

科学技術連携施策群の効果的・効率的な推進

補完的課題 事後評価

「生命科学データベース統合に関する調査研究」

責任機関名：大学共同利用法人情報・システム研究機構

研究代表者名：大久保 公策

(大学共同利用法人情報・システム研究機構

国立遺伝学研究所生命情報 DDBJ 研究センター)

研究期間：平成17年度～平成19年度

目次

I. 研究計画の概要

1. 研究の目的
2. 研究の重要性・緊急性
3. 研究計画
4. ミッションステートメント
5. 研究全体像
6. 研究体制
7. 研究運営委員会について

II. 経費

1. 所要経費
2. 使用区分

III. 研究成果

1. 研究成果の概要
 - (1) 研究目標と目標に対する結果
 - (2) ミッションステートメントに対する達成度
 - (3) 当初計画どおりに進捗しなかった理由
 - (4) 研究目標の妥当性について
 - (5) 研究計画・実施体制について
 - (6) 研究成果の発表状況
2. 研究成果:サブテーマ毎の詳細
 - (1) サブテーマ1(1) 生命科学一般のデータベース統合にかかわる調査研究
 - (2) サブテーマ1(2) データベースの物理的統合にかかわる技術的調査研究
 - (3) サブテーマ2 医学分野のデータベースに関する調査研究
 - (4) サブテーマ3 産業(農林水産業)データベースの統合にかかわる調査研究
 - (5) サブテーマ4 産業(農林水産業以外)データベースの統合にかかわる調査研究

IV. 実施期間終了後における取組みの継続性・発展性

V. 自己評価

1. 目標達成度
2. 情報発信
3. 研究計画・実施体制
4. 実施期間終了後における取組みの継続性・発展性

I. 研究計画の概要

- プログラム名: 科学技術連携施策群の効果的・効率的な推進
- テーマ分類: ポストゲノム—健康科学の推進—
- 課題分類名: 1 ライフサイエンス分野のデータベースの統合化に関する調査研究
- 課題名: 生命科学データベース統合に関する調査研究
- 責任機関名: 大学共同利用法人情報・システム研究機構 国立遺伝学研究所
- 研究代表者名(役職):
大久保 公策(国立遺伝学研究所生命情報 DDBJ 研究センター 教授)
- 研究実施期間: 平成 17 年 12 月～平成 20 年 3 月
- 研究総経費: 総額 261.7 百万円 (間接経費込み)

1. 研究の目的

大規模解析系を用いたプロジェクト型研究は専らデータベースとしてその成果を残す。また多様化再分化を重ねたライフサイエンスの知的集積物も膨大化の一途を辿っている。しかしながらこれらの散在するデータや集積物は一望困難であり、また有機的結合に乏しく個々の研究者による利用は必ずしも容易ではなく研究者に十分利用されているとはいえない。この現状を速やかに是正するために、我が国の知的情報基盤の設計に資する客観的な資料並びに国内主要データベースを中心として統合化に関するフュージビリティ・スタディを通じた実現性検証結果の提示及びそれらの成果を活かした制度設計やロードマップ作成に資する試案を作成することは、極めて重要度が高く緊急を要する課題である

2. 研究の重要性・緊急性

第7回バイオテクノロジー戦略会議における「ミレニアムプロジェクトの成果や、ポストゲノムプロジェクト等の成果を実用化研究に活用できるように、一元的に集約・統合したデータベースの構築が不可欠」という提言に見られるように、ライフサイエンス分野の研究の進展に照らして、本年度着手すべき緊急性・社会的重要度が高い。

統合化の対象となりうるデータベースは、関係府省、関連する独立行政法人、国立大学法人等に散在しており、また関係府省におけるデータベースの統合化に向けた取り組みも進められている。本研究はこれらの取り組みを補完すべきものであることから、単独府省で対処が困難であり、関係府省の枠組みを超えて国内主要データベースの構築機関が連携した総合的な推進体制で実施する必要がある。

3. 研究計画

各対象分野のデータベースの構築経験者および典型的利用者によってライフサイエンスにおいて重要度の高いデータベースの内容、用途・維持管理方法についてつまびらかにすることを目的とした調査を行う。

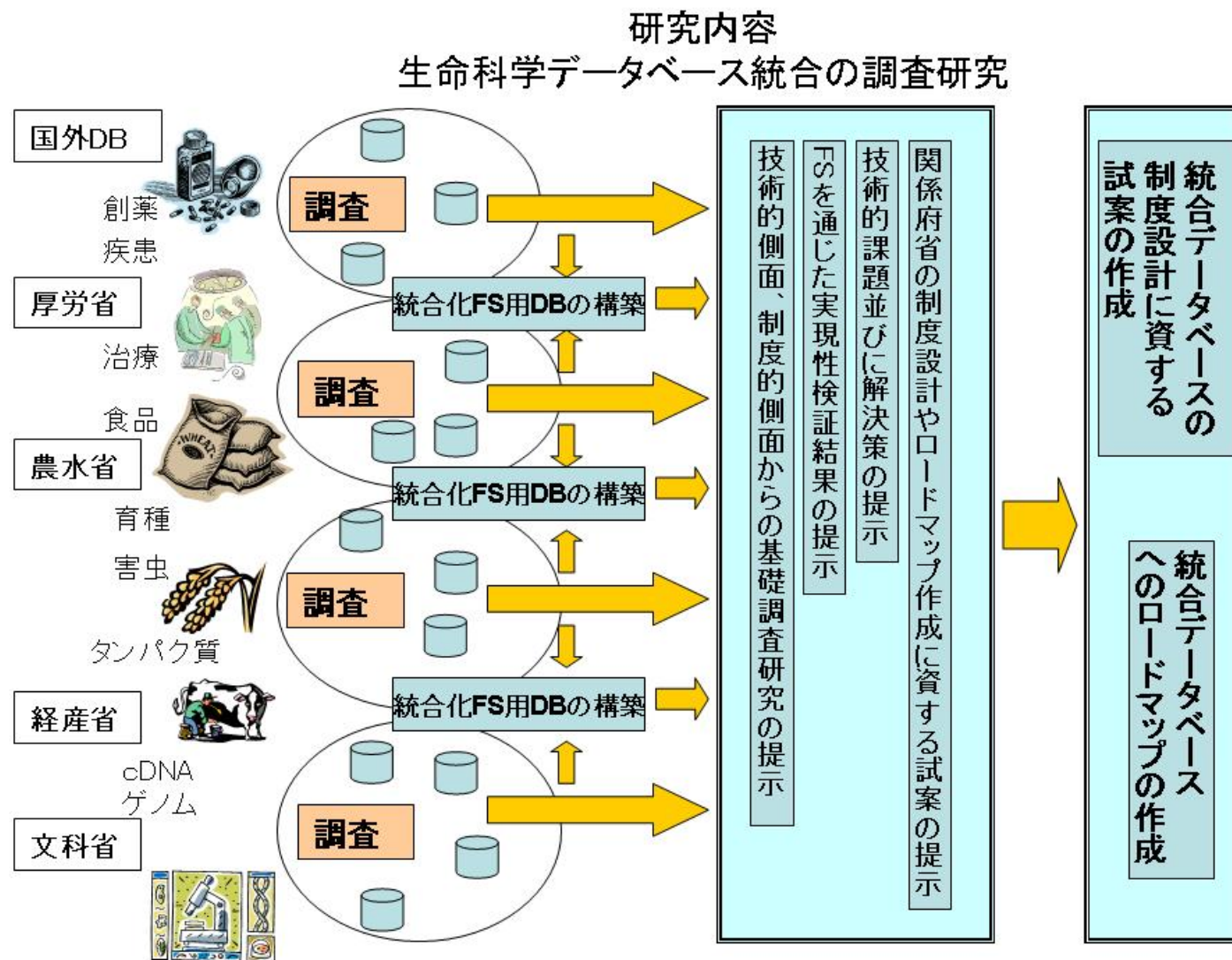
主要データベースの解析によって知識表現、記述的データおよび関係データなど構成データに分解し、個々のデータベースの内容および相互の関係が統一的な表現方法で一覧できるような表現方法を開発し、それぞれのデータベースの特徴が一望できる表現にまとめる。

また複数のデータベース機能を統合する際の技術的制度的問題点を検討することを目的に、少数のデータベースの試験的な統合を試みる。

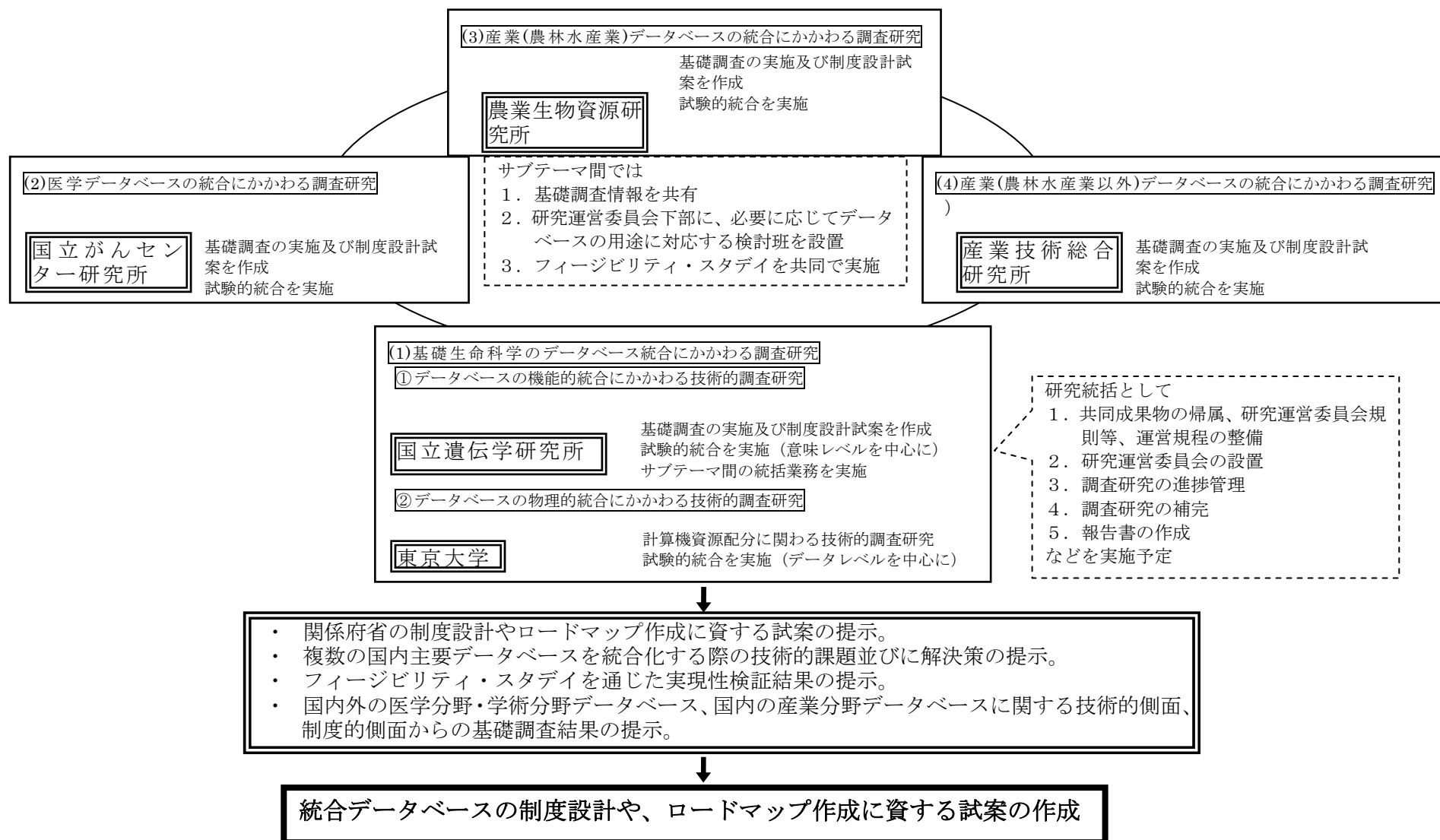
4. ミッションステートメント

- ・①関係府省におけるデータベース統合化に向けた取り組みの補完となる、関係府省の制度設計やロードマップ作成に資する試案の提示。
- ・②複数の国内主要データベースを統合化する際の技術的課題並びに解決策の提示及びフィージビリティ・スタディを通じた実現性検証結果の提示。
- ・③国内外の医学分野・学術分野データベース、国内の産業分野データベースに関する技術的側面、制度的側面からの基礎調査結果の提示。

5. 研究全体像



6. 研究体制



実施体制一覧

研究項目	担当機関等	研究担当者
1. 基礎生命科学のデータベース統合にかかわる調査研究	情報・システム研究機構 国立遺伝学研究所 生命情報・DDBJ 研究センター	◎大久保 公策(教授)
(1) 生命科学一般のデータベース統合にかかわる調査研究	情報・システム研究機構 国立遺伝学研究所 生命情報・DDBJ 研究センター	◎大久保 公策(教授)
(2) 基礎生命科学のデータベース統合にかかわる調査研究	東京大学 大学院 情報理工学系研究科	森下 真一(教授)
2. 医学データベース統合に関する調査研究	国立がんセンター研究所 腫瘍ゲノム解析・情報研究部	吉田 輝彦(部長)
3. 産業(農林水産業)データベースの統合にかかわる調査研究	(独)農業生物資源研究所 DNA バンク&リソースセンター	長村 吉晃(チーム長)
4. 産業(農林水産業以外)データベースの統合にかかわる調査研究	(独)産業技術総合研究所生物情報解析研究センター(JBIRC)	今西 規(チームリーダー)

◎ 代表者

7. 研究運営委員会について

研究運営委員会委員一覧

氏名	所属機関	役職
(研究実施者) ◎大久保 公策	国立遺伝学研究所生命情報・DDBJ研究センター・教授	委員長 研究代表者 サブテーマ1責任者
森下 真一	東京大学大学院情報理工学系研究科・教授	サブテーマ1参画者
吉田 輝彦	国立がんセンター研究所・研究部長	サブテーマ2責任者
長村 吉晃	(独)農業生物資源研究所DNAバンク&リソースセンター・チーム長	サブテーマ3責任者

氏名	所属機関	役職
今西 規	(独)産業技術総合研究所生物情報解析研究センター(JBIRC)・チームリーダー	サブテーマ4責任者
(外部有識者) 楠木 正巳	大阪大学蛋白質研究所附属プロテオミクス総合研究センター蛋白質情報科学研究系・助教授	
黒田 雅子	(独)科学技術振興機構研究基盤情報部バイオインフォマティクス課・課長	
五斗 進	京都大学化学研究所バイオインフォマティクスセンター生命知識システム・助教授	
酒井 明	三菱化学生命科学研究所研究総合推進部・副部長	
林 健志	九州大学生体防御医学研究所遺伝情報実験センター・特任教授	
松尾 泰樹 (～第7回)	文部科学省研究振興局ライフサイエンス課・課長	
菱山 豊 (第8回～)	文部科学省研究振興局ライフサイエンス課・課長	
林 憲一 (～第9回)	厚生労働省大臣官房厚生科学課・研究企画官	
坂本 純 (第10回～)	厚生労働省大臣官房厚生科学課・研究企画官	
牧元 幸司 (～第1回)	農林水産省農林水産技術会議事務局先端産業技術研究課・課長	
高野 浩文 (第2回～)	農林水産省農林水産技術会議事務局先端産業技術研究課・課長	
徳増 有治 (～第9回)	経済産業省製造産業局生物化学産業課・課長	
倉田 健児 (第10回～)	経済産業省製造産業局生物化学産業課・課長	

◎研究運営委員長

運営委員会等の開催実績及び議題

(a) 運営委員会

第一回(平成17年12月28日)

開催場所:独立行政法人科学技術振興機構大会議室(富国生命ビル23階)

- 議題:
- (1)研究運営委員会の設置等について
 - (2)自己紹介
 - (3)調査研究目標及び実施計画等について
 - (4)現在までの調査活動について
 - (5)その他

主な決定事項: 今後の調査研究の進め方の方針など

第二回(平成18年1月26日)

開催場所:独立行政法人科学技術振興機構大会議室(富国生命ビル23階)

- 議題:
- (1)運営委員の交替について
 - (2)第1回研究運営委員会議概要(案)について
 - (3)主要ライフサイエンスプロジェクトとデータベースの役割について(報告)
 - (4)調査研究進捗報告について
 - (5)拡大会議について

主な決定事項: 拡大会議の開催内容

第三回(平成18年3月1日)

開催場所:独立行政法人科学技術振興機構大会議室(富国生命ビル23階)

- 議題:
- (1)第2 回研究運営委員会議事概要(案)について
 - (2)拡大会議における検討等について
 - (3)平成17 年度調査研究報告書について

主な決定事項: 報告書の提出に関する事項

第四回(平成18年4月28日)

開催場所:国立情報学研究所 会議室

- 議題:
- (1)第3回研究運営委員会議概要(案)について
 - (2)平成17年度調査研究報告書(案)について
 - (3)平成18年度実施計画について

主な決定事項: 実施計画の承認

第五回(平成18年7月26日)

開催場所:国立情報学研究所 会議室

- 議題:
- (1)第4 回研究運営委員会議概要(案)について
 - (2)今後の研究内容・計画に係る確認(科学技術連携施策群)
 - (3)科学技術連携施策群ポストゲノムWG への報告について

- (4)各調査研究機関の活動経過報告
- (5)各分野のコーディネーションについて

主な決定事項：本研究の目標と今後の実施計画について、各分野のコーディネーションについての考え方

第六回(平成18年10月26日)

開催場所:情報・システム研究機構 会議室

- 議題：
- (1)第5 回研究運営委員会議概要(案)について
 - (2)各分野のコーディネーションについて(フィージビリティ・スタディの実施計画含む)
 - (3)各調査研究機関の活動経過報告

主な決定事項: フィージビリティ・スタディの実施計画了承

第七回(平成18年12月12日)

開催場所:情報・システム研究機構 会議室

- 議題：
- (1)第6 回研究運営委員会議概要(案)について
 - (2)フィージビリティ研究から制度試案検討への展開について
 - (3)各分野のデータベース及びデータ統合ニーズ等について
 - (4)拡大会議の実施方針について

主な決定事項：拡大会議の実施方針承認、EBMについて今後検討

第八回(平成19年3月9日)

開催場所:情報・システム研究機構 会議室

- 議題：
- (1)第7 回研究運営委員会議概要(案)について
 - (2)EBM の動向等について
 - (3)拡大会議における検討等について
 - (4)各調査研究機関の平成18 年度活動成果の骨子について

主な決定事項：活動成果の骨子確認

第九回(平成19年4月27日)

開催場所:情報・システム研究機構 会議室

- 議題：
- (1)第8 回研究運営委員会議概要(案)について
 - (2)平成18 年度調査研究報告書(案)について
 - (3)平成19 年度実施計画について(各委員)

主な決定事項：平成19 年度実施計画については一部修正のうえ承認

第十回(平成19年7月25日)

開催場所:情報・システム研究機構 会議室

- 議題：
- (1)第9回研究運営委員会議概要(案)について
 - (2)平成19年度実施計画について
 - (3)最終成果取りまとめイメージについて
 - (4)調査研究の経過報告について

主な決定事項：最終成果取りまとめイメージについては今後も継続して議論

第十一回(平成19年10月17日)

開催場所:情報・システム研究機構 会議室

- 議題:
- (1)第10回研究運営委員会議概要(案)について
 - (2)臨床・疫学・治験に係る情報・データとその統合について(中間とりまとめ)
 - (3)調査研究の経過報告について
 - (4)ライフサイエンスDB統合に向けての提言と制度設計について

主な決定事項: ライフサイエンスDB統合に向けての提言と制度設計については引き続き検討
11月29日開催予定のシンポジウム(連携施策群主催)への対応

第十二回(平成19年12月18日)

開催場所:学術総合センター 会議室

- 議題:
- (1)第11回研究運営委員会議概要(案)について
 - (2)遺伝研経過報告～生命科学の基礎・基盤連携群シンポジウムの発表について～
 - (3)調査研究の経過報告と報告書目次案について

主な決定事項: 報告書取りまとめスケジュール等の確認

第十三回(平成20年2月13日)

開催場所:情報・システム研究機構 会議室

- 議題:
- (1)第12回研究運営委員会議概要(案)について
 - (2)本調査研究からの提言(に資する試案)取りまとめについて
 - (3)各調査研究機関の成果報告とりまとめの進捗等について

主な決定事項: 提言に資する試案についてはメールでも議論のほかもう一度会議を開催
報告書取りまとめスケジュールの確認

第十四回(平成20年3月19日)

開催場所:人間文化研究機構 会議室

- 議題:
- (1)第13回研究運営委員会議概要(案)について
 - (2)本調査研究からの提言(に資する試案)取りまとめについて
 - (3)報告書とりまとめのスケジュール等について

主な決定事項: 取りまとめ作業についてのスケジュール、方法等

(b) 研究成果報告会

生命科学の基礎・基盤連携群シンポジウム～ ライフサイエンスデータベース統合への取組
(主催:内閣府)

(平成19年11月29日)

開催場所:浜離宮朝日ホール 多目的ホール(住所:東京都中央区築地5-3-2)

プログラム(下記)の一部として、補完的課題「生命科学データベース統合に関する調査研究」報告の
タイトルで大久保委員長が発表。

・来場者 183 名。

・運営委員会で発表されたシンポジウム全体の評価

背景説明やパネル及び講演者の方々のお話のいずれについても、非常に興味深く聴いていただ

けたようである。ただし、全体的にデータベース統合の向かう方向性が見出せないという意見があった。

プログラム

1. 開会の辞 (13:00~13:05)

本庶 佑 (総合科学技術会議議員)

2. 生命科学の基礎・基盤連携施策群について (13:05~13:25)

五條堀 孝

(科学技術連携施策群生命科学の基礎・基盤連携施策群コーディネータ、国立遺伝学研究所副所長)

3. 補完的課題「生命科学データベース統合に関する調査研究」報告 (13:25~14:10)

大久保 公策 (国立遺伝学研究所 生命情報・DDBJ研究センター 教授)

4. 有識者講演 (14:10~15:50)

座長: 中村 春木 (大阪大学蛋白質研究所・附属プロテオミクス総合研究センター・教授)

- ・「統合データベースプロジェクトの目指すもの」
高木 利久 (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、
東京大学大学院 新領域創成科学研究科・教授)
- ・「イネゲノム情報を中心とした農作物ゲノムデータベース: 基盤から育種への利用をめざして」
佐々木 卓治 (農業生物資源研究所・理事)
- ・「臨床情報データベース - 出る情報、出ない情報、出せない情報 -」
中谷 純 (東京医科歯科大学 情報医科学センター・准教授)

休 憩 (15:50~16:10)

5. ライフサイエンスデータベース統合に関する各省の取組と今後に向けたパネルディスカッション (16:10~17:40)

モデレーター: 五條堀 孝 生命科学の基礎・基盤連携施策群コーディネータ
重藤 和弘 内閣府参事官(ライフサイエンス担当)

パネリスト: 大久保 公策 国立遺伝学研究所 生命情報・DDBJ研究センター 教授
高木 利久 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
中村 春木 大阪大学蛋白質研究所・附属プロテオミクス総合研究センター 教授
文部科学省研究振興局ライフサイエンス課長
厚生労働省大臣官房厚生科学課長
農林水産省農林水産技術会議事務局先端産業技術研究課長
経済産業省製造産業局生物化学産業課長

6. 閉会の辞 (17:40~17:45)

大江田 憲治 (内閣府大臣官房審議官(科学技術政策担当))

シンポジウムプログラム

II. 経費

1. 所要経費

(直接経費のみ)

(単位:百万円)

