統合データベースプロジェクト成果目標と進捗状況（平成22年度）

資料 4

機関名		平成22年3月末進捗状況	平成22年6月末時点の進捗状況	最終年度（22年度）目標
中核機関	DBCLS	**1. 戦略立案・実行評価** ・JST新センター設置準備委員会の場などを通して引き続き統合化の課題の整理を行うと同時に、将来の医療データ統合化のための調査、検討も行った。 ・データ共有、著作権等に関わるシンポジウム「科学における情報の上手な権利化と共有化」を開催した（10/5）。 ・H20年度成果に関わる統合データベースプロジェクト全体を対象としたユーザ評価結果を公開し、コメントへの回答も併せて掲載した。アクセス解析は継続中。 **2. 統合データベース開発** （1）共通基盤技術開発 ・日本語自然文クエリからの MEDLINE 検索システムのプロトタイプを作成し、β公開した. ・DBCLS Galaxy に TogoWS からのデータ取得などの独自ツールを組み込み公開した。 （2）ヒト統合DBの開発・運用 ・構築した辞書やデータの管理システムへのコンテンツの充実、キュレーション情報の付加、DNA DB総覧と遺伝子バンク（GEO）の融合化などを行った。 ・ボディパーツ3Dに関しては、全身の筋肉の3次元モデルを新規作成した。また胸部についても解剖学教科書の記述との差異をチェックし、それにできる限り準拠するようにモデルの精密化を行った。 （3）モデル生物・産業応用生物統合DBの開発・運用(微生物関係) ・2009年6月にb-MiGAPを一般公開した。当初雇用を計画していたb-MiGAP担当者については募集を行ったが適任者を見出すことができなかったため、利用者対応は業者に委託した。利用者のニーズについては、60名ほどの登録利用者のうち比較的利用頻度が高かった12名からアンケートを回収してその把握に努めた。その結果を受けて、スループット向上に取り組み、また、解析対象をバクテリアとアーケアから菌類へも拡張していくこととした。 ・b-MiGAPに比べて解析手法の自由度を高めたs-MiGAPの開発を進めて、一部利用者には公開 **3. 統合データベース支援** （1）ポータル整備・運用 ・プロジェクト成果ポータルサイト「統合ホームページ」の整備・運用は、特に障害無く継続中である。 ・DBカタログは3月末808件の情報を掲載したが、目標1000DBの達成は下方修正する。理由はメタデータ、解説の充実化並びに国内DBのリストアップを優先させるため。 ・横断検索は全インデックスを見直し更新を行った。キーワード表示機能、Blast検索連携機能をLSDBLab.にて公開した。 （広報） ・学会・展示会展示(7回実施、BioEXPO、脂質生化学会、癌学会、生化学会、生物物理学会、分子生物学会、農芸化学会)、文科省/センター主催シンポジウムの開催実施済（2回、2009年6月12日および10月5日） ・広報素材：パンフレット更新終了、ニュース配信の実施（PJサイトおよびJSTサイトでの配信84件、BTJでの配信9件） ・ユーザからのフィードバック：平成20年度成果を中心としたサービス内容に関するアンケートを実施、集まったコメントを整理して公開し、コメントに対する開発者他の回答も併せて掲載した。 （普及啓発） DBCLSで提供しているデータベースの英語版の作成と各種有用データベースの使い方のアップデートを中心に、統合TVをさらに30件（現時点で合計120件）作成した。 （2）データベースの受入と運用 ・生命科学系データベースアーカイブ（旧アーカイブサービス）において以下を行った。 1）以下の9件のデータベースについて、アーカイブを公開した。 ①WorTS②GETDB③Atlas (ISH data base)④KTmAbDB⑤HATODAS⑥Transposon insertion site database – germline⑦EGTC⑧DGSP Database⑨The NAISTrap database 2）11件のデータベースについてアーカイブ公開に向けて準備を進めている。 3）検索機能に、表示件数や項目の表示順・非表示を可変にする等の機能拡張を行った。 4）アーカイブサイトの構築や更新、運用を支援するシステムを構築した。 5）アーカイブサイトの英語化を行った。 ・国内外の蛋白質関連DBの蛋白IDに基づく統合検索システム（TogoProt）について、インタフェースを改良して、累積40DBの収録を終えた。 ・「生命科学系主要プロジェクト一覧（旧名：ゲノム・ポストゲノム主要プロジェクト一覧）」にプロジェクト情報を追加・公開し、21年度分は22件となった。	**1. 戦略立案・実行評価** ・本センターのミッションであるライフサイエンスデータベース統合において文科省の主要5プロジェクトに対して産生されるデータの提供・公開の公募要領への記載を確認すると共に実際に提供されるデータについて各プロジェクトに確認を行った。 ・今年度最終となる本統合データベースプロジェクトの方向付けと最終成果及びその活用について、研究運営委員会を4月に、そして研究運営委員会作業部会分科会を6月上旬にそれぞれ開催し、開発テーマについては本年9月末時点で終了し、10月以降その検証に充てることを確認した。 ・9月末での開発の大まかな完了を受けて、ユーザ評価を10月末から実施できるよう、評価サイト構築の計画中。 **2. 統合データベース開発** （1）共通基盤技術開発 ・Allie検索に分野情報を追加した、TogoDoc推薦機能のβリリースを行った。 ・DBCLS Galaxy に SPARQL endpoint からのデータ取得機能を追加した。 （2）ヒト統合DBの開発・運用 ・DNA DB総覧構築プロセスの整理による簡素化を行い、最新DDBJリリース80を公開した。 ・ボディパーツ3Dに関しては、直感的新UIを構築公開し、モデルの精密さを言葉で表現したデータ仕様を作成公開した。 （3）モデル生物・産業応用生物統合DBの開発・運用(微生物関係) ・遺伝研のPCクラスタへ移植したMiGAPは順調に稼動した。個々の解析に要する時間が短縮されたことに加えて、複数配列の同時解析が可能になったことにより、利用者の利便性が向上したと思われる。 ・利用登録者数、解析投入件数、投入コンティグ数、投入総塩基数、出力CDS数いずれも、99名、346件、1114,004、464Mbp、512,178と大幅に増加した。 ・s-MiGAPについては利用にあたっての自由度が高いため、b-MiGAPの経験をある程度積んだ利用者から申請があった場合に、利用可能とする運用をしている。 3. 統合データベース支援 （1）ポータル整備・運用 ・ポータルサイトの運用は障害無く継続中である。 ・DBカタログについて、最終年度にあわせ各省のデータベース数の調査、DBカタログへの追加を行っている。またカタログ記載内容については、メタデータ項目の検討中。個別レコードのwikiによる運用を検討中。 ・横断検索について経産省との相互参照を実現した。横断検索サーバ管理運用で3月末購入分のクラスタのセットアップ、管理運用ツールの整備を行った。その他、インデックスの更新、追加は随時行っている。 ・シンポジウムサイト、評価サイト開設のサポートを開始した。 （広報） ・学会・展示会展示(2回実施、脂質生化学会、蛋白質科学会)、センター主催シンポジウムを今年度は厚・農・経関連機関と共同で10月に開催するべく計画中。 ・広報素材：本年度のパンフレット更新終了、ニュース配信の実施（センターHPをリニューアルし、センターHPとPJサイトでのニュース配信が連動できるようになった。