NBDC
National Bioscience Database Center

ライフサイエンスデータベース
統合推進事業　事業報告書（概要）

2016

科学技術振興機構

Japan Science and Technology Agency

Ⅰ.事業目的および事業概要

（報告書p2-4参照）

1. 発足までの経緯

背景　平成12年6月 ヒトゲノムの概要配列が解読され、また、その他数多くの生物のゲノム配列が解読されたことを受けて、ポストゲノムシーケンス研究が激しい国際競争の中で本格化した。こうした国際動向の中で、わが国におけるゲノム情報科学に関する戦略が議論され、情報科学の人材養成、研究開発の振興、データベース整備戦略の三つの課題に関する推進方策の検討が行われた。

平成12年11月
「ゲノム情報科学におけるわが国の戦略について」
（科学技術会議 ライフサイエンス部会 ゲノム科学委員会）

H13年度～
バイオインフォマティクス推進センター事業（BIRD　JST）
大学等におけるゲノム情報等生物情報データベースの構築・高度化及び生物情報データベースを活用した研究開発支援

平成18年5月
「我が国におけるライフサイエンス分野のデータベース整備戦略のあり方について」
（科学技術・学術審議会 ライフサイエンス委員会）　*（詳細は次ページ参照）*

H18年度～
統合データベースプロジェクト（文部科学省）
既存のライフサイエンスデータベースの統合化

中核機関：
情報・システム研究機構
（ライフサイエンス統合データベースセンター）

平成21年1月
「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」
（科学技術・学術審議会 ライフサイエンス委員会）
※平成22年度までに2事業を一本化し、JSTに新たな組織を設置して、データベースの統合・維持・運用を図ることを提言。

平成21年5月
「統合データベース タスクフォース報告書」
（総合科学技術会議　ライフサイエンスPT）
※ライフサイエンス分野における我が国全体の恒久的且つ一元的な統合データベースの整備について方針をとりまとめ　*（詳細は次ページ参照）*

独立行政法人科学技術振興機構（JST）が実施していた**「バイオインフォマティクス推進センター事業」（BIRD）**と文部科学省が実施していた**「統合データベースプロジェクト」**が平成23年度より統合されて開始

H23年度～　ライフサイエンスデータベース統合推進事業（JST）/ NBDCの設置

平成23～25年度　3年間の第一段階としてスタート

平成25年1月
「総合科学技術会議ライフイノベーション戦略協議懇談会」
平成26年度以降においても、引き続きNBDCを中心とした現行の体制で推進することを了承

（報告書p2-4参照）

2. NBDCが実施すべき項目、および事業目的

平成18年5月には、ライフサイエンス委員会 データベース整備戦略作業部会で、「日本の生命科学データベースに関して、取り組むべき10の課題」がまとめられた。

日本の生命科学データベースに関して、取り組むべき10の課題

(1) データベースの現状調査、評価、戦略立案機能の充実
(2) 基盤データベースの安定的な支援
(3) データベースの所在情報と利用法に関するポータルサイトの構築と運営
(4) 統合データベースの開発とそのための研究開発の促進
(5) 維持が困難になったデータベースの受入れ
(6) 文献情報との連携
(7) アノテーション（情報解読による実験データの注釈付け）の実施
(8) 新たなデータベース構築への投資
(9) データベースを活用した研究（バイオインフォマティクス）の促進
(10) データベース開発のための人材養成

その後、ライフサイエンスデータベースの統合・維持・運用の在り方が検討され、さらに「統合データベースTF報告書」（平成21年5月）にて、ライフサイエンス分野における我が国全体の恒久的且つ一元的な統合データベースの整備について取りまとめられた。求められる機能としては、以下があげられており、これらを受け、JSTにおいて具体的な設置準備の検討を行い、事業の目的、4本の柱を整理し、これに基づき運営してきている。

統合データベースセンターに求められる機能として、

- データベース統合に必要な調査
- データベースの統合に必要な標準化
- 海外との連携等の実務機能
- システムの構築・維持・管理
- ポータルサイトの構築
- データベースの受入れ・管理・更新
- データベースの品質管理
- 各省等のﾃﾞｰﾀﾍﾞｰｽとのﾈｯﾄﾜｰｸの構築
- データベースの統合化や高度な検索等
  統合的利用のための技術開発（インデックス、辞書、データフォーマットなどの構築）

その他、取り組むべき項目として、

- 国内のデータベース等の整備
- 国際連携
- 人体に由来するデータ等の取り扱い

NBDCの事業の4本柱

（1）戦略立案
（2）ポータルサイトの構築・運用
（3）基盤技術の開発
（4）統合データベース構築
（1）戦略立案

NBDC事業の目的

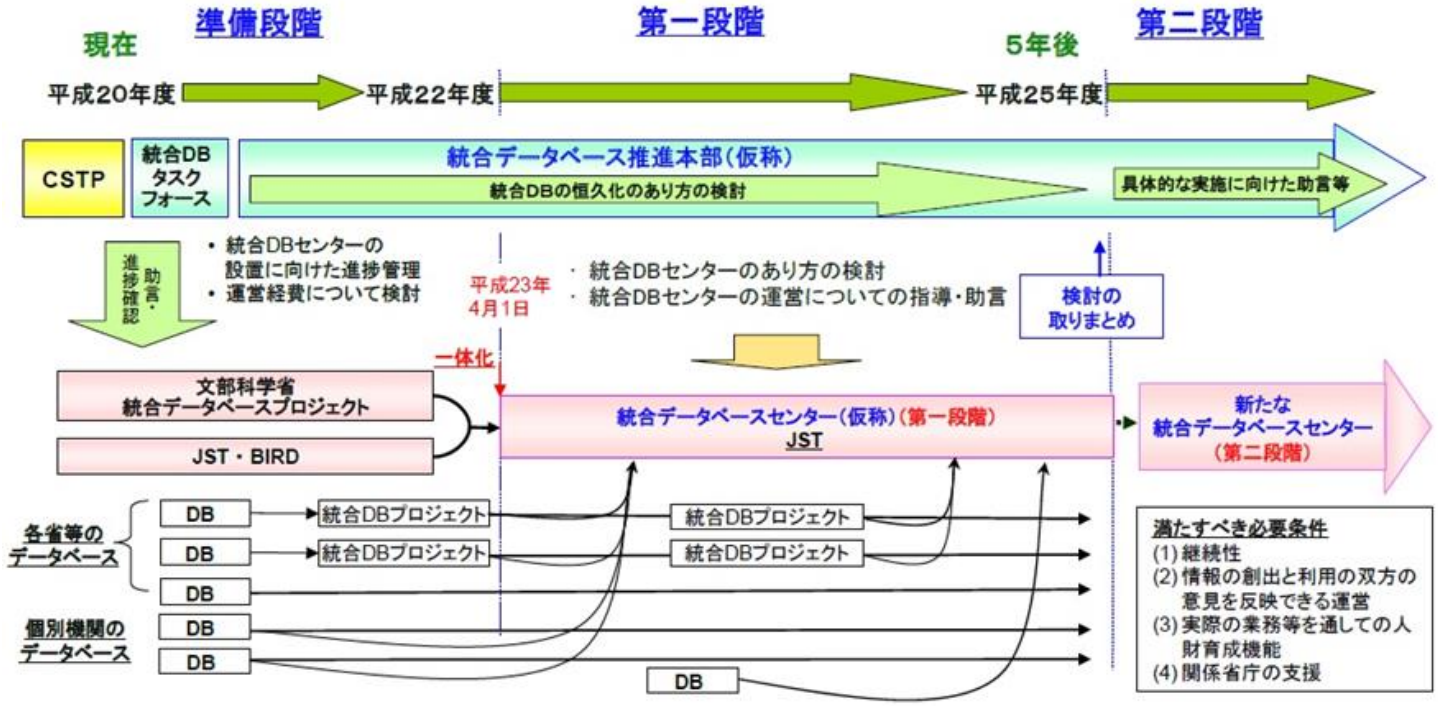
我が国におけるライフサイエンス研究の成果が、広く研究者コミュニティに共有かつ活用されることにより、基礎研究や産業応用研究につながる研究開発を含むライフサイエンス研究全体が活性化されること。

（報告書p2-4,p12-15参照）

3. 事業概要（ロードマップ）

「統合データベースタスクフォース報告書」にまとめられた、NBDCの事業ロードマップは以下の通り。

統合データベース整備のロードマップ



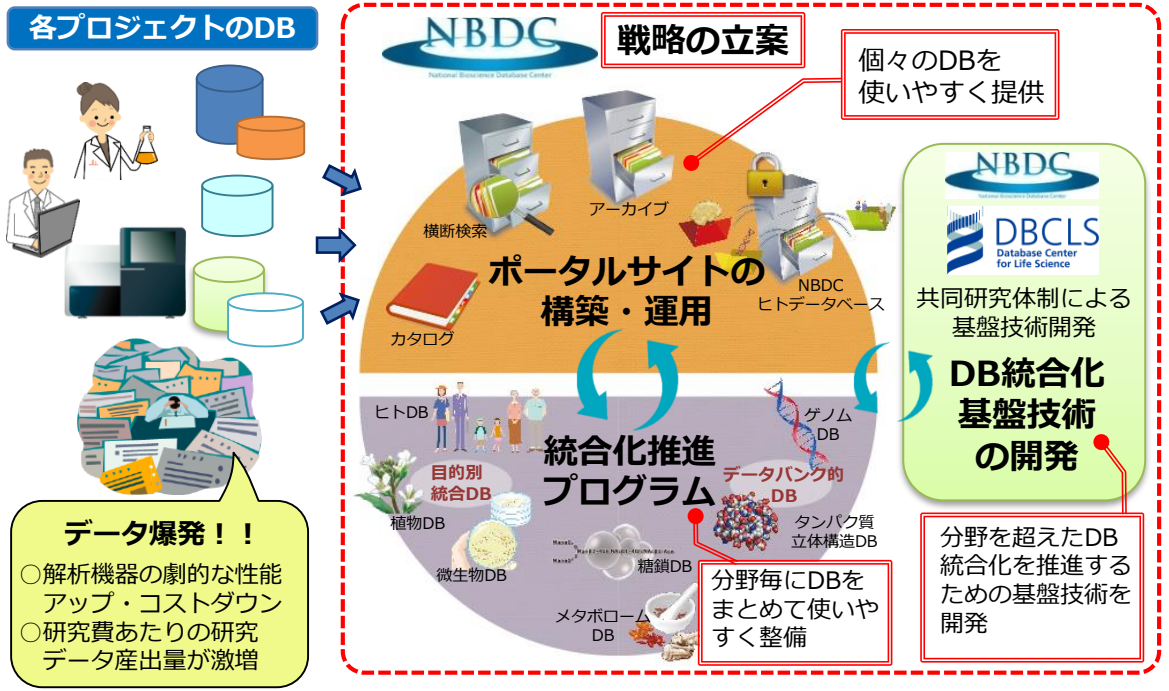
報告書では、下記の通りとされた
- ・統合データベースセンターは平成23年度にJSTに設置
- ・当初3年間で、一元的・恒久的なデータベースセンター設立の検討を行う

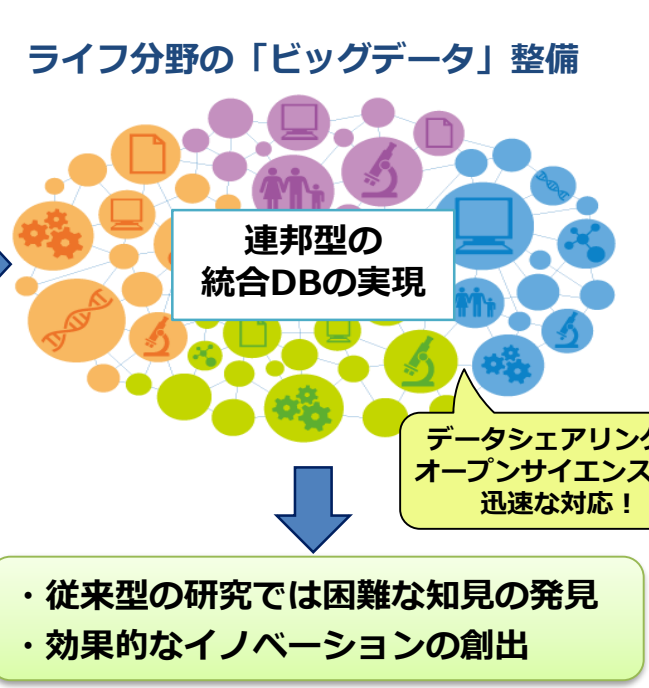
➡ 検討の結果、平成26年度以降においても、引き続きNBDCを中心とした現行の体制で推進することとなった。

（報告書p5-6参照）

3. 事業概要（NBDCの事業の4本柱 ）

NBDCで行っている活動の4本の柱の概要は以下の通り。





1. 戦略の立案：データベース整備･統合化の戦略企画、ガイドラインの策定、国内外との連携構築等を実施。
2. ポータルサイトの構築･運用：データベースのカタログ、横断検索、アーカイブ等のサービス提供を実施。
3. 基盤技術の開発：データベース統合化に向けて基盤となる技術開発とその実装を実施。
4. 統合化推進プログラム：分野毎のデータベース統合化により、国内バイオ関連データベースの統合を実施。

（報告書p12-22参照）

4. 事業計画（平成23年～平成28年度末）

1）戦略の立案

- ✓ 国内外の効果的な連携（4省連携の枠組み立ち上げ、AMEDとの連携開始、国内データセンター間での有機的な連携）
- ✓ ヒトに由来するデータ等の取り扱いへの対応、国際連携 （ガイドラインの策定、審査システムの構築、ヒトデータベースの運用開始、ヒトデータの国際的な共有（DDBJとの連携）、GA4GHへの加入・プロジェクトへの参画）
- ✓ 国際連携（RDFによる統合に向けた国際連携、バイオハッカソン等の開催）
- ✓ 人材育成施策の検討

2）ポータルサイトの構築・運用

- ✓ 4省連携によるポータルサイトの構築・運用
- ✓ NBDCポータルサイトの構築・運用
- ✓ 生命科学データベースカタログの運用（現在国内で公開しているデータベースを網羅する、国際連携）
- ✓ 生命科学データベース横断検索の運用（カタログに掲載されたデータベースを横断検索の検索対象として網羅する）
- ✓ 生命科学データベースアーカイブの運用（既存DBの収録、RDF形式のデータベースを集約して再構築）