研 究 項 目	担当機関等	研 究 担当者	所要経費			
			H17 年度	H18 年度	H19 年度	合計
1. 基礎生命科学のデータベース統合にかかわる調査研究 (1) 生命科学一般のデータベース統合にかかわる調査研究 ①人件費 ②業務実施費	情報・システム研究機構 国立遺伝学研究所 生命情報・DDBJ 研究センター	大 久 保 公策	0 33.5	9.7 56.3	4.8 53.0	14.5 142.8
(2) 基礎生命科学のデータベース統合にかかわる調査研究 ①人件費 ②業務実施費	東京大学 大学院 情報理工学系研究科	森 下 真 一	- -	5.9 0.4	6.5 0.6	12.4 1.0
2. 医学データベース統合に関する調査研究 ①人件費 ②業務実施費	国立がんセンター研究所	吉 田 輝 彦	0 1.6	0 1.4	0 1.5	0 4.5
3. 産業(農林水産業)データベースの統合にかかわる調査研究 ①人件費 ②業務実施費	独)農業生物資源研究所 DNAバンク&リソースセンター	長 村 吉 晃	0.4 1.6	1.1 4.7	1.0 4.2	2.5 10.5
4. 産業(農林水産業以外)データベースの統合にかかわる調査研究 ①人件費 ②業務実施費	(独)産業技術総合研究所生物情報解析研究センター(JBIRC)	今西 規	0 1.9	0 6.0	0 5.9	0 13.8
所 要 経 費 (合 計)			39.0	85.5	77.5	202.0

2. 使用区分

(単位:百万円)

	サブテーマ 1(1)	サブテーマ 1(2)	サブテーマ 2	サブテーマ 3	サブテーマ 4	計
設備備品費	0	0	0	0	0	0
試作品費	0	0	0	0	0	0
人件費	14.5	12.4	0	2.5	0	29.4
業務実施費	142.8	1.0	4.5	10.5	13.8	172.6
間接経費	47.2	4.0	1.5	3.9	4.2	60.8
計	204.5	17.4	6.0	16.9	18.0	262.8

※備品費の内訳(購入金額5百万円以上)

該当なし

III. 研究成果

1. 研究成果の概要

(1) 研究目標と目標に対する結果

1) ミッション① 関係府省におけるデータベース統合化に向けた取り組みの補完となる、関係府省の制度設計やロードマップ作成に資する試案の提示。

試案の検討に当たっては、外部有識者を含む研究運営委員会において議論を行ったが、現状認識のずれから問題意識を共有することに困難があり、十分に合意形成を行うことができなかった。そのため、本調査研究においては研究代表者として以下の試案を提示する。

また試案の重大性に鑑み、これまでの検討を踏まえ、研究代表者として、今後の制度設計に資するものとして留意すべき事項も付記する。

【試案】

1. 「政府資金によるデータ産生型プロジェクト(注1)のデータを我が国の研究社会で早期に共有するためのルール」の作成を検討すべきと考える
2. 公共財としてのデータを保全・管理し、長期にわたるデータの育成と共有を行う公的機関の設置を検討すべきと考える

1. 「政府資金によるデータ産生型プロジェクトのデータを我が国の研究社会で早期に共有するためのルール」の作成を検討すべきと考える

我が国での DB 統合の要請については、「米国立バイオテクノロジー情報センター(NCBI、注2)相当の機関が我が国にも必要である」ということがしばしば言われており、理想的なデータベース統合の例として NCBI でのデータ統合をあげ、我が国の現状と比較することが多い。例えば、文部科学省ライフサイエンスデータベース整備戦略作業部会報告 2007⁽ⁱ⁾や NPO 医学図書館協会による「国立ライフサイエンス情報センター(仮称)」推進準備委員会最終報告 2005⁽ⁱⁱ⁾の中では、予算、人員等の相違が議論されてきた。

本調査では、科学情報一般に関わる「日米の決定的な違い」について「政府資金によるデータの共有利用を保障するルールの差」を最大の原因として指摘する。すなわち「NCBI が必要である」というこれまでの要請は「政府資金によるデータ産生型プロジェクトのデータを早期に国民で共有する為のルールが必要である」と翻訳されるべき内容である。

これまでも研究成果の公開には積極的な取組がなされているが、今後、一定の国家プロジェクトについては、その研究により得られたデータが合理的な期間内に共有されるための仕組みが構築される必要がある。

具体的には、政府資金によるデータ産生型のプロジェクトについては、研究計画の申請時において産生を企図するデータの種別と量につき明確にし、データ共有の為の計画を提案採択時の重要な判断基準とすることなどが考えられる。また臨床医学研究などのプライバシーに関わる研究においても、データ

の匿名化などによる非プライバシーデータ化とその共有のための仕組みを採択条件の一つとしてその質を競わせる等の工夫をするべきである。

注1) データ産生型とはデータ生産に始まりその解析による発見を企図する研究開発である。詳しくはp193)-(B)「データベース統合の背景」を参照。

注2) National Center for Biotechnology Information: 米国国立衛生研究所(NIH)に属する国立医学図書館(NLM)の一部門。医学生物関連のDB構築・運用のほか、関連するソフトウェア、システムの開発などを実施している。

2. 公共財としてのデータを保全・管理し、長期にわたるデータの育成と共有を行う公的機関の設置を検討すべきと考える

契約時の計画に従って研究プロジェクトからデータが国民(もしくは合理的な理由により限定公開を受ける集団)に提供される場合、プロジェクトグループ自体は研究期間終了時に解散となるため、データを維持提供し続ける主体とはなりえない。

加えて科学データは一般に他のデータと組み合わせあって価値が上がるために、データの生産者とは別に、安定に維持・運営できる機関(あるいはその役割を持つ仕組み)がデータの提供にあたるのが合理的である。

この機関には公共財であるデータに対する優先的なアクセスが保障されるので、応用目的研究や企業活動と競合的になってはならない。すなわち集約されたデータの利便性において機関内外及び外部機関間の格差を作らない透明性の維持を義務付けられ、データの集約利用による発見研究や同目的の外部との共同研究を厳格に制限された中立的な性格の機関であるべきである。

一方で上記機関は、データの整理や説明に加えて、データの価値を上げるための組み合わせ等による育成を行うべきであり、その為の研究は認められるべきである。この集められ育てられたデータも、材料として提供されたデータと同様に公共財として利用者と共有するべきである。

また個別研究が独自に公開しているデータベース等を案内する書誌情動的な説明を行い、計算資源を利用して維持困難な情報資源やサービスの保存を代行し、わが国の情報資源やサービスに対する努力を能率よく国民に伝え、データを共有することの重要性を研究者、ひいては国民に説明する極めて重要な役目を果たすべきである。

なお、医学研究等では、試料等の提供者個人毎の情報が、匿名化される等の措置が講じられたとしても、何らかの形で公開され共有される場合には、何よりも国民の理解と協力が不可欠であり、当該機関は社会に対しても事業の透明性を保ちながら、その説明責任を果たすべきである。

我が国の学術情報サービスの原型は1964年米国のWeinbergらがNSFとしてケネディに行った答申であるWeinberg Reportに存在する⁽ⁱⁱⁱ⁾。これは我が国の情報基盤デザインの骨格として利用されているNIST(National Information system for Science and Technology)構想がWeinberg Reportの3年後に学術会議の答申として表された時に言及されている^(iv)。

Weinberg Reportでは、科学技術情報基盤は研究情報源として、また科学政策立案の為の科学管理情

報(原文では科学インテリジェンス)として国家が科学を進めるにあたっての最重要課題であると明言している。加えてそれは図書館業務のような文書やデータの有形物のような保管と手渡し作業ではなく科学分野自体が行うべき情報の選別と濃縮作業を伴うものでなければならないと強調している。そして各専門分野に情報センターを設け、各大学などの図書館やデータセンターと結ぶ木構造の情報系統が提案されている。すなわち上記のような国立データセンターは既に1960年代から計画されていた。

しかし、我が国においては例えば医療・医学情報を中心に多数の情報サービスを使命とする情報サービスセンターは設立運営されているが、研究者を対象にした科学情報やデータの整理濃縮よりも一般市民への解説の著作・翻訳や広報を行うに留まっている。

【留意事項】

我が国では現代科学の直接的で確実なアウトプット(出力)である膨大なデータを共有するか分割所有するかなどの保存管理の方法が長期的に我が国の公的科学全体のアウトカム(成果)に与える影響について、あるいは過剰なデータ共有の強制が研究動機や知財取得機会に与えるマイナスの影響についての現実的な議論(の記録)はこれまでに見当たらない。日本学術会議の2001年の声明に「公的資金(税金)を用いて実験や観測さらには調査から得られたデータは、一定期間、取得者の研究に使われた後は、原則として公開・共有されることが科学の発展に基本的に重要だ」⁽ⁱ⁾という「人類の共同事業としての科学」の立場からの記述があるが、「国のイノベーションの力を持続的に高める為」という見地から、巨大なデータを研究材料とする科学分野において研究者の育成、動機付け、研究資源のすべてに関わるデータの保存管理の方法についての議論は尽くされていない。

したがって、本課題を契機にまずは科学データの取り扱いについてより多角的な視点からの現実的な議論を透明に進め、一般社会や研究者社会での十分な理解に基づく合意形成と、その公正かつ効果的な履行に必要な制度の設計を行うことが望まれる。

公的科学の民営化(後述、p20 参照)後約30年を経て、次々と大型研究を進める米国では過去10年間に急速にデータや論文の公共財化を進めているが^(vi)、情報インフラが充実した環境で価値創生の材料としての公共の電子コンテンツが国家のイノベーション力の増進と発揮に欠かせないとの認識に基づき^(vii)、学術誌のオープンアクセス化とあわせてデータ共有(Data Sharing)についての議論を公開してきた^{(vii)-(x ii)}。

その結果2006年の意見調査では83%の米国民は「納税者は公的科学研究の成果に無料でアクセス可能であるべき」との考えを示している^(x iii)。

現在我が国では、一般的に科学データ(ファクトデータ)や、それを整理したに過ぎない創作性のないデータベースに知的財産権が認められない^(x iv)。この為に生産者(事実上の保有権限者)の多くは当初の研究計画に従った解析と論文発表の後もある種の選別や加工を加えた後のデータを多様な利用規約を設けて個別に「公開」している。^(x v) 統一的でない利用規約を持つデータの統合や利用は極めて煩雑であり、また不明の選別によって虫食い状態になった統合産物も十分な価値を持ち得ないであろう。従ってデータ全体の価値を損なわないために、加工や選別を受けていないデータ(注3)を十分な説明と

ともに保管し共有することが科学的に意味のあるデータベース統合の条件である^(x vi)。英米では既に研究者の動機を損なわずデータの早期公開を競わせるためにデータの占有期限を決めたガイドラインに加えて、公募課題の選択基準への盛り込みや模範行為 (Good Practice) の広報による教育、データ共有履行の調査による評価など科学社会全体での仕組みを作り始めている^{(x vii) - (xxi)}。

さらに、今世紀に入り地球規模で広がる「知識に対するオープンアクセスへの潮流」という文脈の中でデータベース統合を考えることも重要である。知識へのオープンアクセス運動は多国間の非政府運動として、ブダペストオープンアクセスイニシアティブ(2001-)やベルリン宣言(2003)によって知られているが^(xxii)、現在既に11カ国で政府系、民間あわせて 36 のファンディング機関が研究報告論文のオープンアクセス(36/36)、および研究データのオープンアクセス(16/36)に関するガイドラインを設けていることにも留意すべきである^(xxiii)。

アジアでは中国も 2006 年に分野別データセンター新設を含めたデータ共有政策を進める旨を表明している^(x xiv)。

各国の科学政策と歩調を合わせ、2004 年に OECD では我が国を始めとする加盟国および中国他を含めた 30 ヶ国政府による宣言 “Declaration on Access to Research Data from Public Funding” を採択している。この中では、経済の持続的な発展のためにも公的資金によるデータへの自由なアクセスと利用が望ましいことを宣言し^(x x v)、科学データの扱いについての理想的な原則を端的によくまとめている。

(注 3) 科学データは一般に利用目的によってデータの選別や加工などの前処理が異なる。従ってデータの選別や加工はデータの特定目的での利便性を上げる一方で利用目的の多様性を減らすものである。また別々の基準で選別や加工を受けている場合、同種のデータであっても科学的に意味のある統合ができなくなる。米国の規約 (OMB-A130) でも Raw content(生データ) の共有が奨励されている(p26)。

(参考文献)

(i) <http://www.lifescience.mext.go.jp/download/34th/34-02.pdf>

(ii) <http://www.soc.nii.ac.jp/jmla/nlsic/index.html#houkoku>

(iii) 科学と政府と情報－米国政府に対するワインバーク報告－. 日本ドクメンテーション協会, 1966

(iv) http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu4/toushin/06041015/002.htm

(v) <http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-18-k136.pdf>

(vi) <http://grants.nih.gov/grants/sharing.htm>

その他制度は <http://133.11.132.80/intelligence/report/law/index7.html> 参照

(vii) Bits of Power: Issues in Global Access to Scientific Data. (1997) National Research Council. National Academy press (<http://www.nap.edu/catalog/5504.html>)

(viii) A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases (1999) National Research Council. National Academy press (<http://www.nap.edu/catalog/9692.html>)

(ix) The Digital Dilemma: Intellectual Property in the Information Age (2006) National Research Council. National Academy press (<http://www.nap.edu/catalog/9601.html>)

(x) Engaging Privacy and Information Technology in a Digital Age (2007)

- (<http://www.nap.edu/catalog/11896.html>)
- (x i) Privatizing the University--the New Tragedy of the Commons.(2000) Science. Dec 1;290(5497):1701-1702.
- (x ii) Can patents deter innovation? The anticommons in biomedical research(1998) Science. 1998 May 1;280(5364):698-701.
- (x iii) http://www.harrisinteractive.com/harris_poll/index.asp?PID=671
Wall street journal に紹介記事
<http://commerce.wsj.com/auth/login?mg=wsj-users2&url=http%3A%2F%2Fonline.wsj.com%2Farticle%2FBSB114893698047965609.html%3Femail%3Dyes&mg=com-wsj>
- (x iv) 中山信弘「著作権法」有斐閣(2007) p36-38, p119-120、p122-123
- (x v) <http://lifesciencedb.jp/lbdb.cgi?gg=project>
- (x vi) Seeking Consensus on Data Sharing (2002) ScienceNOW 26 February 2002: 3
<http://sciencenow.sciencemag.org/cgi/content/full/2002/226/3>
- (x vii) Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers(2008) PLoS Med. Sep 2;5(9):e183
- (x viii) Data sharing for computational neuroscience.(2008)Neuroinformatics. Spring;6(1):47-55.
- (x ix) Sharing detailed research data is associated with increased citation rate.(2007) PLoS ONE. Mar 21;2(3)
- (x x) Towards effective and rewarding data sharing(2003) Neuroinformatics.;1(3):289-95.
- (x xi) Data sharing and intellectual property in a genomic epidemiology network: policies for large-scale research collaboration (2006) Bull World Health Organ vol.84 no.5 Geneva May 2006
- (x xii) ブダペスト宣言 <http://www.soros.org/openaccess/help.shtml>
ベルリン宣言 http://oa.mpg.de/openaccess-berlin/berlin_declaration.pdf
日本語での解説は「研究助成機関とオープンアクセス情報管理 48(3) 133-143」を参照
http://www.jstage.jst.go.jp/article/johokanri/48/3/133/_pdf/-char/ja/
- (x xiii) SHERPA で作成された世界の研究助成機関の成果に対するオープンアクセスガイドライン
<http://www.sherpa.ac.uk/juliet/index.php>
SHERAは Joint Information Systems Committee (JISC)と CURL (Consortium of University Research Libraries) がファンドした JISC FAIR (Focus on Access to Institutional Resources) programme の一部。UKの研究系大学や British Library などが受託して相互に連携してオープンアクセス e-print レポジトリを準備するための技術制度面の調査研究3ヵ年 (2002-2005)プロジェクト.. 事務局はノッティンガム大。科学雑誌の著作権ポリシーと機関レポジトリに対する対応を網羅した調査 Romeo が有名。我が国でも筑波大学図書館など4図書館が情報学研究所最先端学術情報基盤事業の委託で日本版的 SCPJ(Society Copyright Policies in Japan)を作成している。
- (x xiii) <http://www.scidev.net/en/news/china-unveils-plans-to-boost-scientific-data-sharing.html>
- (x xiv) OECD 2004 共同声明 <http://www.oecd.org/dataoecd/42/12/35393145.pdf>
翻訳は http://133.11.132.80/intelligence/finality/pdf/finality_01.pdf にあり 解説は
An International Framework to Promote Access to Data(2004) Science 19 March, 303,

2) ミッション② 複数の国内主要データベースを統合化の際の技術的課題並びに解決策の提示及びフィージビリティ・スタディを通じた実現性検証結果の提示

DBとは世界の写しである。 計算機の中に作り上げた世界の枠組み(データモデル)に従って世界の中の要素の特徴を記録したレコード群(データ)である。多くの利用者によって世界全体の特徴を解析(データマイニング)したり世界の特定の要素を選び出(検索)したりするための道具である。 同じ対象世界について補完的な関係の対象や特徴を扱うDBがある場合にはそれらのDBを一纏めにしてより多くの要素につきより正確により多角的に記載したDBとすることがデータマイニングや検索利用に対して有利である。これが物理的なDB統合である。

同一世界を記述した別々のDBを一纏めにする場合には、世界の枠組みモデル(要素の特徴群や特徴群同士の関係)についての不一致、同じ特徴を記載する場合に要素の呼び名(ID系列)使われる知識表現(オントロジー)の不一致などがライフサイエンスでも例えば顧客DBでも共通の問題である。

これらの共通問題に関しては情報工学分野が解決方針を準備し様々な道具を開発し解決に牽引してくれている。異なるオントロジーの融合や機械巡回によるID対応テーブルの作成などすぐにライフサイエンス分野でも利用可能である。

本課題のフィジビリティスタディ(FS)でもこれらの技術的課題は情報工学領域で準備された方法が十分に有効であることを確認した。

用語統一の重要性とその解決例は二つのフィージビリティ・スタディによって示した。一つは解剖用語の自動分類機による全植物 EST の統合整理であり、いまひとつは複数の DB における遺伝子 ID 系列の自動更新機である。粒度や分類視点を統一することで全ての植物ESTプロジェクトを、同一プロジェクトのように統合整理することが可能であった。また遺伝子や蛋白質に関する複数のID系列の参照表はDBの統合利用において欠かせないがこの更新は管理者にとって負担となる。比較的安定した欧米の ID 系列(NCBI の ID や SwissProt の ID)を選び、別の ID 系列を用いている DB(例えば H-inv) の外部参照情報をロボットで定期的にアップデートする仕組みと対応関係を管理するサーバを構築することで常にそれらの DB の ID の対照表を維持できることを示した。

DBを質問に対して一定の処理をした答えを返すサービスととらえて別視点DBが供するサービスを組み合わせる新たなサービスを作り上げることは機能的なDB統合(マッシュアップ)と呼ばれる。機能的DB統合の為の技術としてはDBの提供しているサービス(利用者の質問に答えてデータを見せたり、データを加工したりする機能)自体をプログラムでも呼び出せるようにするWeb-API(application program interface)プログラムをサービス提供側に付加する技術が機能的DB統合に有効であると期待されている。例えば Google earthでは衛星写真で作った地図情報の部分的提供とユーザーによる座標へのコメントのマップをWeb-API経由で利用可能にしている。クライアントDB(例えばお店紹介情報)の検索結果をGoogleの衛星画像や地図画像の上にマップして表現することが可能になっている。Web-APIは上記FSで遺伝子発現DBやID対応機に既に実装しその有効性を確認することが出来た。

従ってオントロジーや辞書などによって物理統合は十分遂行可能であることが示された。その条件は統合しようとするDBが表現している領域の内容を熟知し、また、統合過程の情報処理の両方を熟知した研究者の参加、もしくは双方の研究者の密な議論による相互理解である。

【解説】 DB統合による価値の増加程度

「独占されている複数のDBが共有されるとその価値がDB数の組み合わせ数倍になる」Vinton Cerf氏の唱える法則が有名である。Vinton Cerfは通信プロトコル(TCP/IPプロトコル)をRobert Kahn氏とともに開発し現在のインターネットの基本構造を作ったことで「インターネットの父」といわれる。チューリング賞(ACM)を、コンピューターネットワークにおける功績に関して受賞。ICANNの会長職を務めている。いまひとつの技術基盤であるイーサネット(Ethernet)技術を開発したBob Metcalfe(ボブ・メトカーフ)氏が1995年に提唱した「ネットワークの価値はノード数の二乗に比例する」という法則と似ている。

3) ミッション③ 国内外の医学分野・学術分野データベース、国内の産業分野データベースに関する技術的側面、制度的側面からの基礎調査結果の提示

(A) 「データベース統合」の定義

「分野別推進戦略」(平成18年3月28日総合科学技術会議決定)において、ライフサイエンス分野の課題として、新規の医薬品や医療機器の産業化に向けた「実用化研究の基盤」が十分に整備されていないことが指摘されている。本課題にある「データベース統合」はこの「実用化研究の基盤整備」を目的とする事業である。

一方「研究の基盤」とは「情報や材料などの研究成果が①蓄積され②体系化されたものであり、③広く共用されるものである」と3つの特徴で定義されている。(平成19年度4月 知的基盤整備計画 科学技術・学術審議会、技術・研究基盤部会) 従って本課題では研究基盤整備の一部としての「データベース統合」を「データやデータベースを保管し、体系化し、広く共用可能にする行為」と定義する。

(B) データベース統合の背景

a) 巨大データの登場とイノベーションプロセスの変化

1990年代から生命科学では理工学系のデータを集めてミクロな詳細さでマクロに対象を観察することを可能にする分析機器が登場し、生命現象を定量的かつ定性的に観察した巨大観察データが生まれた。これまでの実験データと異なり巨大観察データはデータ生産者以外の多種の研究に役立ち企図しない研究を生み出す新しいカテゴリーのデータである。そのために「実験データ」と区別して「基盤データ」や「基礎データ」と呼ばれる。多くの研究者が同じデータ群を共有し様々な解析やアイデアを競うプロセスはこれまでの生命科学が経験していない研究プロセスである。本報告書では「巨大観察データ」を習慣に従い「基盤データ」と呼ぶ。

b) データ依存を強める生命科学における DB の役割

ゲノムに始まる「基盤データ」を中心にすえた研究として①既存のデータ解析法をあてはめて行う古典研究の推進と②データ中に未知の特徴や制約を求める基礎研究の二つの典型が出現し、さらに③多種類の巨大データを組み合わせて新たな多角観察データとし、①②を助ける統合データ研究も生まれた。このようなデータ中心の研究の登場にもまして重要なのは、生命を扱うあらゆる実験研究者が研究遂行時の参照情報として「基盤データ」を利用することで、研究にブレークスルーを見出したことである。すなわち「実験データ」が誰かの研究の最下流に位置し大多数には無縁のものであったのに対し、「基盤データ」

は多くの研究の遂行上欠かせないものとなった。そして分野全体が共通の「基盤データ」を利用しながら研究を進めるために出現したのが今世紀の生命科学のデータベースである。分子進化学や数理生物学などのごく一部の研究者の机上研究に使われた20世紀のデータベースとは分野に対する重要性が全く異なる。

c) トップダウン研究の登場

科学技術基本法(H7)の成立により科学に対する社会への貢献責務が明確にされた我が国では2000年前後から科学技術基本計画に沿った大型の政府研究開発事業が生命科学でも一斉に開始された。特に生命科学ではヒトゲノムプロジェクト終了時期と重なり「基盤データ」を作る「ポストゲノム」や「オーミック研究」と呼ばれるプロジェクトであった(図1)。分野全体が依存できる「基盤データ」を特定の政府事業が短期間に生産する時代の始まりである。