PJサイトおよびJSTサイトでの配信30件、BTJでの配信3件、文科省記者会を通じてのプレス発表1件[産総研BIRCとの共同プレス]）。 （普及啓発） ・AJACS講習会（2回実施、AJACS湘南3[日大にて]、AJACS & 第22回DDBJing[DDBJと共催、2日間、DBCLSにて]） ・これまで作成したコンテンツを整理しより使い勝手を良くする「統合TV Curated」を作成・公開し、統合TVをさらに40件追加拡充した。 （2）データベースの受入と運用 ・生命科学系データベースアーカイブにおいて以下を行った。 1）以下の5件のデータベースについて、アーカイブを公開した。 ①DIAM – バイオセイフティ関連の法規制②DIAM – バイオテクノロジーの基礎知識③DIAM – 安全性に関する文献 ④RA target genes in the Ciona embryo⑤DIAM – 微生物情報 2）6件のデータベースについてアーカイブ公開に向けて準備を進めるほか、アーカイブ化候補の調査と交渉を多数行っている。 3）検索機能に対し、AND/OR検索やデータ型に合わせた検索フォームの追加などの機能拡張を行った。 4）アーカイブサイトについて、データベース一覧内検索の追加や横断検索活用方法の改善などを進めた。 ・国内外の蛋白質関連データベースの蛋白IDに基づく統合検索システム（TogoProt）について、正式公開を行った。 ・「生命科学系主要プロジェクト一覧」のプロジェクト情報について、19件を更新し、2件を追加した。	**1. 戦略立案・実行評価** ・生命科学分野のデータ共有のあり方を含めた統合データベースの将来像の検討とデータ共有の促進。 ・著作権、個人情報、倫理に関わる問題点を調査、整理し、具体的事例に適用する。 ・プロジェクト全期間にわたる成果についてのユーザアンケートとアクセス解析等により提供サービスの評価を行う。 **2. 統合データベース開発** （1）共通基盤技術開発 ・現行のテキストマイニングツール群の連携による 論文執筆支援環境の充実。 ・データ解析のためのデータ取得解析共有サイバーインフラストラクチャーの統合。 （2）ヒト統合DBの開発・運用 ・ヒト統合DB構築に必要な辞書、オントロジーを整備し、公開する。 ・キュレーションに必要な文献解読システム及びツール群を整備し、DB構築に活用する。 ・配列データ、発現データ、多型データ、構造データ等の統合サイトを構築し、公開する。 （3）モデル生物・産業応用生物統合DBの開発・運用 ・植物及び植物関連微生物については、コミュニティ参加型のゲノムアノテーションサイトを構築し、高等動植物、植物関連微生物を網羅した文献からのキュレート情報を集積し、公開する。 ・産業応用微生物については、オートアノテーション用のパイプラインを構築し、新世代シーケンサでコンティグまたは完全ゲノムを決定した研究グループに、微生物配列に対する標準的アノテーションを提供していく。（補足説明：次世代シーケンサが微生物研究室にも手が届くようになってきたため、決定したゲノム配列からなんとか生物学的意味を取り出したいという要求が近年急速に高まってきたため、MiGAP開発に取り組んだ。2009年6月に公開したMiGAPが事実活用されてきたことから、新規配列に対してMiGAPによって安定したアノテーションを付与することに集中する。） ・アノテータ・キュレータ教育で得られた情報を上記のアノテーション情報に追加し、公開する。 **3. 統合データベース支援** （1）ポータル整備・運用 ・生命科学系DBカタログについては、国内DB並びに国外主要DBを網羅（800DB程度）し、利用しやすくするためにメタデータを多く取り入れ、解説を充実させる。 ・生命科学DB横断検索については、国内外主要DBを網羅（200DB）し、検索の高度化も実現する。 （広報） ・学会・展示会展示(5回)、文科省/センター主催シンポジウムの開催（1回） ・広報素材：パンフレット更新、ニュース配信の実施 ・英語での情報発信（継続） （普及啓発） ・統合TVをさらに100件追加拡充し、これまで作成したコンテンツを整理しまとめるメタ情報を充実させ、より使い勝手をよくする。 （2）データベースの受入と運用 ・我が国のプロジェクトデータベースについて、データベースのタイプに応じた方法で、データベース受け入れを実施する（目標15件）。 ・公開用システ ム「生命科学系データベースアーカイブ」について、検索機能・他サービスとの連携機能等の拡張を行い、運用を行う。 ・我が国で実施された主要な生命科学系プロジェクトの調査を拡げ、その結果を公開する。 ・蛋白質に関する国内外のDB統合検索システム「TogoProt」を拡張するとともに、関連する他機関との連携を進める。 ・これらのコンテンツについて、データ統合に向けた整備を進める。 【9月末時点の判断基準となる目標】 DBCLSとしては以下の公開を目指して進めている。 ・カタログ数850件。 ・横断検索数250件。 ・DB受入数38件（年度内では45件）。
	JST	1. 意見集約システムを継続運用。一覧表自動作成システムも運用。 2. 事業サイト等を更新しつつ継続運用。 3. メタデータエレメントサイトを継続運用。	1. 意見集約システムを継続運用している。 2. 事業サイト等を継続運用している。事業サイトは最終的に「ライフサイエンスの広場」に静的にリンクし保存する。 3. メタデータエレメントサイトを継続運用している。	1. 意見集約システムを継続運用する。 2. 事業サイト等を継続運用する。事業サイトは最終的に「ライフサイエンスの広場」に静的にリンクし保存する。 3. メタデータエレメントサイトを継続運用する。 【9月末時点の判断基準となる目標】 以下の事項を滞りなく遂行する。 1. 意見集約システムを継続して運用する。 Wikiへの参加を誘発するために、（例えばNCBIなど諸外国の）新規DBデータを5DB追加する。 2. 事業サイト等を継続して運用する。 3. メタデータエレメントサイトを継続して運用し、併せて中核機関のアーカイブサイトにサンプルデータを提供をする。
	産総研CBRC	・CBRC情報統合サイトにて11月に公開したタンパク質立体構造予測ワークフローの運用及びプログラム整備を行った。 ・アクティブ・ワークフローとし、遺伝子ネットワーク推定システムであるASIANのプラットフォーム化を実現し、XサーバやWINDOWS上で動作するワークフローを公開した。またその利用マニュアル等のドキュメント作成を行った。 ・CBRCにて独自に開発された解析ツールの標準化/利用推進のため、プラットフォームを用いた解析ツールのSOAP化に着手した。	・アクティブ・ワークフロー上でDBCLSの横断検索と連携するノードの開発を行った。 ・アクティブ・ワークフローとし、ディスオーダー予測であるPOODLE及びRNA二次構造予測CetroidFoldのプラットフォーム化を実現し、WINDOWS上で動作するワークフローを公開した。またその利用マニュアル等のドキュメント作成を行った。 ・CBRCにて独自に開発された解析ツールの標準化/利用推進とし、POODLEやCentroidFoldのSOAP化を実施した。	ワークフロー技術を用いた統合DB環境構築 ・利用者の要望を考慮し、利用者の要求に対し、必要なデータ、解析手法などを国内、海外から選び、情報（知識）提供する解析ツールのワークフローをWEBサービスで実現し、公開する。 ・利用者の要望を考慮し、柔軟性・拡張性に優れたプラットフォーム上で動作する解析要素技術を繋げたアクティブ・ワークフローを開発し、公開する。 【9月末時点の判断基準となる目標】 目標値と内容：ワークフロー技術を用いた統合DB環境構築とし、 1. 3つのタンパク質に関連したWEBワークフローを公開する。 2. CBRCが独自に開発した解析ツールを用い、5つのアクティブワークフローを公開する。 3. アクティブワークフロー上で2つ以上の研究機関（DBCLS等）との連携を行うノードを開発し、公開する。