3）データベース統合化基盤技術の研究開発

- ✓ データ統合のための技術戦略の決定
- ✓ RDFによる統合化のための基盤技術の開発
- ✓ RDF化に関する国際的な標準化活動に着手
- ✓ 「統合化推進プログラム」へのデータベース統合化支援
- ✓ 関連DBの利用の活性化・人材育成につなげるためのエンドユーザ向けのツール等の充実

4）バイオ関連データベースの統合化の推進

- ✓ データベースの統合化を進め、日本を代表するデータベースの構築
- ✓ 主要な分野における、中核・拠点となる統合データベースの構築
- ✓ 個々の統合データベースを統合するため、データのRDF化を進める

（報告書p8-9参照）

1. 事業の推進（1）事業の運営体制

NBDC事業運営体制は以下の通り。

理事長

諮問

答申

ヒトデータベース倫理審査委員会

ヒト由来データの収集・分譲に関することを審査

理事、副理事

バイオサイエンスデータベースセンター

センター長

副センター長（平成26年4月以降は空位）

諮問

答申

運営委員会

倫理分科会（平成23年度）
データ共有分科会（平成24年度～）
人材育成分科会（平成24年度～平成27年度）

ヒトデータ審査委員会

運営委員会及び分科会：
事業の推進に関する事項を調査・審議。
ヒト由来データの取扱いや人材の育成に関しては分科会を設置して、調査・審議。

ヒトデータ審査委員会：
NBDCヒトデータベースのヒトデータの受入れ及び利用に関する審査。

センター長　センター全体の活動を総理。

企画運営室（13名(常勤)）

企画グループ
研究開発推進グループ

研究チーム（10名(常勤)）

基盤技術開発　※平成23年度～平成25年度は委託研究開発で実施

基盤技術開発：
データベース統合化に向けて基盤となる技術開発とその実装。

統合化推進プログラム

研究総括

研究開発課題　研究開発課題　研究開発課題

統合化推進プログラム：
分野毎のデータベース統合化により、国内バイオ関連データベースの統合。

研究総括：
統合化推進プログラムを総括。

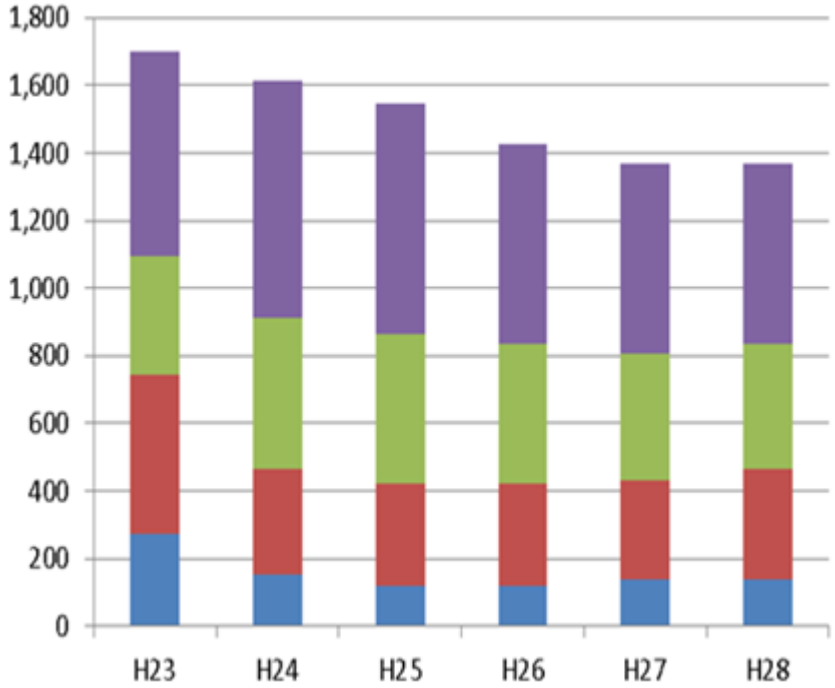
企画運営室　センター運営、予算管理、庶務、委託研究開発及び共同研究の課題管理、

研究チーム　生命科学情報基盤の整備・統合化のための、研究開発動向の調査及び戦略の立案、研究開発並びにポータルサイトの構築運用

（報告書p9参照）

1．事業の推進 (2) 予算

NBDC予算は、センター発足時の平成23年度は約17億円であった。
その後漸減し、平成28年度は約14億円となっている。4本柱毎の
内訳は以下の通り。



予算内訳

統合化推進プログラム	約40%
基盤技術	約30%
ポータルサイト	約20%
戦略立案	約10%

（報告書p10参照）

2. 事業の推進（参考）海外DBセンターとの比較

国・地域	日　本			米　国	欧　州		
所属機関	大学共同利用機関法人 情報・システム研究機構（ROIS）		国立研究開発法人 科学技術振興機構（JST）	国立衛生研究所 （NIH） 国立医学図書館 （NLM）	欧州分子生物学研究所 （EMBL）		スイスバイオインフォマティクス研究所（SIB）
組織名称	ライフサイエンス統合データベースセンター （DBCLS）	国立遺伝学研究所 DDBJセンター （DDBJ）	バイオサイエンスデータベースセンター （NBDC）	国立バイオテクノロジー情報センター （NCBI）	欧州バイオインフォマティクス研究所 （EBI）	ELIXIR （European Life-sciences Infrastructure for Biological Information）	スイスバイオインフォマティクス研究所（SIB）
組織の位置づけ	文部科学省ライフサイエンス分野の統合データベース整備事業開始に伴い、データベースプロジェクトの中核機関として2007年4月に設立。	機構傘下の国立遺伝学研究所の付属施設「生命情報学」の我が国における研究拠点我が国を代表するDNAデータベースを運営	JSTのライフサイエンス分野のデータベース統合を目的として発足。「戦略立案」「ポータルサイト構築・運用」「データベース統合化基盤技術の研究開発」「バイオ関連データベース統合化の推進」の4機能を柱に据える。	NIH傘下であるNLMの付属機関（NLMは生物医学情報の集積をミッションとしており、その中で、NCBIは分子生物学に関するデータ及び生物医学文献に特化した部門）	EMBL傘下の非営利学術機関バイオインフォマティクスの研究とサービスの中心機関	ELIXIR は、組織化されたネットワーク内で、欧州のライスサイエンス関連のデータを共有、保管する枠組みで、EMBL-EBI傘下の独立組織	スイスのバイオクラスターBioAlpsに属し、生物情報科学の研究とサービスに携わる60の組織とスイスの主要な高等教育・研究機関で構成される非営利団体
組織設立の根拠（永続性）	予算の9割近くをNBDCからの時限付委託費により運営	欧州EMBLと米国GenBankから、当時、日本の国際協力事業の参加要請を受け、1987年より本格稼働	総合科学技術会議ライフサイエンスPT での検討を経て、政府予算の認可に伴い2001年4月発足	1998年11月4日に制定されたPublic Law 100-607で、NIHの付属機関としてNCBIを設置することが認可	イギリス政府とEMBLとの協定によって設立	欧州におけるバイオインフォマティクスのインフラへの投資と、欧州内のそれぞれの国における投資を一体化するために2013年に創設	Swiss-ProtとEMBnet node（欧州分子生物学ネットワーク）が共同でスイス政府から長期的資金を得て1998年3月30日に設立
人員規模	25人（2015年度）（非常勤含）	39人（2015年度）（非常勤含）	29人（2015年度）（非常勤含）	287人（2015年度）	513人（2015年度）	2,064人（2015年度）［参加16カ国］	670人（2014年度）
予算の出所	予算の9割近くをNBDCからの時限付委託費により運営	国立遺伝学研究所の運営費交付金により運営	JSTの運営費交付金	NIH	NIH（€7.1M）、Research Councils UK（€5.6M）、European Commission / ERC（€5.5M）、Wellcome Trust財団（€5.4M）、他（2015年度）	スイス政府；SIBを通じて3500万ユーロを投資（2013-2016年） 英国Large Facilities Capital Fundが資金割り当て	NIH傘下NHGRI（米国立ヒトゲノム研究所）をはじめ、米国研究機関、スイス国内外のヨーロッパ研究機関、スイス国立科学研究基金等
予算額	約3億6,473万円（2015年度） 受託研究費含む	約14億6,078万円（2015年度） 受託研究費含む	約13億6,920万円（2015年度）	1億7,584万ドル（2015年度） 〔約181億2,717万円〕	6,720万ユーロ（2015年度） 〔約77億1,250万円〕	221.7万ユーロ（2015年度） ［約2億2,545万円］	679万スイスフラン（2014年度）〔約72億4千万円〕
国際協力	BioHackathonの開催 セマンティックwebによるデータベース統合の国際標準の整備	米国NCBIのGenbank、欧州のEMBL-bank極で国際塩基配DBを構築	BioHackathonの開催	世界に向けてのサービス	世界に向けてのサービス DDBj、NCBIと協力し		2002年、UniProt（protein

日本は欧米に比べ、データベースセンターの人員、予算ともに少ない状況であるが、如何に少ないリソースで、インパクト・プレゼンス、を出せる取り組みを行えるかが、重要な課題となっている。

（報告書p16-22参照）

2. 事業の成果（1）戦略立案機能（第二段階の推進戦略および5ヵ年事業計画）

「統合データベースタスクフォース報告書」では、当初の3年間（平成23年度～平成25年度）で平成26年度以降の体制を検討する、とされていたことから、NBDCでもあり方の検討を行った。
平成24年7月、NBDC運営委員会で「ライフサイエンス分野の統合データベース整備の第二段階のあり方について（報告）」がまとめられた。本報告書は、平成24年8月に総合科学技術会議 ライフイノベーション戦略協議会に報告され、また、平成25年3月と平成25年8月のライフサイエンス委員会において報告・議論され、**平成26年度以降も、NBDCを中心とした現行体制で事業を実施する**こととなった。
これを受け、平成26年3月、NBDCで「ライフサイエンス分野の統合データベース整備の第二段階の推進戦略」をまとめた。なお、より詳細な5ヶ年事業計画を策定し、事業を推進した。

第二段階の推進戦略　骨子

1）戦略の立案
国際的な動向を踏まえたフォーマットの標準化と標準化されたフォーマットの発信、普及に向けた技術支援、効率的で情報の信頼性が確保された統合データベースの構築を目指す。

2）ポータルサイトの構築・運用
文部科学省、厚生労働省、農林水産省、経済産業省の 4 省合同ポータルサイト（integbio.jp）の拡充、研究データ提供依頼対象の拡大、日本医療研究開発機構とも密接な連携を図る。

3）データベース統合化基盤技術の研究開発
ROIS との共同研究開発による、より一体的かつ効率的な研究開発を実施し、RDF による統合化のための基盤技術開発推進、統合化推進プログラム成果等の統合化支援、エンドユーザ向け大規模データ統合解析環境、高度な統合データベース問い合わせシステムの開発を実施する。

4）バイオ関連データベースの統合化の推進
データの公開、補完性、汎用性、国際競争力、標準化の観点を踏まえ、データベースの統合化を推進する。

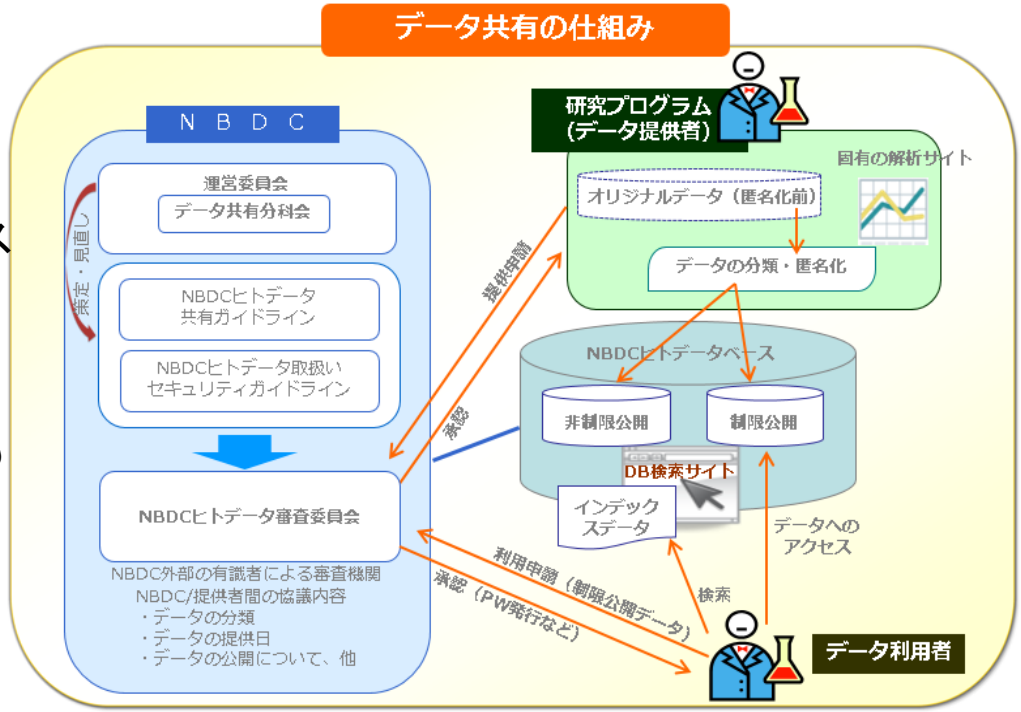
5）研究データ活用支援プログラム
人材育成施策の検討を進める。

（報告書p23-28参照）

2. 事業の成果（1）戦略立案機能（ヒト由来データの公開・共有－1－）

○NBDCヒトデータベース運用開始の背景
- ヒト由来のデータは非常に重要、かつ他の生物種データとは異なる配慮が必要。
- 我が国において、データベースにおけるヒト由来データの取扱いに関する統一的指針が未整備であった。
- 統合化推進プログラムでもヒト由来データを扱う課題を採択。

⇒ ヒト由来データの取扱いについて、NBDCとして積極的に取り組むべきと考え、日本で初めてのガイドラインや審査システムを構築した。
遺伝研／DDBJと協力し、我が国初のヒトデータ共有プラットフォームであるNBDCヒトデータベースの運用を開始した。

