資料1－5

統合データベースプロジェクト成果目標と進捗状況

機関名		本年度成果目標	本年度進捗状況（12月末時点）
中核機関	DBCLS	1．戦略立案・実行評価 ・知財・法律関係：サイトポリシー・著作権方針を策定。科学データ所有権、知財ポリシーに対応する見解提示 ・日本語文献対応：総説誌全文検索の対象拡大（1誌追加）。学協会著作物の扱いに関する方針策定 ・調査関係：新型シーケンサ、医用リシーケンスへの対応、人材養成ニーズに関する調査報告資料を作成 2．統合データベース開発 ①共通基盤技術開発 ・遺伝子名、蛋白質名をPNEの記事に対応付ける技術を開発し、検索システムに組込み、公開 ・遺伝子名、蛋白質名でfull paperを対応付ける技術を開発し、特定生物を対象としたプロトシステムを完成 ・TogoWSの操作性向上のためのツール組合せインターフェイス設計のための調査を行い、報告資料を作成 ②ヒト統合DBの開発・運用 ・文献解読：論文正規化情報の取り出しと2項関係エディターの開発 ・遺伝子名辞書の高度化と辞書のDB化 ・アナトモグラフィ・ボディパーツ3Dの高度化 ・Wired-PDFの開発と普及化 ・バンク目次：DNA DB総覧と遺伝子発現バンク（GEO）の維持更新(3か月毎)と次世代シーケンサデータ受入れ体制整備 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） ・オートアノテーション用パイプラインの構築：標準タイプ決定（9月）後、仕様を決めて開発 3．統合データベース支援 ①ポータル整備・運用、広報、普及啓発 （ポータル整備運用） ・生命科学系DBカタログの拡充（150DB追加）と記載内容の充実 ・生命科学DB横断検索の拡充（目標200DB達成）と検索精度の向上 ・コンテンツの拡充：新聞記事、特定研究報告書、総説誌などがターゲット （広報） ・学会・展示会展示、シンポジウムの開催（年間3回） ・広報素材：パンフレット作成、ニュース配信の実施 ・ユーザからのフィードバック：サービス内容に関するアンケートの実施と結果のサービス内容への反映 （普及啓発） ・統合TV：100件開発(トータル150件） ・講習会：6回開催 ②データベースの受入と運用 ・基盤づくり：メタデータ仕様作成、標準使用許諾作成、一括ダウンロードサイト構築、運用 ・補完課題、関連機関及び個別プロジェクトDBを対象に年間10DBを受入れ、公開 ・蛋白質に関する国内外のDB統合化のための内容検討と仕様書の作成	1．戦略立案・実行評価 ・サイトポリシー見直し完了。英語化準備中。 ・学会誌1誌公開（ポータルの項参照）。医学系総説誌交渉検討中。 ・GWASデータ受入れ、運用方針案策定。 2．統合データベース開発 ①共通基盤技術開発 ・MEDLINE の各タイトルおよびアブストラクトから遺伝子名、タンパク質名、化合物名、病気名、薬剤名、酵素名、症候名を抽出し、日本語に変換し、PNE の検索にかけるシステムを設計し、プロトタイプを作成。 ・J. Biol Chem の論文から、らん藻に関連する87報を選び、文中に書かれた遺伝子名をタグ付けしたコーパスを作成後、作ったコーパスを利用して自動的にらん藻に関連する遺伝子名を抽出するシステムを設計し、プロトタイプを作成。 ・ツール組合せインターフェイス Galaxy に TogoWS を組み込んだプロトタイプを作成し、TogoWS を Galaxy 上で利用する際の問題点の洗い出し実施。 ②ヒト統合DBの開発・運用 ・論文正規化情報の取り出しと2項関係エディターの開発完了。 ・種々のソースから得られた遺伝子名辞書を一元管理するための用語とリソースを一行で格納するシステム用にコンテンツを作成中。 ・辞書とデータの管理システム構築中（著作データ500件）。医師との協業でコメント付けを進めるためのシステムを構築中。 ・キャッシュ機能はベータ版テスト中。累計ダウンロード160,000件。 ・DNA DB総覧と遺伝子バンク（GEO）の融合進行中。 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） ・比較的長いコンティグや完全ゲノムに自動的にアノテーションを付与するパイプライン、なかでも基本的アノテーションを自動付与するいわゆる初級用パイプライン、の年度内公開にむけた開発が順調に進行中。 3．統合データベース支援 ①ポータル整備・運用、広報、普及啓発 （ポータル整備運用） ・DBカタログ、学協会カタログ、動物アイコンのデータを拡充しインタフェースを改良。 ・横断検索の索引作成は165DBまで終了し、インタフェースの改良中。 ・日本語の文献検索については蛋白質核酸酵素の近刊への検索、生物物理学会要旨と論文集の検索、科学新聞掲載の生命科学記事検索を準備中。 ・ポータルサイトの英語化を準備中（3/9公開予定） (広報） ・7/22ゲノムテクノロジー第164委員会第27回研究会を共催[活動紹介およびデモ]。10/15-17BioJapan[展示]。12/9-BMB2008[展示及びシンポジウム]。3/28,29(予定)農芸化学会[ランチョンセミナー及び展示]。 ・LSDB/DBCLS紹介パンフレット（A4サイズ8面）作成。各サービスメニューのリーフレット作成（日英26種類）。ニュース配信（日経BTJ 11件、LSDBウェブサイト[DBCLS, JST 一部ROISと連動] 38件）。英語で公開が可能なサービスや情報について英語版ウェブサイトを整備中（年度末公開予定）。 ・18年度成果評価結果公開ページを作成。19年度成果に対する評価を実施し、結果を公開。寄せられたご意見への対応についてセンター内の検討結果を取りまとめ、12月初旬公開。 （普及啓発） ・統合TV：65件開発(トータル124件） ・講習会：5回開催（日大、DBCLS/東大x2回、北大、九大、長浜バイオ大） ②データベースの受入と運用 ・基盤づくり：データベース受け入れ覚書案作成完了、一括ダウンロードサイト第2バージョン公開準備ほぼ完了。 ・各関連機関等との連携を推進。JSTからは1つ、産総研・糖鎖センターからは1つ、九工大からは2つのDBのダウンロードデータを受け入れ済。現在、利用許諾＆受け入れ覚書について協議中。NEDOプロジェクト産生のFLJ Human cDNA Database及び、Human Gene and Protein Databaseのミラー化を実施。 ・蛋白質統合DBのプロトタイプを作成。 ・受け入れ対象となるデータベースの把握のため、NARに収録された国内のライフサイエンスデータベースについて調査を実施。
	JST	1．意見集約システムの運用 WINGProの公開継続と新規10DB追加 2．広報 本事業全体の広報活動とプロジェクト内サイトの構築・運用 3．データベース受入れ メタデータサイトの2件追加とMouse EmbryoのEST DB受入れ	1．意見集約システムの運用：WINGProの公開継続。新規3DB追加済み。9月末までに追加したものと合わせて10DB追加済み。 2．広報 本事業全体の広報活動とプロジェクト内サイトの構築・運用：継続運用実施中。参画研究者のための情報交換サイト（掲示板）を作成。
	産総研CBRC	1．アミノ酸配列から広範囲な立体構造に関する予測を行うワークフロー（8/E 限定公開） 2．予測とデータ取得により蛋白質に関する網羅的な情報を得るワークフロー（12/E 一般公開） 3．蛋白質の比較情報を提示し、保存部位、変異部位を推定するワークフロー（3/E 一般公開）	1のワークフローは前回記載した通り、8月に完了済。 2のワークフローは開発を完了し、12月26日に一般公開した。 3のワークフローは1月から開発開始、3月末に一般公開する予定。
