

【日時】 平成22年6月4日(金) 15:30~17:20

【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(遺伝研)、中村保一(遺伝研)、黒田雅子(JST)、浅井潔(CBRC)、野口保(CBRC)、福井一彦(CBRC)、富井健太郎(CBRC)、松本裕治(奈良先端大)、大久保公策(遺伝研)、中村春木(阪大)、高木利久(作業部会主査)、岡本忍、川本祥子、坊農秀雅、畠中秀樹、山口敦子、金進東、吉羽洋周、山本泰智、中尾光輝、高祖歩美、坂東明日佳、八塚茂、平井信一、箕輪真理(以上、DBCLS)
(敬称略・順不同)

【議事】

1. 議事要旨について

事前送付済み。修正は2-3日中に連絡していただきたい。

2. 評価結果と進捗状況・成果目標に関する事前配布資料について

事前送付済み。各機関からの発表の際に各コメントなどに対応した内容をご発表いただきたい。

(高木主査より) 今回の会議の趣旨は、9月末をめどとした成果の取りまとめについてを中心に、作業部会委員による評価への対応などご発表いただきたいというもの。まずは、中核から昨年度の対外的活動内容の概要と、今年度の予定をご報告する。

3. 中核機関より(箕輪)

前回の作業部会では、中核機関の開発内容について紹介したので、今回は対外的活動を中心に紹介する。

平成22年度のスケジュール(研究運営委員会、作業部会、行事[講習会、展示会、学会対応、シンポジウム]、進捗報告とユーザ評価)についての説明を行い、特にシンポジウムへの当PJへの参加機関からの協力、9月の成果取りまとめを受けて10月に実施するユーザ評価への協力の依頼があった。続いて、21年度の対外的活動の概要(研究運営委員会・作業部会、シンポジウム、広報活動、BioHackathon、AJACS、ユーザ評価)について報告された。

4. プロジェクトの平成21年度進捗状況及び最終目標について

➤ 国立遺伝学研究所(菅原)

◇資料説明◇ 資料4

本課題は現在、JST BIRD事業の一環として実施している。本年5月末の状態では、受付件数228件のうち27件を公開しており、この1年で、様々なツールを開発改善してきた。本年は、9月末まであるいは年間を通じて実施する項目に分けて、3極での合意に基づいた開発を進める。なおアーカイブの名称について留意願いたい(詳細は資料参照)。また、メタデータ作成ツールMetaDefineについては、DRAからSRA,ERAへソースコードが提供された(採用かどうかは未定)。

作業部会委員による評価への対応としては、DDBJ事業の関連についてのコメントがあったが、すでにあるサービスとの整理統合に取り組んでおり、予算や人材面でのDDBJとBIRD事業の連携を実施している。

◆質疑応答◆

○今後DRAに登録されてくるだろうヒト由来データについて、いずれ公開するという事で扱うのか。

→個人ゲノムについてはネットから切り離すなどの措置を取り、単純に公開することはしない。

○DRAでは個人ゲノムの取り扱いのルールは決められているのか。3極ではどうか。

→DDBJでは検討中。NCBIでは原則公開のものはSRA、コントロールしたいものはdbGaPという使い分けがされているようだ。

○1000人ゲノムデータではSRAの中に公開のものとコントロールされているものがあるようだが。

→SRAがもともと1000人ゲノム対応のためという側面もあるので、一部完全公開になっていない部分がある。dbGaPも公開データとコントロールデータという2つのレイヤーがある。

→3極合意のうえではDDBJでDRAに置くものは原則公開なので、個人ゲノムを置くのは適当でない。かといって、まったく同じ仕組みをコントロールが必要なデータのために別に作るのも無駄なので、だとしたらレイヤーを切って運用するという方法もあると考えている。

○3極での合意もあるから、DDBJだけ別の決まりごとを作ることはできないのではないかと。SRA/DRA/ERA間では従来のGenbank/DDBJ/EBIと全く同じようにデータ交換などがおこなわれるのではないかと。