機関名		3月末進捗状況		最終年度（22年度）目標
中核機関	かずさDNA研	1．高度情報集積DB ・登録ユーザ数80人（2010年3月31日現在）。 ・高度情報集積DBのシステム設計を改善し、検索・表示速度を数倍～100倍程度高速化した。 ・高度情報集積DBに集積されたデータを用いて、ゲノムDB（CyanoBase, RhizoBase）から新規文献情報が検索・参照可能なシステムを構築した。 ・ID変換サービスのための、ゲノムDBへの検索システムの設計と、遺伝子名とIDの対応辞書の調査とプロトタイピングを継続中。 ・植物関連微生物16種類のBiomartシステムへのインストールのためのデータ準備を完了した。 ・シロイヌナズナ（理研）、イネ（農水）、ミヤコグサ（かずさ DNA研）3種の高等植物ゲノムデータ収集を完了した。 ・国内学会やDBCLS講習会で高度情報集積DB、Biomart（KazusaMart）の利用例やデータの解析例を発表し利用を促進した。 2．ゲノムアノテーション情報 ・植物関連生物14種の6468報の文献からGene Indexing型ゲノムアノテーションを163173件抽出蓄積し公開済（2010年3月28日現在）。 ・ゲノム位置情報上に発現情報や論文記載情報など多用なデータを 統合するビューアの開発と講習会での広報意見収集を行った。	1．高度情報集積DB ・DBCLS講習会への積極的な参加を通じ、ひきつづき高度情報集積DB、Biomart（KazusaMart）の利用例やデータの解析例を発表し、普及を促進した。登録ユーザ数95人（2010年6月30日現在）。 ・ゲノムデータマネジメントシステムBiomartによる植物と関連微生物ゲノム情報の統合データの準備をすすめ、クラミドモナス、大豆、ポプラなど、9種の高等植物ゲノムデータ収集を完了した。 2．ゲノムアノテーション情報 ・植物関連生物14種の7,002報の文献からGene Indexing型ゲノムアノテーションを175,192件抽出蓄積し公開済（2010年6月25日現在）。 ・光合成系1，2、チトクロームb6/f複合体など複合体の文献情報（Complex Indexing）を1,891件抽出し蓄積公開済（2010年6月25日）	1．高度情報集積DB ・インタフェースの改善と普及促進。登録ユーザ100人規模。 ・高度情報集積DBシステム運用作業工程の改善。 ・ID変換サービスを統合したゲノムデータベースの利用促進をはかる。 ・ゲノムデータマネジメントシステムBiomartによる植物と関連微生物ゲノム情報の統合。追加8種の構造注釈つき生物種のデータ入力と公開。 2．ゲノムアノテーション情報 ・Gene Indexing型アノテーション情報の2万件追加と既存データの改善。 ・複合体情報（Complex Indexing）を1000件追加。 ・ゲノム位置情報と論文記載情報の統合：ゲノムビューア可視化技術の利用促進をはかる。 【9月末時点の判断基準となる目標】 かずさ担当分の植物の統合のうち、「高度情報集積データベース」については、Gene Indexing型アノテーション情報を、複合体への注釈を含め18万件以上の蓄積を完了し、植物と関連微生物のコミュニティへの周知を徹底するとともに基盤データベースと閲覧環境の改善をすすめ、100名規模のアクティブな利用者を確保することを目標とする。「Biomartによる植物と関連微生物ゲノム情報の統合」については、イネ、シロイヌナズナをはじめとした解析情報の整備された主要植物8種ならびに植物関連微生物16種類のゲノム情報を追加し、統合利用環境の整備を完了し広く利用に供する。
	奈良先端大	1．専門用語辞書システムの開発　専門用語の内部構造による検索、同義語識別子を用いた同義語の検索機能を実現した。MeSHのシソーラスコードを格納し、シソーラスコードを用いた検索機能を実装した。 2．専門用語解析技術の開発　専門用語の内部構造を解析するための構造解析アルゴリズムを設計・実装し、約85％以上の精度での解析を実現した。 3．専門用語抽出ツールの設計と開発　新規の専門用語と類似する用語を一覧表示し、シソーラス上での位置を表示するインタフェースのプロトタイプシステムを構築した。	1．専門用語辞書システムの開発　用語に関する新たな記述項目を柔軟に拡張可能にするため、これまで用いてきたMySQLデータベースからMongoDBへの仕様変更を行った。 2．専門用語解析技術の開発　ライフサイエンス辞書の最新版を用語辞書に格納し、内部構造解析の学習データを約2000語に拡張した。 3．専門用語抽出ツールの設計と開発　専門用語のシソーラス上での位置を表示する機能を専門用語辞書システムの機能として統合するための基本設計を行った。	1．専門用語辞書システムの開発　10万語規模の生命科学用語の辞書を格納し、Webブラウザ上で用語の検索や用語のもつシソーラスコードの表示、および、用語間の上位下位関係を表示・修正することのできる専門用語辞書システムを実現する。 2．専門用語解析技術の開発　用語の内部構造解析の学習データを2000語以上に拡大する。そして、前年度に設計した一般的な統語解析法に基づくアルゴリズムを実装し、90％以上の内部構造解析精度を達成する。 3．専門用語抽出ツールの設計と開発　新規の専門用語にシソーラスコードを付与する支援機能を実装し、専門用語辞書システムの一機能として統合する。 【9月末時点の判断基準となる目標】 ・専門用語辞書システムを、属性情報等の追加に柔軟に対応するため、MongoDBでの実装を完成させる。 ・専門用語の内部構造解析済みデータを2500語以上に拡大する。 ・専門用語辞書システムとシソーラスの階層構造表示システムを統合する。
	九州大	1．JAGQCでのQC機能をPLINKプログラムのモジュールを用いることにより一本化QCパイプラインとする作業を完了し、webページ表示のための作業を完了した。 2．集団内階層構造、実験による系統的誤差の検出のために主成分解析（PCA）を導入する作業を完了し、webページ表示のための作業を完了した。 3．ゲノム4領域応用ゲノムの各研究グループからのGWAS1次スクリーニングデータを受け入れる方法を引き続き検討している。 4．DHaploDBに在る日本人標準確定ハプロタイプ約100個の情報を統合データベースへ移行する方策の一環として、同ハプロタイプ中のコピー数多型を決定し、その結果のDHaploDBで表示する方法を引き続き検討している。	1．ゲノム4領域応用ゲノムの各研究グループからのGWAS1次スクリーニングデータを受け入れる手続き中である。 2．DHaploDBに在る日本人標準確定ハプロタイプ約100個の情報を統合データベースへ移行する方策の一環として、同ハプロタイプ中のコピー数多型を決定し、その結果のDHaploDBで表示する方法を引き続き検討している。	・国内各研究グループからのGWAS1次スクリーニングデータを受け入れ、これを順次JAGQC Pipelineを通すことによりQCを行い、これらの結果を開示する。 ・最近公開されつつあるGWASデータベースと我々がすでに確立している確定ハプロタイプデータを連携させる。 【9月末時点の判断基準となる目標】 9月末には2個以上のGWASのQC結果をJAGQCで公開することを予定している。DHaploDBのデータ移行の完了時期は未定である。