ゲノム・ポストゲノム 主要プロジェクト名	年 度							プロジェクトの概要
	H12	H13	H14	H15	H16	H17	H18	
文部科学省								
ゲノムネットワーク								遺伝子の発現調節機能に関わる複雑的な解析
タンパク3000								主要タンパク質約3000種の基本構造及びその機能解明
遺伝子多型研究								ヒトゲノム遺伝子領域中のSNP関連情報の取得と解析
テラーメイド医療実現化								約30万人のSNPと薬剤の効果、副作用などの関係解明
理研ゲノム、植物、遺伝子多型								ヒト、マウス、植物のゲノム、cDNA解析、遺伝子多型解析
バイオインフォマティクス研究								生命科学分野の基幹データベースの構築・高度化
統合データベース								生命科学分野DB戦略立案支援、ポータルサイト整備
経済産業省								
データベース統合 ゲノム情報統合								国内外の有用なヒトゲノム関連情報、解析ソフトの統合的利用
完全長のcDNA								約3万のヒトの全長cDNA配列情報の取得と解析
生物システム制御基盤技術								創薬支援のためのゲノム、タンパク、化合物・異性体解析技術
生体分子立体構造								膜タンパク質及び関連複合体の立体構造・機能解明
蛋白質機能解析								完全長cDNAの遺伝子発現量など多方面からの機能解析
遺伝子多様性モデル解析								ヒトのモデル疾患に関わる遺伝子多型情報の取得と解析
複雑SNP解析								日本人集団768人に関するSNP15万種のアレル頻度の解析
厚生労働省								
疾患ゲノムデータベース								がん等5疾患のゲノムワイドなSNP解析などのデータベース構築
トキシコゲノミクス								遺伝子発現解析によるゲノムレベルでの毒性発現機構解明
疾患関連蛋白質								主要疾患を対象とした疾患関連タンパク質の探索、同定
農林水産省								
イネゲノム								イネゲノム配列の解析および遺伝子の機能解明
落花生ゲノム								ブタのcDNA配列情報、発現頻度、マーカー情報の取得と解析
大豆ゲノム								大豆のゲノム、cDNA配列情報、連鎖地図情報の取得と解析
農林水産生物ゲノム情報統合DB								イネその他の農林水産生物統合ゲノムデータベースの整備

図1 現行および終了間もないライフサイエンスの政府科学技術研究事業。

黒ぬりは「基盤データ」生産が目標に含まれる事業。

d) 知財立国政策と科学の民営化

公的資金による科学的研究開発で研究者から生まれたアイデアに知的財産権(工業所有権や著作権)を認めることは研究の動機付けになると同時にアイデアの円滑な流通利用も促す。また科学が経済的生産性を持つことは社会の持続的発展には欠かせない要件かも知れない。この科学社会と市場経済の連携は科学技術基本法(1996)の精神の一つの柱であり、産学連携、技術移転、知的財産保護をキーワードにTLO制度(H10(1998))日本版バイドール法(H11(1999))知財基本法(H14(2002))大学独法化(H16(2004))と大学市場連携の為の制度改革が続いた。米国で20世紀の最後の4半期に起こったこのような「政府資金による科学」の「民営化(privatization)」(以下 科学の民営化)が我が国では少し遅れる形で過去10年間に急速に導入された。(図2)

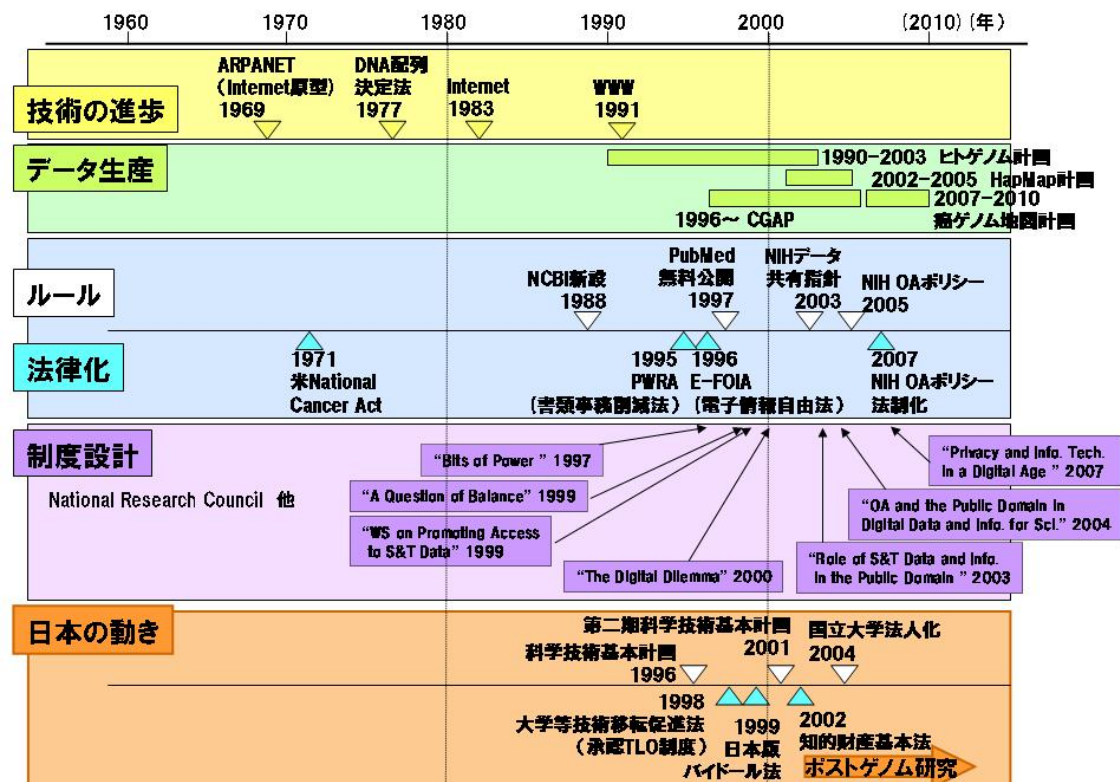


図2 統合を可能にした制度整備(米国)

(C) データベース統合の現状

a) 我が国の生命科学 DB の状況

DBの増加は1990年から急速に生じており我が国でも250を数える。DBの増加は我が国においてもライフサイエンスがデータ依存を強めていることを物語っている。我が国で公開されているデータベース群には、構築方法、及び所有権や公開統合の考え方で区別すべき6群に分類できる。我が国では国際バンク事業への協力を欧米に次いではじめたことやプロジェクト型、知識モデル型のデータベースの先駆けを1990年台半ばまでに行うなど、データベース研究時代をリードして進めてきた歴史がある。しかしながら、現在我が国で、米国のNCBIやEUのEBIなどのようなデータベースセンター的な統合がなされてきていないことは、データベース統合は、我が国と米英との情報工学における技術の差ではないことを意味している。

データベースの6分類

①データバンク(全DBの5%):

不特定多数の著者(分野全体)からデータをドキュメンテーションとともに受け付ける。著者数が利用者数に近い「皆が作って皆で使う」ことで誰も私物化できない古典的な公共データバンクである。登録時に著作権の主張や利用制限を課さないと宣言している事業であり、知財権の対象にはならない。科学雑誌のような著作権の移譲契約を結んではない。

②プロジェクトDB(31%):

特定の著者が大型予算でバンクに匹敵する量の分析や観察をした記録。資金が公的な場合データとして共有化すべき部分が主体である。通常公共データ(主に米国NCBI)と組み合わせて出来るだけ

統合を行っているので、独自データと公共データを組み合わせてサービスを構築している。それぞれに論文公開にともない末端利用を許す公開をしている場合が多いが、競合サービスが出来ないようにデータの完全な共有はしていない場合が多い。独自データには知財権が与えられず、統合DBも統合に使用した原材料のデータの利用条件によって商業利用は制限を受ける場合が多い。

③プログラム型 (28%):

バイオインフォマティストによる①②の再加工品。オリジナルデータにバイオインフォマティックス研究者の開発した一定の処理をした結果を載せたDB。類似目的を研究者間で競うために専門家以外には区別の付かないDBも多い反面、特定の用途の標準となっているDBも見られる。素材データとちがいで個人のアイデアによる改変物であり著作物に準じた著作権が生じる。ただし原材料のデータの利用条件によって商業利用は制限を受ける場合が多い。プログラム自体は新規な部分については知財権が生じる。

④キュレーション型 (12%):

典型的なドメインの視点を導入してキュレーターによって①-③を総合したデータを解釈レベルに高めたデータベース。人的な努力と判断の結果が主であり③同様著作物に準じた扱いが必要。

⑤知識モデル型 (9%):

測定データではなく特定分野の標準的知識(解釈)の要素の対応関係や要素同士の上下関係、木構造などによる形式的な表現データ。代謝ネットワークやシグナル伝達ネットワーク、遺伝子機能のオントロジーや解剖のオントロジーなどである。情報源である辞書や図譜を一度理解して人的に表現しなおした場合には著作物である。ただし情報源として他の著者の編纂物である辞書や図譜を使いそれらを単近年は電子的情報源と公共用語集を材料に単にプログラムの共起などで単純に関係を作る場合も多く、その場合にはプログラムDBと同様の扱いを受けるとされる。

⑥総説型 (2%):

特定分野の標準的知識(解釈)を読めば分かる程度の表などの形式にまとめたもの。総説記事と同様の著作権により保護される。

【解説】DBの公開度; 末端使用と転用許可

DB利用には2種類のタイプを区別すべきである。ひとつは「末端使用(エンドユース)」であり利用者の質問への答えを求めるものである。Googleなどの検索はこれにあたる。いまひとつの利用は「転用(デリベーション)」である。ここではDBサービスの材料であるデータを別サービスやデータマイニングと呼ばれるデータからの発見に使うことを指す。たとえば「末端使用」は許すが「転用」を許さないDBには「電子辞書」がある。商用の電子辞書は購入すれば末端利用は可能であるがそこから用語データを抜き出して翻訳ソフトを作る派生利用は禁止されている。

DBの「転用」を許すことは、大学や企業において①新しい利用の創出②価値付加の競争③間違いの発見と訂正④新たなデータの組み合わせによる解析による発見 ⑤第3者による自発的データアップデートとエンドユースの多様化など様々な実益がある。一方「公開」してエンドユースを許し「派生利用」を許さない理由は通常付加価値産物が生まれ第3者によって流通させられることで構築者の利益回収が妨げられる場合に限定される。しかしながら国家プロジェクトはもとより研究成果公開促進費などでつくられた公開用DBでも「エンドユース」を無償で認めるが派生利用を許さないものも多く見られる。分かりやすい例に学会と文部科学省が著作権を持つ「学術用語集」DBがある。研究成果公開促進費によって情報学研究所で電子化検索利用が許されているが「全データのダウンロード」は許されていない。

b) 純粋科学的視点からの巨大観察データの取り扱い方針

気象データ、衛星データ、などの地球科学データをはじめ今世紀に可能になった巨視的スケールの分析データは全て巨大な観察データであり「基盤データ」である。生命科学同様天体や地球環境など込み入った現象を扱う科学はこれらのデータへの依存を高めている。人類の生活の向上や福祉への貢献という科学の純粋な役割を考えて世界中の政府資金によって作られる生命科学や地球科学などの基盤データを人類共有の財産とすることの重要性を国際科学会議・科学技術データ委員会(CODATA)、が繰り返し指摘している。(Bits of Power NRCpress 1997) 我が国の公的資金による科学研究事業の人類への貢献と国内への貢献がどのようなバランスにあらうとも巨大なデータが民間資金により競争的に市場に提供されない性質のものであることを考えると政府資金による観察データの保管や流通の重要性に関する主張には傾聴すべきものがある。

(注) CODATA(Committee on Data for Science and Technology) は国際科学会議(ICSU: 各国アカデミーの協力を進めるNPO)によって1966年に設立された学際的な科学委員会である。我が国の窓口は学術会議

(注) ICSU(International Council for Science) は、人類の利益のために、科学とその応用分野における国際的活動を推進することを目的として、1931年に設立された国際的な非政府組織である。各国の科学アカデミーが参加主体となっている。ICSUの活動目標は次のとおりである。

- * 科学や社会において重要な主要問題の識別とそれらへの対処
- * あらゆる分野や国家の科学者間の関係の促進
- * 国際的な科学の目標におけるあらゆる科学者の参加の促進
- * 科学界や政府、市民社会、民間部門の間の建設的な対話を促進するための独立した権威あるアドバイスの提供

c) 経済的視点からのデータへの考え方

OECD加盟国が目指す持続的発展が可能な経済の為に政府資金による科学データの扱いが重要であると2003年のOECD科学技術閣僚レベル会議の共同声明で述べられている。我が国からは稲葉文部科学副大臣(当時)と野依理化学研究所理事長の出席したこの会議の声明では「政府資金に基づく科学技術データへのオープンアクセスは研究開発への投資以上の利益をもたらし、データの下流に豊かな商業的成果を生み出すのみならず、科学リーダーが複雑な将来の問題への正しい判断を行うにも欠かせない」と述べている。(Science 19 March 2004, 303, An International Framework to Promote Access to Data)

【解説】データ公開で国益を損なうか

データの公共への納品(公共財化)に抑制的に働く「受益範囲を国内にとどめるための非公開」という思想は、慎重に検討されるべきである。仮に「我が国の科学技術力が国際競争力において保護すべき段階にある」としても非公開独占利用によって生じる不利益——たとえば自国の科学事業が相互にシナジー効果を生まない、科学技術データに価値付加(add value)したり実用化探索(mining)したりする企業や大学研究が育たない、事業の質(データの質)に対する批判機会が失われ向上しない等——の社会におよぼす不利益について十分検討されていない。非公開を続けることは幾つかの知財の獲得と引き換えに我が国のライフサイエンスから国際競争力を奪う可能性があることを意識すべきである。

d) 米国でのデータベース統合のケース研究

米国では各地に存在していたデータベースがヒトゲノムプロジェクトの開始直前にNIHに新設されたNCBI(国立医学図書館バイオテクノロジーセンター)に全てのヒト生命科学データが統合され、いわゆるポストゲノム研究はNCBIでのデータ公開と連携して行われてきた。ゲノム研究以前には、DNA配列バンクはエネルギー省の国立研究所であるロスアラモスにあり、ゲノム地図は私立のエール大に、遺伝病の原因遺伝子地図は私立大学の教授のライフワークとして編纂が続けられていたが、これらが全てNIH管轄の政府機関であるNCBIに移管されすべて統合され転用可能な形で公開された。我が国のプロジェクト型やプログラム型のデータベースは例外なくこれらのどれかを利用することで統合サービスを提供している。NCBIはデータ統合と共有の為に創設された国立データセンターでありまさにその機能を果たしたといえる(図3)。

米国NCBIにおけるDB統合の実例

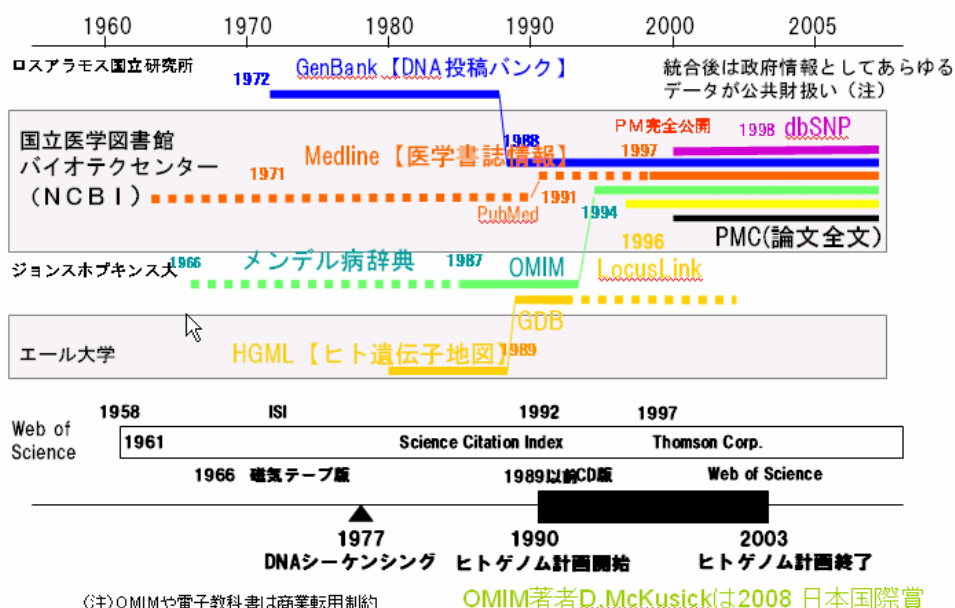


図3 米国 NCBI における DB 統合の実例

e) 米国でDB統合を可能にした制度

このような統合を可能にした背景には米国における政府情報およびその一部である公的事业由来の科学技術データの厳格な公開制度があると考えられる。当時米国では1995年の書類事務削減法と1996年の電子情報自由法の成立により情報スーパーハイウェイを介しての政府情報の公開義務付けが厳格化された。例えばMedlineもNIHの国立医学図書館(NLM)で100年以上続けられてきた医学文献の抄録作成事業であるが、電子化抄録が政府情報と看做されたためにインターネット経由で米国民にむけて無償で完全公開されたのである。現在年間1億人以上に利用されているMedlineの公開事業をNCBIが担当したことがデータベース統合の成功を決定付けたとも言えるかもしれない。NCBIのスタッフは連邦職員であり彼らは最も厳格に市民の利益に相反する活動を禁じられておりデータは完全に市民と共有されると同時に十分に説明され最低限の整理統合が施されている。他の分野でも気象情報におけるNCDC (National Climatic Data Center) 国土地理情報におけるUSGS (US Geological Survey) や化学情報でのNIST (National center for Standard and Technology)など国立データセンターはNCBIと同列の法的制約を受けて運用されているために政府機関として大学などの非営利団体に比べ厳格な法制度下でその事業内容に制約をうけていることで大学や企業での付加価値サービスと競合せず、かれらへの高品質なデータの提供者として機能している。このためにエンドユーザーは第3者が競争して生み出す多彩なサービスから最適なものを選択することが出来る。

それ以外にNCBIでのデータ統合と公開を可能にしたのは個々のプロジェクトのデータ公開ポリシーである。例えばヒトゲノムのバーミューダ会議やHapMap計画でのデータ公開ポリシーは極めて厳格なものであり、独占的利用が許されるのはデータ取得後1年のみであった。これらのポリシーは国際協力プログラムで他国からのデータを速やかに集めることが解析力に自信のある米国を利するための政策的側面もあったと思われるが2003年には国内プロジェクトに対しても同様に50万ドル(5000万円)を越えるグラント請求ではデータ共有の計画を明示することと堅持することがNIHポリシーで義務付けされた。

データの流れを作る米国の規則体系は明快である。政府科学事業ではデータは連邦調達規則 (Federal Acquisition Regulation) に従い残らず一旦政府に納品され政府情報となる。一部の例外を除いて内部で作られたもしくは納品された政府情報は情報公開法 (Freedom of Information Access) により速やかに無制限に市民に提供される。しかも OMB-circular A-130 (Office of Management and Budget) によって特に付加価値生産物が作成可能な原材料 (raw content) の配布を強く奨励されている。情報企業や大学はユニークなアイデアと少ないコストでこの生データに対して様々な発見努力や付加価値競争 (個人プロジェクト) を行い、結果として豊富な知財が生み出されエンドユーザーは多様な商品、情報、サービスから好みのものを適正な価格で手にいれることができる。国家情報由来の知財化は全く認められない一方公共データを用いた「個人プロジェクト」では政府研究助成金 (NFS や NIH からの grant) を利用していてもバイドール法によって知財が保護されるために一層発見開発は活気を帯び科学の進歩とともに産業化が保障されている。すなわち、材料は政府が提供し、価値付加競争を知財化して保護することで活発化するという投資価値の最大化を約束する二つの合理がここで成立している。

【解説】米国における科学技術データに関わる制度群 ①精神

米国は世界で最大の科学技術情報 --- 連邦記録 (federal record) と科学技術データベースを含む --- の creator, user, disseminator であり、それは価値の高い国家の資産 (assets) であると考えられている。

米国の情報関連法で貫かれている基盤精神は「情報の創造と伝播と利用の多様さを育むことで民主主義は繁栄し情報の社会的恩恵が最大化される」というものである。もっというと、政府の情報資産から、最大の経済的社会的な恩恵を引き出すために、政府情報資産は市民全員に最も効率的でタイムリーで公平な方法で入手可能にされるべきである、という考えである。

米国の法や政策には原則としてこの考えを盛り込んである。プライベートセクターの財産権を守ることを強調するのと対照的に連邦レベルの米国内の情報政策は ①strong freedom of information law, ②no government copyright, ③fees limited to recouping the cost of dissemination, and ④no restriction in reuse とまとめることができる。

【解説】米国における科学技術データに関わる制度群 ②政府情報

米国の法は明示的に政府の省や機関が著作物に著作権を主張することを禁じこれらの著作を public domain においている。1976年の著作権法(Copyright act)には「以下の著作権保護は如何なる連邦政府の仕事(any work of US government)にも当てはまらない」と記されている。(17USC sec105)理由はいくつか存在し

- (1) 市民が政府の行為や職権の濫用を監視するための open access (情報公開)という考えに著作権が反するということ
- (2) 「市民は同じ情報に対し、2度対価を払わない (taxpayers should not have to pay twice for the same information—once for the cost of generating work, and second time to obtain it)
- (3) 政府の情報を批判するための情報入手を保障し、言論の自由を奪わない
- (4) 市民は public good (例えば公的科学技術データや情報)から利益を引き出すことが可能であるべきで、データや情報へのアクセス機会を増すことで、常に個人の権利保護に有益である教育機会の改善を享受することが可能であるべきである

という4つが主である。

従って連邦議会のポジションはこれまでつねに、政府情報を利用した個人や企業の2次市場の形成を奨励し、また有用な政府情報の配布に注力するものであった。

【解説】米国における科学技術データに関わる制度群 ③受託契約事業について

連邦調達規則(Federal Acquisition Regulation)ではすべての政府機関外の団体との契約に適用される。「データの権利と著作権」の項(27.4 節)には政府との契約に基づいて作成されたデータについての利用や複製、公開についての権利と義務について書かれている。規則(proposition)の主題は政府が契約の履行において作成されたほとんどのデータについて無制限の権利を所有し、受託者がうける権利がいくらかあるとしても非常に限定的なものであるということである。公的なまたは公的資金を受けたS&Tデータの入手性(availability)と交換に的を絞ったオープンアクセスポリシーは、機関や機関間のプログラムや国際プログラムについて、特に環境データと地球科学データについては最近整備された。(ポリシー集は www.gcdis.usgcrp.gov/policies にあり)

【解説】米国における科学技術データに関わる制度群 ④研究助成金について

高等教育機関、病院、非営利団体に対する補助金(Grant)と合意(Agreements)の統一的事務的取り扱い(OMB-circular A-110, article 36(a))には「対象機関は補助金によって開発された著作権主張可能な作品(any work that is subject to copyright)や購入した著作権所有権にたいしては著作権を主張できる」と書かれている。但し、連邦補助金機関(FAA; Federal awarding agency)は公的目的(Federal purpose)には無料にて非独占的に取り消せない(irrevocable)コピーしたり出版したりその他の利用についての権利を有するとされている。つまり政府は自由に使用可能である。また Article24(h)には「FAAの規則や補助金の条件に禁止されていない限り、補助金受賞者は連邦政府に対して著作権由来のライセンス料や著作権、特許、特許応用、商標や発明に対する使用料に関して全く責務を負わない(no obligation)」とされている。

同様に National Science Foundation (NSF)の補助金の手引き(Grant Policy Manual)には 732.1 節に NSF のグラントで作られた copyrightable material についての扱いが書かれており

- a. NSF は以下の権利に限り獲得する: copyrightable material が連邦政府のポリシーの目的を達したり、要求を満たすために必要である権利
- b. インセンティブを保つために NSF は copyrightable material からの収入を要求しない
- c. 例外的な場合には NSF は補助金受賞者の権利を限定したり排除したりする。プログラムやグラントの目的にそうすることが最適であるとNSFが判断した場合はそれにあたる

共同研究開発合意(Cooperative Research and development Agreements, CRADAs)はさらに database やサービスが政府の要求に見合うような法的メカニズムを提供する。CRADA 規則は一方で FIAの例外を作る。CRADA の元で作られたDBは FOIA のに基づく公開要求から取り下げられる可能性がある。CRADAとは 15 U.S.C. sec.3710a に定める「連邦研究所から人員や施設が提供されて、研究所の使命にかなう研究を行うために交わされる一つもしくは複数の連邦研究所と外国の研究所の間の研究契約。

(2) ミッションステートメントに対する達成度

ミッションステートメント

1) 関係府省におけるDB統合化に向けた取り組みの補完となる関係府省の制度設計やロードマップ作成に資する試案の提示。

試案の検討に当たっては、外部有識者を含む研究運営委員会において議論を行ったが、現状認識のずれから問題意識を共有することに困難があり、十分に合意形成を行うことができなかった。そのため、本調査研究においては制度設計のみの研究代表者試案を提示するに留めた。