→3機関の連携の中に、SRA/DRA/ERAも含まれているという考え方なので、Genbank/DDBJと同じ関係になる。また、未来永劫公開しないデータをDRAに置くことはできない。

○別件で海外機関とのデータベースの交換について、文科省に事前に連絡してほしいという話があったが、MetaDefineのソースコード提供については同様の手続きを踏む必要はないのか。

→本件は文科省の直接の委託ではなく、JSTの委託分なので、文科省への直接の手続きは必要ない。

○年度末までのデータ増加の見込み(200TB)の根拠は。

→示したデータ量の推移はDRAのみなので、今後(NCBIの)SRAと(EBIの)ERAとのデータの交換もでてくることからその位が必要になると考えている。

○SRAはDRAの100倍くらいとのことだったが、その場合は足りなくなるのではないかと。

→実務者会議で得られた情報ではNCBIでのデータ増加量は1日1T(圧縮前)だが、これの1年分を圧縮すると大体200Tと考えている。

○全体の予算規模と費用内訳を教えてください。

→(正確な数値は手元に資料がないが)昨年度は3000万x85%(△削減分)、今年度は3000万x75%で考えており、プログラム開発費が約2000万(MiGAPの開発も含む)+JST研究員雇用1名分、ハードの運用や空調電源等についてはDDBJの既存施設を利用している。また、DDBJで雇用しているアノテーター1名がデータのメンテナンスに参画している。

○ダウンロードの実績はどのくらいか。

→手元に資料がないが、件数は多くない(分析するほどない)。受入は128件だが、公開は27件である。

▲9月末時点の到達目標、要望について▲

- ・本委託事業で実施したツールなどは再利用可能なようにする(DBCLSが調査する)。
- ・新サービスの公開時には、中核からの広報活動による紹介や使い方説明として統合TVを活用する。
- ・MiGAPについても微生物コミュニティ以外にもいろいろなコミュニティへ紹介する。

➤ 科学技術振興機構(黒田)

◇資料説明◇ 資料5

既開発のものを継続運用するのが今年度の計画なので、9月末の区切りといったものは無い。運用を継続しているのは、WINGpro(DB登録、意見集約サイト)、事業サイト(最終的には「ライフサイエンスの広場」に静的リンクを張る予定)、MDeR(メタデータ要素リポジトリ)。このほかにROISとの受託研究契約に基づく生物情報DBの統合・維持・運用を図る。文科省→JSTへ費用の拠出元が変わることから先ほど指摘があったような成果公開に関する手続きについても留意したい。

◆質疑応答◆

○WINGproの意見集約システムで集約された意見の実績は。

→Wiki の仕組みを利用して DB を登録・修正できる仕組みなので、意見よりはそのような登録内容が主である。ユーザからメールで修正の依頼も来る。情報が年間 10 件くらい。リンク切れなどは年に 1 度くらい JST で一括して修正している。

○利用状況・アクセスについての分析はできているか。

→報告書の利用状況の欄にて報告しているが、総数は 700 万位(3 年間)である。JST では net か jp か海外か程度の分析しかできていない。co、ac までは判らない。ログはすべて DBCLS に渡しているので、必要であれば解析をお願いしたい。

○WINGpro のねらいは何か。

→自分の作った DB があれば登録できる仕組みである。年間 10 件、全体 500 件。日本語なので、海外からの直接の登録は無い。リクエスト内容からみると、授業などで使われているようだ。統合 TV とのリンクもリクエストによって入れている。

○当 PJ に参画している長浜バイオ大でも授業でアクセスしていると聞いている。

○無くなると困るサービスか。DBCLS のサービスで代替可能なものか。DBCLS でやっているとの関係は。

→(DBCLS) 経産省や農水省でも特徴ある WINGpro 様のものを作っているが、そちらからもできれば全体で 1 つにしたいという意見があるので、いずれ全体を見直して 1 つにしていくのがいいのではないかと思います。ただし、今年度すぐに対応するのは難しいと感じている。新センターの課題かもしれない。