	東京大学	＜講義状況＞4月～12月の間に4名の受講者に対して24回の講義を行った。4月～5月はスクリプト言語によるプログラミング・サーバー管理・データベース管理等のバイオデータベース構築に必要な基礎技術の講習を行い、6月～7月はJava言語によるプログラミング演習を行った。7月～10月はEnsemblゲノムブラウザーを演習用サーバーにミラーする演習を行った。10月～11月はサーバー障害の対策としてデータのバックアップと復旧について演習した。12 月はUTGBゲノムブラウザーのインストールとゲノムブラウザー開発のためのウェブプログラミングについての解説を行った。また10 月～12月の演習では、昨年度の受講者を講師として、研究で使用しているデータベース・ライブラリ・解析ツールの紹介と演習課題の解説を行った。＜受講者の進捗状況＞OSのインストール・ネットワーク接続・セキュリティアップデートの設定・ウェブサーバーの設置・動的ウェブページの設置・データベースサーバーの設置・データベースの作成・Javaプログラミング演習・Ensemblゲノムブラウザーの基本部分のインストール・サーバーバックアップと復旧・UTGBゲノムブラウザーのインストール・ゲノムブラウザー開発のためのウェブプログラミングの解説を行った。 Ensemblゲノムブラウザーのインストールについては、受講者4名中3名が基本部分のインストール	東京大学大学院新領域創成科学研究科情報生命科学専攻の大学院生3名と研究員1名を対象に、Sun Grid Engineによる並列計算と次世代シーケンサーのデータ解析について演習した。具体的には、シェルスクリプトプログラミングの解説・Sun Grid Engineによる並列計算の基礎・依存関係のある並列ジョブのスケジューリング・計算機リソースの指定方法を解説し、さらにSun Grid Engineを用いて次世代シーケンサーで解読された大量のリード配列をゲノムに高速にアラインメントする演習を行った。	大規模計算のためのクラスター利用技術と独創的バイオDBサーバーの構築技術を指導し、次世代シーケンシング情報のデータベースを構築できる人材を育成する。ウェブで公開中の演習ノートも改善する。本年度の受講者数は東大新領域4名。 【9月末時点の判断基準となる目標】 4名の受講者が次世代シーケンサー解析技術の基礎を習得することを9月末時点での目標値とする。
	お茶の水女子大	社会人(主婦3名)を含む5名及び学生27名が修了。加えて社会人5名が部分的に受講。基礎コースとしてデータベース、ネットワーク、遺伝学、及び、専門コースとしてネットワークによるデータ利用、Java言語によるプログラミング、基礎統計、データマイニング技術を開講。統合DB事業へ従事するものとして、本年度の受講生の内1名がDBCLSでテキストマイニング関連の開発に従事、3名がTA及びサーバ管理者として統合DB事業に従事している。また、昨年までの受講生から、2名が統合TVに従事、講義・演習のTAとして2名が統合DB事業に従事している。	社会人4名、学生2名が「ネットワークを通じたライフサイエンスデータ利用」を受講中。以下の日程で、講習を実施予定。社会人8名、学生11名を予定。ネットワーク基礎 7/26、現代遺伝学 7/28、データベース基礎 8/6、ネットワークを通じたライフサイエンスデータの利用（2回目）8/9-12、18、データ解析の基礎 8/19、20、23-25。データマイニング技術 10月～毎週木曜。統合データベースセンターで2名、テキスト処理事業に従事。お茶の水女子大学で2名が統合データベースのサーバ管理に従事。	DB高度利用者を養成する。本年度は社会人2名以上を含む7名以上の修了者を輩出する。また統合DB事業へ従事する人材を3名以上輩出する。 【9月末時点の判断基準となる目標】 9月時点で社会人2名を含む7名以上への講習を行うと共に、3名以上が継続して統合データベース事業に従事する事が目標となる。
	長浜バイオ大	1．初級アノテーション教育（250名）　環境由来メタゲノム配列からの「健康に貢献する遺伝子発掘」の実習を行い、シニア研究者が精査完了したデータとテキストの更新版を統合DBから公開。 2．2回生のアノテーション教育（270名）　実習授業「バイオ分野のデータベースの統合的な利用」を実施、テキスト最新版のWEB公開。 3．中級アノテーション教育（50名）　次世代シーケンサーで決定した約1000万件の微生物ゲノム由来の大量断片配列データを対象としたアノテーションを実施。次世代シーケンサー由来配列解析用の学内版テキスト作成。同学外版の作成着手。 4．2名の4回生の人材養成として、自己組織化マップによる約2000件の遺伝子の生物系統を推定し、そのアノテーション結果を統合DBセンターより公開。 1．で得られた結果を対象に相同性によらない生物系統の推定を行った。 5．シニア研究者と学部生の共同作業によるtRNAのデータベース（tRNADB－CE）に、ウイルスゲノムやイネゲノムを追加し、機能拡充を行った更新版を公開。 6．1名をDBCLSの統合TV開発で活用。 7．ミニシンポジウムの開催。テーマは、「自分が興味を持つ遺伝子がどの範囲の環境微生物に存在するのかを探索する方法；環境微生物ゲノム配列からのお宝遺伝子発掘の学部や高校教育における活用」。 8．近隣の3高等学校で総計180名の高校生を対象とした土曜講座、出前実習の開催。テーマは、「バイオ分野のデータベースの利用　他」。	1．2回生のアノテーション教育（270名）　実習授業「バイオ分野のデータベースの統合的な利用」を実施し、テキストの最新版のWEB公開を行った。 2．3回生のアノテーション教育（270名）　実習授業「病気に関係する遺伝子の探索」を6月から開始しており、7月に終了する。既に作成ずみの学内版テキストに追加等を行った後に、学外版を作成し公開を行う予定。 3．キュレータ育成を目的として、シニア研究者と学部生の共同作業によるtRNAのデータベース（tRNADB－CE）のデータ更新と機能の拡充を完了し、6月より更新版の公開を行っている。 4．1名をDBCLSの統合TV開発で活用し、2作品を完成させたので、公開を準備している。 5．環境微生物生物ゲノム配列からのお宝遺伝子発掘についての、他大学での出前実習の実施については、一大学について、7月内での開催日程を決定した。 6．中学生・高校生・高専生を対象にした講習会の開催準備については、一高校について、テーマと9月中での日程を決定した。	1．初級アノテーション教育（270名）バイオサイエンス分野におけるデータベースの統合的な利用の実習と教材の更新版の公開。 2．初級アノテーション教育（270名）多様な環境由来メタゲノム配列から健康や持続可能型社会に貢献する遺伝子発掘と教材更新ならびに公開。 3．中級アノテーション教育（50名）次世代シーケンサーで解読した微生物ゲノム配列を対象にしたアノテーションの実施と教材の作成。 4．コンピュータとバイオ実験に習熟した人材の育成を目的に、「バイオ技術のデータベース」と教材作成。 5．シニア研究者と学部生の共同作業によるtRNAのデータベース（tRNADB-CE）のデータと機能の拡充とキュレータ育成。 6．一名をDBCLSの統合TV開発で活用。 7．環境微生物ゲノム配列からのお宝遺伝子発掘についての、他大学での出前実習の実施。 8．中学生・高校生・高専生を対象にした講習会の開催。テーマは「バイオ分野のデータベースの利用～他」。 【9月末時点の判断基準となる目標】 1．1 回生、2回生、3回生の実習に関しては、すでにカリキュラムとして設定しているので、完全に実施される。 2．tRNAデータベースの更新と機能拡充に関するキュレータ養成に関しては、すでに作業を完了しているが、この新バージョンに 関しての紹介論文がnucleic acid researchのデータベース特集号に採択されることを目標にしている。 3．他 大学での出前実習については、9月までには1大学程度を想定しているがそれが達成できれば十分と考える 4．中 学生、高校生、高専生を対象にした講習会については、学内での開催を含めて3件程度が達成出来れば良いと考えてい る。