2) 複数の国内主要 DB を統合化する際の技術的課題並びに解決策の提示及びフィージビリティ・スタディを通じた実現性検証結果の提示。

達成した。

3) 国内外の医学分野・学術分野 DB、国内の産業分野 DB に関する技術的側面、制度的側面からの基礎調査結果の提示を行う。

達成した。

(3) 当初計画どおりに進捗しなかった理由

研究計画では課題の技術的側面および技術的なフィージビリティ・スタディに人的にも時間的にも過剰なウェイトを置いたため、それ以外の調査研究について、内容の具体例を伴う詳細分析に至るには十分な時間がなく、研究分担者との理解の程度の差を生むと共に、研究運営委員会において委員との現状認識等の差を生み、共通の認識に至らない結果となった。

(4) 研究目標の妥当性について

目標として設定されたミッションについてそれぞれ述べる。

1)の試案の提示に関しては、前提となる関係府省におけるDB統合化に向けた取組自体がまだ固まっていない部分もあり、現状認識に多くの時間を割かざるを得ず、関係府省における取組の補完となるべき試案としては、制度設計に関する研究代表者の試案を提示するにとどまった。前提となるべき状況が達せられていない段階は、ロードマップの提示の目標は、時期尚早であったと考える。

2)の実現性検証に関しては、現状に即した個別の問題を扱うものであったので、具体的な目標を立てながら実施できており、妥当であったと考える。

3)の基礎調査に関しては課題の実施の過程において、当初の想定よりも制度面での調査が必要であることが明らかになってきたため、途中で重点となる調査項目を変更しながらの実施であったが、おおむね妥当な目標設定であったと考える。

(5) 研究計画・実施体制について

研究計画では課題の技術的側面および技術的なフィージビリティ・スタディに人的にも時間的にも過剰なウェイトを置いたことは適当ではなかった。

また、DB統合における問題点として、当初想定していたよりも技術以外の側面である機構と管理(Institution and management,)、予算と経済、法と政策、文化行動、の4側面に対する調査研究が重要であるということが明らかになってきたが、調査研究の担当者の中にはこれらの諸問題について専門に扱うメンバーを当初想定していなかったことから、実際の調査は事務局と遺伝研が担当することとなった。手法としては、主に机上の調査と有識者との面談で行ったが、主に時間的な制約からその程度には不十分な面が残った。また、これらの調査の実施が、主に最終年度に集中していたため、研究分担者の多くが深く理解する機会や時間が限られており、研究運営委員会において委員間の現状認識等の差を生む結果となり、又、研究運営委員会の改組も時間的余裕がなかったため、困難であった。

(6) 研究成果の発表状況

4) 研究発表件数

	原著論文発表 (査読付)	左記以外の誌面発表	口頭発表	合計
国 内	該当なし	4 件	8 件	12 件
国 外	5 件	1 件	該当なし	6 件
合 計	5 件	5 件	8 件	18 件

5) 特許等出願件数: 該当なし

6) 受賞等: 該当なし

7) 原著論文(査読付)

【国内誌】(国内英文誌を含む)

該当なし

【国外誌】

1. Budrul Ahsan, (36 authors) and Shinichi Morishita. UTGB/medaka: genomic resource database for medaka biology. *Nucleic Acids Research* Advance Access published on October 11, (2007)
2. Taro L. Saito and Shinichi Morishita. Efficient Integration of Structure Indexes of XML. *Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA)*, April 2007, Bangkok, (2007)
3. Taro L. Saito and Shinichi Morishita. Amoeba Join: Overcoming Structural Fluctuations of XML Data. *Ninth International Workshop on the Web and Databases (WebDB)*, (2006)
4. Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, and Ito T. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A*. 103(47):17846-51, (2006)
5. Tomoyuki Yamada, Haruhiko Souma, Shinichi Morishita. PrimerStation: a highly specific multiplex genomic PCR primer design server for the human genome. *Nucleic Acids Research* 34: W665-W669, (2006)

8) その他の主な情報発信(一般公開のセミナー、展示会、著書、Web等)

①Web 公開:「生命科学データベース統合に関する調査研究」, <http://133.11.132.80/intelligence/>(暫定公開サイトのため、今後変更する可能性あり), 2008.03.31 現在

主に、遺伝研事務局が作成した調査資料、委員会資料を掲載

研究成果:サブテーマ毎の詳細

(1) サブテーマ1(1) 生命科学一般のデータベース統合にかかわる調査研究

(分担研究者名: 大久保 公策、所属機関名: 国立遺伝学研究所)

1) 要旨

ライフサイエンス分野における知的情報基盤の設計に資する情報資源についての客観的な資料となるデータベースの現状については、特に国内を中心として調査を進め、その分析も行った。また、統合化を進めるにあたって考慮すべき、データベース関係者(インフォマティクス関係者並びにユーザーである研究者、特に創薬分野の研究者)からの意見を収集し、今後のデータベース統合化においてひとつの方向性を示す情報を得た。

技術的課題を洗い出すための統合化のモデルとしてのフィージビリティ・スタディでは、統合対象となる分野ごとの分類情報や、情報の構造化するための基本モデルの必要性を指摘した。

一方で、統合の素材となる学術情報の流通に関して、国内の様々な施策等を調査し、現行ルールでは情報の自由な流通が妨げられているケースもあることを指摘した。このような状況を改善するために国外で行われている施策についても調査し、今後、国内でどのような対策を行うべきかについても提言に資する試案としてまとめた。

2) 目標と目標に対する結果

本サブテーマでは、基礎生命科学分野の観点からデータベース統合化のための調査を行うために、基礎生命科学分野における遺伝子・蛋白質・遺伝子発現・文献などの国内主要データベースについて調査を行い、その結果に基づき試験的統合の対象を決定し、統合の技術的な検討を続け実際の統合を開始し、技術的問題及び維持管理上の問題を探ることを目的の一つとしている。リストについてはかなり網羅性の高いものを作成することができ、データベースの型分類や、統合作業の容易さなどを検討したので、今後具体的な統合作業において役立つ情報をまとめることができた。

また、我が国のライフサイエンス分野における知的情報基盤の設計に資する当該分野の情報資源についての国内海外の状況についての客観的な資料を提示するという目標に関しては、様々な視点からの調査を実施したが、調査の過程で、データベース統合に必要なデータの流通に関してより詳細な調査を行う必要が生じてきたことから、後半では主に、制度やルールといった社会的な仕組みについての検討を行った。

さらにこれらの成果を活かした統合データベースの機能や制度についての試案を作成し、実現に向けたロードマップ作成に資する試案を作成するという点については、分担研究者としてのではあるが、根本的なルール修正の必要性を指摘することにより、今後のデータベース統合に向けて必要な考え方を提示することができたものと考えている。

3) 研究方法／調査方法

国内のデータベースの現状を把握するに当たっては、文部科学省関連については、すでに実施されていた国内でのデータベースの包括的な調査であると思われるいくつかの調査や、(独)科学技術振興機構(JST)、文部科学省科学研究費特定領域等のホームページを情報源とした。その他の省庁関連のデータベースについては、本調査研究に参画している各分野の研究者からの情報を加えて、さらに足りな

い部分については、個別の公開情報を集めて用いた。

各分野の専門家からの聞き取り調査は、事務局へそれぞれの研究者を招いて、あるいは、合宿形式でこちらの提示した資料をベースに議論していただくことで、意見を収集した。

統合における技術的な検討課題の洗い出しのためのフィージビリティ・スタディとしては、モデルケースとして、植物 EST データを基礎生物学的なデータベース作成の手法を用いて整理することを試みた。また、情報整理の手法検討のため、3次元に情報をマップする際のモデルとして人体モデルを用いたときの問題点を検討した。

制度や政策に関する調査では、主に、WWW や白書、報告書の形で公表されているものをベースに情報をまとめ、具体的な問題点については、現場で実際に問題を感じている研究者等にヒアリングしてまとめた。

4) 研究結果

(A) 国内のデータベースに関する状況

a) 調査データベース活動の現状

現在、データベースと総称されるなかに明確な型が区別される。これらの異なる型のデータベースは異なる工程によって作られ異なる方法で維持運営されている。従って統合化や援助を考える上で、国内のデータベース活動も構築法からみた型分類を行って対処すべきである。表 1.1 に構築法からみたデータベースの型別に構築後の同時性と網羅性を保つための維持作業をまとめた。構築維持の共通の特徴は専門的知識を背景に一定の作業を行う“アノテーター”の役割が大きいことである。アノテーターの業務にも、電子化情報の書式などの構造を理解して主に電子化情報の形を変える処理をする情報アノテーターと論文や総説などの内容を理解して記述を行う生物アノテーターが区別される。

表 1.1 構築法からみたデータベースの分類

型名称	情報源の種類	処理方法	処理主体	データ形式	例
バンク型	測定器と登録者	-	不特定多数	構造化テキスト	DDBJ
プロジェクト型	測定器と実験者	-	特定人間	構造	BodyMap
プログラム型	データベースレコード	機械的処理	マシン	構造	UniGene
キュレーション型	データベースレコード	高度情報処理	特定人間	構造	SCOP
知識モデル型	読み物(電子・紙)	高度情報処理	特定人間	構造	KEGG
総説型	読み物(電子・紙)	高度情報処理	特定人間	構造化テキスト	OMIM

国内のデータベース構築活動をより詳細につかむために、分子生物学施設または歴史のある複数のカタログサイト(NAR Database Collectionをはじめとする安定した分子生物学系の代表的カタログサイトを15サイトリストアップ、表1.2)に掲載されている国内のデータベース情報に加えて、国内データベースに関する過去の省庁主体の調査結果および本調査の分担者の調査を統合し、日本のデータベース構築活

動の特徴を抽出した。

文部科学省関連の基礎生物学分野のデータベースをリストアップするにあたっては、過去の調査としては国内でのデータベースの包括的な調査であると思われる、知的基盤整備委員会平成16年11月見直し(うち生物・成体に関する計測などと分類されたもの)、平成16年度学術情報データベース実態調査、JST(独立行政法人 科学技術振興機構)のBIRD(バイオインフォマティクス推進センター)のWINGホームページ、さらに文部科学省科学研究費特定領域「ゲノム4領域」および「発生システム」のホームページを情報源として用いた。これらに加え、他の省庁に関して担当分担者によるもの【農水省】、分担者調査にかずさDNA研を加えたもの【経済産業省】、事務局独自調査【厚生労働省】による調査結果を追加した。網羅性は保障されていないので型別の相対的な関係以外省庁間の差には意味がない。

表 1.2 調査対象としたカタログサイト一覧

ディレクトリサイト	構築機関	掲載 DB
NAR DB Collection	Oxford Journals	853
ExPASy	Swiss Institute of Bioinformatics	1, 109
Genexpress	CNRS (France)	1, 109
bioinformatik.de	Dr. Andrea Ramge (private)	172
MetaDB (neurotransmitter)	Shawn Thomas (private)	1, 162
bionetbook	Institute Pasteur	8, 191
DBcatalog	Infobiogen (France)	487
NIH	NIH	58
Sanger Institute	Sanger Institute	119
EBI	EBI	5
Biology Links	Harvard Univ.	358
TIGR related links	TIGR	81
Baylor College of Medicine	Baylor College of Medicine	99
Stanford Univ.	Stanford Univ.	205
DDBJ	DDBJ	300
計 15 ディレクトリ		14, 308

表 1.3 国内データベース型分類

	文部科学	経済産業	厚生労働	農林水産	計
バンク型	8	0	0	5	13
プロジェクト型	62	21	6	11	100
プログラム型	40	8	0	0	48
キュレーション型	1	1	0	1	3
知識モデル型	15	4	2	1	22
総説型	6	0	0	0	6
分類・稼動確認不能	58*	0	0	0	58
	190	34	8	18	250

以上により得られた国内のデータベースを対象に型分類すると、8 割弱がプロジェクト型とプログラム型である(表 1.3)。これらは維持するうえで研究者負担の少ない型であり、統合による維持作業の軽減の考えやすい対象である。

*分類・稼動確認不能の 58 件には、データベース構築の報告等があるが、上記の分類には該当しないもの、今後公開予定であるためあるいは何らかの原因で調査当時にアクセス不可能であったもの、一般には非公開であるためアクセス不可能なものが含まれる。

現行の大型測定プロジェクトは重要な分子データを生産しているものが多い。総合科学技術会議配布資料、関係省庁の予算関連公開資料、理化学研究所およびバイオ産業情報化コンソーシアムのホームページを情報源に、現行(および終了直後)の遺伝子・蛋白質に関わる大型(1 億円以上)測定プロジェクトを列举して(図 1.1)、今後公開されるであろうデータについて分析を行った。その特徴は総ての大型プロジェクトが独自のプロジェクトデータベースを計画もしくは構築している点である(図 1.2)。生産されるデータでは立体構造データおよび遺伝子発現データ、立体構造データ多型データなど外部に受け皿となるデータベースがあり集約管理可能なものも多いが、蛋白質の相互作用、薬剤との相互作用などプロジェクト内での管理法の開発を要するデータも少なからず見られる。

プロジェクト名	H12	H13	H14	H15	H16	H17	主な産生データ
文部科学省							
ゲノムネットワーク							転写開始に関わる情報
タンパク3000							蛋白立体構造情報
遺伝子多型研究							SNP及びアレル頻度
テラーメイド医療実現化							SNPと薬剤、疾患に関わる情報
バイオインフォマティクス研究							
経済産業省							
データベース結合							
ゲノム情報統合							
完全長cDNA							ヒトcDNA配列情報
蛋白質機能解析							遺伝子発現、蛋白間相互作用
生体高分子立体構造							蛋白立体構造
標準SNP解析							SNPおよびアレル頻度
厚生労働省							
疾患ゲノムデータベース							疾患ゲノムデータベース
トキシコゲノミクス							遺伝子発現と薬剤に関わる情報
疾患関連蛋白質							疾患と蛋白発現に関わる情報
農林水産省							
イネゲノム							イネゲノムに関わる総合情報
家畜ゲノム							ブタのcDNA配列情報
蚕ゲノム							蚕のゲノム、cDNA配列情報

図 1.1 生命科学分野の主要国家プロジェクト

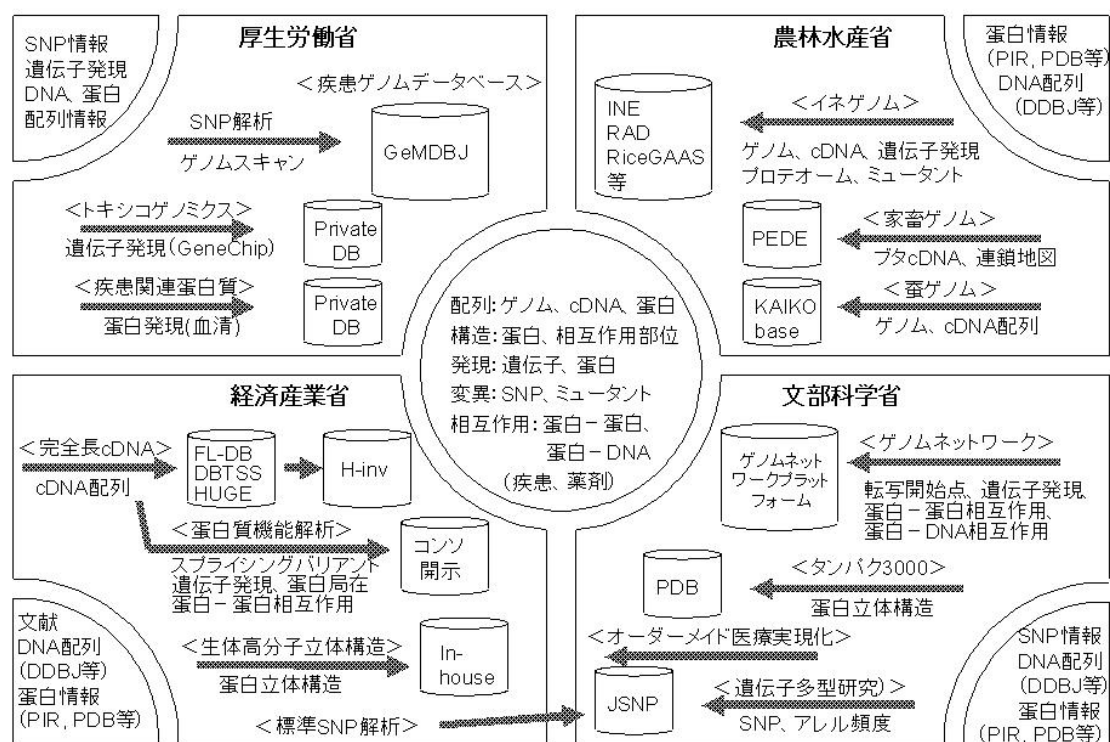


図 1.2 生命科学分野の主要国家プロジェクトとデータの出口

これらの調査により、国内データベースの全体図がほぼ明らかになり、今後統合化を進めるにあたって、対応方針を検討するための基礎となる貴重な資料を提供することができた。

b) データベースの作成あるいは利用に関する聞き取り調査

生命科学分野のデータベースの抱える問題点を多角的に把握し、データベース統合化への方向性を探るために、システム生物学、分子生物学、創薬、機器開発分野からデータベースの作成者もしくは利用者たる研究者を事務局にお呼びして、研究代表者が面談し意見を聴取した。協力者は総計23組27名である。

また、合宿形式で参加者の意見聴取を行う会議を実施した。データベース統合化の調査が分野と離れたものにならないように情報共有する為、現状の改善を常に考えており最新の動向にも詳しいと期待されるバイオ情報処理分野の中堅若手を中心に、実験生物学に精通しておりさらに情報リテラシーの高いベテラン実験研究者も参加者として招待した。分子生物学2名、動物学の1名、医学・医療4名および創薬3名、分子進化2名、人類遺伝1名である。会議の方法としては、事前アンケートであらかじめ質問に回答していただき、意見のまとめを全員に示しながら以下の項目につき議論を深めた。

- ①データベースの現状や統合化に対する意見
- ②データ駆使による研究のイメージ
- ③ゲノム配列利用を促進するためには
- ④発現情報の利用を促進するには
- ⑤その他の情報の利用促進
- ⑥現状のDBのここがいけない
- ⑦分散か集中か
- ⑧統合化調査に対する自由意見

これらの聞き取り調査の結果を踏まえた生命科学分野の研究者の意見に分析を加えると以下のようにまとめられた。

・生命科学分野の価値は多様である

生命科学では多細胞生物の個体を構成する体細胞は総て同一のゲノム配列を持つという法則と免疫担当細胞ではゲノム配列の組み換えが起こる“例外”が同程度に重要である。即ち生命科学は、詳細を排し抽象化を進める自然科学としての側面に加えて、人の健康や福祉を増進させるための実学としての側面を持つ。したがってデータ(ベース)にも法則の発見のための基礎データとしての性格と、特定の事例を参照するためのアーカイブとしての性質が要求される。

・構築者と利用者の深刻な断絶

データベースの課題や統合化に対する意見は「現状に満足で統合化に反対」という構築者側と「現状に不満で統合化を望む」という応用分野の研究者の間で開きが大きい。また応用分野の研究者間では要望や将来像に異なる意見を持つ場合が多い。情報に親和性の高い研究者は生命科学の個別の問題に疎く過度に単純化されたモデル理解に基づいて思考しており、個別の問題に取り組んでいる研究者は情報に対するリテラシーが充分でなく処理上の問題を過小評価し現状に不満が強いことが、意見が異なる原因のように感じた。つまり思考思想の違いによる断絶よりは、相互の知識の不足による意見の違いであると思われる。両者の背景知識をより広く共通なものにすることが、データベースをより役立つ統合されたものへと変えてゆく解決法である。これまでも指摘されている課題であるが、データベース統合に関しては一層積極的な方策が必要である。

さらに、本研究の対象としているデータベースの主たる利用者である創薬関連の研究者からの聞き取り調査を行い、統合データベースの主要な利用現場の1つである創薬のプロセスと情報環境についてまとめた(図 1.3、1.4)。

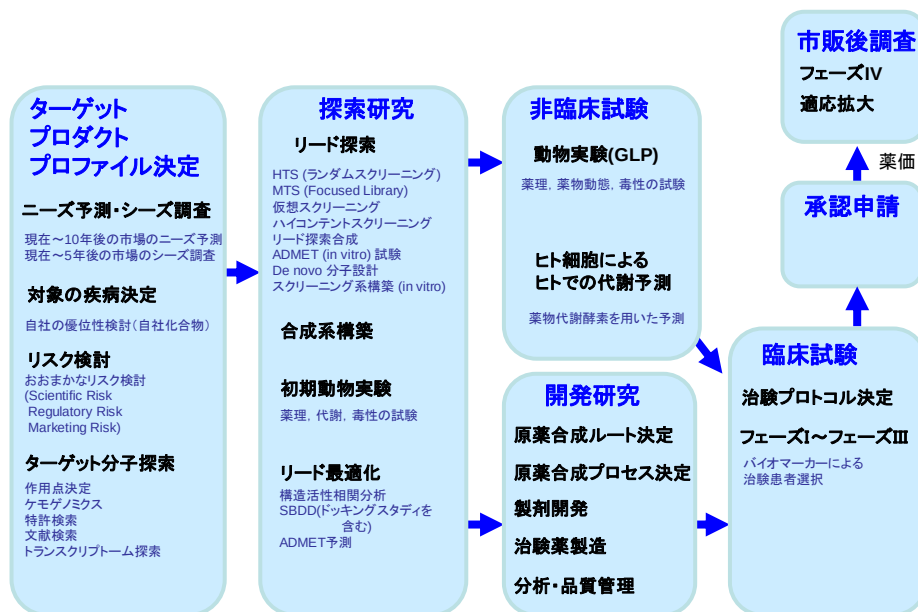


図 1.3 新薬開発のプロセス

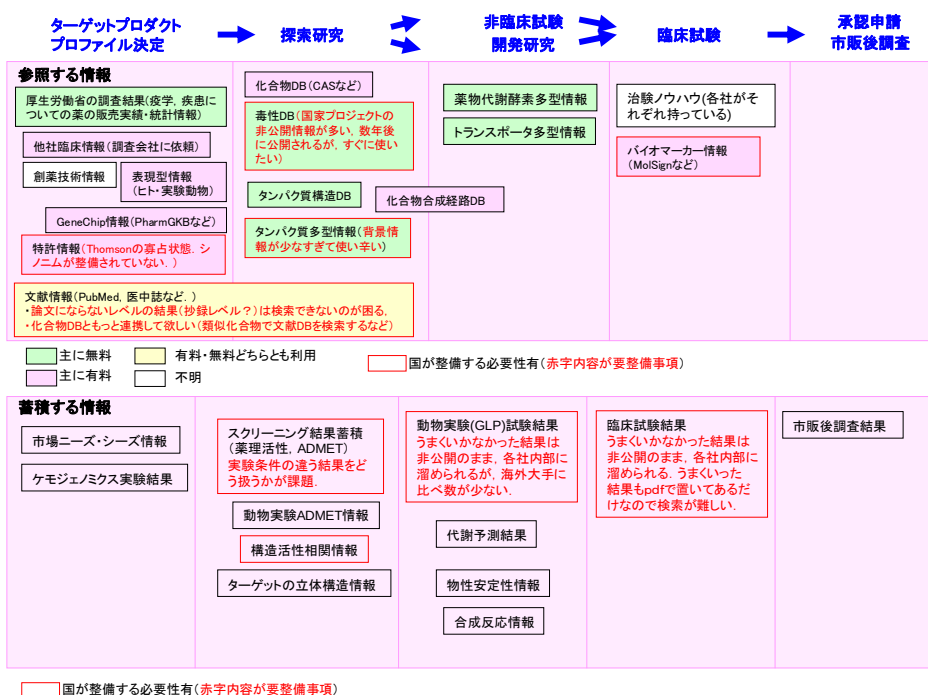


図 1.4 新薬開発のプロセスで参照する情報と蓄積する情報

創薬は「判断・開発・判断・治験」のプロセスからなっており、開始の判断は、効果のデザインであり、「どの病気にどんな効果を(TPP:開発薬剤プロファイル)」を決定する。次に、効果の実現である薬剤開発、前臨床を行う。ここでは、デザインされた効果を生む為の分子の設計と合成を行う、実験室での創薬プロセスである。次に第二の判断である中止の判断を行う。ここでは毒性の予見を行う。前臨床試験までのプロセスで落第を予見して開発中止を判断し、コストのかかる失敗治験をしないようにする。この判断を通