	かずさDNA研	1．高度情報集積DB インターフェイスの改善と利用者の拡大（100人規模） イネを対象にネームサービスの充実 2．ゲノムアノテーション情報 GeneIndexing型アノテーション4万件蓄積 ゲノム位置情報と論文記載情報の統合	1．高度情報集積DB インターフェイスにソート可能なテーブルビューを追加するなど使い勝手を改善した。アクティブユーザは現在68人。DBを活用した解析例を国内外学会等で発表し周知をはかり利用を促進している。発表は国際学会2回: Genome Informatics workshop (CSHL/Sanger Inst), The 5th Rice Annotation Project Meeting (AIST)、国内9回: 進化学会、植物細胞分子生物学会、明治大学ワークショップ、植物微生物研究会、育種学会, 統合DB講習会4回の、計11回に達している。 2．ゲノムアノテーション情報蓄積 ゲノム位置情報と論文記載情報の統合のための情報蓄積をすすめている。Gene Indexing型ゲノムアノテーションを85563件蓄積し公開済（2008年12月17日現在）。
	奈良先端大	1．専門用語辞書システム 同義語に同じ識別子を持たせ、表記ゆれや別表記を検索し表示する機能実装 複数語から成る用語のタグ付け機能と内部構造の表示機能実装 2．専門用語解析技術 用語の内部構造を90%以上の精度で解析、500語の内部構造解析データ作成 70%以上の精度で複数の用語の並列構造を解析 3．専門用語タグ付け手法 MeSHオントロジーに従った用語分類システムのプロト構築	1．専門用語辞書システム 予定していた項目（同義語に同じ識別子をもたせることにより、表記ゆれや別表記の語の検索を可能にする機能、および、複合語の内部構造のタグ付けと構造の表示機能）の実装は完了。 2．専門用語解析技術 病名および体の部位に関する約800語について内部構造解析データを作成。用語の内部構造解析のプロトタイプツールを構築した。並列構造解析については、設計した解析手法の実装を行い，60%弱の解析性能を確認した。 3．専門用語タグ付け手法 専門用語の意味クラス分類手法の設計を行い，プロトタイプシステムを実装中。
	九州大	データ生産者とは独立して各データソースの倫理規定にしたがった3段階の共有形態による多型情報提供を行う。 1．アレルタイプと頻度情報 2．国際水準に則ったQuality control解析結果 3．多型タイピング一次データ	1．ゲノムワイド関連解析（GWAS）での基本的なデータQCを目的としたパイプラインを作成し、これを用いて「応用ゲノム」から提供されたジェノタイプ生データを解析してWEB上に提示するデータベースを作成した。2．さらに進んだQCを行うために、世界で広く用いられているGWAS解析ソフトウェアであるPLINKのQC機能を取り入れて、上記パイプラインの機能強化とweb表示の拡大を行っている。

機関名		本年度成果目標	本年度進捗状況（12月末時点）
中核機関	東京大学	DB構築技術を習得した人材を育成する。本年度はDBCLS2名、自治医科大1名、東大新領域8名	＜講義状況＞4月～12月の間に21回の講義を行った。4月～7月は9名の受講者に対してバイオデータベース構築に必要な基礎技術の講習を行った。9月からは8名の受講者に対してEnsemblゲノムブラウザーをインストールする演習を開始した。さらに10月からはEnsemblのインストールや設定を進めながら、6名の受講者に対してUTGBゲノムブラウザーに新しいデータを表示するトラックを作成するためのプログラミング講習も行っている。＜受講者の進捗状況＞受講者の進捗状況については、バイオデータベース構築に必要な基礎技術の習得までは受講者8名が進み、さらにEnsemblゲノムブラウザーの基本部分のインストールを進めている。受講者中2名はEnsemblゲノムブラウザーの基本部分のインストールを終了しトップページの表示を確認するところまで進んでおり、今後は遺伝子の検索などの詳細機能のインストールや複数種のデータのインストール等を進めていく。