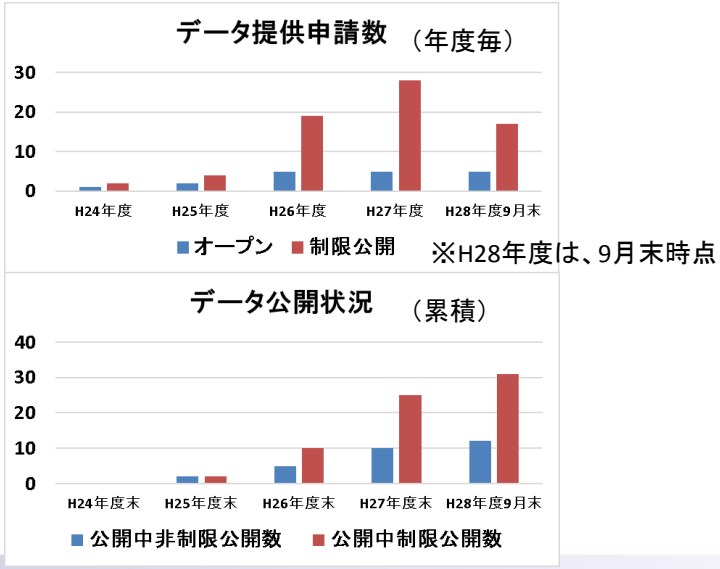


○国際的な取組みであるGA4GHへ参画し、Beacon Projectへ参加するなど、ヒトデータの利活用促進も図っている。

※Beacon Project：
ゲノム上の特定の1塩基の変異について、世界の200以上のデータセットを検索できるシステム。検索の方法と結果情報を限定することで、データ利用審査なしに迅速な情報入手が可能。
GA4GH（Global Alliance for Genomics and Health）がデータ共有の意義を示すために取り組んでいる4つのプロジェクトのうち1つ

（報告書p23-28参照）

2. 事業の成果（1）戦略立案機能（ヒト由来データの公開・共有－2－）

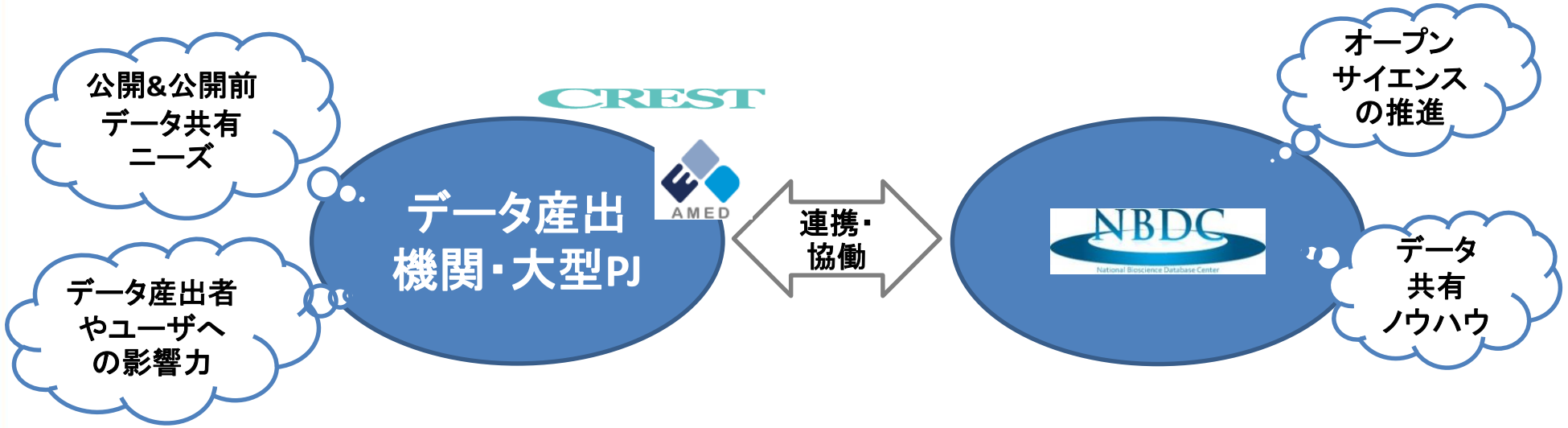


※ヒトデータについては、第5期科学技術基本計画のオープン＆クローズ戦略の観点から、グループ間共有の取組みを始めることを決定。
当面はグループ間の共有なるも、将来は公開のカテゴリーに移行することが見込まれるデータを受け入れる。
平成28年度中にシステムを立ち上げる予定。

（報告書p29-32参照）

今後の取り組み　他機関、他プロジェクトとの連携　－1－

利用ニーズを把握し、より利用される統合データベース構築を行うため、ユーザにより近い機関・大型プロジェクトとの連携関係を構築していくこととしている。
それぞれの強みを生かした役割分担と協働により、データの共有と利用を促進していく。
（AMEDやCREST「植物頑健性」領域から着手）



データ利活用の働きかけ

- ユーザニーズの把握
- データ利活用の企画
- データ共有ポリシーの作成　など

システム基盤整備

- データベース運営
- 既存データ資産との連携
- データ共有ガイドラインの作成　など

（報告書p29-32参照）

今後の取り組み　他機関、他プロジェクトとの連携　-2-

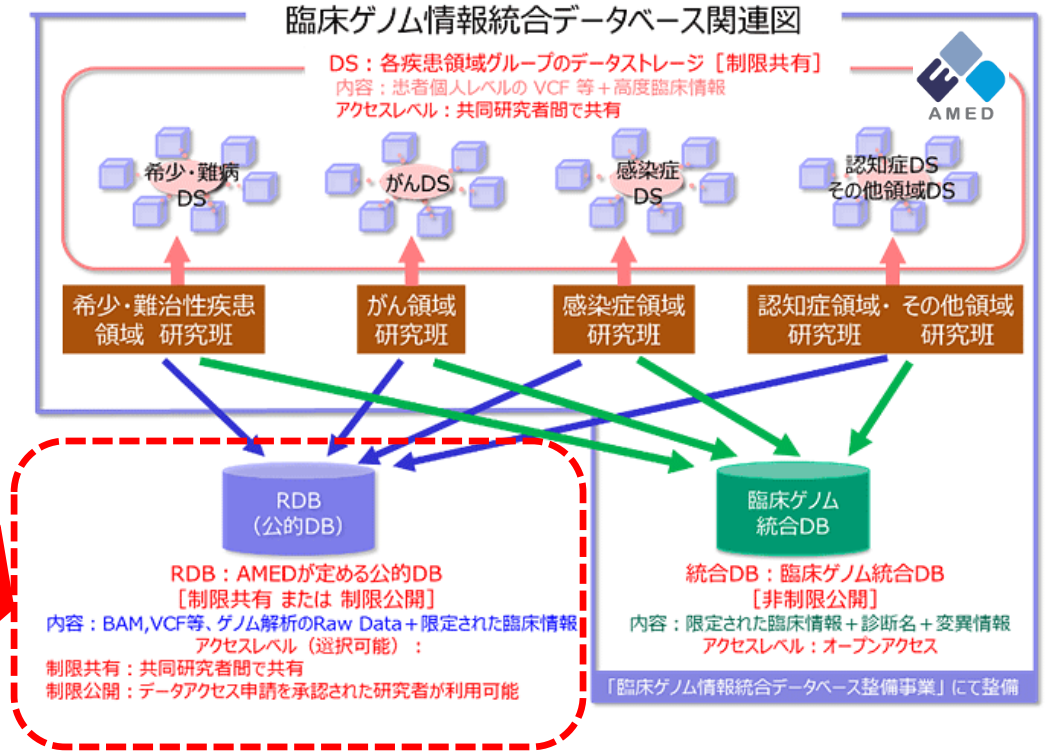
他機関、他プロジェクトとの連携の例

1）AMEDと基本連携協定を締結
「ゲノム医療の実現に資するデータシェアリングに係る基盤整備及び利活用」を推進するため、AMEDと連携協定を締結した。

2）グループ共有データベース立ち上げ
AMEDがデータシェアリングポリシーで規定している制限共有に歩調を合わせ、NBDCでもグループ間共有のためのデータベースを立ち上げる。公開系のヒトデータベースに続き、グループ間共有のデータも、NBDCに集約する見込み。

3）AMED ゲノム情報基盤推進分科会への参加
分科会に参画し、バイオバンク一元検索システムの検討等で、NBDCの知見を共有。

AMED臨床ゲノム情報統合データベース事業のヒトデータの流れ



AMED　臨床ゲノム情報統合データベース整備事業のWebサイトより抜粋。
http://www.amed.go.jp/program/list/04/01/047.html

（報告書p35-38参照）

2. 事業の成果 (2) ポータルサイトの構築・運営 (integbio.jp)

ライフサイエンスデータベースに関連する文科省、厚労省、農水省、経産省で連携して、各省のデータベース統合に取り組んだ。情報発信のWebサイトを束ね、4省合同ポータルサイト「integbio.jp」をNBDCが取りまとめて立ち上げ、運営を実施している。

第一期中に実施

（1）カタログ *データベースの所在情報を提供* ・・・ 連携して、データベースの探索。

（2）横断検索 *複数のデータベースを横断的に検索* ・・・ 連携して、横断検索用INDEXファイルを相互持ち合い、相互参照。

（3）アーカイブ *「統一フォーマット」でのダウンロードの実現* ・・・ 連携して、アーカイブデータの作成。

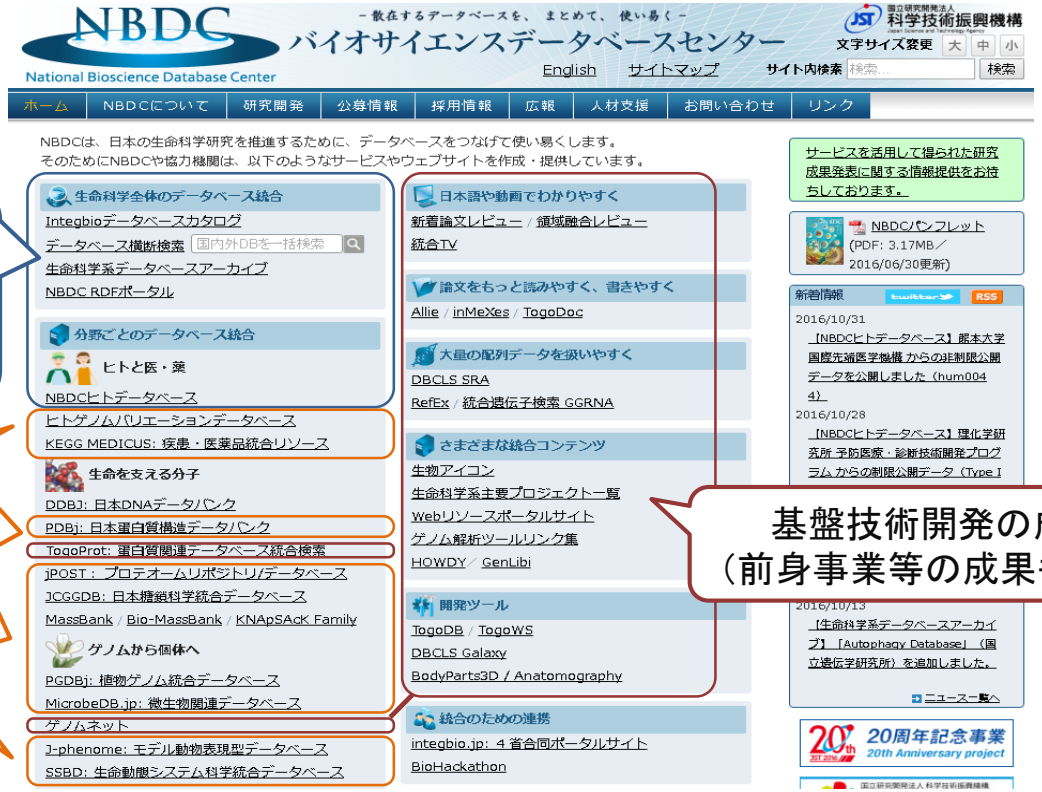
第二期以降を想定

（4）データベース再構築 *「データベースの再構築」による高度な検索の実現* ・・・ まずは、RDFで再整備されたデータベースを収載したポータルサイトを構築。

4省データベース統合の段階的進展

（報告書p35-38参照）

2. 事業の成果 (2) ポータルサイトの構築・運営 (biosciencedbc.jp)



NBDCで開発・提供するサービス

統合化推進プログラムの成果

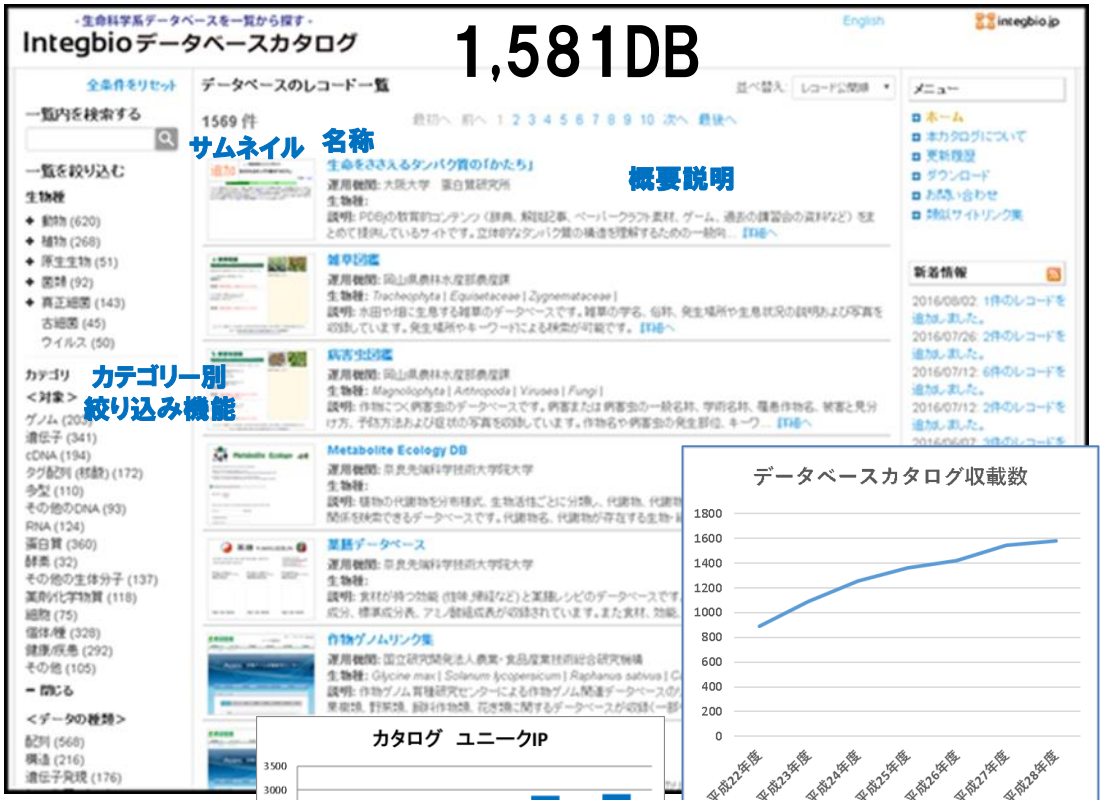
基盤技術開発の成果
（前身事業等の成果も含む）

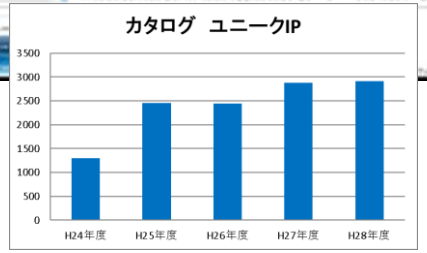
NBDCの事業に関する成果物（サービス、ツール類）にワンストップでアクセスできるポータルサイトを構築し、運用している。