たものだけが、臨床試験(治験)に進むことになる。開発や治験の成功率、即ち【判断力】が企業の力の本態である。製薬企業は化学企業ではなく情報処理企業であり、外注可能なのはむしろ開発と治験である。ここでは、情報の質・量が判断の質を決めることになる。

開発開始の判断には、対象疾患決定のための情報である疫学情報、臨床治療情報、他社開発状況、作用点の決定のための情報である文献情報(病理病態研究)、特許データベース、分子生物データベース(遺伝子発現データベース、パスウェイ・相互作用データベース、遺伝子カード、マウス遺伝子情報などが重要な役割を果たす。治験中止の判断には、前臨床試験(動物・培養細胞実験)の結果が重要で、ここでは、PubChem、トキシコゲノム情報への期待、文献情報全般への期待が大きい。また、自社内の過去の治験帰結情報や、一部は文献情報から入手可能な他社の治験帰結情報が必要である。すなわち、分子データの利便性向上だけでは効果の期待が薄く、疫学情報・診療情報治験情報などが莫大な費用を要する治験の開始/継続/中止の鍵となる。

c) 医学・疫学系データベースの調査分析

方法

疫学調査についてのガイドとなるいくつかのウェブサイト(下記、参考ウェブサイト)を中心に、「疫学」「コホート」などをキーワードとしてWWW 検索を広げ、目的、対象地域・人数、取得するデータの種類、取得されたデータの公表・活用状況などについて調査した。

＜参考ウェブサイト＞

Tsubono Report(世界と日本の大規模コホート研究) <http://www.metamedica.com/news2002/howto10.html>

自治医科大学「疫学の研究の倫理に関するページ」内「疫学と社会に関するリンク集」
<http://www.jichi.ac.jp/ethics/link.html#10>

循環器疫学サイト

<http://www.epi-c.jp/>

結果

主に、国内の機関が携わっている(いた)45件のコホート研究、疫学調査についての情報を収集し、プロジェクト名称/開始・終了(予定)時期/調査対象集団/対象者数(有効)/予算元/中心機関(協力機関)/目的・成果発表の状況/公開・データ外部利用の状況/収集データ及びサンプルの項目・種類について整理した。

対象者数が明確な(含む計画)調査40件について、開始年と人数規模の関係、ゲノム関連の情報取得の有無をグラフにした(図1.5)。

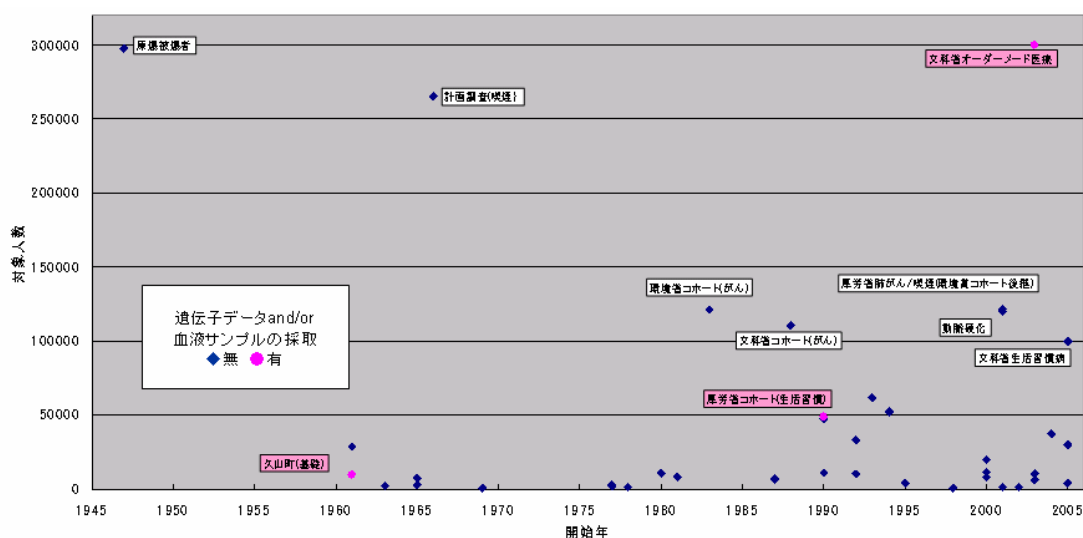


図 1.5 疫学調査の開始年と対象人数

10 万人以上および遺伝子データや血液サンプルを採取しているスタディについては名称や調査対象がわかるようラベルした。いわゆる大規模コホートは、省庁あるいは国立の機関主導で実施されている。遺伝子データの取得されている(できる)スタディは少ない(久山町は最近になって一部のゲノム情報の取得を開始したらしい)。

まとめ

- ・疾患別などに一部まとめられているが、疫学調査全体を体系的に調査できる資料はなかった。
- ・いわゆる大規模コホートは、省庁あるいは国立の機関主導で実施されている。
- ・調査結果のDBの構築・公開については、あまり行われていない。
- ・ゲノム情報については取得されている(取得できる)ケースがほとんど無い。
- ・結果の公表は、論文等統計的な処理を施されたものがほとんど。生データへのアクセスは不可。
- ・データの再利用については、研究班に新たに参加することが必要と思われる。
- ・データの再利用については、倫理面でのコンセンサスが必要(文部科学省・厚生労働省からH14.6.17に「疫学研究に関する倫理指針」が、日本疫学会からH14.10.25に「疫学研究を実施するに当たっての倫理指針」が出されているが、過去に取得されたデータの扱いを含め、実施者間での共通理解が必要である)。

(B) 統合化におけるフィージビリティ・スタディ

a) 農業・基礎生物統合フィージビリティ・スタディ

ここでは、農業分野と基礎生物学分野のデータベースの統合の際に必要なと想定される、有用な生物分類の必要性の実証を目指した。農業側の問題意識としては、植物ではいくつかのモデル植物以外公的なデータ整理が進んでいない、特にモデル生物情報の育種への応用が未着ということがある。また、ヒトをターゲットとした基礎生物学分野から見ると、植物は分類方法によってはカビ・シダなども一緒になっているので、種ばかり多くて整理に手をつけにくい、といった問題意識があった。そこで、植物 EST データの透明化による農業・基礎生物統合フィージビリティ・スタディを試みた。ここでは、植物データを非植物研究者や育種関係者が理解できるレベルにすることをターゲットに、比較的イネ麦などに近縁の植物(被子植物)と裸子、シダ、カビの分離、裸子植物に関して横断整理に必要な材料の提供、植物解剖にも基づいた植物材料自動分類機の設計を行った(図 1.6)。なお、本フィージビリティ・スタディの成果は、文部科学省統合データベースプロジェクトで植物ボディーマップとして事業化すると同時に、農林水産省データベース統合プロジェクトのイネゲノムアノテーション更新時に発現データとして利用可能とした。

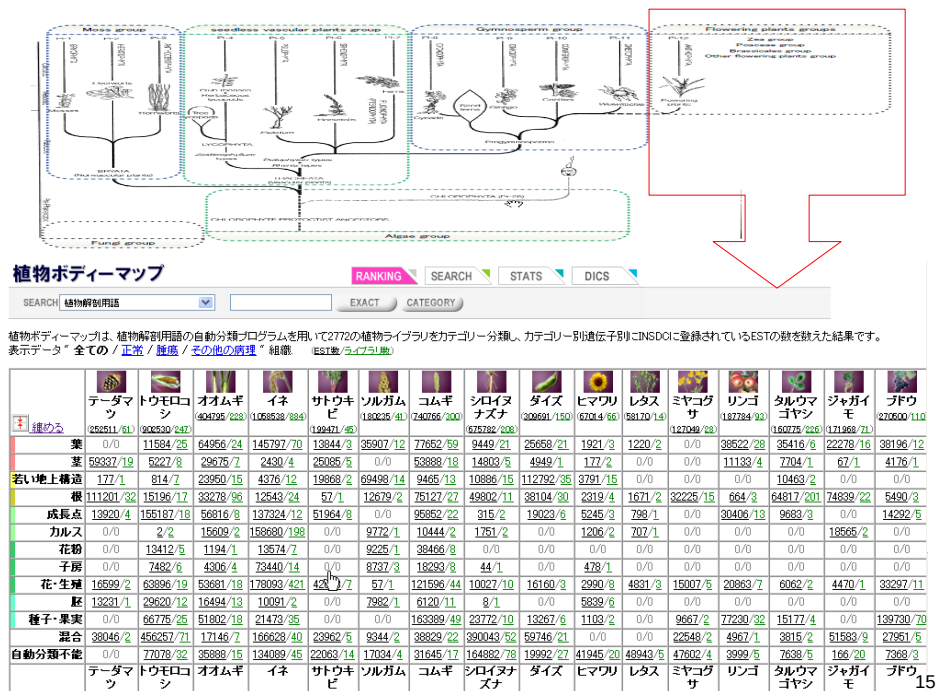


図 1.6 農業・基礎生物統合フィージビリティ・スタディ:植物 EST データの透明化

b) 解剖学用語 3D 辞書の構築

既存の解剖学オントロジーの問題点として、(i)オントロジー表現は形、位置情報が欠落している、(ii)解剖学用語間の関係は、単純な階層構造ではない、(iii)人手でひとつひとつ知識を入力する必要がある、といったことが挙げられる。これを 3D 表現にすることにより得られる効果として、(i)可視化により形、位置関係を直感的に理解可能、(ii)解剖学用語を座標でシンプルに表現可能、(iii)座標間の計算による知識の取得、が期待される。本フィージビリティ・スタディでは、数値人体モデルデータベース((独)情報通信研究機構が、北里大学、慶應義塾大学及び東京都立大学と共同開発した電磁波影響計算のための Voxel モデルファントム、分類組織数約 50)を 3D データの基盤とした。解剖学用語リストは、京都大学で開発されたライフサイエンス辞書から解剖学用語約 3,000 語を抽出し、PubMed アブストラクト出現頻度に基づく追加優先順位決定作業を行った。得られたボクセルマンの画像を図 1.7 に示す。本フィージビリティ・スタディで 3D 化の有効性は示されたが、いくつかの問題点も指摘されたため、今後のその改良は、文部科学省ライフサイエンス統合データベースプロジェクト内で検討されることとなった。

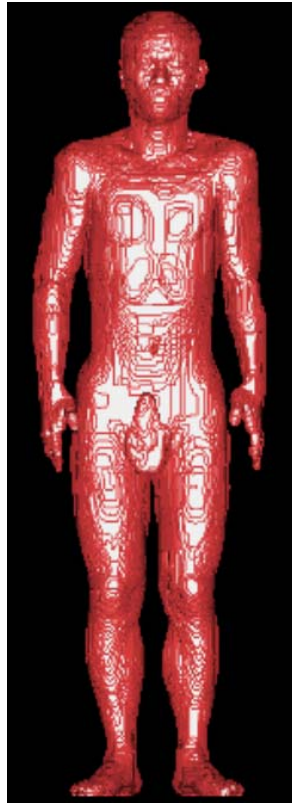


図 1.7 本研究により得られた人体モデル3D 画像

(C) 現行社会制度の課題検討

a) 日本の学術情報関連政策(米国との比較を含めて)

我が国の学術情報関連政策の概要を図 1.8 に示した。

我が国の科学情報政策は、戦後まもなく図書館の整備が行われ、また、政策の具体的な内容を検討するための日本学術会議・科学技術行政協議会が設置されたことから始まった。その内容の主要な項目の一つである科学技術情報の全国的流通システム(NIST)構想については、1963 年に米国で提出された Weinberg 報告書に少なからず影響されたものと考えられる。

80 年代は、米国からの情報流入が多く、一方的な状態であったが、近年では科学振興政策が功奏して、徐々に外国誌への投稿も増加している。現在では、外国誌への投稿そのものが国内の研究費の流出に当たるという批判も出てきている。

1986 年に出された「科学技術政策大綱」(92 年に一部改正)を経て、1995 年には科学技術基本法が施行され、その後 5 年ごとに科学技術基本計画を策定し、見直しをかけることで、科学技術政策の方針を決定している。

国内の学術情報収集の要となる、国立国会図書館(NDL)、科学技術振興事業団(JST)、国立情報学研究所(NII)が次々に設立された。これらに関する科学技術情報関連の施策については、国の政策や計画の中にたびたび盛り込まれ、近年では、大学図書館(「外国雑誌センター館」を含む)も交えて、協調して行動を起こすようにもとめられて来た。

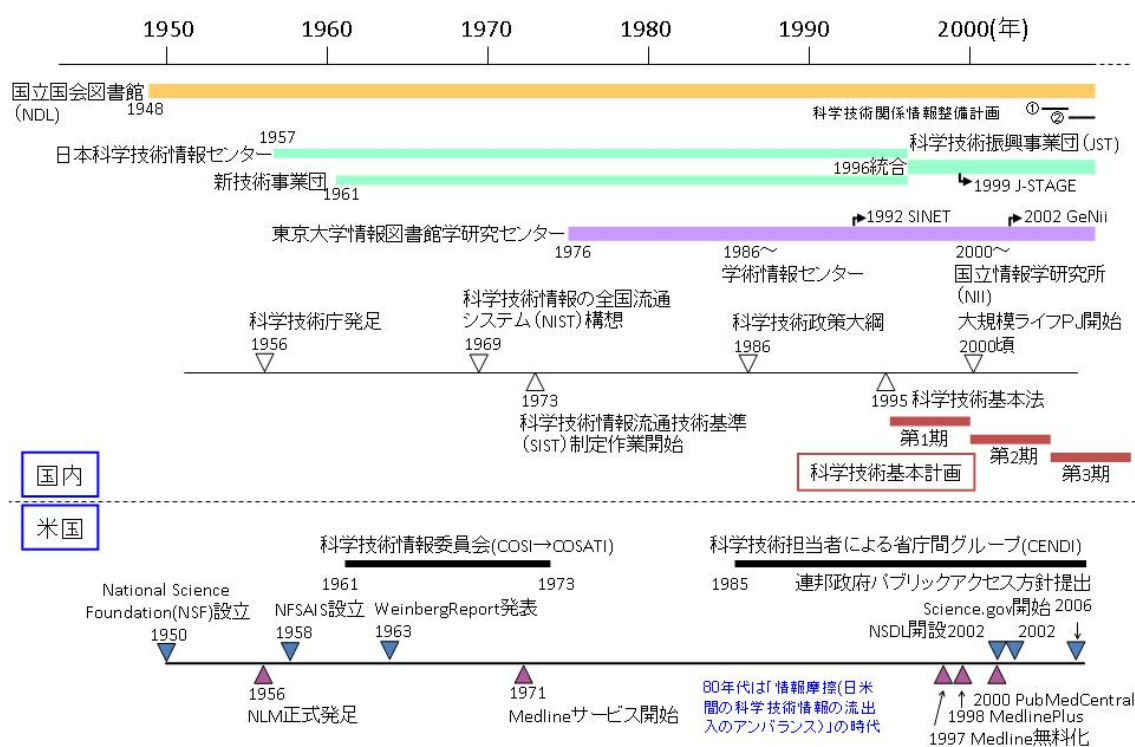


図 1.8 我が国の学術情報政策の概要

一方、米国では、戦後まもなくに全米科学財団 National Science Foundation(NSF)が設立され、学術情報の普及サービス the National Federation of Science Abstracting and Indexing Services (NFSAIS)が開始された。

科学技術情報委員会 (COSATI)は、連邦政府予算による科学研究や技術開発の結果が増大してきたことを受けて1961年に設立された。COSATIの活動期間中に出されたWeinberg報告書はアメリカ大統領の科学諮問委員会の報告書であり、情報の生産者である科学技術界、情報を管理する側である政府諸機関それぞれに対して、研究開発と情報との問題を明示し、将来のビジョンを示して、それに対する各界の協力を求めたものであった。当初の目的を遂げたCOSATIが70年代に自然消滅し、1985年には新たにCENDIとして発足、後に多くの省庁の情報部門が参加して、連邦政府資金による研究成果を公表する場所として、science.govが公開された。

その後、パブリックアクセス法案が上院に提出され、連邦政府資金による研究成果等をすべて公共のものとする努力が払われている。

一方、ライフサイエンスに限れば、既に母体のあったNational Library of Medicine(NLM)が正式に1956年に発足している。医学文献の検索手法もMedlarsのオンライン化(1971)、Web上無料化(1997)、一般向け情報サイトが公開(1998)、論文の全文提供サイトの公開(2000)が実施された。NLMは米国内でもかなり進んだ情報統合の例であると、Weinberg報告書でも述べられ、その後も情報統合のモデルケースとして先頭を走ってきたことは間違いなく、米国内でもMedlinePlusを含むNLM全体をお手本として各分野での情報統合が進められているようである。

b) 情報流通に影響を及ぼすその他の制度やルール

我が国の知的財産資源の有効活用を図る目的で実施されてきた一連の産学連携施策のうち、平成11年に発効した「産業活力再生特別措置法(いわゆる日本版バイドール法、後に関連条文は平成19年に「産業技術力強化法」へ移管された)」は、国の委託研究成果を実施機関へ移転することにより、研究成果の実用化を推進することを目的としている。しかしながら、研究成果の一部としてのデータの流通の観点から考えると、研究成果の権利化のためとして、研究終了後も一定期間は実施機関の中だけでデータが活用される、という状況を生み出すことにもなりかねない。研究者のアイデアが盛り込まれた知財と、科学的な事象を観察したデータを明確に切り分け、他者も利用可能な観察データに関してはその流通が円滑に速やかに図られるよう、実施機関に対する知的財産の優先権を確保しながらも、何らかのルール付けによる具体的な取り組みが必要と考えられる。

また、特に疫学研究において問題になっている個人情報保護法の影響に関しては、調査実施時に各人からデータやサンプルが収集される時点と、一度集められたデータやサンプルが再度解析される時点の2つのケースで問題となっている場合がある。前者では、同意を得る必要があることはもちろんであり、これについては現行の様々なプロジェクトにおいてもルールが遵守されている。後者の場合は、原則本人の同意が必要である、とされているものの、一方で個人情報を除外するという条件のもと再利用可能とされているが、実際には最初にデータを収集した研究者側がその後の影響を懸念して情報(およびサンプル)を2次利用者へ提供しないと決定するケースがあり、研究資源の有効活用の点から考えても何らかの改善が必要である。この改善に関しては、ルールの見直し以外にも、研究者や情報及びサンプルの提供者である一般国民の理解が必要であり、個人情報保護の原則を遵守しながら、理解推進のための活動をしていく必要があると考えられる。

c) 米国国立衛生研究所(NIH)におけるデータ共有ルール

政府資金によるデータの共有化ルールを明確にしているNIHの例を以下に挙げる。NIHでは、データ共有(Data Sharing)についていくつかのポリシーが挙げられており、内容を見てもGrant(研究補助金)とContract(契約)が一緒になっているのでわかりにくいですが、どのような案件にどの条例やポリシーが適用されるかを判断するためのチャート(表1.4)も用意されている。

表 1.4 NIH 提供資金によるデータ共有に関する規則/ポリシー/ガイド等

(<http://grants.nih.gov/grants/sharing.htm>よりData Sharing Regulations/Policy/Guidance Chart for NIH Awardsチャートから抜粋・翻訳)

適用規則等	対象者(範囲)	適用内容
OMB Circular A-110	高等教育機関、病院、非営利団体は適用	Grant等の手続きにおける連邦機関共通の統一基準の提供が目的 連邦政府の権利: grantによる作成データについて取得、複製、出版、あるいは使用 連邦政府が第三者に与える権利: 上記のデータを授与、複製、出版、あるいは使用させること
	民間は非適用 州あるいは地方政府は非適用	
OMB A-110適用の場合		
Shelby Amendment (for OMB A-110)	Grantその他の財政援助を連邦政府から受けた高等教育機関、病院、非営利団体に適用 以下の“データ”に適用 1. 全部あるいは一部を連邦政府資金により援助されたプロジェクトで産出、かつ 2. 法律と同等の実効性を持つ連邦機関により定められている場合	本Amendmentは、連邦資金により産出されたデータが情報自由法(FOIA)に定められた手続きにより公になることを、連邦資金提供機関が保証するよう、要求している。 連邦機関がデータの要求者に対して費用を請求することを許可する
NIH Data Sharing Policy	単年度\$500K以上の申請あるいは公募情報に特に指定されている申請(2003年10月1日以降)	申請にデータ共有についての計画を盛り込むこと NIHがプラン等を容認したことを資金提供の契約文書にもりこむ
NIH Grants Policy Statement: Access to Research Data	すべての申請	NIHより情報共有に関して期待されるものが提示
OMB A-110非適用の場合		
	民間、州/地方政府	州/地方政府に対してはOMB A-102 “Grants and Cooperative Agreements with State and Local Governments” が適用 民間企業等にはFAR 31.2 (48 CFR) “CONTRACT COST PRINCIPLES AND PROCEDURES” が適用
NIH Data Sharing Policy	単年度\$500K以上の申請あるいは公募情報に特に指定されている申請(2003年10月1日以降)	申請にデータ共有についての計画を盛り込むこと NIHがプラン等を容認したことを資金提供の契約文書にもりこむ
NIH Grants Policy Statement: Access to Research Data	すべての申請	NIHより情報共有に関して期待されるものが提示

それによると、単年度の直接経費が\$500k以上かどうかで適用されるポリシーも変わり、\$500k以上の案件についてはデータ共有についてのプランを提案時に盛り込むことを要求されている。

なお、OMB A-110(*)の対象となる高等教育機関、病院及びその他の非営利機関に対しては「公的資金による産生データが連邦情報公開法(FOIA)指定の手順により公開されることを連邦資金提供機関が保証する」というShelby Amendment(Provision of Public Law 105-277 – Interim Final Rule effective April 17, 2000)が適用される。

なお、本項にかかわる資料として、文部科学省/研究費の不正対策検討会/(第4回、平成18年10月26日)配付資料「米国NSF及びNIHの競争的資金制度について」に詳しい調査結果があるので、以下にURLを記載しておく。http://www.mext.go.jp/b_menu/shingi/chousa/gijyutu/008/shiryo/06110108/002.htm

(*) 大統領府所属機関の行政管理予算局(OMB) (Office of Management and Budget)から出された高等教育機関、病院及びその他の非営利機関とのグラント及び契約に関する統一的な管理上の要件に関する通達

d) 技術の進歩とプロジェクトの実施、および制度設計の関連

学術情報は、解析技術の発展に伴って、その技術を利用した大型の国際的なプロジェクトが実施されることにより、その量が飛躍的に増大していった。一方で、それらの情報の流通に関しては、社会的インフラ整備が同時に発達してきたことにより、さらに加速された。欧米では、主要な資金配分機関である米国国立衛生研究所(NIH)、Howard Hughes Medical Institute, The Wellcome Trustなどが、公的資金による研究成果を広く公開すべきというポリシーを、具体的な公開方法を示しながら提示しており、一部は強制力

を持つ法制化まで実施されている。学界からの意見も数多く表明されて、これらのルールに生かされていると思われる。ルール作成は、大型プロジェクトの進行と同時に進められており、プロジェクトから出てくるデータの扱いについて今後適用されていくことになろう(図 2、前出)。

我が国においても 現在ではまだ知財活用に重点を置いたルールが先行している感はあるものの、2003 年以降、学会等から学術情報の有効活用に関する報告等が次々に出されているので、それらを参考にしながら、今後は知的基盤を支える学術情報についてのルール作りを、国民への成果還元という視点から考えていく必要がある。

5) 考察・今後の発展等

当初、統合 DB に関する課題としては、国内の DB の現状に関する俯瞰的な資料が無いことから、客観的な評価方法の開発も含めて、データベースの網羅的なリストを作成する必要があるが、このリストについては、今後のデータベース統合化の方針検討に役立つ基本的なものを作成することができた。また、データベースの作り手と使い手それぞれの意見や、具体的な利用状況についての理解を進めることができたので、これらの情報も、データベース統合に際して考慮すべき点であると考えられる。

一方、技術的な課題の洗い出しとして、実施したフィージビリティ・スタディからは、モデルケースとして実施した内容が現行のデータベースの1機能として盛り込まれることになり、今後もデータの理解に役立つ内容となったと同時に、新たな問題点もあぶりだされてきたので、それらの検討を進めることにより、データベースの機能アップに役立つものと考えている。