	お茶の水女子大	DB高度利用者の養成。本年度は20名を対象。2名をDBCLSの統合TV開発で活用	延べ30名を対象にDB高度利用者の養成プログラムを実施中。演習でTogo Web Serviceの成果を利用。DBCLSの統合TV開発で受講者2名を活用中。
	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健康に貢献する遺伝子発掘と教材公開 2. 中級アノテーション教育（50名） 新規ゲノム配列を対象に実際にアノテーションを実施 3. 自己組織化マップによる養成（卒研生数名） 1. の結果をもとに相同性によらない生物系統の推定 4. シニア研究者と学部生の共同作業によるtRNAのデータベース化 5. 1名をDBCLSの統合TV開発で活用	1. 初級アノテーション教育（260名）「環境由来メタゲノム配列から健康に貢献する遺伝子発掘」を6月に完了、作成した教材を8月に公開開始。この教材を滋賀医科大学1回生の講義に使用。現在 DB構築中で3月中に公開予定。 2. 中級アノテーション教育（50名）は新規微生物由来のドラフトゲノム配列を対象に、平成20年11月より実施中。 3. 自己組織化マップによる相同性によらない生物系統の推定については、2名の卒研生が約600件の遺伝子の生物系統を推定した。 4. シニア研究者と学部生の共同作業によるtRNAのデータベースに約14万件のtRNAを収録し公開を行い、Nucl. Acids Res.のDatabase Issue用の論文が公開された。2008年度更新用データとして、原核生物429種、25,232遺伝子についての精査が完了。 5. 5件の統合TVコンテンツを開発し、公開を開始。
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 ・化学構造比較の高速化：SIMCOPMベースに高速化を実装し公開（7/E） ・化学反応ネットワーク予測：2つの化合物構造を入力し、その間の反応経路を予測するシステムプロト ・酵素番号自動割り当て：化合物ペアからその間の反応を触媒する酵素番号を割り当てるシステムの改良 ②ウェブ技術開発 高速・高機能な検索エンジンを開発し、構造検索システムと統合 2. 統合データベース開発・運用 ①医薬品・化合物データベース開発・運用 ・JAPIC添付文書の更新、DB間のリンク情報の更新作業を行う。併せて、付属情報の検索機能も検討 ・各種化合物DBのキーワード検索機能の公開（7/E） ・脂質DB、病原性に関わるDBを統合し、ゲノムネット化合物DBとして提供 ②LinkDB開発運用 随時追加、変更、更新を行い、併せて、キーワード検索などの機能拡張を行う	1. 共通基盤技術開発 ①知識処理技術開発 ・化学構造比較の高速化：部分構造検索プログラムSUBCOMPにおける原子や結合の認識方法におけるバグを修正。クエリ構造に含まれる部分構造の検索への拡張を実現。 ・化学反応ネットワーク予測：システムの効率化は引き続き検討中。化学構造比較アルゴリズムの改良によるキラルの認識や反応中の基質一生成物ペアの抽出アルゴリズムを実装中。 ・酵素番号自動割り当て：公開に向けたシステムのテスト中。ゲノム情報とのリンク方法について開発中。 ②ウェブ技術開発 引き続き、構造検索システムとキーワード検索との統合を検討中。 2. 統合データベース開発・運用 ①医薬品・化合物データベース開発・運用 ・JAPIC添付文書の更新、DB間のリンク情報の更新作業は引き続き順調に進めている。副作用情報など付属情報については、ライフサイエンス辞書のシノニム情報とKEGG BRITEの階層分類情報を使う方式を開発中。 ・脂質DB等に関しては、LipidBank、LIPIDMAPS、KNApSAcKとCOMPOUND、DRUGとの対応関係を引き続き更新中。LipidBankに関してはキーワード検索ファイルの提供について交渉中。病原性に関しては、引き続き関係する生理活性物質のデータベース化を検討中。 ②LinkDB開発運用 随時追加、変更、更新作業は順調に進んでいる。キーワード検索などの機能拡張については引き続き検討中。