	機関名	3月末進捗状況		最終年度（22年度）目標
分担機関	京都大学	1．共通基盤技術開発 ①知識処理技術開発 ・化学構造検索の高機能化：類似構造検索、部分構造検索システムをユーザーがローカルで使えるように、データベース作成などのプログラムとセットでパッケージ化した。ドキュメントを整備した後に公開する予定である。 ・酵素番号自動割り当て：データの整備とシステムの精度改良に応用するために、化合物ペアのアライメントを改良するためのオプション指定ができるテスト環境を構築した。 ・化学反応ネットワーク予測：生物種ごとに既知の反応データを用いて反応経路を計算する PathComp のインタフェースの改良を進め、テスト中である。 ②ウェブ技術開発 ・LinkDBの等価エントリーの情報を利用し、各種化合物・医薬品データベースのキーワード検索結果から、対応するKEGGのエントリーを取得できるようにした。 2．統合データベース開発・運用 ①医薬品データベース開発・運用 ・JAPICデータの更新は引き続き順調に進めており、3月末で医療用医薬品13,514件、一般用医薬品12,313件の情報がデータベースに登録されている。副作用情報に関連して医薬品相互作用情報を抽出した。 ②化学情報データベース開発・運用 ・反応データの知識として反応パターンの意味付けを行うために、酵素番号自動割り当てシステムの結果と専門家の割り当て結果を比較した。 ③LinkDB開発・運用 ・エントリー表示時にLinkDBの結果リストを表示する仕組みをKEGG以外のデータベースにも拡張した。 ・随時追加、変更、更新作業は順調に進んでいる。	1．共通基盤技術開発 ①知識処理技術開発 ・化学構造検索の高機能化：複数化合物を MOL 形式または KEGG ID で指定して入力できるようにし、ターゲットの化合物と全対全で類似構造検索するためのウェブインタフェースを構築した。現在はテスト環境でテスト中である。パッケージは9月までに公開する予定。 ・化学反応ネットワーク予測：予測結果からゲノム情報へと結びつけるために、対応する酵素反応の予測や KEGG PATHWAY の参照を容易にできるようにして、公開した。反応経路を計算する PathComp は継続して改良しており、9月末に公開する予定。 ②ウェブ技術開発 ・各種化合物情報からパスウェイや反応などを経由して遺伝子情報へと関連づける仕組みを構築中。一部は医薬品データベース開発にて実装済み。 2．統合データベース開発・運用 ①医薬品データベース開発・運用 ・JAPIC 添付文書情報から LinkDB を用いた関連情報の検索が出来る仕組みは既に実装し公開している。KEGG DRUG 以外にターゲット、輸送タンパク質、代謝経路を検索できる。相互作用に関しては引き続き整備を進めている。 ・化学構造検索と個人用の医薬品リスト作成用のインタフェースを構築して公開した。 ②化学情報データベース開発・運用 ・ウェブ技術開発における統合検索の仕組みを実現するために、化合物と化学反応パターンの関連づけを整備し RCLASS データベースとして公開した。 ③LinkDB開発・運用 ・LinkDB 自動更新の仕組みを構築して、内部的に利用している。 ・ユーザーが定義したリンク情報を利用する仕組みはコマンドラインベースのものを開発した。ウェブインタフェースを9月末までに構築する予定。 ・随時追加、変更、更新作業は順調に進んでいる。	1．共通基盤技術 化合物、医薬品データベースに対してキーワード検索から構造検索まで様々な方法で検索するための技術を開発する。また、化合物や医薬品が生体内外でどのように使われ、変換されているかを調べることができる技術も開発し、検索から解析までウェブ上でできるデータベース基盤技術を確立する。 2．統合データベース開発・運用 各種化合物、医薬品データベースの統合化を、データ間のリンク情報を中心として実現し運用する。さらに各化合物の生体における作用や病原性など付随する情報を統合化することにより、研究者から一般のユーザーまでが有効利用できるデータベースをウェブを通じて広く公開する。 【9月末時点の判断基準となる目標】 1．共通基盤技術開発 ①知識処理技術開発 ・化学構造検索の高機能化：ターゲットの化合物と全対全で類似構造検索するためのウェブインタフェースを構築した。現在はテスト環境でテスト中である。パッケージは9月までに公開する予定。 ・化学反応ネットワーク予測：反応経路を計算する PathComp は継続して改良しており、9月末に公開する予定。 2．統合データベース開発・運用 ③LinkDB開発・運用 ・ウェブインタフェースを9月末までに構築する予定。
	医科歯科大Gr	【医科歯科大】 1．疾患・臨床医科学データベースに特化した統合技術開発に関する研究：セマンティック検索エンジンの高度化として、外部要素データベースとのHTTP/XMLを利用した統合検索を開始した。また、分子情報に関するデータを取り扱うためのインタフェースに遺伝子発現情報を手がかりとして関連する臨床情報の検索や個別症例を辿ることを可能にするなどの高度化を行なった。 2．疾患データベース（がん、神経難病を中心として）の高度化と実証的統合に関する研究：倫理規定案に準拠の上、東京医科歯科大学の統合検索システムを利用する一般研究者、登録研究者各々のアクセスレベルでの公開を行なった。加えて各ユーザより本検索システムの評価を得るためのアンケート集計システムを構築した。 3．プロジェクトの統合的推進：利用者が必要とする情報への効率的なアクセスに向けた仕組み作りの1つとして日本肝癌研究会、大腸癌研究会への関連情報のリンク付けを行った。 【阪大】 1．疾患・臨床医科学データベースに特化した統合技術開発に関する研究 ： 医科歯科Grが開発した統合検索エンジンと阪大Grのパーキンソン病DBの統合検索について、医科歯科−阪大間の打ち合わせで調整した詳細仕様に基づいて、統合検索の為のインタフェースを開発した。 2．疾患データベース（がん、神経難病を中心として）の高度化と実証的統合に関する研究：同意取得作業を継続し、公開データ数を400データから500データまで増加させた。 3．プロジェクトの統合的推進 ： 検索機能の高度化、および関連情報へとの連携について、利用者の使い勝手向上を中心に次年度以降の機能拡張の方針を検討した。	【東医歯大】 1．疾患・臨床医科学データベースに特化した統合技術開発に関する研究：最終年度であることも考慮して、実用に耐えられる最小限の項目に限り、実用に耐えられる完成度をもった各データベース間の横断的検索システムの整備を開始しており、９月末までに開発を完了する。 2．疾患データベース（がん、神経難病を中心として）の高度化と実証的統合に関する研究：９月末を目処に下記の作業を開始した。特に統合的臨床オミックスデータベース（iCOD）の「分子から臨床情報を統合して症例データを検索・表示するシステム」について、仮想的な利用の類型を設定して、データベースの意義、使い方を理解しやすくするように各画面の説明、流れのガイド、結果の表示を充実させる。