- 40種類以上のサービス
- 生命科学のDB関連
- 登録不要
- 無料
- どこからでも、誰でも

（報告書p39-40参照）

2. 事業の成果 (2) ポータルサイトの構築・運営（生命科学系データベースカタログ）





（平成28年度9月末時点）

国内外の生命科学系データベースの所在情報、データベースについての説明、生物種などの様々な属性情報（メタデータ）をまとめたリストを提供。

散在するデータベースから、利用したいデータベースを容易に探し出すことができるシステム。

メタデータを付与しており、対象を絞り込んだ検索も容易。

4省連携のもと、各省の研究機関のDBの調査、ファンディングプログラムの報告書の調査などを行い、平成28年9月末現在、収録レコード数1,581件。国内主要データベースをほぼ網羅。

海外の取組み（biosharing.org）との国際連携により、約800のデータベースを順次掲載予定。

（報告書p41-43参照）

2. 事業の成果 (2) ポータルサイトの構築・運営 (生命科学系データベース横断検索)

（平成28年度9月末時点）

※一部改修のため(平成27年7月～)、平成27年度は4月～6月の平均。
従前データとの比較ができないため、平成28年度のデータは割愛。

散在する生命科学系のデータベースを特許や文献と一括して検索できるシステム。

平成28年9月末時点、596DBを対象とした横断検索を実現。

データベースの中身を深く網羅的にクローリングすることで、一般的な公開Web検索サービスでは難しい、網羅性を実現。さらに、ライフサイエンスと無関係な検索結果の排除を可能にした。

4省連携での相互横断検索を実現。
検索インデックス更新作業の自動化や省力化への取り組み、また検索レスポンス向上の取り組みを実施。

（報告書p44-45参照）

2. 事業の成果 (2) ポータルサイトの構築・運営 (生命科学系データベースアーカイブ)

国内のライフサイエンス研究者が生み出したデータセットを長期間安定的に維持保管し、多くの人が容易にデータへアクセスしダウンロードを行えるシステム。データの死蔵・散逸の防止の一端を担っている。

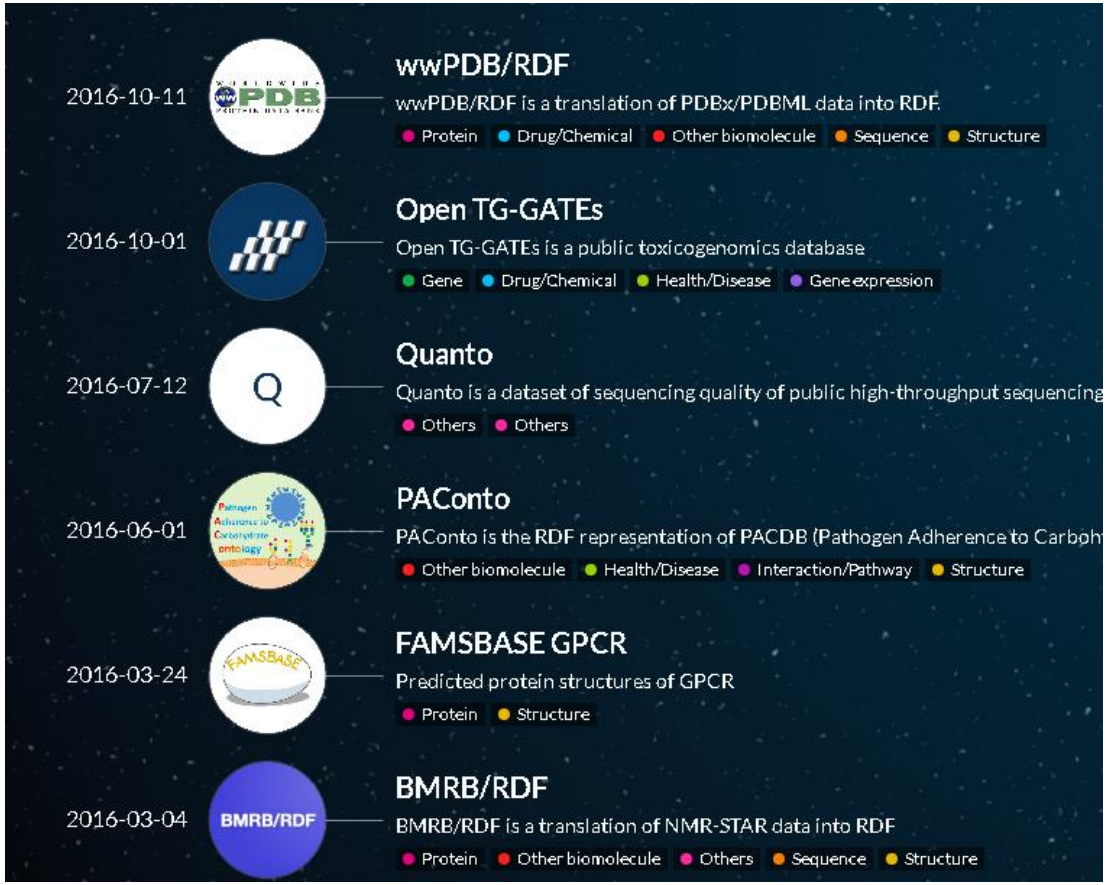
CCライセンス（Creative Commonsライセンス）を適用し、権利関係も処理されたデータベースを搭載している。

アーカイブ内横断検索機能やRDFダウンロード機能等のシステム改良。レスポンス向上、J-GLOBAL等とのリンク、DOIの付与など、機能・内容ともに拡充。

（平成28年度9月末時点）

（報告書p49-51参照）

2. 事業の成果 (2) ポータルサイトの構築・運営 (RDFポータルサイト)



様々な研究機関が作成したRDF形式データを掲載したポータルサイト。

13件のデータセット（wwPDB/RDF、微生物、糖鎖構造、遺伝子発現、化学物質[日化辞]等）を公開。（平成28年9月末現在）

ユニークIP

平成27年度	1,342	（11月～3月までの月毎のユニークIP合計）
平成28年度	781	（ 4月～9月までの月毎のユニークIP合計）

RDF化の状況

主要なデータベースの多くはすでにRDF化されており、RDF形式は国際標準となっている。

RDF化された国際的な主要DB

- PubChem、医学用語集MeSH （米国、NCBI）
- Ensembl、ExpAtlas、UniProt （欧州、EBI）
- GenBank/DDBJ/ENA （日米欧、INSDC）
- UniProt （欧州、SIB）
- PDBj （日本、大阪大学）

（報告書p5-6, p52-55参照）

3. 基盤技術開発の推進（1）目的

ライフサイエンス分野のデータベース統合を実現するため、国内の基盤的データベースや、統合化推進プログラムで構築される分野別統合データベースを統合するために必要な技術開発を行う

背景

統合データベース タスクフォース報告書　H21.5 CSTP ライフサイエンスPT

4．体制整備（3）「統合データベースセンター（仮称）」の整備②求められる機能
具体的には、データベース統合に必要な調査、データベースの統合に必要な標準化、システムの構築・維持・管理、ポータルサイトの構築、データベースの受入れ・管理・更新、データベースの品質管理、各省等のデータベースとのネットワークの構築、海外との連携等の実務機能に加えて、データベースの統合化や高度な検索等、統合的利用のための技術開発（インデックス、辞書、データフォーマットなどの構築）の機能とする。

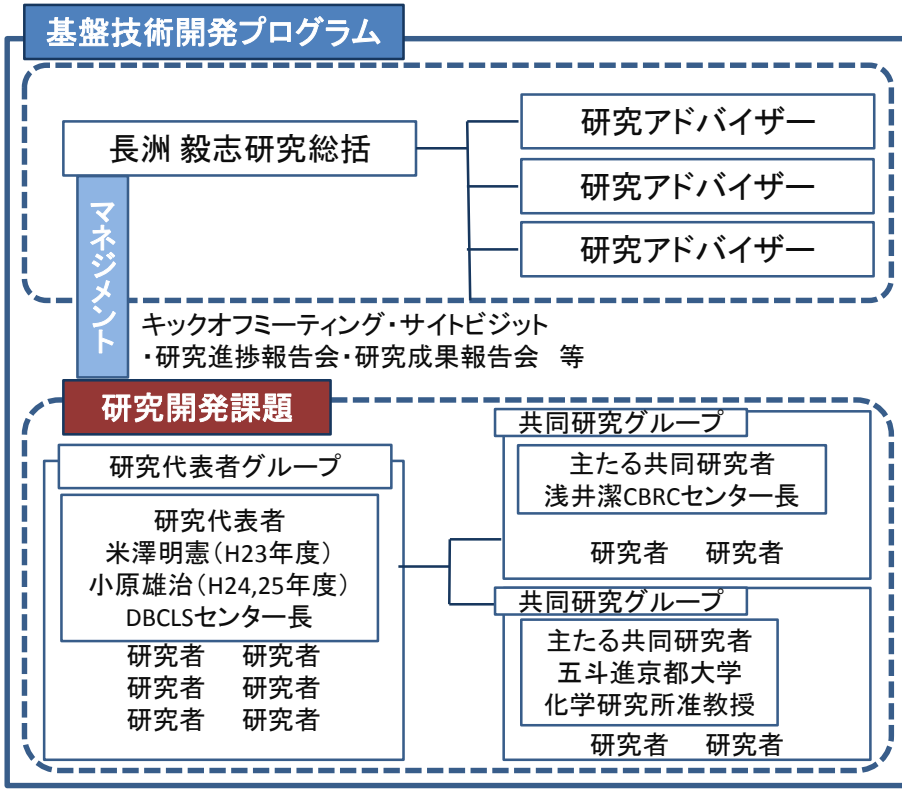
⇒ JSTライフサイエンス分野統合データベースセンター制度検討ワーキンググループにて推進方法を検討
統合DBに必要なトータルな技術を1つの機関が責任を持って開発する。提案を公募して行う。

公募にあたり募集要項で提案に求めたこと
・国内の基盤的データベースおよび本事業で構築される分野別統合データベースのRDF化を実現するための、標準フォーマット、オントロジーの提供、RDF化の支援を行い、RDFコンテンツを公開すること
・他　高度検索技術開発、統合利用環境の整備、文献、画像等多様なデータの利用技術開発、オントロジー、辞書、コーパス、標準化技術、DBやツールの使い方、チュートリアル動画等利用のためのコンテンツ作成、アノテーション支援等

（報告書p53-54参照）

3. 基盤技術開発の推進 (2) 推進体制等

H23～H25　公募　直接経費3億円程度／年



H26～　NBDCとの共同研究　直接経費3億円程度／年

共同研究開発機関：
情報・システム研究機構
ライフサイエンス統合データベースセンター

相手方研究者：
小原雄治
情報・システム研究機構
ライフサイエンス統合データベースセンター　センター長

研究推進体制：
- ✓ 毎年度、運営委員会において開発計画の妥当性確認・承認と実施報告
- ✓ DBCLS内月次ミーティングへのNBDC研究員の参加

（報告書p56-64参照）

3. 基盤技術開発の推進 (3) 研究実施 (研究計画)

下記の研究・技術開発に取組み、機械可読なデータ規格の適用とデータ統合に必要な各種技術開発を通じて、データの統合・利活用を容易に行うための環境づくりを目指している

①データベースのRDF化とRDFによる統合化のための基盤技術開発

各種機能開発・利用環境整備、リファレンスデータセットの整備、新規分野データ活用技術の開発、Keyとなるデータを付与しながらのデータベースのRDF化、ヒト関連情報を安全に共有・取り扱うためのセキュリティの検討等

②国内外のDBについての統合化支援、およびそれらに必要な連携等の推進

データベース統合化の国際的標準化とその実装を目指すワークショップBioHackathonや、統合化推進プログラム課題との連携によるデータベース統合化支援のためのワークショップSPARQLthonなどの開催、実務者が情報交換しながら開発を進める機会を設けるとともに、必要なツールの開発、補完的なDBのRDF化など

③データの統合解析環境、質問応答システム、日本語コンテンツの充実

リファレンスとなる遺伝子発現データのRDF化や、大規模塩基配列検索技術を利用した遺伝子機能解析実験支援ソフトウェア、解析ワークフローを共有するシステム等の開発、RDFデータへ検索をかける言語であるSPARQLを自然文から生成するシステムなどの、RDFデータに問い合わせをかけ答えを得る質問応答システムの開発、データベース利用方法等の普及のためのチュートリアル動画の作成、日本語コンテンツの作成等

④統合的運用を目指した効率的で安定したDBの分散運用の実現

各DBサービスの運用の効率化、UniProt/EBI/NCBI/DDBJなどの大規模公共RDFデータ提供機関との分散的な連携など

（報告書p58参照）

3. 基盤技術開発の推進 (4) 研究実施 (セマンティックウエブ技術について-1-)

ゴール：次世代生命科学データベースの実現

・次世代生命科学データベース＝データ駆動型サイエンスを実現するデータベース

・データ駆動型サイエンスにおいては、新規データ生成も必要ながら、膨大に蓄積されたデータを効率的・効果的に再利用する必要がある
→データインフラの整備

- そのためには、データのセマンティクス（データの意味）を扱うことが不可欠

- また、データ処理の大幅な省力化も必要

（報告書p58参照）

3. 基盤技術開発の推進 (4) 研究実施 (セマンティックウエブ技術について-2-)

