題になるので、そこを変換する Wrapper が必要だが、それがクリアできれば可能。
○産総研のネットはセキュリティが厳しいと聞いているが、WF の利用については問題ないか。
→本サービスは産総研の外の LAN 上(産総研 FW の外)で稼働しているため、問題なし。

- 奈良先端大

◇資料説明◇
用語辞書のシステムと解析技術について、大項目 2 つ(小項目 3 つ)について実施中。辞書システムについては京大ライフサイエンス（LS）辞書をコンテンツとして管理システムを構築。複合語について人手で編集するための IF を備えている。専門用語解析技術としては上記で作成したデータを用いて複合語の内部構造解析、並列構造解析を実施。それぞれのツールは作成したが、現在精度を上げるためチューニング中。専門用語抽出ツールとしては、シソーラスの拡張を目的として新規登録対象の専門用語の意味クラスの推定するための IF を開発中。今年度は、辞書システムとして同義語も考慮、専門用語の内部構造解析のためのデータ拡張、各ツールの高性能化を計画している。本開発で用いている辞書は LS 辞書をベースにしており、金子先生に公開の許可をもらっているので、ユーザー登録すれば中を見られる。DBCLS で利用されるのであれば、著作権のあまり絡まない形で使うことは可能。
○中を見るためにはどのような手続きが必要か？
→登録してもらえばいい。ユーザーとしては中を触れるユーザーと閲覧のみのユーザーがある。具体的な公開条件については金子先生と現在検討中。
○LS 辞書をソースとしているということは全部英語対応が付いているか？
→一部（20000 語）しかついていない。辞書エントリ総数は 90000 語くらいある。

○金子先生は今でも五斗先生のところのメンバのはずであるので、ある意味プロジェクトの一員である。
○並列句の同定の処理スピードはどのくらいか？Medline 全件ではどのくらいかかるか？
→定量的な情報は持っていないが、そんなにかからないと思う。
○フルパージングした結果があることが前提か？
→Yes。ただ、パージング自体はそんなにかからない。
○ライフサイエンスのワードに特異的な手法か？
→手法としては特化しているわけではない。言語（日英）に依存する部分もない。ただ、複雑な係り受けはライフサイエンスに特異なもの。新聞などでは見られない。

- 総合討論

全体を通じての議論は特になし。

最後に高木主査より、1)前回の議事録確認依頼、2)シンポジウム(6 月 12 日)への協力御礼と依頼があり、会を終了した。

（１６：３０終了）