	医科歯科大Gr	1. ターミノロジー、シソーラスを肝細胞癌、パーキンソン病から、大腸癌、舌癌及びGeMDBJに搭載されている癌疾患を対象に拡充する。 2. プロトシステムの機能向上 ・平成19年度分を含め、癌200例、神経疾患400例を公開予定 ・GeMDBJとの統合：平成20年度は、特にがん症例を追加することによって、癌と神経疾患では分かりにくかったセマンティック検索の効果をより明確にユーザに認識させることを期待する。 ・検索GUIの高度化、セマンティック検索エンジンの高度化：検索GUIの高度化では、直感的（分かり易く）で、複数の臨床・病理所見から類似症例を検索可能とする機能を追加する。 ・最終年度のシステム化に向け、より効率的に他のデータベースとの連携を可能とするため、検索エンジン等の標準化をプロトタイプレベルで先行して検証を行う。 3. 倫理規定草案の作成	1. 肝細胞癌、大腸癌に特化し、ターミノロジー、シソーラスなどを手作業にて収集整理作成している。自動作業化のためのハードルを明確化している。 2. プロトタイプシステムの機能向上に向けて、以下の作業中である。 ・癌、神経疾患の症例を増加中である。 ・GEMDBJとの統合に向けたハードルをクリアーすべく、具体的な連携方式の策定を行っている。 ・検索GUIについては、より直感的な操作を可能とする新たなパネル方式を試験中である。 ・検索エンジン設計の要素化、各要素技術のISO, WHOの場での国際標準化などは、順調に計画進行している。 3. 国際的で広範囲な倫理規定の基本調査を行っている。草案に向けたステップ項目を策定し、各項目ごとの必要案を検討してた。現在は、具体的な基準内容の検討を行っている最中である。
	東大医学部Gr	1. 標準SNP DBの構築 ← 08/07から一部のデータで公開 新たなplatformデータに対応した品質管理のための基準作成。 2. GWAS（ゲノムワイド関連解析）DBの構築 ← 08/08から一部のデータで公開 CNV対応：検出、標準、及びケースケースコントロールのデータベース化、可視化 DB拡張：新規データ受入れ用の計算のパイプライン化、第2ステージデータ受入れ対応 3. リシークエンスによる臨床情報・ゲノム情報DBの構築 パーキンソン病について、臨床情報と変異情報を搭載したリシークエンスデータベースを構築	1. 最新のplatformデータに対応した品質管理法を作成し、DBに追加。 2. CNVのデータベースの基本機能を実装。CNVを用いた関連解析結果のGWAS-DBへの搭載は実装中。また、GWAS拡張機能としてSNP間相互作用表示を実装中。計算のパイプラインは基本的な遺伝統計値については完了。GWAS-DB登録数は着実に増加。 3. パーキンソン病のDBの実装はほぼ完了。症例数追加中。
補完課題	理研	1. シロイヌナズナの発現、表現型、リソースに関する計6DBを統合化して公開。 2. 高等動植物等由来の蛋白質構造データを付随する実験データを含めて30件公開。 3. 微生物由来蛋白質に関わる試料調整（発現プラスミド構築実験 1万、培養実験 5千、精製実験 3千）、結晶化実験データ（結晶化条件：90万件、観察1000万件）と200件の回折実験データ公開。 4. 変異導入蛋白質に関わる実験データ150件、重原子導入蛋白質に関わるデータ500件を公開。 5. アノテーションシステムの開発運用と変換データの中核機関への提供（契約締結後）。	1. シロイヌナズナの発現、表現型、リソースに関する計6件DBの内、発現に関する4件はゲノムブラウザに統合して公開した。表現型とリソースはセマンティックウェブ化が終了し下記外部公開サーバから公開を予定（豊田）。 2. 高等動植物等由来の蛋白質構造データに付随する実験データ30件分について、各々の位相決定に用いた回折実験データを中心に集積中である（横山）。 3. 微生物由来蛋白質に関わる試料調整・結晶化実験データ・回折実験データ・重原子導入蛋白質に関するデータについて公開に向けたデータクレンジング作業を行っている（国島）。 4. アノテーションシステムを使って上記以外の理研データベースの統合化を推進している。一方、外部公開用サーバについては、本委託事業費に計上していなかったため、理研内の仮サーバを臨時に用いて、年度内には公開化を完了する（豊田）