RDFを採用する利点

- フォーマットが共通になる
  - ☑ 統合に有利、システム構築のコスト削減
- データの意味が明確になる
  - これまでの大半のデータベースは、意味が不明確
  - 従って人の介在なしにデータをつなげることは事実上不可能
  - ☑ 統合に有利、データ処理の自動化によるコスト削減
- W3C標準
  - ☑ 標準規格があるので、勝手にルールが変更されないため、統合に有利
- 様々な分野で、また国際的にも、利用されている
  - ☑ 統合したデータ活用の可能性が著しく高まる
  - ☑ 国際連携により競争力を維持できる

（報告書p58参照）

3. 基盤技術開発の推進 (4) 研究実施 (セマンティックウエブ技術について)

○RDFの利点　国際的な潮流

2001 セマンティックウェブの提唱
ティム・バーナーズ・リー

2006 **UniProt**

2008 **BIO2RDF**

2011 **PDBj**

2013 **TOGO GENOME**

2014 **EMBLE-EBI**
PubChem　MESH

2015 **RDF portal**
Ensemble

- ✓ 主要な国際的DBが次々とRDF化を行っている。
- ✓ 1000万人規模の大規模医療情報データベース構築を目指すCDISC:（Clinical Data Interchange Standards Consortium）が、RDFバージョンの仕様策定を行っている。
  ※PMDA（Pharmaceuticals and Medical Devices Agency)は、CDISCによる申請の義務化を行っている。

↓

将来、RDF形式で大規模な医療情報が利用できる可能性がある。

（報告書p65-83参照）

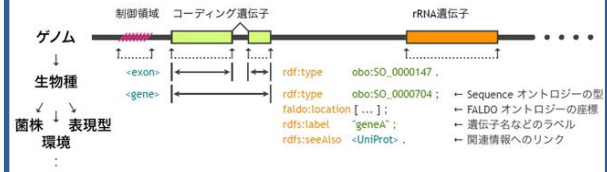
3. 基盤技術開発の推進 (5) 研究実施 (研究成果-1-)

データベース統合化の実現のために重要な基盤技術の開発・実装を目的とした研究開発を実施

①データベースのRDF化とRDFによる統合化のための基盤技術開発
各種機能開発・利用環境整備、リファレンスデータセットの整備、新規分野データ活用技術の開発、Keyとなるデータを付与しながらのデータベースのRDF化、ヒト関連情報を安全に共有・取り扱うためのセキュリティの検討等

TogoGenome
- RDFによるゲノム情報の統合DB

【成果】
- ✓ 生命科学の白地図であるゲノムにモノをマッピングし、IDと型と座標をつける。
- ✓ ゲノムを中心にあらゆる情報をRDFで足しこんでゆき、関連情報にリンクする。
- ✓ 既存の知識階層（オントロジー）を利用した、絞り込み検索や、高速な塩基配列検索、ID検索などさまざまな検索を行なうことができる。システム内で利用しているデータセットや可視化コンポーネント（TogoStanza）は、相互利用できる。



GA4GH - Beacon Project

機微情報であるヒトデータの利用にはデータ利用申請が必要である場合も多いが、入手して初めて利用したいデータではないことがわかるのでは非効率である。Beaconは、研究者が入手したい変異のあるゲノムデータの有無を、各国のどのゲノムセンターが持っているか一括サーチできるデータ共有支援ツールである。

例：12番染色体の34567番目の塩基がAのデータはあるか → yes or no

【成果】NBDCヒトデータベースのオープンデータのうち、変異情報を持つものをBeacon化 Beacon Network (Beaconサーバーの統合検索サイト)に登録。

【今後の計画】ヒトデータベースのオープンデータに適応可能なデータが追加され次第、順次追加する。

制限公開データについては制度／セキュリティ面について検討中。

※GA4GHとは
ヒトに関わるデータのような機微データについて、共有・流通の国際標準化動向把握の観点から、ヒトゲノムデータ等の共有に関する国際的な共通枠組み検討の取組。各国のもつヒトデータの利用申請前に、求める配列情報の有無の検索を可能とするBeaconプロジェクト等、具体的取り組みが実施されている。
42か国446機関参加(平成28年11月時点)
NIH（米）、EBI（欧）、BGI（中国）などの公的機関、医学系大学・病院・研究所、製薬企業、IT系企業などが参加している。

Global Alliance for Genomics & Health

（報告書p65-83参照）

3. 基盤技術開発の推進 (5) 研究実施 (研究成果-2-)

データベース統合化の実現のために重要な基盤技術の開発・実装を目的とした研究開発を実施

②国内外のDBについての統合化支援、およびそれらに必要な連携等の推進
データベース統合化の国際的標準化とその実装を目指すワークショップBioHackathon（毎年度開催）や、統合化推進プログラム課題との連携によるデータベース統合化支援のためのワークショップSPARQLthon（毎月開催）などの開催、実務者が情報交換しながら開発を進める機会を設けるとともに、必要なツールの開発、補完的なDBのRDF化など

BioHachathon
- ✓ 生成されたデータを統合的に利用するために必要な標準化の試みとして，ゲノムアノテーションにおける位置情報の共通オントロジーFeature Annotation Location Description Ontology（FALDO）整備や、糖鎖構造の標準RDFフォーマットと糖鎖オントロジーの仕様策定等を実施。
- ✓ ヒトゲノム情報の変異を含むRDFモデルとオントロジーの整備、大規模なゲノム情報のファイルをSPARQLで検索するミドルウェアの開発を実施。
- ✓ RDFデータを活用した研究、とくに機械学習や人工知能への応用を今後進めていくにあたり必要とされる開発を実施。

RDFサミット
- ✓ BioHackathonで形成した人的ネットワークを生かし、ゲノムRDF規格化の会議開催に繋げた。
- ✓ DBCLSとEnsemblで、ゲノム配列の共通RDFモデルを考案し、採用した。
- ✓ GA4GHで議論が進められているRGG（Reference Genome Graph）を用いたゲノム情報をRDFで格納することで議論がまとまった。

RDF化ガイドライン/RDFポータル
RDF化ガイドラインを策定、それに準拠したRDF化DBを集約し、アクセスできるポータルサイト「RDFポータル」を実現した。

【成果】
- ✓ RDF化したDBを NBDC RDF Portal から公開（15DB）
  - RefEx FANTOM5 RDF (発現)
  - NBDC NikkajiRDF (化合物)
  - wwPDB/RDF (タンパク質立体構造)
  - MBGD RDF (微生物ゲノム)
  - GlycoEpitope (糖鎖)
  - Linked ICGC (がん変異データ)など

【今後の計画】
- ✓ 新規DBの追加とデータ
- ✓ 更新システムの構築

（報告書p65-83参照）

3. 基盤技術開発の推進 (5) 研究実施 (研究成果-3-)

データベース統合化の実現のために重要な基盤技術の開発・実装を目的とした研究開発を実施

③データの統合解析環境、質問応答システム、日本語コンテンツの充実
本項目は、エンドユーザー向けのツール・インターフェイスの研究開発となっており、データ統合の直接的な技術開発ではないが、関連DBの利用の活性化・人材育成につながることを目的として実施している。

リファレンスとなる遺伝子発現データのRDF化や、大規模塩基配列検索技術を利用した遺伝子機能解析実験支援ソフトウェア、解析ワークフローを共有するシステム等の開発、RDFデータへ検索をかける言語であるSPARQLを自然文から生成するシステムなどの、RDFデータに問い合わせをかけ答えを得る質問応答システムの開発、データベース利用方法等の普及のためのチュートリアル動画の作成、日本語コンテンツの作成等

〇NGS解析ワークフロー（統合DB活用環境）
NGSデータを解析するにあたって、以下のような問題がある
・コマンドラインのツールが多い　・環境によってセットアップ方法が異なる
・膨大な計算資源が必要　・ソフトウェアのバージョンが異なる場合の再現性
【成果】
一般的な解析ワークフローをパッケージとして提供した
対象：RNA-seq, ChIP-seq, BS-seq, Variant calling
Dockerファイルをダウンロードすることで、ローカル環境やクラウド環境で実行可能。
【今後の計画】
メタゲノム、非モデル生物等の解析ワークフローの構築
ワークフローを利用したDRAデータの再解析・提供

〇日本語文献、画像、動画コンテンツの拡充
・100～150本程度/年の「新着論文レビュー」*および、12本程度/年の「ライフサイエンス融合領域ビュー」*の掲載を実施。毎月のユニーク訪問者数は約3万件。
・150本程度/年の動画を公開。生命科学分野の有用なデータベースやツールの使い方を動画で紹介する「統合TV」では、新たに150本程度/年の動画を公開した。毎月のユニークな訪問者は1万人を越えている。
※「新着論文レビュー」：（トップジャーナルに掲載された日本人を著者とする生命科学分野の論文について，論文の著者自身の執筆による日本語の解説
※「ライフサイエンス融合領域レビュー」：日本分子生物学会，日本蛋白質科学会，日本細胞生物学会，日本植物生理学会との協力のもと，生命科学において注目される分野・学問領域における最新の研究成果について，第一線の研究者の執筆による日本語の総説

自然言語処理に基づく質問応答システムの開発
・PubDictionary 、PubAnnotation
文献アノテーション用辞書の収集プラットフォームと、
文献アノテーション収集プラットフォーム
・LODQA
SPARQL検索を基礎とした質問応答システム
【成果】
世界的な生命科学オントロジーのレポジトリBioPortalを活用して、辞書と文献アノテーションはPubDictionariesとPubAnnotationという独自のプラットフォームを開発した。一部のアノテーションはRDF化。また、RDFデータの活用のため必須的なSPARQLクエリ言語を使いこなせなくても検索できるよう、自然文からSPARQLクエリを自動生成するシステムであるLODQAを開発した。
http://lodqa.org
【今後の計画】
アノテーションのRDF化およびクオリティコントロール
既存のRDFデータセット、PubAnnotationを利用した
一般ユーザ向けの質問応答システムインターフェイスの提供

④統合的運用を目指した効率的で安定したDBの分散運用の実現
セマンティックウエブ技術に基づく各DBサービスの運用、分散DB環境に基づいた統合的な検索インターフェースの開発、安定的な運用と　ユーザ対応等を実施している。

（報告書p6,p84,p87-91参照）

4. 統合化の推進（統合化推進プログラム）（1）目的

生命科学分野のデータベースを、生物種や個々の目的やプロジェクトを超えて幅広い統合化を実現することを目指す。その前段階としては、生物種別、分野別、目的別またはデータ種類別などのデータベース統合化を実現する。

背景

統合データベース タスクフォース報告書　H21.5 CSTP ライフサイエンスPT

4．体制整備（3）「統合データベースセンター（仮称）」の整備②求められる機能
具体的には、データベース統合に必要な調査、データベースの統合に必要な標準化、システムの構築・維持・管理、ポータルサイトの構築、データベースの受入れ・管理・更新、データベースの品質管理、各省等のデータベースとのネットワークの構築、海外との連携等の実務機能に加えて、データベースの統合化や高度な検索等、統合的利用のための技術開発（インデックス、辞書、データフォーマットなどの構築）の機能とする。

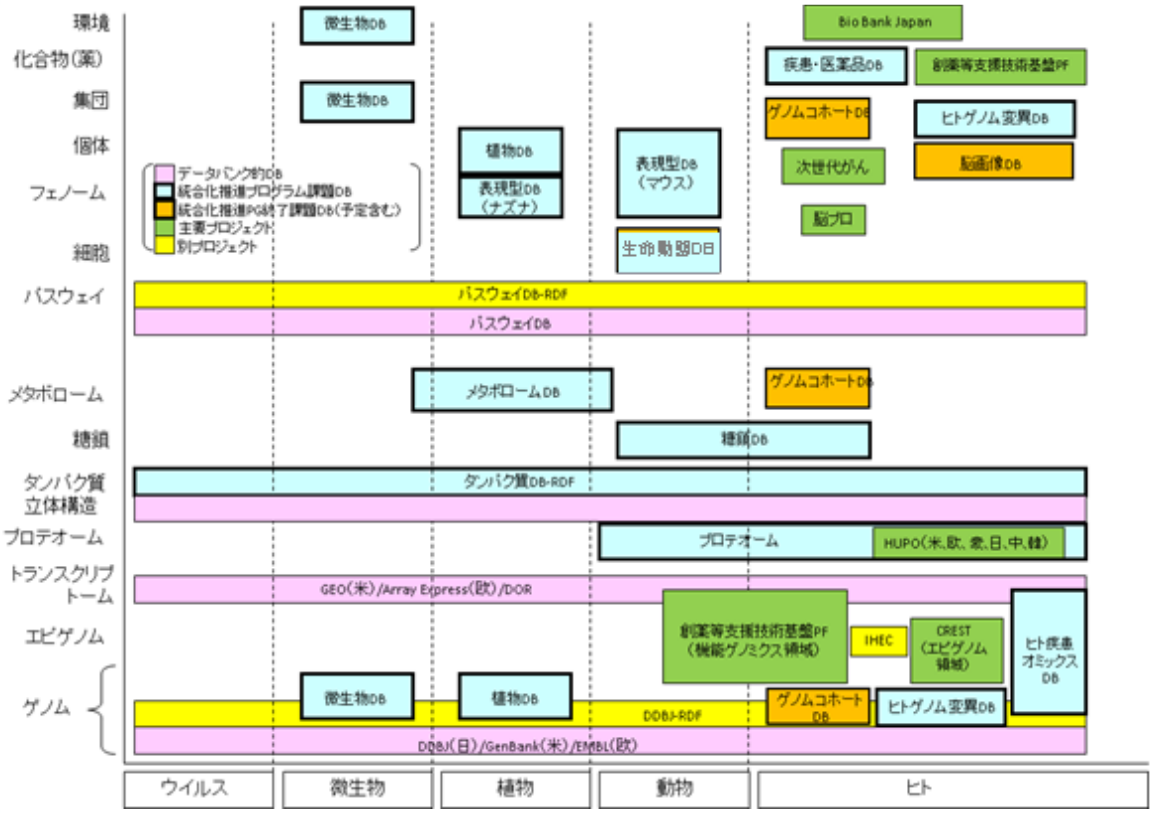
⇒ JSTライフサイエンス分野統合データベースセンター制度検討ワーキンググループにて推進方法を検討
当初は、統合化推進プログラムの実施により、それぞれの分野において、日本を代表するとともに、中核、拠点となる統合データベースが構築されることを目指した。網羅性があること、コミュニティの了承等があること、コラボに近い形でインタラクティブに実施すること、プロジェクト終了後のDBの維持・メンテナンスについて見通しがあること等の条件を付すこととした。