特に分子・病理・臨床情報を統合的に表示する二次元三層マップを標準化し、他の「分子から臨床情報を統合する疾患データベース」に適応可能な一般化を図る。 また、倫理規定については、既に作成した我々の疾患症例データベースの倫理規定ガイドラインに、ゲノム情報関連のガイドラインで決められた公開情報の階層化や、利用者の類別に従ったアクセス権限などの詳細な使用に関する倫理規定の統合化作業を進めた。 3．プロジェクトの統合的推進：阪大グループとともに、現状の統合検索エンジンについての検討、倫理規定ガイドラインについての検討を行い、その結果を含め、文科省および中核機関に作業の方向性を確認した。 【阪大】 1．疾患・臨床医科学データベースに特化した統合技術開発に関する研究 ：昨年度に医科歯科大と共同で開発した統合検索エンジンを用いた阪大のパーキンソン病DBへの統合検索について、改善すべき問題点および、開発した「統合方式」を本事業の成果として、どのような形で公開するかの方針を議論し、公開するプログラム等の整備を進めている。 2．疾患データベースの高度化と実証的統合に関する研究：複数機関から収集した臨床データを統合する仕組みを本事業の成果として公開するための方針を議論した。 （９月時点で達成すべき目標）1．で述べた異なる疾患間の統合、2．で述べた他機関から収集したデータの統合、それぞれについて、技術的・倫理的課題をどのように解決したかのロールモデルを（ドキュメントの形で）を示すと共に、再利用可能な形のプログラムを公開する準備を整える。	1．がん疾患、神経疾患を対象にした臨床・疾患データベースの統合化技術開発 疾患DBの統合化に必要となる技術要件をまとめ、疾患DBを統合するための標準様式と標準APIのモデルを開発・整備する。 2．実証的統合 東京医科歯科大学のがんDB300症例、大阪大学の神経疾患DB500症例の実証的統合に加え、GeMDBJなどがん、神経疾患に特化してDBの統合を進め、統合データベース検索システムをプロトタイプ的に構築し、公開する。 3．オントロジー、ターミノロジー、シソーラスの整備 がん、神経疾患を典型例として、オントロジー、ターミノロジー、シソーラスを整備する。 4．倫理面の検討結果 臨床疾患データベースを統合公開するにあたり必須となる倫理規定案を作成し、中核機関へ提案する。 【医科歯科大、阪大共に同じ】 【9月末時点の判断基準となる目標】 異なる疾患間の統合、他機関から収集したデータの統合、それぞれについて、技術的・倫理的課題をどのように解決したかのロールモデルを（ドキュメントの形で）を示すと共に、再利用可能な形のプログラムを公開する準備を整える。
	東大医学部Gr	1．標準SNP DBの構築 健常者約1,300検体のデータを追加登録し、公開を待っている。 2．GWAS（ゲノムワイド関連解析）DBの構築 CNVの標準DBにおいて、個体ごとのCNV表示機能の追加、クラスタリング有り・無しの表示機能などの機能追加を行った。一方、GWAS DBにおいても、データダウンロード時のオプション追加などの、詳細な機能追加を行った。 更に、多系統萎縮症のGWASデータを新たに登録した。また、Epistasis効果・遺伝子間相互作用解析を目的として、変数としてのSNPおよび疾患の関連構造を、グラフィカルモデルを用いてネットワークグラフとして表すツールの精度向上を行った。 3．リシークエンスによる臨床情報・ゲノム情報DBの構築 新たに、パーキンソン病のリシークエンスDBを公開するとともに、各種会議にて変異データの登録を広く呼びかけた。また、ALSリシークエンスDBの実験データの追加登録と家族性痙性対麻痺と副腎白質ジストロフィーのリシークエンスDBの文献データの追加登録を行った。 4．中核機関との連携の下、倫理検討委員会で開催・決定したデータの共有に関する方針を広く外部へアナウンスするとともに、実際のデータ配布の運用を開始し、データ利用申請者への対応を審議した。	1．標準SNP DBの構築 追加登録した健常者約1,300検体のデータについて、依頼グループの論文発表を待って公開する。 2．GWAS（ゲノムワイド関連解析）DBの構築 GWASのDBにおいて、量的形質の表示を行えるように機能追加するとともに、CNVのDBにおいてクラスタリング後のP-valueを表示できるように機能追加を行った。また、各種会議でデータ登録を広く呼びかけた。 3．リシークエンスによる臨床情報・ゲノム情報DBの構築 家族性痙性対麻痺と副腎白質ジストロフィー、パーキンソンのリシークエンスDBの実験データの追加登録を行うとともに、家族性痙性対麻痺と副腎白質ジストロフィーのリシークエンスDBを公開した。また、国内外の各種会議でデータ登録を広く呼びかけた。 4．中核機関との連携の下、倫理検討委員会で開催・決定したデータの共有に関する方針を広く外部へアナウンスするとともに決定した方針下での運用を開始した。その他、gen2phenのグループと連携方法の協議を行うとともに、新学術領域・支援活動によって産出される日本人のゲノム変異データの蓄積のためのHuman Variation DBについて、新学術領域の支援班と打合せを開始した。	1．標準SNP DBの構築 健常者700検体以上（Affy 500K, 6.0, ilumina 515Kなど）。 2．GWAS（ゲノムワイド関連解析）DBの構築 ゲノム4領域、厚労省、CRESTなどのデータ登録。Epistasisの解析手法開発。 3．リシークエンスによる臨床情報・ゲノム情報DBの構築 ALS,、パーキンソン病、副腎白質ジストロフィー、家族性痙性対麻痺について、疾患関連・原因遺伝子のmutationデータの網羅的な収集とDB構築。 【9月末までの判断基準となる目標】 （1）標準SNP、GWAS、CNV、リシークエンスのデータベースについて、合計40プロジェクト（studies）以上のデータを登録すること。 （2）これらのデータの受入れ・登録およびデータの提供が円滑に行われるよう、倫理社会的手続きおよび作業手順を整備すること。

機関名		3月末進捗状況		最終年度（22年度）目標
補完課題	理研	1. シロイヌナズナオミックス情報の注釈付けと公開 新しいデータベース「シロイヌナズナ変異体情報」を作成し、12月末までに整備した新しい表現型定義情報と、既存のデータベースの対応するレコードへのリンクを設定し、これらを公開した。ここまでで、明確な表現型情報が公開されている既存のレコード全てについて新しいアノテーション情報が関連付けられたことになる。 2. 高等動植物等由来蛋白質構造のアノテーション 高等動植物タンパク質の試料調製に関わる発現検討実験（無細胞タンパク質発現用発現系構築実験4万8千件、スモールスケール発現実験8万件）に基づいて X 線結晶構造解析のための大量調製実験（大量合成および大量発現実験4千件、精製実験5千件）を行い、結晶化実験（結晶化条件 138 万件、結晶観察 1517 万件）へと進め、タンパク3000プロジェクトで解明されたタンパク質構造データのうち平成21年度分としては1万1千件の回折実験データ（画像データ枚数）を公開可能とした。 3. 微生物由来蛋白質構造のアノテーション 微生物由来蛋白質（変異導入蛋白質を含む）に関わる実験情報および重原子導入蛋白質に関わる情報について、追加データのキュレーション作業を完了した。 4. アノテーションシステムの開発運用とデータ変換 検索システムの更なる改良により、より高速かつ精密に表現型情報等をキーとした検索が実行でき、結果もわかりやすく表示されるようになった。	