○利用者の分析もサービス側の責任ではないか。サービスの提供に際しては利用者プロフィールを意識してほしい。

→(DBCLS) DBCLS でデータを貰っていると思うので、こちらで解析してみたい。

▲9 月末時点の到達目標、要望について▲

- ・利用者プロフィールについて情報を提出 (DBCLS がサポート) する。
- ・本委託事業で実施したツールなどは再利用可能なようにする (DBCLS が調査する)。
- ・新サービスの公開時には、中核からの広報活動による紹介や使い方説明として統合 TV を活用する。

➤ 産業技術総合研究所生命情報工学研究センター (福井)

◇資料説明◇ 資料 6

メンバーについて、今年 4 月から藤博幸が参加している。22 年度はアクティブ・ワークフロー(WF)に向けた要素技術の開発、WF によるプラットフォーム(PF)開発を計画している。特に注力したいのは Local PC +SOAP 通信による PF である。要素としては 5 月に DISORDER 解析を公開し、今後、RNA 配列解析(6 月)、立体構造モデリング(7 月)、配列解析(8 月)を公開予定で、これらを総括する配列解析 PF と構造解析 PF を 9 月末に公開したい。他機関との連携も含めて開発する。

◆質疑応答◆

○CBRC 以外の他機関のツールはどうやって載せていくのか。外部の人が搭載できるような構造になっているか。CBRC の人が取りこんで何かしないとだめか。

→CBRC で取りこむと時間がかかる。比較的簡単なのは API の利用、さらに SOAP 化されているものがあればそれを利用するのが最も簡単。今回の開発に関しては SOAP 化されているものを利用することを前提としている。ただし、ソースを公開していただければ、CBRC での加工も可能。

○長時間かかる解析については、SOAP は対応可能か。例えば DDBJ では、長時間かかる解析への対応として、SOAP にくわえて REST も提供している。

→CBRC で受けるジョブについては SOAP で対応可能。CBRC から外部に投げたジョブについては外部の機関の SOAP の仕様に依存するので、不明である。

○PDBj でも SOAP→REST への移行中である。

→この中でもいくつかは REST で行っているが、引数の複雑なものなど SOAP 対応が必要である。

○SOAP の問題としては、型のマッチングもあると思うが、それについては対応しているか。

→CBRC 側で開発して鑄型を公開しているものについては対応を保証できる。

○Web サービスの問題点として、互換性の問題(プログラム言語も含め)が常にあると思う。SOAP では対応範囲に限界があり、そのため、SOAP から REST へという移行の流れもある。本開発については SOAP で対応できるある程度の範囲のみに限定するのか、より大きな範囲まで取りこむ方針か。

→本開発の趣旨は、ユーザがそのつど言語を意識してプログラムを書かなくてもよいように鑄型を用意するというもの。開発の最初の段階でまずは SOAP 化という方針が決まったので、全体の方針が変わるのであれば、方式については変更可能である。

○こういった取り組みは、どこまでをカバーしていくのか (PJ 後も含め)。例えば蛋白質の研究者がやりたいこととしては、スクリーニング、その前段階の複合体の計算があるがこれらへの対応は予定されているか。

→有償のものにそういった機能に対応するものがある。CBRC でも独自の DockingTool はあるが計算負荷も大きいので、現在は組み込みを予定していない。

○CBRC としては外部向けのサービスをミッションとしているのか。アルゴリズムを開発して、ツールを配布するのがミッションか。

→解析サービスがメインの目標ではない。ただし、技術的にできることを示すと運用が変わってくることもある。ファンドがどこから取れるかにもよる。サーバーリソースの情報も含めて、何ができるかを示すことが重要だと考えている。