	産総研糖鎖	1. 糖鎖データ統合への参加機関を確定し（12/E）、各機関と相談の上、統合化の手順を決定。 2. 検索アイテム（糖鎖構造、遺伝子名など）の統一と中核機関へのデータ等の受け渡し。 3. キーワード並びに糖鎖構造による横断検索機能の開発。 4. 糖鎖科学統合DBの検索機能追加：糖鎖構造全体の推定や抗体、レクチン等の部分構造推定情報表示。 5. ノックアウトマウスを題材にモデル生物の情報を集約するカテゴリーの構築を行う（名古屋大と連携）。 6. 別途開発のAPIを用いた統合検索用プロタイプの構築。	1.参加確定した機関（他3機関と交渉中） 理化学研究所のシステム糖鎖生物学研究所の糖鎖コンフォメーションDB 野口研究所の有機化学による糖鎖合成DB＋化合物DB 創価大学のショウジョウバエのGlycoGeneDBとフェノタイプ情報 2.GPDBのデータを提供。HyperEstraierのポートをDBCLSに開放。GlycoForumを横断検索に追加。糖鎖構造検索用のXMLの整理中。 3.構造による横断検索の仕組みを開発中。 4.保留中。2)と3)が完成しないと取り掛かれない。 5.プロトタイプ版をより柔軟にドキュメント登録して公開できるシステムを開発し、KOマウスの公開用として使い勝手をヒアリングする予定。 6.GGDBのAPIが完成（経産省側Proj.）。GlycoEpitopeのAPIが完成。多次元HPLCDBは2月中旬完成予定。
	遺伝研	1. トレースデータ用FTPサーバ構築。キーワード検索、ダウンロード、統計情報閲覧サイト作成。 2. トレースデータ登録処理システム、波形表示システム開発。	1. プロトタイプ開発をもとに、DB構築、DBへのデータ登録システム、Web検索システム、 波形表示システムを開発中。 2. データベース構築 → データベースのスキーマを仮に構築し、実データを投入して検証中。 3. web検索システム → 検索・結果表示・個別データダウンロード機能は実装済み。データベース構築と共に性能評価やテストを元に引き続き開発中。 4. 波形表示システム → 既存プログラムをベースに修正中。より良い表示方法がないか引き続き検討中。 5. 統計情報表示ページについても表示項目の詳細を検討中。
	九工大	1. 蛋白質と変異体の熱力学データ1000件と構造データの対応表作成。 2. 蛋白質と核酸の相互作用の定量的な熱力学実験データ1,300件を対象に構造データの対応表作成。 3. 蛋白質・蛋白質相互作用データ格納用、熱力学データ用XMLフォーマット、データ抽出プロトを作成。	1. 蛋白質と変異体の熱力学データと構造データの対応表については、これまでに約1，100件のデータについて作成した。 2. 蛋白質と核酸の相互作用の熱力学データの対応表については、これまでに約700件のデータについて作成した。 3. 蛋白質・蛋白質相互作用データについては、試験用のデータベースシステムとWebページを作成中。熱力学データ用XMLフォーマットは、蛋白質と核酸の相互作用の熱力学データについて暫定版を作成し公開した。熱力学データについて使用しているボキャブラリーを整理している。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、DBセンターが開発したテキストマイニングシステムを試験中。 その他、プロジェクトのホームページを作成してWeb上に公開した。