当初フィージビリティ・スタディの検討項目の一つとして挙げられていたデータの流通に関する制度等の影響調査については、実際にはかなり根本的な問題点であることがわかってきた。そのため、後半の調査研究においては、制度等の調査にも重点を置いてきたが、欧米の先行しているルールのよき点は見習いながら、我が国の実情に合わせた柔軟なルール作りを今後検討する必要がある。また、我が国の国民感情も考慮しつつ、学術情報の流通に関して広く理解を求めていくことも、今後統合データベースの計画を推進するためには必要なことであるので、データベースの統合を進めていく機構(あるいは制度)の中には、理解推進のための具体的な方策も盛り込んでいく必要がある。

6) 関連特許

該当なし

7) 研究成果の発表

(成果発表の概要)

1. 原著論文(査読付き)	該当なし
2. 上記論文以外による発表	国内誌: 2 報
3. 口頭発表	招待講演: 5 回、主催講演: 1 回
4. 特許出願	該当なし
5. 受賞件数	該当なし

1. 原著論文(査読付き):該当なし

2. 上記論文以外による発表

- 1) 大久保 公策:「生命科学データベースの現状と課題」,科学(岩波書店) ,77(4),364-369, (2007)
- 2) 大久保 公策:「生命科学データベースの統合化の背景」,蛋白質核酸酵素(共立出版) ,52(9),1027-1031, (2007)

3. 口頭発表

招待講演

- 1) 大久保 公策:「ライフサイエンスにおけるゲノム情報の高度利用に向けた生命知識の構造化」,東京大学(本郷),知の構造化ワークショップ, 2006.12.4
- 2) 大久保 公策:「知的生産性向上のための情報処理」,名古屋国際会議場,日本分子生物学会 2006 フォーラムシンポジウム“プロジェクト型研究次代の生命科学の課題“, 2006.12.8
- 3) 大久保 公策:「オントロジーや辞書は役に立つのか」,名古屋国際会議場,日本分子生物学会 2006 フォーラムバイオテクノロジーセミナー, 2006.12.8
- 4) 大久保 公策:「ライフサイエンスのデータベースの現状と課題」,学術総合センター(東京・千代田区),情報とシステム 2007, 2007.3.1
- 5) 大久保 公策:「生命科学データベース統合に関する調査研究」,中央合同庁舎 4 号館,総合科学技術会議ライフサイエンス PT(第 5 回) , 2007.4.19

主催講演

- 1) 大久保 公策:「生命科学データベース統合に関する調査研究」報告,浜離宮朝日ホール,生命科学の基礎・基盤連携群シンポジウム ライフサイエンスデータベース統合への取組, 2007.11.26

4. 特許出願:該当なし

5. 受賞件数:該当なし

(2) サブテーマ1(2) データベースの物理的統合にかかわる技術的調査研究
(分担研究者名:森下 真一 所属機関名:東京大学)

1) 要旨

本研究項目では、ライフサイエンスデータベースを統合する際の課題を、ハードウェアおよびネットワーク周辺の物理的レベルから考察する。つづいてユーザーとの窓口である問合せ言語のレベルからも考察する。そして限りある計算資源を最大限に活かして、ライフサイエンス研究者の満足度を充足させるための統合化の道筋を提示する。

ポストゲノム研究の中心的な課題は、ゲノム解読によって明らかになった遺伝子およびその制御領域が、形態等の表現型を生み出す際にどのような機能を発揮するかを明らかにすることである。既にゲノムが解読された出芽酵母、ショウジョウバエ、線虫では遺伝子を網羅的に破壊・阻害もしくは強制発現することにより、形態にどのような変化が起こるかという情報を蓄積し、形態変化と遺伝子の対応関係をオントロジーとして統合化する萌芽的研究が展開されている。森下研究室ではこのような先駆的な研究を過去行っており、出芽酵母およびショウジョウバエを対象として国際的な研究成果を挙げている。今後は、様々な生物種で形態変化と遺伝子の因果関係がオントロジーとして統合データベース化されてゆくと考えられる。

平成 18 年度は生物種によらない一般化の可能性を検討した。その研究調査を踏まえ、平成 19 年度は、具体的な種を決めてプロトタイプを作成した。

2) 目標と目標に対する結果

目標 多様な生物データを統合化する実験を進めるにあたっては、データ収集が進んでいる生物種を対象に考察を進めることにより、広範囲のデータを統合化の際の問題点を把握できる。このようにあらかじめ多様なデータ型を対象に研究を進めてゆくことにより、生物データの収集が進んでいない生物種に対してさえ、方法論を部分的に適用することが可能である。そこで平成 18 年度は、以下の2つの視点からフィージビリティ・スタディを行うことを目標とした。

- ① 適切な計算資源配分に関する技術調査研究:既に運用されているライフサイエンスデータベースを例に、計算資源のバランスを調査し、適切な資源配分を調査研究する。
- ② 異種データベースを統合するデータベース問合せ言語に関する技術調査研究:異種データ結合型(マルチモーダル)問合せ言語により統合化可能なライフサイエンスデータベースについて研究調査する。

平成 19 年度はフィージビリティ・スタディを踏まえ、具体的な生物種を対象にプロトタイプを作成することを目指した。我々は対象生物種として出芽酵母およびメダカを選択した。なぜなら出芽酵母ではゲノム、ヌクレオソーム構造、ゲノムタイリングアレー、転写開始点、完全長 cDNA、ゲノムのメチル化部位、遺伝子発現量(マイクロアレー, SAGE タグ)、蛋白質の相互作用、遺伝子破壊株、表現型情報(生育速度, 細胞周期等)等の異なるデータ型が豊富に蓄積されており、これらのデータ型は他の生物種で集められているデータ型をほぼ包含している。またメダカは我国でゲノム配列を決定した生物種であり、我国発の発生生物学のモデル生物として世界的な広がりを見せている。そのためわれわれの調査研究の成果であるプロトタイプが現実には十分に利用可能かを試すためには十分な実験対象となった。

目標に対する結果

平成 18 年度のフィージビリティ・スタディの成果概要は以下のとおりである。

① 適切な計算資源配分に関する技術調査研究

計算資源を適切に配分するためには、バイオインフォマティクスで頻繁に利用される計算に適切であるかどうかということを踏まえながら、計算機のもつ演算能力を調査する必要がある。今回の調査では、ヒトゲノム解析センターが保有するスーパーコンピュータである SunFire15K と、森下研究室で購入したオプテロン計算サーバを、それぞれ大きな主記憶をもつ対称型多重処理 (SMP) サーバと、小さな主記憶をもつ計算機クラスターのモデルとして比較した。比較に利用した SunFire15K は UltraSparcIII cu(1.2GHz)を 72 個、主記憶を 295GB 搭載しているスーパーコンピュータ(サーバ名:tako)であり、オプテロン計算サーバは AMD opteron 248 (2.2GHz)を 2 個、主記憶を 16GB 搭載したマシン20台からなるクラスター(サーバ名:hx)である。オプテロンクラスターは 2 個の CPU を持つマシンが複数台という構成になっている。そのため、クラスター全体での大規模な処理を行うためのソフトウェアが必要である。Hx サーバでは、各計算ノードの状況を監視しておき、計算可能なノードに仕事を振り分けて効率よく実行するためのバッチ処理システムとして Sun Grid Engine をインストールして利用している。バイオインフォマティクスで頻繁に利用される計算の代表的な例として、遺伝子配列のゲノムへのアライメント、ショウジョウバエの翅画像の翅脈や細胞抽出処理について調査を行った。その結果、対称型多重処理 (SMP) サーバに比べ、オプテロンクラスターを用いることによって高速に処理できることが明らかになった。今後の計算機資源を投資してゆく際の指針が得られた点で重要な技術調査研究となった。

② 異種データベースを統合するデータベース問合せ言語に関する技術調査研究

生命科学データベース統合に関する調査研究として、特にゲノムブラウザを中核とした統合について調査検討をすすめてきた結果の骨子を報告する。我々はメダカ、カイコ、ショウジョウバエ、酵母、ヒト等のゲノム解析、情報統合を行う必要からゲノムブラウザを開発してきたが、この過程で、「いかにして生物基礎データを作成し収集するべきか」、「いかなる生物情報がデータベース利用者に提示されるべきか」、「いかにして生物データベースをメンテナンスするべきか」の三点をデータベース統合における課題として設定し改善してゆく方法を検討した。特に、基礎データを集積するという点からすると、Ensembl, UCSC, NCBI のように大量の基礎データ(ゲノム配列、遺伝子配列、遺伝子発現量)をアーカイブしている拠点が存在するため、基礎データを作成し、集積する必要性が本当にあるのかという見方もあった。しかし、調査の結果、大規模な新規のデータ(ゲノムマーカー、ゲノム配列、転写開始点周辺配列、ヌクレオソーム情報、メチル化、変異体表現型)を処理してゆく際には、これらの拠点が提供しているサービスが充分であるとはいえず、新規の基礎データを加工して集積するシステムは必須であることが明らかになった。

平成 19 年度の成果は以下のように整理できる。

項目①ゲノムブラウザトラックの作成、および 項目②遺伝子への情報付与:

- ・ 当初の計画通りゲノムブラウザ上にヌクレオソーム構造、タイリングアレー、転写開始点、完全長 cDNA の情報をトラックとして表示する機能を作成した。これらの成果は論文[1,4]で報告したウェブサーバに取り込まれ、ソフトウェアは以下の web site から公開中である。

UT Genome Browser (Yeast) <http://yeast.utgenome.org/>

UT Genome Browser (Medaka) <http://medaka.utgenome.org/>

また解説記事 [7]にて調査研究活動を報告している。

項目③異種トラックへの一括問合せ、および 項目④ニーズ指向の問合せ:

当初の計画通りに、異なるトラックについて文書およびデータ量を検索する問合せを並列に実行し、返される情報を統合する操作(具体的には論理和、論理積、否定を組合せた一階述語論理程度の表現力をもつ問合せ演算)を実装した。問合せ操作は技術的ニーズ指向であり、高度な機能を持たせることは出来るが現実には利用されない場合がありうる。そのため、生物学者との緊密な共同研究

を実施し、現実のニーズを把握して適切な問合せ機能を絞り込んだ。これらの成果は論文[2,3,8]で報告し、論文[5]で報告したウェブサーバにもその成果が取り込まれ、ソフトウェアは上記2つの web site から公開中である。また解説記事 [6,7]にて調査研究活動を報告している。

3) 研究方法／調査方法

バイオインフォマティクスにおいては、ゲノム配列、遺伝子情報などの既存のデータと、新たな解析によって得られたデータをいかに素早く統合するかが重要となる。ただし、これは単一のデータベースシステムですべてのデータを管理するという意味ではない。なぜなら、個々の研究のニーズ、環境に応じて、必要とされるデータ形式、使用可能なシステムが多岐に渡るからである。データベースシステムとしては、既に、Oracle、Microsoft、IBM に代表される企業らによって、豊富な機能を備えたデータベースシステムが開発されているものの、これらの商用製品は、“One-Size Fits All (ひとつのシステムで全ての用途を担う)” の精神で開発されており、たとえば、細粒度のトランザクション機能など、研究用途には少々大げさな機能も含まれている。また、そのような商用のシステムでは、テーブル形式のデータを用いるのが通常であるが、実際には、簡単なテキスト形式で十分な場合、バイナリ形式に変換して効率を重視する場合など、データベースシステムを導入してテーブル形式を使う方法が、必ずしも最適解とならない場合が多く存在する。本質的に重要なのは、個々の研究ニーズに合わせて構築されたデータベースをつなぎ合わせることであり、この効率化において、まだ多くの仕事が残されている。そのため、本調査研究では、種々のデータの結合をサポートするためのシステムに必要な要件を調査してきた。そのシステムの利便性を高めるためのパイロット研究として、現在の生命科学研究の根幹となるゲノム配列情報と、それに関連する付加情報を統合する「ゲノム情報 + α 」のためのウェブ・ブラウザを作成した。

4) 研究結果

本調査研究では、データベース作成を支援するために、ゲノムブラウザのフレームワークを作成した。このフレームワークは、ウェブ・ブラウザ側と、ウェブサーバ側が、それぞれ独立した構成となっている。この独立性により、ブラウザ側では、UTGB のインターフェースを活用することができ、サーバ側では、プログラミング言語、実装の種類を問わず、データを配信するために、任意の方法で実装されたものを利用することもできる。サーバ側の実装も、UTGB のフレームワークを活用すると、データベースへの接続、データ形式の変換などを手軽に行える。

(A) ゲノムブラウザの実装

UTGB の実装にはプログラミング言語である Java を使用しており、このプログラムは、オープンソースとして公開することを前提に開発されており、今年度末の公開に向けて、ドキュメントの整備なども進めている。本調査研究に関連する開発事項として、ブラウザ側では、以下に挙げる開発を行った。

- 多数のウェブデータを個々のトラックとして統合して表示するためのインターフェース
- 最先端の Web 技術である Google Web Toolkit を導入することにより、ブラウザ上でのドラッグ&ドロップ、トラックの部分的な更新をサポートし、快適な操作性を提供
- 画面スクロールなど、ボタン操作等もトラックとして実装できるようにしたことで、プログラムの再利用性を高め、新しいトラックを追加する際の拡張性を実現

また、サーバ側での開発支援のために以下のライブラリを開発した。

- 種々のデータベースシステム(PostgreSQL, MySQL, SQLite など)との接続をサポートするデータベースアクセスライブラリ

- データベースの検索結果、XML、JSON、テキストなど様々な種類のデータを、最小限のコーディングでオブジェクトにマッピングするための、BeanUtil ライブラリ。
- BeanUtil ライブラリを活用し、ウェブリクエストのデータの型変換を自動化し、新しい Web API の作成を簡易化するライブラリ

ブラウザ側、サーバ側での統合データベースブラウザの開発支援のためには、

- ローカルとリモートにあるデータを統合する、ポータブルウェブサーバ(UTGB Portable)
- 新しいトラックプロジェクトの作成を簡易化するユーティリティである UTGB Shell

などを作成した。これらの開発成果により、データベースの内容を表示するためのトラック開発の効率化とともに、データベースを表示する入り口部分の実装が整備されてきた(図 2.1)。

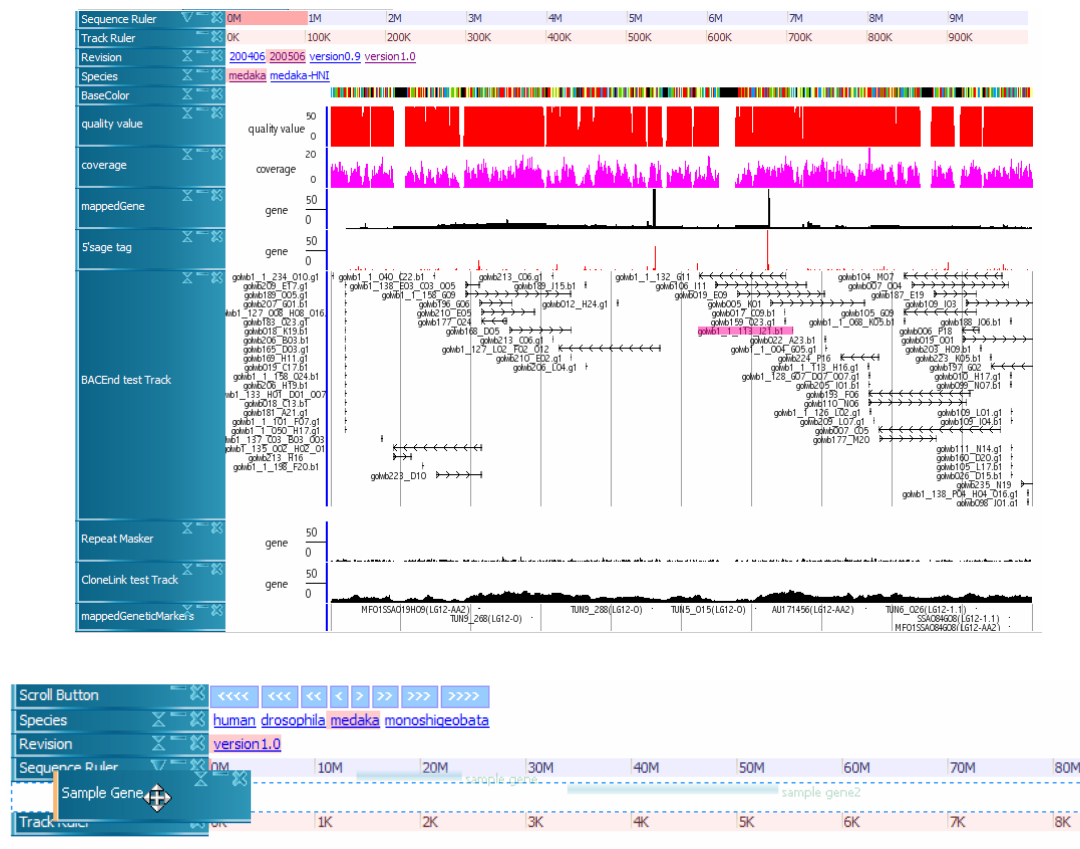


図 2.1 Google Web Toolkit を使った GUI アプリケーションの操作性を持つゲノムブラウザ

(B) 様々な生物種に対応するためのフレームワーク

本研究で開発したゲノムブラウザの仕組みを様々な生物種へ容易に応用できるように、ソフトウェア全体をフレームワークとしてまとめた(図 2.2)。この結果、より一般的なデータリソースの表示に対応できるような仕組みを整えている。ブラウザ側は、トラックと呼ばれる単位でデータを表示し、個々のトラックでは、任意のウェブ上のデータベースを表示させることができる。また、各種アダプタを適宜追加していくことで、様々なデータベースサイトのデータを混合して表示することができるようになっている。この結果、データベースを実際に統合してゲノムブラウザで表示するため、実際にデータベース側を整備し、データの扱いを簡便にするためのライブラリの開発が容易になり、次に挙げるような研究開発を行うことができた。

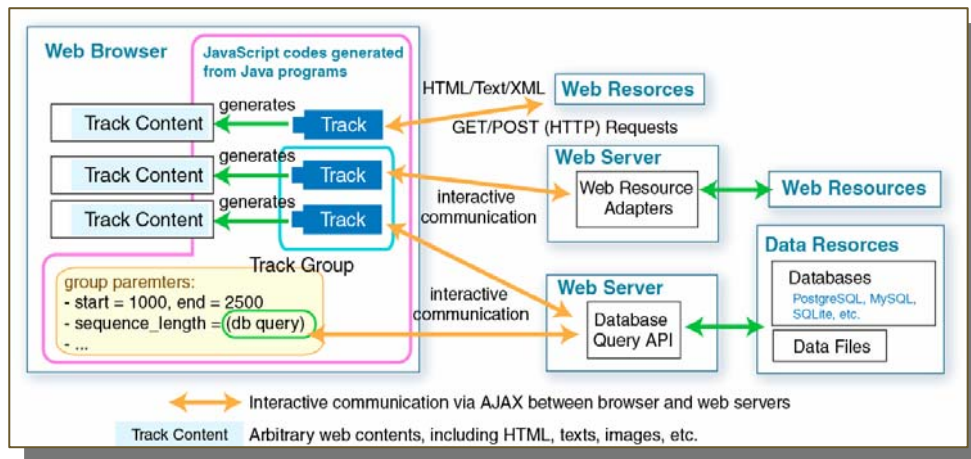


図 2.2 ゲノムブラウザのフレームワーク

(C) ゲノムブラウザのポータブル化

従来のゲノムブラウザでは、データベースをサーバマシンに設置してウェブ上で公開しなければ利用できなかったが、データの統合という観点からいうと、サーバ管理者に依頼し、データをアップロードして公開してもらうまで、データベースを統合できなかったため、複数のデータを組み合わせて利用する上での大きな障壁となっていた。そもそも、システム管理上、他者のデータを安定して公開し続けられる組織はそう多くは無いし、データが適切に統合されるとも限らない。

そこでデスクトップ上で、インストール作業なしに、クリックするだけで、ゲノムブラウザ(ローカルサーバ)が起動し、ウェブ・ブラウザ経由で UTGB を利用できるソフトウェア(UTGB Portable)を研究開発した。今回の UTGB Portable 開発により、精査前の実験データや、非公開とすべきデータでも、手元のマシンで、UTGB のフレームワークを利用して統合して表示する仕組みを提供できる枠組みができた。また、トラックの表示を手元でデバッグするための開発環境としても現在活用し性能を改善している。

(D) 分散問い合わせの実現

UTGB Portable の実装は、通常のウェブサーバとしても使えるように設計されている。ヒト、メダカ、酵母(yeast)など、各々のデータソースごとに分散してサーバを設置してある場合、各々が管理するデータベースの内容を、互いに通信する機構の開発を進めた(図 2.3)。この実装の利点は、例えば、各々のデータベース情報を収集しているデータベースセンターに、「メダカゲノム配列の QV (quality value)」の情報が欲しい、という命令を出すことで、そのデータを持つサーバのアドレスを受け取り、当該のサーバから直接データをダウンロードできることである。UTGB Portable をローカルマシンで動作させていれば、手元のデータと、外部データの結合も可能になる。また、ネットワーク地理的に遠い場所にあるが、使用頻度の高いデータを、手元の UTGB Portable にキャッシュして、データの検索や表示を高速化するなどの活用法も考えられる。

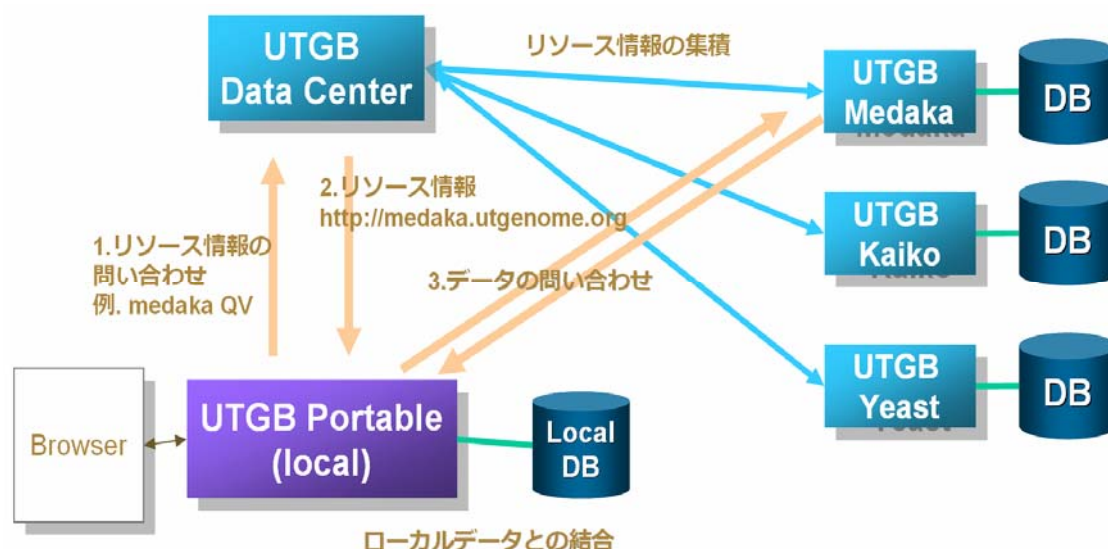


図 2.3 分散したデータリソースの問い合わせ

(E) データベースセットアップ

本調査研究では、データベース統合に伴う問題点を洗い出すために、手法の考案と実践を交互に行ってきた。実際に、パイロット研究として、ゲノムに関連する各種データを統合するゲノムブラウザ(UTGB)を、メダカゲノム(UTGB Medaka), 出芽酵母ゲノム(UTGB yeast) 関連データについて開発した。この研究開発の過程で、データベース統合における、いくつかの問題点が浮かび上がってきた。

データベースブラウザの開発は、言葉の上では、「データベースの内容をウェブ上で表示する」と、簡単に言い表すことができるが、その内容には、ウェブサーバを動かすための設定ファイルの記述、必要なプログラムライブラリのインストール、開発用(テスト用)サーバの構築、データベースへの接続設定、データベースの検索結果をプログラム上で扱える形に変換(たとえばオブジェクト形式にする)、オブジェクトの内容をテキスト・画像などの形式に変換して出力するなど、非常に多くの設定項目や、実装内容が含まれている。