また、平成26年度以降はさらに、データがより多くの分野の研究者、開発者、技術者に簡便に利活用できるようにして、データの価値を最大化することを目指した。基盤技術開発チームと協働で、データのRDF化にも取り組んだ

（報告書p84-85参照）

4. 統合化の推進（統合化推進プログラム）（2）成果概要

データベース俯瞰図と統合化推進プログラムの対象データベース



- ✓ 散在していたライフサイエンス分野の研究データが、日本の中核となるデータベースへ統合
- ✓ 微生物、植物、ヒト等の生物種別、さらにゲノム、プロテオーム、メタボローム、フェノームなどのオミクス単位統合
- ✓ データ形式はRDF規格を採用

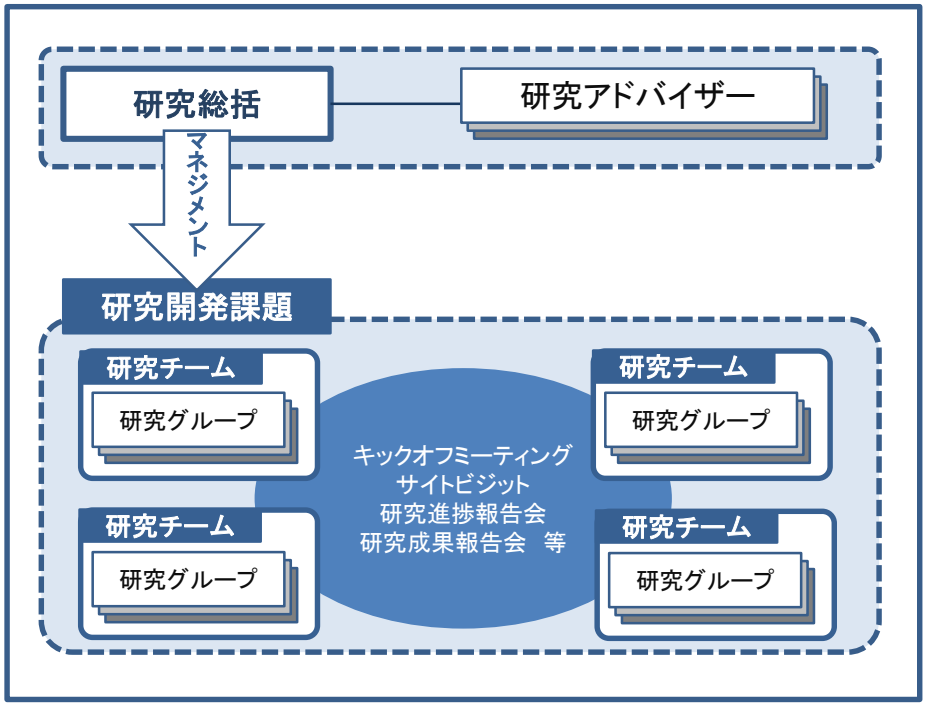
分野を超えた統合に向けての流れを作り出すことに繋がった

プログラムの対象は　統合化推進プログラムDB
統合化推進プログラム終了課題DB

（報告書p85-91参照）

4. 統合化の推進（統合化推進プログラム）(3) 推進体制等

研究推進体制



○研究総括・領域アドバイザーによるマネジメント
年に一度以上、キックオフ・ミーティングもしくは研究進捗報告会もしくはサイトビジットを開催し、研究開発の進捗状況を研究代表者が報告し、研究総括、研究アドバイザーと意見交換する機会を設けた。

研究開発費、研究開発期間

採択年度	研究開発費（直接経費のみ）	研究開発期間
平成23、24年度	3千万円～7千万円／年（研究期間総額9千万円～2.1億円）	3年以内
平成26、27年度	3千万円～5千万円／年（全研究開発期間総額9千万円～1.5億円）	

（報告書p92-93参照）

4. 統合化の推進（統合化推進プログラム）（3）研究開発課題一覧

研究代表者	所属・役職	研究開発課題（採択年度）
岩坪　威	東京大学 大学院医学系研究科・教授	ヒト脳疾患画像データベース統合化研究（H23）
金谷　重彦 有田　正規	奈良先端科学技術大学院大学 情報科学研究科・教授 理化学研究所　環境資源科学研究センター・チームリーダー	メタボローム・データベースの開発(H23)/ 生物種メタボロームモデル・データベースの構築（H26）
金久　實	京都大学 化学研究所・特任教授	ゲノム情報に基づく疾患・医薬品・環境物質データの統合(H23)/ ゲノムとフェノタイプ・疾患・医薬品の統合データベース(H26)
黒川　顯	東京工業大学　大学院　生命理工学研究科	ゲノム・メタゲノム情報を基盤とした微生物DBの統合（H23）/ ゲノム・メタゲノム情報統合による微生物DBの超高度化推進（H26）
田畑　哲之	かずさDNA研究所 ・所長	ゲノム情報に基づく植物データベースの統合（H23）/ 植物ゲノム情報活用のための統合研究基盤の構築（H26）
徳永　勝士	東京大学 大学院医学系研究科・教授	ヒトゲノムバリエーションデータベースの開発(H23)/ 個別化医療に向けたヒトゲノムバリエーションデータベース(H26)
豊田　哲郎 桝屋　啓志	理化学研究所　情報基盤センター・統合データベース特別ユニットリーダー 理化学研究所　バイオリソースセンター・ユニットリーダー	生命と環境のフェノーム統合データベース(H23)/ 生命と環境のフェノーム統合データベース(H26)
中村　春木	大阪大学 蛋白質研究所・所長／教授	蛋白質構造データバンクの国際的な構築と統合化(H23)/ 蛋白質構造データバンクの高度化と統合的運用(H26)
成松　久	産業技術総合研究所　糖鎖医工学研究センター・センター長	糖鎖統合データベースと研究支援ツールの開発(H23)/ 糖鎖統合データベースおよび国際糖鎖構造リポジトリの開発(H26)
松田　文彦	京都大学　大学院医学研究科　附属ゲノム医学センター センター長・教授	大規模ゲノム疫学研究の統合情報基盤の構築(H23)
大浪　修一	理化学研究所 生命システム研究センター・チームリーダー	生命動態システム科学のデータベースの統合化(H24)/ 生命動態情報と細胞・発生画像情報の統合データベース(H27)
菅野　純夫	東京大学 大学院新領域創成科学研究科・教授	疾患ヒトゲノム変異の生物学的機能注釈を目指した多階層オミクスデータの統合(H26)
石濱　泰	京都大学 大学院薬学研究科・教授	プロテオーム統合データベースの構築(H27)

※所属は採択時のもの

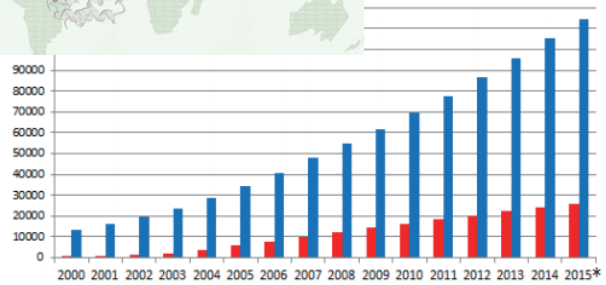
（報告書p101-120参照）

4. 統合化の推進（統合化推進プログラム）（5）顕著な成果例－1－

○タンパク質構造データバンク（大阪大学　中村 春木）

課題概要

PDB（タンパク質構造データバンク）とBMRB（NMR実験情報データバンク）を日米欧の国際協力により継続的に構築・公開し、日米欧3極で構成するwwPDB のメンバーとして、主にアジア地区からのタンパク質立体構造の登録データについて登録処理を行い、さらなる高度化を目指す研究開発課題。
生命科学研究の世界的な基盤データベースとしての地位を獲得している。



PDBのデータ登録数

・データ駆動型研究の推進に寄与するため、データの質の更なる向上（データ品質自動検証の精度向上）を図った。
・ PDBjによる登録作業数が世界全体の約22%を占めている。

○糖鎖統合データベース（産業技術総合研究所　　成松　久）

課題概要

国際協働体制を拡大し、国際糖鎖構造データリポジトリシステムを開発するとともに、全糖鎖構造データの標準化作業を進める。並行して糖鎖関連データベースの開発と標準化対応開発を進めることによって、他分野データベースとのより高度な統合を目指す。

・これまで、日米欧の糖鎖分野を代表する研究者との議論を主導・調整してきた結果、日本発の国際的な糖鎖構造データリポジトリシステムとして“Glycan Repository”を開発した。現在4万件以上の構造情報が登録されている。
・さらに、論文投稿時における本レポジトリへの糖鎖構造情報の登録について、MIRAGE（糖鎖実験についての論文を執筆する際の標準）を推進している団体の支援を既に得ており、引き続き、糖鎖分野における世界標準を目指していくこととしている。

4. 統合化の推進（統合化推進プログラム）（5）顕著な成果例−2−

○プロテオーム統合データベース（京都大学　石濱 泰）

課題概要

国内外に散在している種々のプロテオーム情報を標準化・統合・一元管理し、多彩な生物種・翻訳後修飾・絶対発現量も含めた横断的統合プロテオームデータベースの開発を目指す研究開発課題。

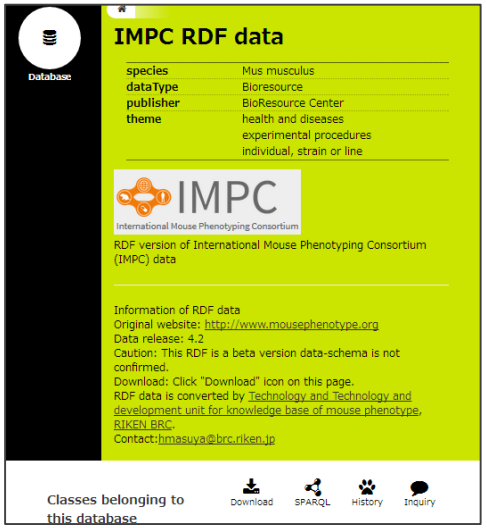
・デポジットされたデータは、様々な実験手法・解析手法で得られており単純に統合解析できないため、統合解析するための処理方法を決定した。
・有力ジャーナルの推奨レポジトリとなっているプロテオームの国際コンソーシアム（ProteomeXchange）にアジア地域初として加入し、プロテオームのレポジトリを公開。

○表現型データベース（理化学研究所　桝屋 啓志）

課題概要

・ヒトやモデル生物の表現型情報を収集し、生物種や階層性の垣根を超えて標準化・統合化・体系化して公開し、横断的な解析研究を可能とする事を目的とする研究開発課題。

・メダカ、マウス、その他の細胞リソース等の表現型データをRDF化して公開した。
・国際マウス表現型コンソーシアム(IMPC) が保有する多量のマウス網羅的表現型データもRDF化して公開した。
・多生物種の表現型データベース国際横断プロジェクト（Monarch initiative）へ、本研究開発課題で収集しているデータを提供することについて合意した。
・IRUDプロジェクト（未診断疾患イニシアチブ、Initiative on Rare and Undiagnosed Diseases）へ、本研究開発課題で収集しているデータを提供することについて合意した。ヒト疾患研究でデータベースが利活用されることが期待できる活動を進めている。



IMPC保有データをRDF化し、RIKENメタDBで公開

（報告書p101-120参照）

4. 統合化の推進（統合化推進プログラム）（5）顕著な成果例−3−

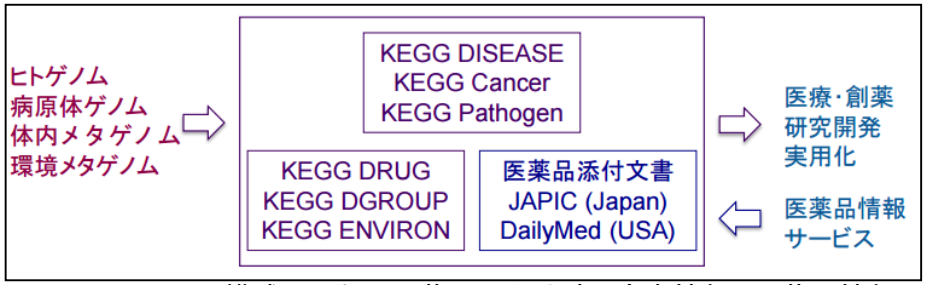
○ゲノム・疾患・医薬品の統合データベース（京都大学　金久 實）

課題概要

個々の遺伝子や、複数の遺伝子から構成された機能モジュール、さらには遺伝子、タンパク質、環境因子、医薬品等から構成された相互作用ユニットに関する知識をデータベース化し、ゲノム情報を有効利用するための統合データベースリソースを開発する研究開発課題。

・遺伝子、タンパク質、また代謝やシグナル伝達などの分子間ネットワークに関する情報を統合したデータベースKEGG（Kyoto Encyclopedia of Genes and Genomes）のうち、ヒトゲノム、病原体ゲノム、様々なメタゲノムなどのシークエンス解読と有効利用を促進する統合リソースKEGG MEDICUSを開発し、提供。

・KEGG MEDICUSの月間平均アクセス数は20万件以上（IPアドレス別）。研究者のみならず、一般市民からも広く利用されていると考えられる。



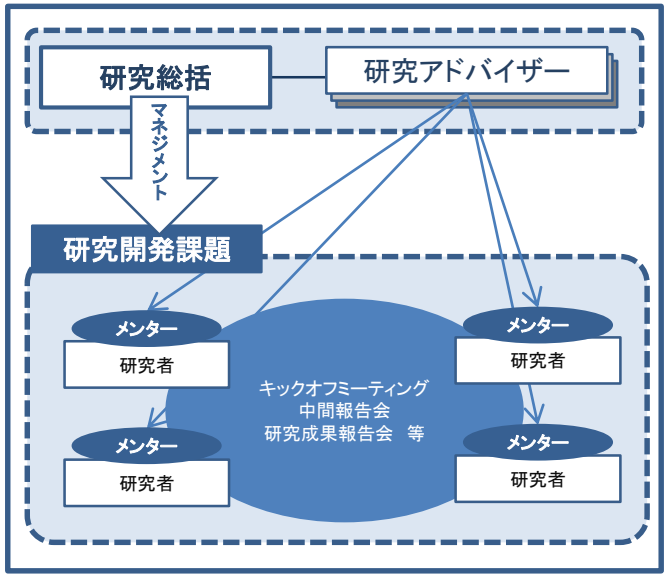
KEGG MEDICUSの構成。日米の医薬品添付文書、疾患情報、医薬品情報を整備し、ゲノム、メタゲノム情報と統合して提供している。