1. シロイヌナズナオミックス情報の注釈付けと公開（豊田） シロイヌナズナの表現型変異に言及している論文を広くサーベイし、それぞれの論文にプレインテキストで記述されている表現型情報を、PATO,PO,GO,TO,EO,Chebiなどの公開オントロジーを用いて標準化した（たとえば「植物の成長が遅い」という形質であれば' [PATO:0000304] decreased speed', '[PO:0000003] whole plant', '[GO:0040007] growth'の組み合わせで表現できる）。このようにして、約380種類の表現型をリストアップした。 2. 高等動植物等由来蛋白質構造のアノテーション 高等動植物タンパク質の試料調製に関わる発現検討実験（無細胞タンパク質発現用発現系構築実験4万8千件、スモールスケール発現実験8万件）に基づいて X 線結晶構造解析のための大量調製実験（大量合成および大量発現実験4千件、精製実験5千件）を行い、結晶化実験（結晶化条件 138 万件、結晶観察 1517 万件）へと進め、結晶観察画像についても公開へ向けての整理を開始した。タンパク3000プロジェクトで解明されたタンパク質構造データのうちこれまでに 3 万1千件の回折実験データ（画像データ枚数）を公開および公開のための準備を進めた。 3. 微生物由来蛋白質構造のアノテーション 微生物由来蛋白質（変異導入蛋白質を含む）に関わる実験情報および重原子導入蛋白質に関わる情報について、追加データを横浜の公開基盤に送付し、アノテーション作業を始めた。 4. 理化学研究所のデータベース統合化のためのモデルケース構築（豊田） 公開したデータをクライアントサイドプログラムからウェブ経由で利用できるAPIとして、Semantic-JSONを構築、公開した。このAPIは、ユーザがデータリクエストを URIで指定し、その結果をJSON形式データとして返すもので、様々なプログラミング言語で利用が可能となった。 次に、これまで構築したアノテーションシステムに、クライアントサイドのウエブブラウザから理研内外を問わず誰でもプログラムを書いてデータベースを利用できる機能を実装し、プログラムやその処理結果を交換、公開できるオープンな参加型システムが実現した。	1. シロイヌナズナオミックス情報の注釈付けと公開（豊田） トランスクリプトームデータを集積したリポジトリを公開し、それらの詳細な解析結果に基づいたアノテーションデータを公開する。また、理研で観測されたフェノタイプデータを、理研以外で観測されたフェノタイプデータと統合化し、共通のオントロジーで整理し公開する。具体的には、表現型情報を理研サイネス上のデータベースに登録し、対応する遺伝子へのリンクを設定する。9月末までに少なくとも約330の遺伝子コードへの対応付けが可能となる見込みである。これらを利用した予備的な解析を試み、研究者からみた同データベースの有用性を検証し必要な改善を進める。これにより、共通の表現型にかかわる遺伝子のリストを効率的に取得（フェノーム・ゲノム検索）できるよう体制を整える。 2. 高等動植物等由来蛋白質構造のアノテーション（横山） タンパク3000プロジェクトで解明された高等動植物等由来の蛋白質構造データのうち、3 万1千件の回折実験データ（画像データ枚数）を公開し、それに関わる実験情報についても関連づけを完了させ、可能なものについては公開する。 3. 微生物由来蛋白質構造のアノテーション（国島） 平成21年度に引き続き、これまでに公開したデータについてアノテーション作業を行い、アノテーションデータを公開する。またそれ以外のデータについてもデータ公開を進める。平成22年度末までに、理研播磨研究所における微生物由来の蛋白質構造データに付随する実験データに関し、統合データベースの一部として生命科学分野の幅広い研究者に役立つ体制の確立を完了する。 4. 理化学研究所のデータベース統合化のためのモデルケース構築（豊田） シロイヌナズナで実施してきた統合化をモデルケースとして、シロイヌナズナ以外の生物種（マウス等）にも適用することで、データベース統合化の範囲を拡大する。また、本事業において公開したデータをクライアントサイドからウェブ経由でプログラム的に利用できるAPIとツールを構築して公開する。さらに、これまでに構築した、アノテーション作業を行うためのシステムを運用・改善し、アノテーション等のデータ交換やプログラム交換を理研外部者でも実施できるオープンな参加型のシステムとして公開する。 【9月末時点の判断基準となる目標】 1. シロイヌナズナオミックス情報の注釈付けと公開（豊田） 文献情報から収集したシロイヌナズナ変異体の遺伝子および表現型の変異情報に基づき、表現型をオントロジーを用いて整理すると同時にそれぞれの表現型変異に対応する遺伝子（TAIRのagiコード）を列挙する。9月末までに400種類の表現型を登録し、330の遺伝子コードへの対応付けを行い、これらすべてを理研サイネスを通じて提供する。 2. 高等動植物等由来蛋白質構造のアノテーション（横山） タンパク3000プロジェクトで解明された高等動植物等由来の蛋白質構造データのうち、累計3 万1千件の回折実験データ（画像データ枚数）を公開し、それに関わる実験情報についても関連づけを完了させ、9月末までに可能な 50 万件の結晶観察データについても提供する。 3. 微生物由来蛋白質構造のアノテーション（国島） BacpediaのオリジナルDB（実験データを有するタンパク質約8700種類）を公開し、一般の研究者にとって使いやすい環境を整える。
	産総研糖鎖医工学センター	1. 九州大学の伊東ラボがグリコシダーゼのDBの協力を表明。理研の糖鎖関連遺伝子のマイクロアレイを用いて解析した情報を随時DBに登録して公開したいと要望がある。更に他の機関にも協力を打診中。来年度も立命館大学がオンラインプロトコルの整備に協力することの約束を得る。野口研も合成DBを来年度も継続して協力して頂くことを了解して頂いた。糖鎖合成DBと化合物DBに岐阜大学・木曽研に協力して頂く。 2. 糖鎖科学のコミニュケーション(ACGG)で口頭発表。国際複合糖質学会で口頭発表。JCGGで進捗報告と協力の呼びかけ。糖鎖の統合DBとLSDBの普及活動。 3. 統合検索の連携のためのAPIを構築した。野口研究所、理化学研究所、創価大学のAPIを完了した。 4. 立命館大学と共同で実験オンラインプロトコルの整備に取り掛かった。45種類のプロトコルの原稿をまとめ体裁を整えているところ。 5. 糖鎖関連疾患とその糖鎖関連遺伝子の関連した情報を集め（77の病気）、DBに格納した。病態の情報に関する項目に関して著作権者に許可をもらう作業を終え公開した。 6. 糖鎖と感染に関する情報を収集。DB化してPACDBとして公開した。構造DBと連携するところは完成。オンラインリポートに関して、糖鎖の合成分解のパスウェイのミニレビューを作成し、DBとのリンクを行った。その他、糖鎖技術を産業に応用している方からのオンラインレポートとして原稿依頼した。 7. 横断検索用インデックスの更新システムの構築を完了。 8. 専門用語の整備を継続。統合検索技術開発の目処がたった。	1. プロトコルの画像に関する著作権の許諾作業も8割終了。今年度のプロトコル収集のための組織体制をセットアップ中。 2. グリコシダーゼの基質特異性などの情報を収集し始めた。特に哺乳類の情報を重点的に行っている。 3. 糖鎖関連遺伝子のDNAマイクロアレイのデータを公開するシステムの仕様を決め、現在開発中。 4. 糖鎖と病原菌のデータベース(PACDB)の更新作業を一時中断。 5. 糖鎖関連のマーカーの情報を収集し終わり、公開に向けて整備中。 6. 糖鎖関連の特許情報をDBに格納。