これらの設定等を、最初からミスなしに行うことは非常に難しく、わずか 1 行の間違いでも、ウェブサーバが正しく動作しない原因になり、究明に数日を要することもある。1つのデータ配信用トラックならまだしも、生命データベース統合というテーマのように、ここに挙げたウェブサイトの構築の手順をデータリソースの種類だけ繰り返さなくてはならないというのでは、非常に効率が悪く、時間コストを要するだけでなく、プログラム開発を専門としていない研究者にとって、データを公開するための敷居が高いままになってしまう。

このような初期的な部分に注意を払わない限り、データベース統合のために欠かせない「情報を集める」という、大前提が崩れ去ってしまう。そこで、我々は、前回の報告後、これらのセットアップをサポートするプログラム、UTGB Shellを開発した。UTGB Shellは、ブラウザ側、サーバ側のライブラリをすべて統合して扱えるようにしたものである。UTGB Shellを使うと、「データベースの内容をウェブで表示する」ウェブサーバプログラムが、この言葉の短さとおりの速さで開発できるようになる。

例えば、UTGB Shellのインストール後、create というコマンド1つで、新しいトラックプロジェクトの雛形が作成される。そのトラックをウェブサーバとして動かす際に必要なライブラリ群は全てウェブからダウンロードされ、手作業によるインストールや設定作業を省くことができる。また、データベースへの接続もアドレスを数行設定ファイルに記述するだけで良く、プログラム中では、BeanUtil とデータベースアクセスライブラリ

の機能により、データベース検索の結果を1行のコードで、オブジェクトにマッピングすることができる。また、UTGB Shellにはポータブルウェブサーバ(UTGB Portable)が同梱されているので、デスクトップ上でサーバを起動することができ、デバッグ用のウェブサーバを本番用のサーバと別途に用意する必要がなく、即座に開発と試験運用が行えるようになっている。UTGB Human (図 2.4) の例は、UTGB Shell を用いて開発されたものであるが、表示するトラックの種類を記述した設定ファイルが数十行、遺伝子情報や発現量のデータを画像描画部分が数百行程度でしかない。画像描画部分もライブラリ化すれば、将来的には、データベースのありかだけを記述し、プログラミングなしで、ブラウザを作成することも可能となるであろう。

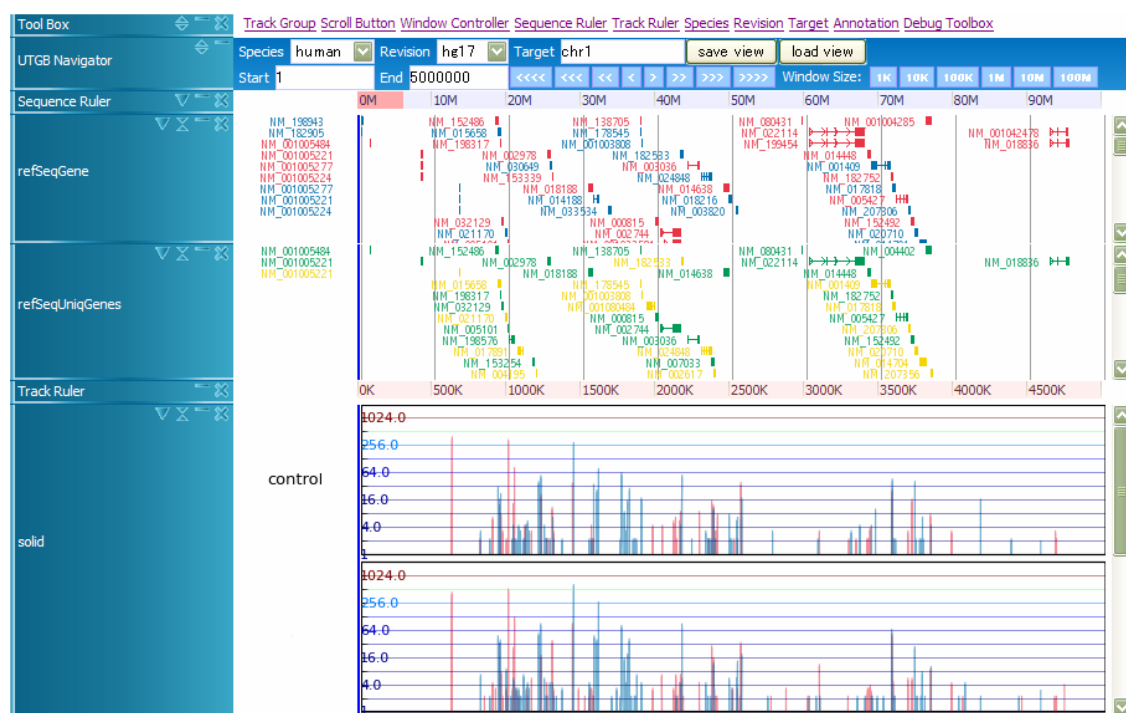


図 2.4 UTGB Shell で作成した Solexa タグの表示例

5) 考察・今後の発展等

本調査研究では、データベース統合のために必要なソフトウェアツール群を調査し、その結果をふまえてプロトタイプを実装してきた。個々のデータベースの視覚化や、ウェブサービス化をサポートするユーティリティかつライブラリ群である、UTGB フレームワークを開発した。このフレームワーク中の、データベース接続、ウェブインターフェース、オブジェクトへのマッピングなどの機能を、UTGB Shell を用いて統合することで、データベースブラウザの作成を効率化できるようになった。これにより、生命の謎に迫るため、個々のデータベースをウェブ・ブラウザという形で視覚的に統合し、比較検討するための土台は整備されてきた。

今後の展望としては、視覚的な統合にとどまらず、ブラウザ上で見えている複数のデータベースを横断して検索し、データの絞込み検索、統合検索を行うことが課題となるであろう。例えば、ブラウザ上に、遺伝子と、様々な条件化での遺伝子発現量の違いが見えていた場合、その差異は、どの遺伝子と密接に結びついているかを検索し、対象遺伝子がどのような機能を持っているか、という検索が可能となれば、複数のデータベースを渡り歩きながら、ブラウザ上で「ゲノム情報 + α 」のデータを次々と表示できるようになるだろう。そのような統合検索を実装する際にも、本研究で開発した、データベースシステムの差異や、データの構造の違いを吸収する BeanUtil などのライブラリが十分活用できるであろう。

ただし、遺伝子機能を解明するためには、遺伝子発現を調べるだけでなく、遺伝子機能を制限したとき

の表現型の変化や、比較ゲノムなど、様々な方向からのアプローチがあるため、それぞれが目的に特化したデータ形式を使うのは自然の流れであり、各機関が常に共通のデータ形式を使用するとは考えにくい。たとえば、出芽酵母遺伝子破壊株の表現型のデータと、既知の遺伝子機能のデータ(GO: Gene Ontology)を統合して、未知遺伝子の機能を調べるという解析を行ってきた経験では、GO データ提供側と共通なデータ形式を用いたわけではない。実際には、遺伝子機能に関連するデータに目印をつけ(+ α の情報を付加することに対応)、独自のデータ形式を用いてデータベースの統合を行っている。すべての事象やデータを表現できるデータ形式の仕様の策定は困難であるし、共通目的がない機関同士では、データ形式を統一するための強い動機も生じないためであろうと分析している。このように、データベースの統合によって新たな情報を見出したいと考える研究者にとっては、人の目と、手作業によって、種々のデータ形式を結合する操作がどうしても必要となってくる。この結合作業の効率化を図ることが、データベースの統合を真に促すものであろう。本調査研究では、複数のデータベースからデータを取得し、ゲノム+ α の情報を付加していくことで、手元で容易に複合データベースの検索、合成が行えるシステムの開発を進めてきたが、今後の研究開発においても引き続き持続発展させる必要がある。

また、2007 年に入り、Solexa (ソレクサ。現在はイルミナ社が買収) などの高速シーケンシング技術が登場し、20~30 塩基程度の断片的なゲノム配列が、短い実験期間で大量に得られるようになった(1 日で3~5 億塩基の配列情報という規模)。そのため、研究者各々が関心を持つ対象や、様々な環境要因下で得られた断片的ゲノム配列が、今後加速的に生成されていくことが予想される。Solexa などのように高収率シーケンサから出力される膨大なデータを全てダウンロードするのは現実的ではない。分散問い合わせの実現には、データの規模に合わせて、転送量を押さえるための仕組みが必要である。既存のデータベース問い合わせの技術を組み合わせることで、UTGB に適した分散問い合わせの仕組みを探っていく必要があるであろう。超高速シーケンシングにとどまらず、今後、多数のデータリソースが容易に得られるようになることを念頭に入れると、ゲノム+ α の情報の種類も豊富になり、検索条件に指定する属性(パラメータ)の数も増えてくるであろう。このような多属性データベースにおける検索を効率よく行うためには、多次元索引が必要となる。ただし、単純に既存の多次元索引を用いるだけでは、データ量があまりにも多いため、データを圧縮しつつ多次元の索引を構築する方法を模索する必要があると考えられる。

また、若干軽微な問題に感じるかもしれないが、データの表記ゆれの問題は、軽視することのできない問題である。たとえばアノテーション中の染色体情報一つをとっても、その表現方法は、chr1, ch1, Chromosome 1 などと異なっていることが多い。(DAS では、配列に global ID をつけることで、これらを統一する試みがなされているが、それに従わないデータは世の中に数多く存在する。) たとえテキストや、テーブル形式で複数のデータが得られたとしても、それらを結合するための、マッピングテーブルのような中間テーブルが必要となってくる。このように、データを複数取り出すだけでデータベースの統合につながるわけではなく、解決すべき問題は多々残っている。

6) 関連特許

該当なし

7) 研究成果の発表

(成果発表の概要)

1. 原著論文(査読付き)	5 報 (筆頭著者:3 報、共著者:2 報)
2. 上記論文以外による発表	国内誌:2 報、国外誌: 0 報、書籍出版: 1 冊
3. 口頭発表	招待講演: 0 回、主催講演:0 回、応募講演: 1 回
4. 特許出願	該当なし
5. 受賞件数	該当なし

1. 原著論文(査読付き)

1. Budrul Ahsan, (36 authors) and Shinichi Morishita. UTGB/medaka: genomic resource database for medaka biology. *Nucleic Acids Research* Advance Access published on October 11, (2007)
2. Taro L. Saito and Shinichi Morishita. Efficient Integration of Structure Indexes of XML. *Proceedings of the 12th International Conference on Database Systems for Advanced Applications (DASFAA)*, April 2007, Bangkok, (2007)
3. Taro L. Saito and Shinichi Morishita. Amoeba Join: Overcoming Structural Fluctuations of XML Data. *Ninth International Workshop on the Web and Databases (WebDB)*, (2006)
4. Miura F, Kawaguchi N, Sese J, Toyoda A, Hattori M, Morishita S, and Ito T. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proc Natl Acad Sci U S A*. 103(47):17846-51, (2006)
5. Tomoyuki Yamada, Haruhiko Souma, Shinichi Morishita. PrimerStation: a highly specific multiplex genomic PCR primer design server for the human genome. *Nucleic Acids Research* 34: W665-W669, (2006)

2. 上記論文以外による発表

1. 曲 薇・バドルル アーサン・森下 真一 「超高速シーケンサの原理と情報処理」 森下 真一／阿久津 達也編 実験医学 2008 年増刊号「生命研究への応用と開発が進む バイオデータベースとソフトウェア最前線」2008 年 4 月
2. バドルル アーサン・山田 智之「ゲノムブラウザ」 森下 真一／阿久津 達也編 実験医学 2008 年増刊号)
3. Taro L. Saito, Shin Sasaki, Budrul Ahsan and Shinichi Morishita. UTGB SHELL: AN OPEN-SOURCE BROWSER FRAMEWORK FOR THE INTEGRATION OF BIOLOGICAL DATA. The biology of Genomes. Cold Spring Harbor Laboratory. 2008 (in press)

3. 口頭発表: 該当なし

4. 特許出願: 該当なし

5. 受賞件数: 該当なし

(3) サブテーマ2 医学分野のデータベースに関する調査研究
(分担研究者名:吉田 輝彦、所属機関名:国立がんセンター)

1) 要旨

医学分野のデータベースに関する調査研究として、①フェノタイプ情報の統合に関する調査、②フェノタイプ情報－分子情報の統合に関する調査、③プロジェクト型 DB の試験的統合という具体的目標を掲げた。これらの目標を、医学分野データベースの大づかみの俯瞰図の中で配置すると図 3.1 の通りである。

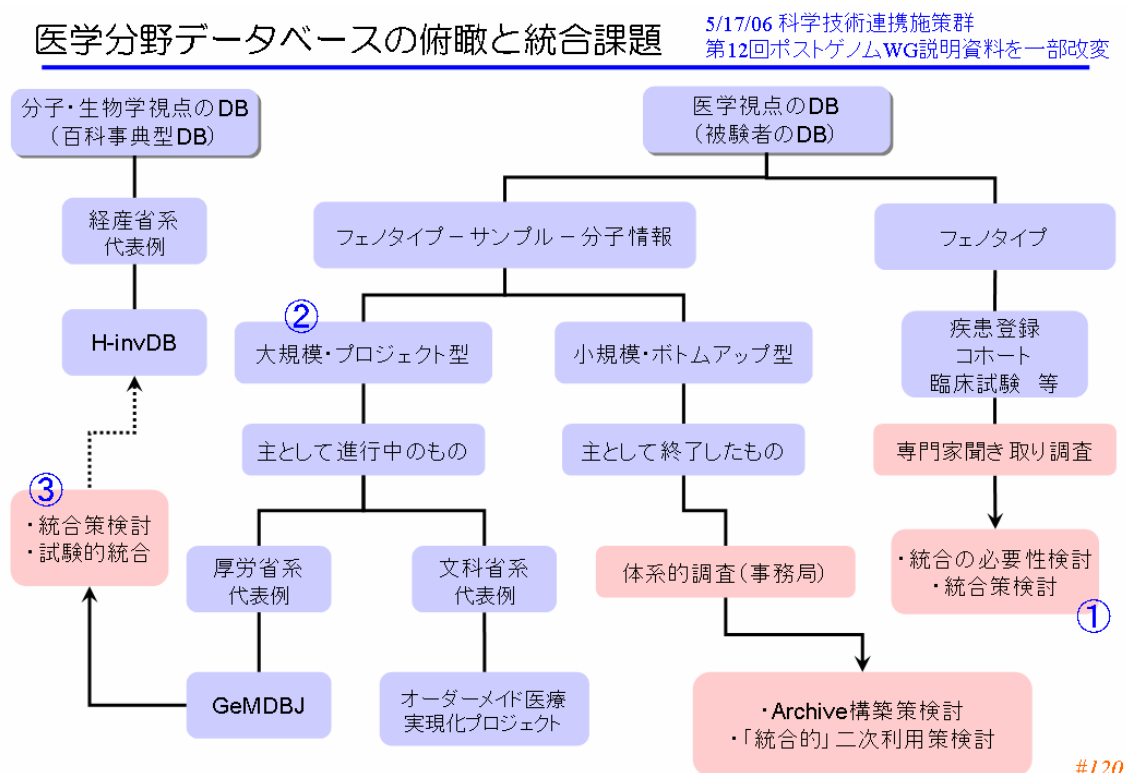


図 3.1 医学分野データベースの俯瞰と統合課題

これら研究テーマ①～③毎の成果については後述の「成果の詳細」で述べる。それら調査研究に基づき、以下のように成果を総括する。

- (ア) 医療情報は極めて複雑であり、たとえば疾患登録の根幹となる疾患分類にしても、どのように病態を把握し、疾患を定義し、体系的に分類するかも疾患研究そのものであり、疾患研究の専門家が長年、取り組んできたところである。すなわち、医療・医学情報の統合は、疾患研究そのものとも言える。
- (イ) 今、どのような「統合データベース」が欠けているのか、それをどのような順番で補完していくのか、そのためにはどのような資源が必要なのかを明らかにする必要がある。
- (ウ) 上記の議論においては、その「統合データベース」の目的をあらかじめ明確にすることが特に重要である。すなわち、その目的に応じた、データの集め方と、そのデータを解析する方法・体制をあらかじめ明確にしておくべき。特に医療情報は bias (系統誤差) やエラー (偶然誤差) に充ち満ちており、単に既存のデータをかき集めておけば、あとは第三者が高度な統計学・情報学的

手法を適用することで画期的な知見が得られるというのは疾患研究に対して楽観的と言わざるを得ない。

- (エ) たとえば、疾患の分子レベルの研究のために想定されるデータベースの目的として、質・量に優れた臨床試料の網羅的分子情報と、それに付随した臨床・病理・生活習慣情報のデータを集積し、そこから新たな仮説を探索したり、あるいはそれぞれの研究者が自らの個別研究の個別の知見を検証したりする、という目的であれば、新たに標準化された方法で、そのようなデータ公開を目的として新たなデータ取りも行うプロジェクト型研究が必要である。例えば厚労省においては、疾患ゲノムデータベース GeMDBJ (<http://gemdbj.nibio.go.jp/>) としてそのような取り組みを進めている。
- (オ) 一方、bottom-up 型の個別疾患研究においても、それぞれの結論を導いた直接の raw data の公開・保管も今後、進めるべきである。すなわち、研究者が日々生み出す、garbage に満ちたデータではなく、あくまでも peer review journal に受理された結論の根拠となる、信頼性の高いデータをデータベースにアーカイブ化し、将来の検証やメタ解析に供する。これはすでに網羅的発現解析等においては世界の標準となりつつある。
- (カ) また、重篤な薬物有害反応を発症した症例については、その臨床情報と血液やその他の試料、さらに可能であれば家族の血液試料なども全国から集める恒常的な組織が欲しい。
- (キ) 上記(カ)のような例に限らず、個人情報保護法や各種研究のための倫理指針において、個人の権利(特にプライバシーの保護のみでなく、自己情報コントロール権を含む人権部分)の保護と並んで、社会・集団の福祉の適切なバランスの確保が必要である。これは taboo 視される傾向があり、研究者側からだけでは言いにくく、またたとえそうしたとしても社会に理解してもらいにくい。総合科学技術会議の指導力が必要である。
- (ク) 具体的には、臨床試料・情報の研究への利用に関するいわゆる「包括同意」や、説明・同意に関する opt out 方式の適切な活用も検討されるべきである。そのためには社会の信頼と理解を得る努力が必須である。
- (ケ) 疾患の分子レベルの研究においては、データは試料そのものに詰まっている、と言える。データ解析の手法は日進月歩であり、最終的には試料の適切な収集・保管が最も重要。データベースとリソースバンクは本来は並行してその議論を進め、構築に当たるべきである。

2) 目標と目標に対する結果

・ 全体の目標

極めて多種多様、かつ膨大な数の医学におけるデータベースの鳥瞰的調査等を通して、複数のデータベースを有機的に関連づけ、あるいはさらに「統合」することが新しい医学・医療の発展に有効に寄与すると期待され、またそのような統合へのニーズとフィージビリティがある領域を同定する。ついで、その統合のために克服すべき課題、開発すべき技術やオントロジー等の知識体系、組織されるべき研究事業の在り方及び制度設計等について提言を行う。さらに、統合データベースのシステム設計を行い、実際に、一部の既存のデータベースの試験的統合に向けた作業に着手する。

・ 具体的研究目標

① フェノタイプ情報の統合に関する調査:

(ア) 臨床試験の情報の統合に関する調査

(イ) コホート研究に関する調査(事務局と共同)

(ウ) データ公開への、個人情報保護や倫理的側面の障壁に関する提言

② フェノタイプ情報－分子情報の統合に関する調査:

(ア) 国際的な疾患ゲノム研究とその成果のデータベース化に関する調査

③ プロジェクト型 DB の試験的統合:

(ア) H-InvDB 開発者に協力して、データレベルでの連携を組み込み、最終的にユーザーが2つの DB を行き来していることを意識せずに、virtual に一つの DB の中で知識探索を行う環境構築を目指す。

(イ) その他の国内のデータベースとの同様の連携を進める。

3) 研究方法／調査方法

各種文献や web 上に公開されている情報の検索・調査に加えて、専門家・有識者への訪問インタビューや、会議形式での質疑・意見交換を行い、さらに自らの考察を加えて解析・研究結果をとりまとめた。

4) 研究結果

① フェノタイプ情報の統合に関する調査:

(ア) 米国を中心に着々と進められつつある臨床試験の情報に関する標準化と統合の動きについて、調査を行った。具体的には、CDISC (Clinical Data Interchange Standards Consortium)、BRIDG (Biomedical Research Integrated Domain Group)、caBIG (cancer Biomedical Informatics Grid) 等のプロジェクトである。医学研究において、しばしば最後の段階としての臨床試験が重要であるのは、医療情報は様々な bias を含むためであり、比較可能性を可能な限り厳格に担保し、高いレベルのエビデンスを生むことが出来るのが臨床試験であるため。企業治験では世界同時開発の流れの中で、EDC (Electronic Data Capture)、CDISC/HL7 の強い影響が予想される。

(イ) コホート研究に関して専門家への聞き取り調査等を行った。概要は以下の通り。

- A 循環器領域で有名な Framingham study (1948 年開始) や Seven Countries Study (1958 年開始) の最初のコホート集団はほぼ全員死亡か。循環器は罹患率が高いので、比較的少数のコホートを詳細に調査する方法が採られる。
- B 一方、がんは個々のがんとしては罹患率が低いので数万人以上の規模が必要。古い研究は規模が小さすぎ、1970 年代より大型コホートが開始され、その成果が今、出てきているところ。
- C 我が国における主なコホートは以下の通り:

1. 平山研究

- 26 万人、6地域、1965 年開始、ほぼ終了している。
- 追跡調査は死亡のみ、人口動態を使って全死亡を見ている。
- 質問票は葉書一枚程度で、喫煙や緑黄色野菜摂取など。

2. 被爆者コホート

- 93,000 人の被爆者を含む約 12 万人のコホート、1950 年より死亡追跡している。
- 広島・長崎にがん登録システムを作っている。

3. 環境庁コホート(3府県コホート)

- がん登録のある宮城・愛知・大阪で 1983～1985 に開始。住民基本台帳を用い、40 才以上を対象、一地域あたり3～4万人。
- 大気汚染と肺がんの関係など。栄養調査はあまり詳しくない。1997 頃で解析は中断している。

4. JACC Study

- 文科省がん特定領域大規模コホート研究、1988 年開始。全国から12万人、40～79才。質問票に回答した11万人が追跡調査の対象。全地域で死亡調査、地域によっては罹患も調査。3.9 万人から血清も収集。
- 10年が経過し、結果・論文報告が出だしているところ。

5. JPHC Study

- 1990 年より、死亡及び罹患(ほぼ全員について)を追跡調査。全国から約10万人。
- 厚生労働省がん研究助成金による指定研究班「多目的コホートに基づくがん予防など健康の維持・増進に役立つエビデンスの構築に関する研究」(主任研究者 津金昌一郎 国立がんセンターがん予防・検診研究センター予防研究部長)→詳細は事務局招へい聞き取り。

6. J-MICC Study(日本多施設共同コホート研究)

- 文科省ゲノムコホート。2005～2009 年に約10万人を目標に参加者募集中。我が国で recruit 中のコホート研究はこれのみ。
- 二次調査(中間調査)を行い、生体指標の調査も行う。エンドポイントは全死亡、がん罹患など。

7. その他、小規模コホート

- 数多くある。久山町コホート、宮城コホート、高山コホートなど。

D アジア諸国

1. 韓国

- 国ががん対策に注力している。国民背番号制度があるのが強み。病院のネットワークから診療に関するデータが中央に集められる。Central Cancer Registry があり、健康保険の情報をを用いてがん死亡・罹患が把握できる。
- 66万人規模の研究や、J-MICC に相当する約25万人のゲノムコホート(KoGES: Korean Genome Epidemiological Study、2001 年～)も開始しており、血液・尿検体を収集、日本を軽々と追い抜きつつある。

2. 中国

- 最近、がん登録に力を入れている。以前からがん登録を行っていた上海にて、米国 NCI と協力して行っている上海 Study をはじめ、新しいコホート研究を数多く立ち上げている。

3. 米国

- 5～6万人規模のものは数多く走っている。10万人規模の研究も多い。血液(ゲノム解析含め)検体収集は当たり前となっている。
- Harvard Professional Follow-Up Study、Harvard Nurses' Health Study、Iowa Women's Health Study、California Teachers Study、Cancer Prevention Study II、Prostate Cancer Prevention Trial 等々。