（報告書p121-123, p132-133参照）

5. 統合化の推進（統合データ解析トライアル）（1）目的、および推進体制

○目的

統合化推進プログラムで統合されたデータベースについて、データ解析を行うツールなどの開発を推進するため、また若手の研究人材がこうした研究開発に取り組む契機とするため、統合化推進プログラムの一環で実施した。



○研究総括・領域アドバイザーによるマネジメント

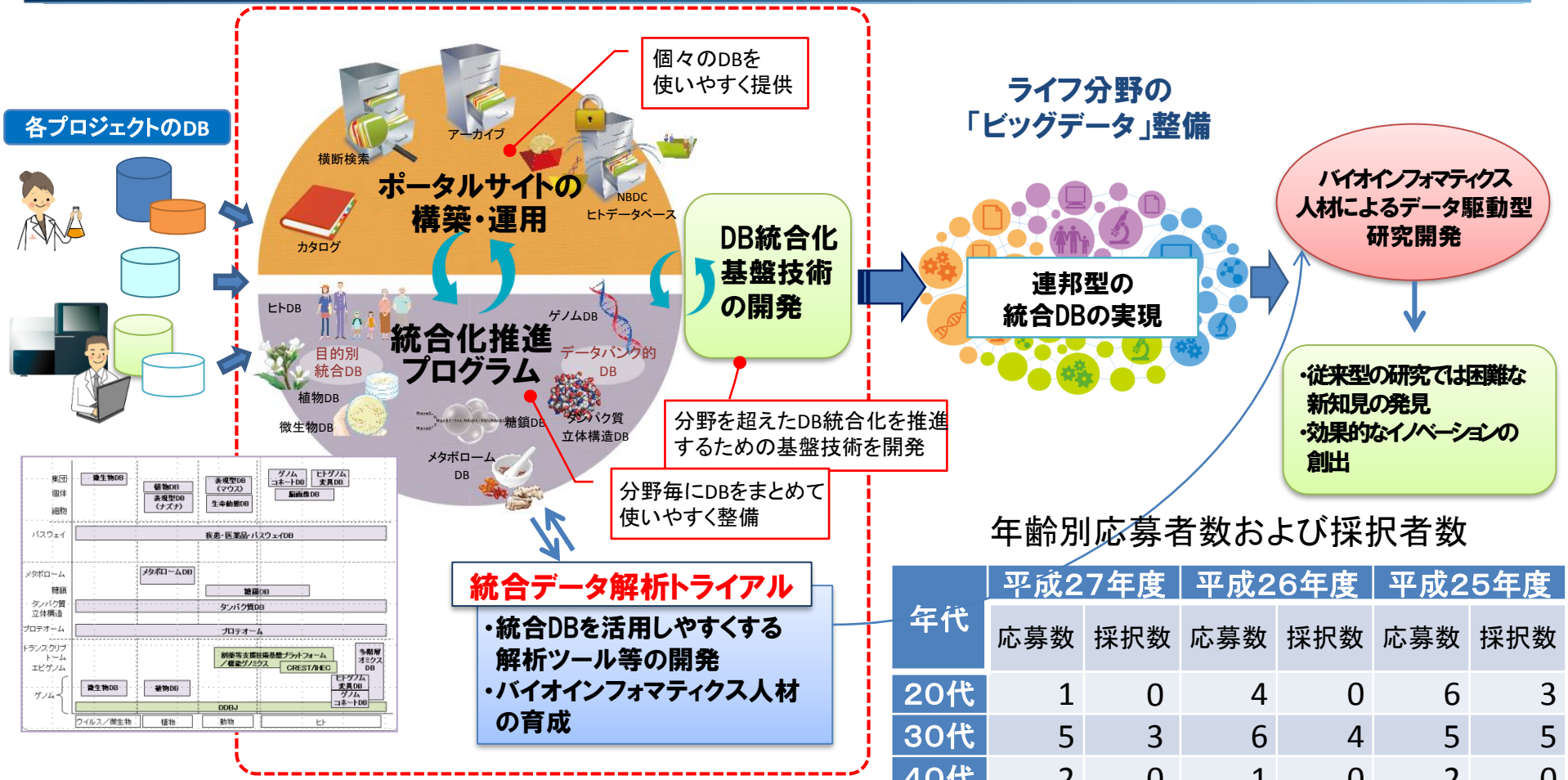
研究進捗を把握し、助言を行うため、キックオフミーティング、中間報告会を開催した。また、成果を広く一般に周知し、一般からの意見を今後の研究開発にフィードバックするため、研究成果報告会を開催した。

若手人材育成の観点から、各研究実開発課題には、1課題ずつ1名メンターを配置し、個別に助言等を行った。また、バイオインフォマティックス分野の研究者が参入しやすいように具体的な「解決したい課題」の提示も行った。（平成27年度）。

年度	研究開発費（直接経費のみ）	研究開発期間
平成25年度	80万円以内	4ヶ月以内（平成25年9月1日～12月31日）
平成26年度	80万円以内	6ヶ月以内（平成26年9月1日～平成27年2月28日）
平成27年度	100万円以内 (研究チームを編成する場合、200万円以内)	10.5ヶ月以内（平成27年5月15日～平成28年3月31日）

（報告書p84,p121-126,p150参照）

5. 統合化の推進（統合データ解析トライアル）（2）位置づけと応募採択状況



※継続的な実施が必要

- 戦略的重点分野、新興分野DBの統合化
- 新分野のオントロジーの開発　等

年齢別応募者数および採択者数

年代	平成27年度		平成26年度		平成25年度	
	応募数	採択数	応募数	採択数	応募数	採択数
20代	1	0	4	0	6	3
30代	5	3	6	4	5	5
40代	2	0	1	0	2	0
合計	8	3	11	4	13	8

想定通り若手中心の採択となった。

様々な工夫を行ったものの、応募者の増加が認められなかった。

（報告書p127-131参照）

5. 統合化の推進（統合データ解析トライアル）（2）研究開発課題一覧

採択年度	研究開発課題	研究代表者(所属・役職)
H25	大規模なタンパク質データ解析のための高速な局所配列特徴抽出法の開発	蝦名 鉄平 (理化学研究所脳科学総合研究センター ・研究員)
	マルチオミクスデータを用いたゲノム規模代謝モデリングのためのネットワーク解析システムの開発	西田 孝三 (理化学研究所生命システム研究センター・テクニカルスタッフ)
	KNApSAcKを用いた植物の効能メカニズム解明のための基盤構築	西村 陽介 (京都大学化学研究所・大学院生)
	機械学習を用いたタンパク質－リガンド結合部位予測ツールの自動生成パイプラインの開発	番野 雅城 (東京大学大学院農学生命科学研究科・大学院生)
	植物代謝物プロファイリングデータベースAtMetExpressの開発とオミックスデータ統合化の推進	福島 敦史 (理化学研究所環境資源科学研究センター・研究員)
	共起関係解析によるタンパク質の機能モジュール探索法の開発	藤井 聡 (九州工業大学大学院情報工学研究院・助教)
	タンパク質－糖鎖間の糖鎖結合部位の解明のためのツール改良及び解析	細田 正恵(創価大学大学院工学研究科・大学院生)
	MicrobeDB.jpデータを用いたメタゲノム解析Webアプリケーションの開発	森 宙史(東京工業大学大学院生命理工学研究科・助教)
H26	PDBjタンパク質をゲノムにマップしたpdbBAMの作成	城田 松之(東北大学大学院医学系研究科・助教)
	RDFストア間データ連結フレームワークの開発およびオーソログ解析への適用	千葉 啓和 (自然科学研究機構基礎生物学研究所ゲノム情報研究室・研究員)
	生化学反応ネットワーク統合解析環境の拡張	西田 孝三 (理化学研究所生命システム研究センター・テクニカルスタッフ)
	HLA遺伝子完全配列決定パイプラインの構築	細道 一善 (情報システム・研究機構国立遺伝学研究所人類遺伝研究部門・外来研究員)
H27	ChIP-seq SRAの統合的可視化とバイオデータベースとの連携	沖 真弥 (九州大学大学院医学研究院・助教)
	配合生薬の横断検索のためのソフトウエアツールの開発	桂樹 哲雄 (豊橋技術科学大学大学院工学研究科・助教)
	高精度全電子計算に基づくレクチン－糖鎖間相互作用解析	中野 祥吾(*) (静岡県立大学食品栄養科学部・助教)

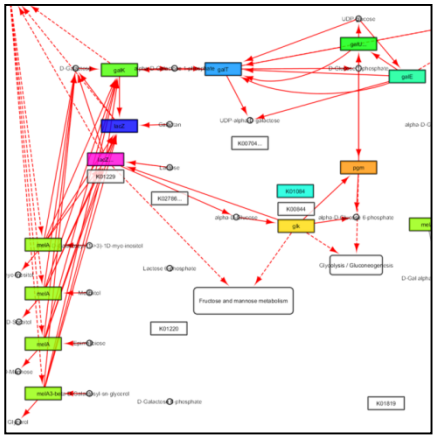
（報告書p134-149参照）

5. 統合化の推進（統合データ解析トライアル） (3) 主要な成果例

○西田 孝三(理化学研究所生命システム研究センター・テクニカルスタッフ)

KEGGScape
http://apps.cytoscape.org/apps/keggscape

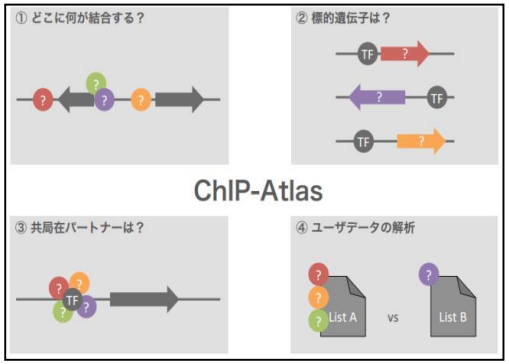
遺伝子ネットワーク分析などに広く用いられているネットワーク可視化プラットフォーム「Cytoscape」のアドイン。生体分子の相互作用ネットワークの定量的な測定データをKEGGScapeを用いてKEGGのパスウェイデータと比較する事で、ゲノム規模の代謝モデリング技術による薬剤応答、耐性機構予測に役立てる事ができる。



○沖 真弥(九州大学大学院医学研究院・助教)

ChIP-Atlas
http://chip-atlas.org/

SRAのChIP-Seqデータを、特別な技術や環境を要さず利活用できるWebサービス。個別に解析する事なく、塩基配列中のどこに何の因子が、何と一緒に結合し、どのような機能を持っているかを容易に閲覧できる。



（参考）国際連携まとめ−1−

ヒトに由来するデータ等の取り扱いへの対応

ヒトデータベースのうち、アクセス制限データベース（JGA)でも連携を実施

NBDC DDBJ

NBDCヒトデータベース（英語化）
（DDBJと連携して運用）

JGA
(Japanese Genotype-Phenotype Archive)

相互インデックスのためメタデータの概要情報を交換

DDBJ
INSDC
NCBI EMBL-EBI
International Nucleotide Sequence Database Collaboration

dbGaP
（Database of Genotype and Phenotype)

EGA
European Genome-phenome Archive

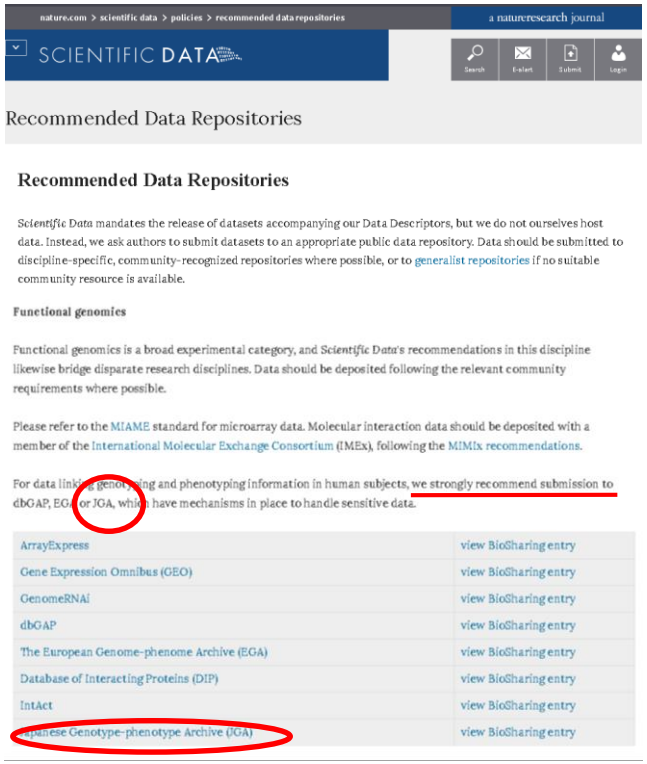
GA4GHへの加入、および国際プロジェクトへの参画

Global Alliance for Genomics & Health （平成26年5月 加入）

Beacon Project

Beacon Network （平成28年9月参画）

NBDCヒトデータベース（JGA)が、
Scientific Data誌の推奨レポジトリに掲載



Human Phenotype Onthology(HPO)の日本語化

人間の病気の原因となる表現型について標準化された用語を提供するHPOの日本語化への取り組み。（平成27年度）

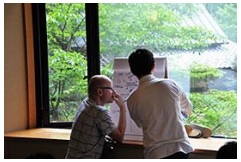
（参考）国際連携まとめ−2−

RDFによる統合に向けた国際連携

BioHachathon, RDFサミットの開催等による、RDF国際準化の推進

BioHachathon

✓ 共通オントロジーの整備等(平成24年9月）



BioHackathon 2016 in Yamagata

RDFサミット

✓ DBCLSとEnsemblで、ゲノム配列の共通RDFモデルを考案し、採用(平成26年11月）

✓ GA4GHで議論が進められているRGG（Reference Genome Graph)を用いたゲノム情報をRDFで格納することで議論がまとまった（平成28年2月）

データベース統合化のための基盤技術の開発

BioNLP（Natula Language　Processing) communityとの協力により、PubAnnotationを開発

（平成28年1月）

ポータルサイトの構築・運用

生命科学データベースカタログに関し、海外の同様のサイト（biosharing.org ）との連携により、データベース情報の拡充および、データベース情報に関するメタデータの国際標準化を狙う。