全文検索用のindexに格納。日本語の特許用キーワードの抽出も順調。表記のゆらぎをどこまでカバーできるかが今後の課題。 7. オンラインリポートの原稿の10人に依頼済み。原稿の内容を決めて更に依頼する予定。 8. JCGGDB関連のコンテンツや論文から専門用語を抽出するところのプログラム開発をほぼ終えた。頻度計算の処理に移行。	1. 平成21年度と平成22年度の2年間で100以上のプロトコルを収集する。年度末までに公開する。 2. グリコシダーゼのデータベースを構築する。 3. 糖鎖関連遺伝子のDNAマイクロアレイのデータを公開するシステムを構築 4. 糖鎖と病原菌のデータベース(PACDB)の更新 5. 糖鎖関連のマーカーの情報を収集しDB化する 6. 糖鎖関連の特許情報を集め整理しDBとして公開する。 7. オンラインリポートの原稿を20前後にする 8. 直観的なポータルサイト実現のための統合検索システムの構築を実現する。できるだけDB間の連携を高められるようにする。 【9月末時点の判断基準となる目標】 ・マーカーDB(TuMarDB)の公開 ・グリコシダーゼの基質特異性の情報整備が公開間近になっていること ・糖鎖関連遺伝子の阻害剤DBが公開間近になっていること ・糖鎖関連特許DBが公開間近になっていること ・オンラインプロトコルDB（GlycoPOD）の公開、今年度分の原稿待ち ・オンラインリポートの今年度分の原稿待ち ・有機合成DBのデータ移行のシステム完了（検索システムを開発中になっていること） ・糖鎖関連遺伝子の発現情報解析システムの構築完了（データを入力して仮運用になっていること
	遺伝研	1. トレースアーカイブへの取り組み ・ 受け付けた1件のデータについて、登録用ファイルを作成中。 2. ショートリードアーカイブへの取り組み ・ 研究手法の進化、次世代シークエンサから出力されるファイル形式の変化に対応すべく、日米欧3極（DRA/ERA/SRA ）共同でメタデータの XML スキーマを、バージョン 1.0 から 1.1 に更新した。また、これに伴い、DRA の全システムをスキーマバージョン 1.1 に対応させた。 ・ 登録受け付けの仕組みを確立して3月31日までに191件を受け付け、23件を公開して、利用者によるダウンロードを可能とした。 ・ ユーザのフィードバックを登録ツールに反映し、登録者がより早く、より正確なメタデータを作成できるように年間を通して改善を続けた。、2010年3月8日にはFlex 技術を利用したウェブベースのメタデータ作成ツール MetaDefine の運用を開始した。MetaDefine はオンラインでのデータチェック、過去の登録データの再利用、既定値の自動埋め込みなどの入力支援機能を提供しており、登録者のデータ作成にかかる負担を軽減している。	1. トレースアーカイブへの取り組み ・ 受け付けた1件のデータについて、登録用ファイルを作成中。 ・ 新たに1件の登録申し込みがあり、対応中。 2. シークエンスリードアーカイブへの取り組み ・ データ受付公開実績：受付件数230件、公開件数28件（6月末時点） ・ 登録者のデータ登録作業の効率をあげるために、Project ID 申請フォームを登録受付システム D-way へ組み込んだ。また、ユーザ評価をもとにしたメタデータ作成ツール MetaDefine の改善を行った。 ・ DRA と NCBI SRA との間で DRA/ERA/SRA から公開されている全てのメタデータと fastq ファイル（約 30 TB）のデータ交換を実現した。全データを DRA から公開した。 ・ 5月18-21日に EBI で開催された国際実務者会議に DRA の担当者が参加し、DDBJ/EBI/NCBI 間でデータベース運営上の問題について話し合った。 ・ ユーザからの要望や研究技術の進歩に対応するため DRA/ERA/SRA 共同で新しいメタデータの XML スキーマ version 1.2 を作成し、会議で承認した。 ・ 第3世代シークエンサから出力されるデータに対してどのように取り組むか会議で話し合った。 ・ EBI と NCBI には MetaDefine のようなメタデータ作成のための簡便なツールがなく、登録者はメタデータを XML 形式で作成しなければならず、大きな負担となっている。そのような状況のなか、EBI と NCBI から MetaDefine のソースコードの提供可否について打診があった。会議後、EBI と NCBI にソースコードを提供した。	1. トレースデータ用データベースの開発 ID発行とアーカイブデータベースの作成、および外部公開データベース作成。 2. トレースデータ用登録システムの開発 アノテーションプロトコルの確立と一般からのデータ受付の開始。 3. トレースデータ活用のためのソフトウェア開発 公開用FTPサイトの整備とwwwサイトの公開。 【9月末時点の判断基準となる目標】 新世代シーケンサ由来データのアーカイブは、自らデータを生成する活動ではないこと、また、欧米と協力して日々継続していく事業であることから、目標を以下のように設定する。 1. 登録データから SRA Toolkit により SRA ファイルを作成し、データリストとともに DRA から公開する（9月末時点） 2. メタデータの検索系、及び、それと連動した fastq ファイルのダウンロード提供システムを完成させる（3月末時点） なお、次世代シークエンサの普及に伴い登録件数は増えている上半期の動向を外挿すると、年度内に 300 件以上の登録が予測される。
	九工大	1. 蛋白質と変異体の熱力学データと構造データのクロスレファレンスを約1,000件のデータについて作成した。 また、蛋白質と核酸の相互作用の熱力学データと構造データのクロスレファレンスを約1,300件のデータについて作成した。蛋白質・蛋白質相互作用データについては、格納用のデータベースのプロトタイプを作成した。 2. 蛋白質と核酸の相互作用の熱力学データはXMLフォーマットに変換し、テキストフォーマットでのデータとともに公開した。また、熱力学データのオントロジーについては、海外の研究者と意見交換しながらすすめており、熱力学データのControlled Vocabularyの作成を行った。 3. 文献検索と文献からのデータ抽出の自動化方法については、統合DBセンターが開発したTogoDocなどのツールの評価を行った。	1. 蛋白質と変異体の熱力学データと構造データのクロスレファレンスを約300件のデータについて作成した。また、蛋白質と核酸の相互作用の熱力学データと構造データのクロスレファレンスを約300件のデータについて作成した。蛋白質・蛋白質相互作用データについては、格納用のデータベースのプロトタイプの試験を行った。 2. 蛋白質と核酸の相互作用の熱力学データの更新分についてXMLフォーマットに変換し公開した。 3. TogoDocを用いて熱力学データを含む文献を自動的に収集するため、ポジティブとネガティブのサンプルを用意し、システムの学習と評価を行った。	1. 蛋白質熱力学データベースおよび蛋白質・核酸相互作用データベースのそれぞれ新規700件程度の熱力学データと構造データのクロスレファレンスを作成する。蛋白質・蛋白質相互作用データは格納用データベースシステムを完成させる。 2. 蛋白質・核酸相互作用データをXML化し公開する。 3. 統合DBセンターと連携して熱力学データの文献収集とデータ抽出の自動化技術の開発に必要な解析やツールの評価を行う。 【9月末時点の判断基準となる目標】 9月末までに実用的な自動収集が行えるようにすることを目標とする。