○産総研全体との連携については。

→(資料 6 の 17 ページに示した) RIO-DB は、産総研内の各センターでばらばらに開発したものを蓄積するという研究支援。産総研の中の仕分け的な活動で一時期止めそうになったが、最近復活の方向で、リソースをきちんと管理しようとしている。ライフを超えた取り組みなので、情報基盤などの仕組みも入っている。これらのリソースからも取り込める有用なツールは取り込んでいきたい。

▲9 月末時点の到達目標、要望について▲

- ・開発したアクティブ WF を統合・総括した PF (配列解析 PF, 構造解析 PF) を公開する。
- ・本委託事業で実施したツールなどは再利用可能なようにする (DBCLS が調査する)。
- ・新サービスの公開時には、中核からの広報活動による紹介や使い方説明として統合 TV を活用する。

➤ 奈良先端科学技術大学院大学 (松本)

◇資料説明◇ 資料 7

統合 DB 開発に役立つ専門用語のための、1. 辞書管理システムの開発、2. 解析技術の開発、3. 辞書拡張支援ツールの設計と開発を実施している。9 月末には 1. 10 万語以上の辞書格納・ブラウザ上での表示修正機能、2. 2000 語以上の内部構造解析用学習データの搭載、3. 類似度判定の精度向上等を実施する予定。年度末に向けても別途目標を立て実施していく。

◆質疑応答◆

○9 月末に到達するレベルは実用レベルといえるか。

→辞書システムについては使えるものになると思う。研究項目としては、類似度評価などは実用のレベル(何を指標にするかという問題もあるが)というのは難しい。用語解析については 90%位に達している。全体としては、研究要素も強いので“実用”というのは難しいかもしれないが、辞書システムとしては使えるものになると思う。

→ (DBCLS) ここから出てきた辞書の活用の点では、中核も開発の一部を担っていると思う。

○研究要素については理解しているが、外部の人に向けてサービスするという面で、9 月末時点での開発・公

開したものが、どの程度評価されるか。例えば Google 辞書は使えないが、Google サービスは使える。参考になるという使われ方もあると思う。

→ (DBCLS) ご質問の趣旨はエンドユーザ向けという観点だろう。内部利用的な側面が強かったため、これまで公開されてこなかったのが、評価の観点としてはこれがサービスにどの程度生かされているかが重要である。一方で、おっしゃるように確かにあまり完成していなくてもエンドユーザが使いたいということもあるかもしれない。

→看護学の先生に外部で使ってもらおうとしたらインストールが大変だったので、そのあたりも少し改良したいと思っている。ちなみにその方は、こちらのサーバーにデータを置いて使ってもらっている。

→ (DBCLS) その意味では PJ 外のユーザで、辞書を構築したいという方にも使ってもらえるもの。

→用語の意味についてはこのシステム上には情報がないが、MeSH との対応付けや Wikipedia の情報が参照できる場所などは利用できる。

▲9 月末時点の到達目標、要望について▲

- ・本委託事業で実施したツールなどは再利用可能なようにする (DBCLS が調査する)。
- ・新サービスの公開時には、中核からの広報活動による紹介や使い方説明として統合 TV を活用する。

➤ 全体として

○各機関で開発したものの公開に際して、各機関からのお知らせに加えて、中核からのアナウンスにも反映させてほしい。また、統合 TV のエントリーとしても代表的な使い方 (たとえば CBRC の WF など) について紹介してもらおうとありがたい。

→ (DBCLS) 公開の際にはぜひこちらにもお知らせいただき、統合 TV 作成のリクエストをいただきたい。

○人材育成の観点から、長浜とかお茶大の方に見てもらって作成してもらおうというのはいかがか。

→実際に長浜の方にはそうしていただいているものもある。ただ、長浜で構築した DB などが優先になっているので、各機関で9月末にまとめていただいたものについても作成の対象としていただくことも検討する。

○MiGAP についても微生物コミュニティ以外にもいろいろなコミュニティへの紹介を希望する。PJ 全体として成果物について広く宣伝していただきたい。

(17 : 20 終了)