（平成28年9月～）

バイオ関連データベースの統合化の推進

主要な分野における中核・拠点となる統合データベースの構築

国際連携による、タンパク質立体構造データバンクを運用

プロテオームの国際コンソーシアムに加入 (平成28年6月）

日本発の国際糖鎖構造データリポジトリの開発（平成27年3月）

IMPCのデータをRDF化して公開（平成28年9月）

（報告書p33-34,p151)参照）

6. その他（人材育成）（1）戦略立案機能（バイオインフォマティクス人材育成）

不足しているバイオインフォマティックス人材の育成、キャリアパス形成などを誘引する施策として企画を立案・予算要求を行った。

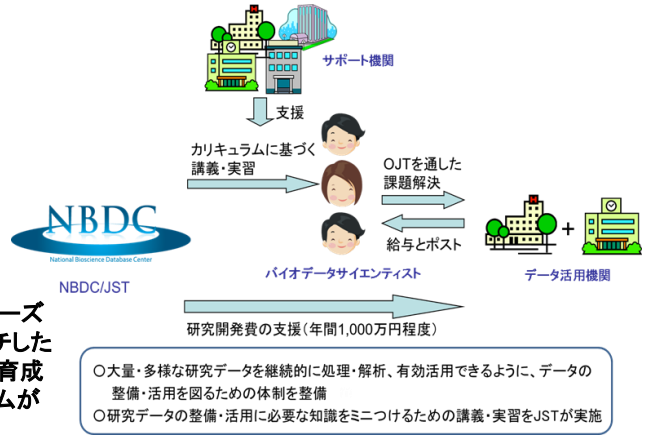
課題：

①大学等における人員枠の問題により、新分野であるバイオインフォマティクスに若い人材を取り込めていない。

②産業界においても、バイオインフォマティクス人材は必要といいつつ、実験系研究者が多数を占める。

③安定した就職先が無く、バイオインフォマティクスを学んだ人材が他業界（プログラミング系、情報系）などに進んでいる。

人材ニーズにマッチした新たな育成システムが必要



期待される効果：

- ライフサイエンス研究現場で、最新の解析技術を利用した大量・多様なデータを有効活用。
- データ共合的な利用を意識した形での効率的なデータベース作成。
- 育成された人材が、各研究現場で活躍することにより、認知度が向上し、社会的地位を確立。人材基盤の強化。

概要

大量・多様な研究データを継続的に処理・解析、有効活用できるように、データの整備・活用を図るための体制を整備。さらに、研究データの整備・活用に必要な知識を身につけるための講義・実習をJSTが実施。

⇒平成25～28年度、
予算要求をしてきたものの認可には至らず。本施策は実行できていない。
検討した速習カリキュラムに基づく講義・実習について実施した。*（講習会については、次ページ参照）*

（報告書p151-157参照）

6. その他（人材育成）

バイオインフォマティクス人材が不足している中、特に喫緊の課題となっている次世代シーケ ンサー（NGS）の取扱いに関する基礎技術や個別解析技術について、2週間連続の集中講習会を実施。
毎年の受講生アンケート等を元に、講義の改良を加えて平成28年度まで3回実施している。

○H26年度NGS速習コース講習会（参加86名）

平成25年度にNBDCで策定した「速習コース」用カリキュラムに基づき実施。
「体系的に学べた点が良かった」等好評。
受講者ニーズを取り入れ、次回以降、座学（講義形式）からパソコンを用いた実習にシフト。

○H27年度ハンズオン講習会（参加109名）

「講義は今後役立つ内容である」等好評。
受講１年後の追跡調査では、「受講によりスキルアップした　9割」、
「受講内容を活用している　7割」との結果。

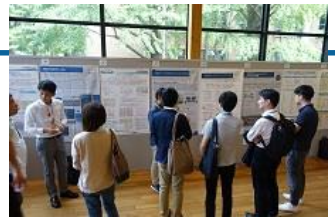
○H28年度ハンズオン講習会（参加126名）

アンケート結果：118名中113名が、
「この講義は今後の役に立つ」、と回答。
講習会の一部は、東京大学農学生命情報科学特論I・IIの
単位取得が可能なものとなっている。

（報告書p158-165参照）

6. その他（広報活動）

（１）トーゴーの日シンポジウム



NBDCではライフサイエンスのデータベースの統合に向けた活動をしていることに鑑み、平成23年度から、毎年10月5日を「トーゴーの日」とし、ライフサイエンス分野におけるデータベース統合の成果を報告するシンポジウムを開催している。（参加者　平均250名程度）

（２）講習会

①AJACS講習会
生命科学系のデータベースやツールの使い方、データベースを統合する活動を紹介する初心者向けのオーダーメイドのハンズオン講習会として実施。
開催希望機関を公募し、毎年6回程度、国内各地で開催。（参加者　のべ1,200名超）

②合同講習会
データセンター間連携の一環として、NBDC, DBCLS, PDBj,およびDDBJ と合同で、講習会を企画し、2回（平成27, 28年度）実施した。基本的なデータベース利用法を紹介する講習会として実施。（参加者　各35名程度）

（３）学会展示等

主に研究者へのNBDC活動の普及のため、関連学会において、ブースを設ける等により、NBDCの活動の紹介（主に提供しているサービスの紹介）を実施している（4～7回/年実施）。特に、分子生物学会においては、統合化推進プログラム採択課題等と合同でブースを設けるなど、NBDC全体としての広報活動も実施している。

（参考資料）

3. 基盤技術開発の推進　開発成果例-1-

主な取り組みの該当	名称	概要/URL	URL	公開開始日
i) 分散型のDB連携を目指したRDF統合化のための技術開発	Allie RDF Data	Allie のSPARQL エンドポイント。	http://data.allie.dbcls.jp/	2011/12/1
	OntoFinder/OntoFactory	データをRDF化する際に適したオントロジーの検索と推薦、マッピングをするシステム。	http://ontofinder.dbcls.jp/	2012/10/1
ii) 国際的な標準化をリードする統合化支援と分散統合DB環境の実現	TogoWiki	国内版バイオハッカソンの情報交換ならびに成果を公開するためのサービス。	http://wiki.lifesciencedb.jp/mw/	2009/8/1
	BioHackathon	最先端の研究開発者を招聘した国際的なソフトウェア開発会議の情報交換ならびに成果を公開するためのサイト。	http://www.biohackathon.org/	2011/1/1
	TogoTable	データベースIDを含む表形式のデータに対して、IDを検索キーにして複数のトリプルストアからデータベースをまたいだアノテーション情報を取得するシステム。	http://togotable.dbcls.jp/	2012/3/9
	TogoGenome	ゲノムの決定した生物種を中心に、遺伝子・生物種・環境・表現型などの情報をRDFで整備したセマンティック・ウェブによるゲノムデータベース。	http://togogenome.org/	2013/10/5
	SPARQL Builder	ユーザが欲しいデータを取得するSPARQL文を生成するサービス。	http://sparqlbuilder.org/	2014/9/26
	SPARQL support	ブラウザ上でのSPARQLクエリの記述を補助するツール。コードの補完、クエリの記憶が可能になる。	http://web.kuicr.kyoto-u.ac.jp/supp/moriya/piero/edit/sparql-support.html	2015/10/1
	D2RQ Mapper	関係データベース（RDB）を簡単にセマンティックウェブに対応させられるウェブアプリケーション。D2RQ Mapperを利用することで、現在利用中のMySQLなどのRDBのデータを効率良くRDF化できる。	http://d2rq.dbcls.jp/	2015/10/5

（参考資料）

3. 基盤技術開発の推進　開発成果例-2-

主な取り組みの該当	名称	概要/URL	URL	公開開始日
iii) 大規模データの円滑な利用のためのエンドユーザ向け利用環境構築	BodyParts3D	人体各部位の位置や形状を3次元モデルで記述したデータベースです。3Dレンダラー上のウエブAPIを使うモデルエディター(アナトモグラフィー)を使って、BodyParts3D から解剖概念を選択して自由に人体のモデル図を作成、交換でき,利用者の情報もモデル上にマップ表現できる。	http://lifesciencedb.jp/bp3d/	2007/10/5
	統合TV	生命科学分野の有用なデータベースやウェブツールの活用法を動画で紹介するウェブサイト。	http://togotv.dbcls.jp/	2007/7/19
	Gendoo （Gene, Disease Features Ontology-based Overview System）	文献情報をもとに、遺伝子、疾患について、関連する疾患、薬剤、臓器、生命現象などの特徴をキーワードでリスト表示するツール。	http://gendoo.dbcls.jp/	2008/12/12
	DBCLS galaxy	生命科学データに特化したウェブベースの対話的ツール組み合わせインタフェース。DBCLSで開発されたツール群も組み込んでいる。	http://galaxy.dbcls.jp/	2009/10/1
	ライフサイエンス 新着論文レビュー	Nature、Science、Cell などのトップジャーナルに掲載された日本人を著者とする生命科学分野の論文について、論文の著者自身の執筆による日本語によるレビューを、だれでも自由に閲覧・利用できるよう、いち早く公開するオンラインジャーナルサービス。	http://first.lifesciencedb.jp/	2010/9/1
	togo picture gallery	ライフサイエンス分野のイラストをだれでも自由に閲覧・利用できるようWeb 上にて無料で公開しているウェブサイト。	http://g86.dbcls.jp/togopic/	2011/4/1
	RefEx （Reference Expression dataset）	EST、GeneChip、CAGE、RNA-seq の4 種類の異なる手法によって得られたヒトおよびマウス、ラットにおける遺伝子発現データを並列に表示し、遺伝子発現解析を行う上で基準となるリファレンス（参照）データベースとして利用することを目的とした遺伝子発現データベース。	http://refex.dbcls.jp/	2011/10/1
	DBCLS SRA	公共データベース（SRA [NCBI]、ENA [EBI]）、DRA[DDBJ] ）に登録された「次世代シーケンサ」データについて、目的別、機器別、生物種別等、さまざまな統計情報から閲覧、比較、データのダウンロードができる目次サイト。論文からのデータの検索も可能。	http://sra.dbcls.jp/	2011/1/5
	統合遺伝子検索GGRNA	遺伝子や転写産物をさまざまなキーワードからすばやく検索し、その結果をわかりやすく提示することができる遺伝子検索エンジン。遺伝子名や各種ID、タンパク質の機能や特徴などのキーワードだけでなく、短い塩基配列やアミノ酸配列から遺伝子を高速に検索することも可能。	http://GGRNA.dbcls.jp/	2011/5/18
	ライフサイエンス 領域融合レビュー	生命科学において注目される分野・学問領域における最新の研究成果について、第一線の研究者の執筆による日本語のレビューを、だれでも自由に閲覧・利用できるよう、公開するオンラインジャーナルサービス。	http://leading.lifesciencedb.jp/	2012/9/1
	高速配列検索GGGenome	塩基配列を高速に検索するウェブサーバー。	http://GGGenome.dbcls.jp/	2012/7/4
	CRISPRdirect	CRISPR/Cas9ゲノム編集法に用いるガイドRNAを設計するためのソフトウェア。	http://crispr.dbcls.jp/	2013/12/31
	AOE (All Of gene Expression)	公共遺伝子発現データベースの目次。	http://aoe.dbcls.jp/	2014/10/1
	ChIP-Atlas	論文などで報告された ChIP-seq データを閲覧し、利活用するためのウェブサービス。データ処理の知識やスキルがない方でも簡単に利用可能。データソースは、公開 NGS データレポジトリ (NCBI, EMBL-EBI, DDBJ) に登録されたほぼ全ての ChIP-seq データ。	http://chip-atlas.org/	2015/12/3
	LODQA	自然言語による質問応答システム。自然文で書いた質問からSPARQLを自動生成しRDFデータの検索が可能。	http://lodqa.org/	2013/10/1
	PubAnnotation, TextAE	Webによる文献アノテーションのための統合環境。PubMed, PMCのオープンアクセスな文献に関しては文献データのファイルフォーマットを標準化し、文字列の絶対番地を提供することによって、異なったグループの文献アノテーションを一元的に取り扱うことができる。遺伝子名辞書などにより自動で固有表現抽出が可能。	http://pubannotation.org/	2012/4/1
	PubDictionaries	辞書データをウェブで共有し、辞書に基づくテキストアノテーションが行えるシステム。	http://pubdictionaries.org/	2015/5/11

（参考資料）

3. 基盤技術開発の推進　開発成果例-3-

主な取り組みの該当	名称	概要/URL	URL	公開開始日
iv) 既存サービスの拡充と運用	OReFiL	オンライン上に存在する多数の生命科学系の資源(データベースやソフトウェアなど)を効率的に見つけるための検索システム。	http://orefil.dbcls.jp/	2007/8/1
	Allie	MEDLINE を対象とし、出現する略字とその正規系のペアを検索するシステム。略字を入力することで、それの使われ方を一覧表示する。	http://allie.dbcls.jp/	2008/4/1
	DBCLS OpenID	一つのID で複数のサイトを認証できるシステム。各サイトで認証サービスを用意する必要がなく、サイト間のユーザ情報の集約が容易に行える。	http://openid.dbcls.jp/	2008/4/1
	TogoDB (旧)	エクセルなど表形式のデータを簡単に読み込み、DB 化し、自動的に共通のウェブ検索インタフェースを生成するシステム。	http://togodb.dbcls.jp/	2008/4/1
	TogoWS	国内外のウェブサービスを共通のAPIで利用できる仕組みと、サービス間の連携に必要なデータ形式変換機能、サービスの稼働状況の監視等を提供するシステム。	http://togows.dbcls.jp/	2008/4/1
	inMeXes	MEDLINE を対象とし、利用者が入力した文字列とマッチする表現を逐次的(1文字入力毎)に検索するシステム。論文中の英語表現を容易に検索できる。	http://docman.dbcls.jp/im/	2009/7/1
	TogoDoc (Suite)	文献情報及び論文PDF を管理し、また、特定の文献情報群に関連する論文情報を提示するシステム。TogoDoc Client と連携して文献を管理することも可能なほか、スマートフォンにも対応。	https://docman.dbcls.jp/pubmed_recom/	2009/12/1
	TogoDB	TogoDB の機能に加え，アップロードしたデータを半自動的にRDF へ変換する機能をもつシステム。	http://togodb.org/	2011/12/1
	RDF化したライフサイエンス辞書のSPARQLエンドポイント	ライフサイエンス辞書プロジェクトにより編纂された辞書をRDF化し、SPARQLで問合せ可能なサイト。	http://purl.jp/bio/10/lsd/sparql	2013/1/2
	Colil	PMC OAサブセットに含まれる論文について、本文中で引用されているPubMedに索引付けされている論文との関係を、引用文脈とともに検索可能なデータベース。	http://colil.dbcls.jp/	2013/4/6