4. 欧州

- 小国多く、国ががん登録を持っているところが多い。北欧・英国で盛ん、南欧は少ない。
- 北欧のように、出生から死亡までの情報を一箇所に集中しているところでは、容易に大規模疫学研究が可能。
- EPIC (European Prospective Investigation into Cancer and Nutrition) では食事についてきちんと評価している。がん以外の慢性疾患も解析対象。

E 展望

1. 住民コホートでは、複数のエンドポイントで、数万～10万人規模、ゲノム解析は当然となりつつある。
2. 一方では、より小規模でも、特定の地域・特定の集団(たとえば乳がん検診受診者など)に焦点を当てたコホートも必要。
3. エンドポイントの正確な把握が鍵。国民総背番号、疾患登録などの社会システムの整備の有無が成否の鍵となる。
4. 少し罹患率の低い疾患や、small effect size の因子の解析、生活習慣・環境要因等の曝露範囲の狭さなどを考えると、10万人規模でも足りずに、meta-analysis の他、pooled analysis が実施されていたりする(Harvard の Pooling Project of Prospective Studies of Diet and Cancer、EPIC など)。あるいは100万人コホートが必要なのか、コホート研究は、どこまで小さな effect を解析しようとしているのか、どこまで行こうとしているのか、研究者も模索中。
5. 追跡調査段階に入っているコホートをいつまでやるかが今後の検討課題。

(ウ) データ公開への、個人情報保護や倫理的側面の障壁に関する提言

- A 個人等の権利(人権・著作権)と学術貢献・公共福祉の適切なバランスを確保する様々な工夫・取り組みが必要である。具体的には、以下の点について国レベルでの議論が必要である。

1. 新たな指針作成や既存の指針・法律等の改訂
2. 倫理審査委員会の整備・強化・標準化

3. 研究(特に人への介入研究)の監査体制の確立
4. 研究協力者(被験者)を保護する制度の確立
5. 社会への理解を得る努力
6. 権利を侵害せず公益目的利用を阻害しない工夫

B 包括的な同意の適切な活用・浸透の必要性

1. 個人情報取扱・保護の基本的理念は以下の通りである:
 - プライバシー保護
個人に関する情報を漏洩無く、本来の使用者以外が使用しないように管理する。
 - 個人情報の目的外利用の禁止
個人情報を本来の目的以外に利用しない。これには情報の操作・集計なども含む。
 - 個人情報の開示
情報源である個人は、希望するときにはその情報を見ることができる。

C 考察:

1. 「個人の同意」に対する考え方が変遷しているのではないと思われる。従来は、試料等提供者に研究内容を全て説明し、全て理解してもらい、その上で同意を得ることが原則。今は研究が高度化して一般人には十分な理解が難しいこと、また説明を受けることに対する患者の負担もあり、研究者がどのように責任を持ち、個人情報等を保護しているか、提供者の身体的・経済的損失を防いでいるか、またデータを他の研究者と共有するためにはどのようなシステム・データ保護対策を確保しているかという点の説明がより重視されている。
2. 研究の手法が変化していくと、その研究内容に対して同意をとっていないサンプルが無駄になってしまう。研究がこれから5～10年の間でどのように変わっていくかを見極めて、どのような説明・同意の枠組みが必要なのかを考えるべき。包括同意以外では、将来のことを考えずに、今ある枠組みの中でインフォームド・コンセント(IC)を得ざるを得ないところが問題。
3. 一方、包括的な同意について、医学全般の研究という言い方では行き過ぎだが、疾患の原因となる遺伝子の探索・解析をしたい、というレベルで認められるかどうか課題。
4. 同意書には論文発表、学会発表に使用するとは記載しているが、DB 公開については必ずしも書かれていないことがある。
5. 公開(自由アクセス)部分のデータについても、利用目的の制限を謳う DB もあるが、利用者がその DB が定めるポリシーに違反したとしても、公開部分のデータ利用については実効的な監視・抑止力は無い。
6. 提供者には自分の情報をコントロールできる権利がある。例えば、本人にとって役に立ちそうもない情報でも、見せてほしいという要請があったら見せる、等。

7. DBにおける個人別タイピングデータ公開のICがない時にデータを出せるかどうかは、国民的議論が必要であると思われる。
8. 遺伝性疾患等について、論文の中で家系について発表するのは広く行われている。但し、昨今は、プライバシー保護(匿名化)や、人権(症例報告への同意)への配慮が強化されている。仮に、誰でもアクセス可能な論文でそれが許される状態となっているのであれば、データベースで公開する多因子疾患の多型データ等は、そのような症例報告の内容よりも「粗い」データなので、問題ないと考えられる。
9. ELSI(Ethical, Legal and Social Issues、倫理的・法的・社会的問題)側の主張としては、HUGOミーティングにて、どんなに個人情報情報を隠したとしても、理論的にはSNP数10程度で全人類の中で特定の個人を同定することができ($3^{21}=100$ 億、現在の世界の人口は約66億)、実際に警察や司法のデータと付き合わせる事が可能となると言っている。匿名化の意味が問われている。
10. 共同研究契約を結ぶことにより、現在でも個人別タイピングデータを他施設に提供できるが、これはそのたびに研究計画書を作り、両方の倫理委員会を通さなければならない。
11. 三省合同倫理指針があるが、時代に即していない部分がある。個人情報保護法もかなり学術目的のデータ共有に重荷。

② フェノタイプ情報－分子情報の統合に関する調査:

(ア) 米国のGAINプロジェクトに関する調査

- A 欧米では対照群のGWAS(genome-wide association study)の個人別タイピングデータは公開が始まっている。
- B 症例については、米国でdbGaPの中にGAINが組織され、急速にGWASのデータの共有が始まりつつある。
- C GAINでは2段階公開を採用し、頻度データは一般公開し、個人別タイピングデータはDACが審査した上で提供している。
- D 対照群と疾患群でデータ公開制限が異なるのは、試料等提供者保護の点の他、研究者の権利保護、という点があると思われる。
- E GAINでは、データ取得・品質検査後の公開は即座に行うが、研究者のincentiveと研究発表の質の担保等のため、公開されたデータを用いた文献発表についてはデータを公開した研究者に12ヶ月間の独占実施権を認めている。
- F 元々の同意文書の内容は公開されていないことが多いようであるが、データ利用が当初の同意の範囲外(目的外使用)に当たるかどうか、という点はあまり厳しく問題視せず、むしろ研究開始前の審査に加えて、研究開始後の監査等により、データの適切な使用を担保することで試料等提供者の権利を守りつつ研究を促進する、という考え方ではないかと思われる。

(イ) その他の国際ゲノムプロジェクトに関する調査

③ プロジェクト型DBの試験的統合:

(ア) H-InvDB 開発者に協力し、GeMDBJ 及び H-InvDB 双方のデータの更新や、双方が共通で

参照している外部ゲノムリファレンス基本情報の更新管理を自動化することにより、データの正確性と最新性を保つシステムを実稼働させる。

- (イ) H19 年度に、GeMDBJ に蛍光二次元電気泳動 (2D-DIGE) によるプロテオームデータを追加し、当該プロテオームデータにおいて同定されたタンパク質から H-InvDB のキーワード検索を通して様々なデータとの連携を組み込んだ。
- (ウ) H-InvDB が提供するデータマイニング等の手法に基づく疾患遺伝子探索ツールを、GeMDBJ のデータに適用し、両 DB 連携解析例として検討した。具体的には GeMDBJ の胃がんのゲノムスキャンのデータを用いて、H-InvDB の研究者が情報学的手法を用いて胃がん関連候補遺伝子の抽出を行った。
- (エ) H-InvDB 開発者との協議を通して、疾患研究のために求められる解析及びアノテーション機能を提案し、連携する2つのデータベースの、疾患研究基盤としての強化を図った。
- (オ) その他の国内のデータベースとの連携を進めた。具体的には東京医科歯科大学田中博教授を研究代表者とする科学技術振興調整費重要課題解決型研究「網羅的疾患分子病態データベースの構築」による東京医科歯科大学のトランスレーショナルリサーチのためのデータベース CODS (Clinical Omics Database based on Systems Pathobiology) と連携して、CODS 内に、トランスクリプトームデータを対象にしたクラス予測問題を解くための定型的な解析パイプラインを構築した。その解析ツール群が取り扱うデータとして GeMDBJ から公開中のトランスクリプトームデータに関して新たに、CEL ファイルの公開を開始し、CODS での解析を可能にした。

5) 考察・今後の発展等

① フェノタイプ情報の統合に関する調査:

- (ア) H20 年度以降に継続する他の研究事業として継続する下記③の課題の中で適宜フェノタイプ情報の統合に関する調査を継続する。特に、欧米などのデータベースの動向に注意して、国際的水準での DB 公開・統合を目指す。

② フェノタイプ情報－分子情報の統合の関する調査:

- (ア) H20 年度以降に継続する他の研究事業として継続する下記③の課題の中で適宜フェノタイプ情報－分子情報の統合に関する調査を継続する。特に、欧米などのデータベースの動向に注意して、国際的水準での DB 公開・統合を目指す。

③ プロジェクト型 DB の試験的統合:

- (ア) GeMDBJ の蛍光二次元電気泳動に関するプロテオームデータから H-InvDB の連携を引き続き強化する。具体的には GeMDBJ で同定・公開されるタンパク質データから H-InvDB のプロテオームデータ、トランスクリプトームデータ等の有用と思われる情報を直接取得できるような連携を組み込む。
- (イ) その他、H-InvDB との間で、H19 年度までに構築を開始した様々な連携を可能にするシステムの維持・更新・強化を図り、ユーザーが2つのデータベースを使っていることを特に意識することなく知識探索・検証に専念できる環境を作る。
- (ウ) 東京医科歯科大学のデータベース CODS (Clinical Omics Database based on Systems

Pathology)との連携を継続する。具体的には GeMDBJ で公開しているがん等の臨床試料のトランスクリプトームデータの.CEL ファイルを CODS からシームレスにアクセス・ダウンロードし、CODS が提供する様々なデータやツールを用いた統合的解析を行う仕組みを作る。

6) 関連特許

該当なし

7) 研究成果の発表

(成果発表の概要)

1. 原著論文(査読付き)	該当なし
2. 上記論文以外による発表	該当なし
3. 口頭発表	該当なし
4. 特許出願	該当なし
5. 受賞件数	該当なし

(4) サブテーマ3 産業(農林水産業)データベースの統合にかかわる調査研究

(分担研究者名:長村 吉晃、所属機関名:独立行政法人 農業生物資源研究所)

1) 要旨

大規模解析系を用いたプロジェクト型研究は専らデータベースとしてその成果を残しているが、散在するデータや加工物のデータベース群は有機的結合に乏しく、利用者に十分利用される形で提供されていないものが多い。このような状況を是正するため、国内の主要データベースの調査・解析を実施し、データベースの試験的な統合化を検討した。本課題では、農林水産省傘下研究機関のデータベース群について調査・解析及び試験的統合化を実施した。データベース調査の結果、多様な生物種の研究が実施されている農林水産省関係のデータベース群において、統合化が可能と思われるゲノム視点データベースは34個であった。生物種としては、イネ関連のデータベースが最も多く、イネの主要なデータベースで試験的統合化を試みた結果、使用されているRDMSの違い等により移植の必要性が生じたが、技術的には概ね統合化は可能と思われた。一方、研究機関をまたぐデータベース統合の問題、また多くのデータベースが時限的なプロジェクト予算(競争的資金)で作成されているためプロジェクト終了後の継続的公開のための予算(保守経費や情報更新経費)が確保出来ない等の問題が明らかとなった。今後、作成されているデータベースを有効に活用するためには、継続的な公開・情報更新が必要であり、そのための予算措置や維持体制の検討が不可欠である。

2) 目標と目標に対する結果

近年の大規模解析プロジェクト型の研究は、得られる情報量が非常に膨大であり、それら成果はデータベース(デジタル情報)として成果物を公表、公開している場合が極めて多い。特にゲノム研究が開始されて以来、ライフサイエンス研究のデータベース(知的集積物)は増加の一途をたどっている。しかしながら、これらのデータベースは一望困難であり、また有機的結合に欠けているため利便性が高いとは言えず、研究者に有効に利用されているとは言い難い。これらの状況を踏まえ、これまでに作成されているデータベースを調査し、有機的かつ効率的に活用するための方策の策定(データベース統合化)を目標とする。本目標を達成するため、本研究課題では農林水産省配下の研究機関における主要なデータベースの構造・機能、利用法及び維持運営方法等の情報の収集・調査・分析を実施し、データベース統合化に利用可能なゲノム視点の34個のデータベースを選抜した。これらデータベースの便覧を作成するとともに、データベース統合化のフィージビリティ・スタディとしてデータベースが充実しているイネについて実施し、データベース統合化ブラウザ(育種家、生命科学研究者の両者が利用可能なブラウザ)を開発し、データ入力を行い、プロトタイプを作成した。この作成過程において、複数種のリレーショナルデータベースの統一の必要性、欠落データベース(特に代謝関連データベース)の補完、統合化における独立行政法人間の評価及び情報共有化の問題等が明らかとなった。このような現状を踏まえ、今後は農林水産生物ゲノム情報統合データベースプロジェクトの中で、ユーザーに利便性の高い統合化データベース構築に取り組む。以上、本課題の目標は概ね達成できたと考えている。

3) 研究方法／調査方法

農林水産省傘下の各研究機関に散財すると予想される多種多様なデータベースの調査・分析を効率的に実施するために、調査項目をリストアップ・精査し、オンラインで入力可能なユーザーインターフェースの作成を実施した。また、Web インターフェースより入力された調査内容を特定の管理者のみが閲覧可

能な表示機能を装備した。作成した本調査システムを用いてデータベース調査・情報収集を実施した。収集したデータベース情報を分析し、その分析結果に基づき、データベース便覧を作成した。散在しているデータベース情報を有機的に連結し、統合化を図った場合の利用者を、大学・独法・都道府県の公的研究機関及び民間において基礎研究を実施している研究者(生命科学研究者)及び大学・独法・都道府県の公的研究機関及び民間において育種研究や実用化研究を実施している研究者(育種研究者)と想定し、多様なデータベースの統合化に向けたフィージビリティ・スタディを実施した。本フィージビリティ・スタディを通して、データベース統合化のあり方、可能性を検討し、また解決すべき課題(技術的、制度的問題点)を抽出した。それを踏まえて、今後のデータベースの維持・管理、情報の公開・提供について考察する。

4) 研究結果

(A) 農林水産省所管の研究機関が公開しているデータベースの調査と主要ライフサイエンスデータベースの分析

農林水産省所管の研究機関が対象とする生物種は、植物(農作物、果樹、林木等)、動物、昆虫、水産生物等多種多様である。最終目標のデータベース統合化を目指すため、農林水産省傘下の研究機関で公開されている個々のデータベースの調査実施(作成目的、データの質と量、データ内容の把握)した。本調査に先立ちインターネットを利用し農林水産省配下研究機関に散財するデータベースの事前調査を実施した。農林水産省所管の研究機関が公開している複数のデータベースについて事前調査を実施した結果の取りまとめは、連携施設群ポストゲノム「ゲノムネットワーク及びデータベース」小 WG の中間報告意見を参考にし、A)ゲノム視点のデータベース等と(B)その他の視点のデータベース等到大別した。125のデータベースを分類した結果、A)ゲノム視点のデータベースが30、(B)その他の視点のデータベース等と考えられるデータベースが95個存在した。(B)のデータベースの内、一部アノテーションに利用可能と思われるデータベースが7個存在した(表 4.1)。

表 4.1 農林水産省配下研究機関のデータベース調査(事前調査)

番号	データベース名	研究機関	ゲノム 視点 の DB	その他 の視点 の DB	備考
1	論文記事検索	農林水産政策研究所		○	一部アノテーションに利用可能
2	KTSOS 害虫関連文献データベース検索システム	農研機構・中央農業総合研究センター		○	一部アノテーションに利用可能
3	Web 版農業生物用語集-お試し版	農研機構・中央農業総合研究センター		○	
4	ダイズ草姿・葉形ファジィ検索	農研機構・中央農業総合研究センター		○	
5	農業機械 農作業・作業能率データ概要	農研機構・中央農業総合研究センター		○	
6	農業事例の概念検索	農研機構・中央農業総合研究センター		○	
7	捕食・寄生性昆虫ダニ類文献データベース PPRBASE 検索システム	農研機構・中央農業総合研究センター		○	
8	有用植物の病害診断/防除総合システム	農研機構・中央農業総合研究センター		○	
9	資源作物見本園	農研機構・作物研究所		○	
10	イネ品種・特性データベース検索システム	農研機構・作物研究所		○	一部アノテーションに利用可能
11	果樹園における生育調節剤とその使用法	農研機構・果樹研究所		○	
12	果樹害虫天敵リスト	農研機構・果樹研究所		○	
13	果樹関係新聞記事等所在情報検索システム	農研機構・果樹研究所		○	
14	果樹研究所育成品種情報検索システム	農研機構・果樹研究所		○	
15	昆虫の多様性	農研機構・果樹研究所		○	
16	昆虫病原糸状菌データベース	農研機構・果樹研究所		○	
17	主要研究成果情報検索システム	農研機構・果樹研究所		○	
18	職員の発表論文の検索	農研機構・果樹研究所		○	
19	ニホンナシ育成品種の系統図	農研機構・果樹研究所		○	

番号	データベース名	研究機関	ゲノム 視点 のDB	その他 の視点 のDB	備考
20	肥料取締法及び関係検索システム	農研機構・果樹研究所		○	
21	果樹品種情報検索システム	農研機構・果樹研究所		○	
22	果樹気象災害対策集(強風害・潮風害・ひょう害編)	農研機構・果樹研究所		○	
23	平成15年度果樹農業生産構造に関する調査報告書ー果樹農業に対する気象変動の影響に関する調査ー	農研機構・果樹研究所		○	
24	農業総合気象観測装置(金谷)	農研機構・野菜茶業研究所		○	
25	野菜茶業研究所 Q&A 検索システム	農研機構・野菜茶業研究所		○	
26	写真で見る外来雑草	農研機構・畜産草地研究所		○	
27	飼料作物病害図鑑	農研機構・畜産草地研究所		○	
28	草地植生ファクトデータベース	農研機構・畜産草地研究所		○	
29	畜産環境問題等に関する研究技術情報の検索システム	農研機構・畜産草地研究所		○	
30	飼料イネ技術情報	農研機構・畜産草地研究所		○	
31	監視伝染病	農研機構・動物衛生研究所		○	
32	NIAH 病理アトラス	農研機構・動物衛生研究所		○	一部アノテーションに利用可能
33	獣医細菌学関係細菌名一覧(IJSEM 準拠)	農研機構・動物衛生研究所		○	
34	家畜衛生研修会(病性鑑定:細菌部門)データベース	農研機構・動物衛生研究所		○	
35	家畜伝染病予防関係法規集	農研機構・動物衛生研究所		○	
36	家畜中毒情報	農研機構・動物衛生研究所		○	
37	写真で見る家畜の有毒植物と中毒	農研機構・動物衛生研究所		○	
38	電子顕微鏡で見た病原体	農研機構・動物衛生研究所		○	
39	おもしろい糸状菌	農研機構・北海道農業研究センター		○	
40	乾田直播の栽培適地判断データ	農研機構・北海道農業研究センター		○	
41	作物病害の画像情報	農研機構・北海道農業研究センター		○	
42	札幌市羊ヶ丘の気象データ	農研機構・北海道農業研究センター		○	

番号	データベース名	研究機関	ゲノム 視点 の DB	その他 の視点 の DB	備考
43	積雪統計データ検索	農研機構・北海道農業研究センター		○	
44	微生物の画像情報	農研機構・北海道農業研究センター		○	
45	マメハモグリバエ寄生蜂の図解検索	農研機構・北海道農業研究センター		○	
46	大曲の気象	農研機構・東北農業研究センター		○	
47	水稻冷害早期警戒システム	農研機構・東北農業研究センター		○	
48	直播関連研究事例集	農研機構・東北農業研究センター		○	
49	東北地方 1km メッシュ気温データ表示・検索システム	農研機構・東北農業研究センター		○	
50	畑作研究部気象観測システム	農研機構・東北農業研究センター		○	
51	九州沖縄のアメダス気象図	農研機構・九州沖縄農業研究センター		○	
52	九州沖縄の市町村別気象図	農研機構・九州沖縄農業研究センター		○	
53	農業生物資源ジーンバンク	農業生物資源研究所		○	一部アノテーションに利用可能
54	NIAS DNA BANK 農林水産DNAバンク	農業生物資源研究所	○		
55	イネゲノムリソースセンター	農業生物資源研究所	○		
56	Rice Genome Research Program (イネゲノム研究プログラム)	農業生物資源研究所	○		
57	ミュータントパネルデータベース (Tos17)	農業生物資源研究所	○		
58	RMOS(Rice Microarray Opening Site)(イネマイクロアレイサイト)	農業生物資源研究所	○		
59	KOME(イネ完全長 cDNA データベース)	農業生物資源研究所	○		
60	イネプロテオームデータベース	農業生物資源研究所	○		
61	作物学データベース	農業生物資源研究所	○		
62	イネ発現遺伝子データベース: RED	農業生物資源研究所	○		

番号	データベース名	研究機関	ゲノム 視点 の DB	その他 の視点 の DB	備考
63	RAP-DB:イネゲノム配列情報& アノテーション	農業生物資源研究所	○		
64	イネ第1染色体塩基配列情報	農業生物資源研究所	○		
65	RiceGAAS;イネゲノム自動アノ テーションシステム	農業生物資源研究所	○		
66	RAD:イネゲノムアノテーションデ ータベース	農業生物資源研究所	○		
67	イネ遺伝地図情報 Ver.1-Ver.4	農業生物資源研究所	○		
68	イネ物理地図情報 Ver.1-Ver.2	農業生物資源研究所	○		
69	イネ発現遺伝子地図情報	農業生物資源研究所	○		
70	イネゲノム専用 BLAST 検 索 (RiceBLAST)	農業生物資源研究所	○		
71	イネゲノム統合データベース:IN E	農業生物資源研究所	○		
72	植物転写因子結合シス配列デ ータベース:PLACE	農業生物資源研究所	○		
73	イネミトコンドリアゲノム情報	農業生物資源研究所	○		
74	イネ遺伝子機能予測データベ ース:Gfselector	農業生物資源研究所	○		
75	イネタンパク質構造データベー ス:RPSD	農業生物資源研究所	○		
76	ブタ EST 情報データベース: PEDE	農業生物資源研究所	○		
77	ブタ連鎖地図情報	農業生物資源研究所	○		
78	カイコ EST 情報データベース	農業生物資源研究所	○		
79	カイコ遺伝地図&物理地図情報 データベース:BombMap	農業生物資源研究所	○		
80	カイコホールゲノムショットガン情 報	農業生物資源研究所	○		
81	カイコゲノムアノテーションデー タベース	農業生物資源研究所	○		
82	カイコゲノムデータ BLAST 検 索 (KAIKO_BLAST)	農業生物資源研究所	○		
83	Ecosystem Database	農業環境技術研究所		○	
84	宇宙から見た日本の農業	農業環境技術研究所		○	
85	自然植生の純一次生産力と農 業気候資源の分布図	農業環境技術研究所		○	
86	植物細菌病の診断と病原細菌 の同定に関する情報	農業環境技術研究所		○	

番号	データベース名	研究機関	ゲノム 視点 の DB	その他 の視点 の DB	備考
87	農業環境用語事典	農業環境技術研究所		○	
88	アジア・太平洋外来生物データ ベース(APASD)	農業環境技術研究所		○	
89	微生物インベントリー (microForce)	農業環境技術研究所		○	
90	日本野生植物寄生・共生菌類 目録	農業環境技術研究所		○	
91	日本産糸状菌類図鑑	農業環境技術研究所		○	一部アノテ ーションに 利用可能
92	農環研標本館所蔵タイプ標本 データベース	農業環境技術研究所		○	
93	日本産オオアブラムシ属のチェ ックリスト及び種の検索表	農業環境技術研究所		○	
94	マメハモグリバエ寄生蜂の図解 検索	農業環境技術研究所		○	
95	植物の金属元素含量に関する データ集録	農業環境技術研究所		○	
96	施設栽培向け計測データセンタ ー	農業工学研究所		○	
97	SEICA[青果ネットカタログ]	食品総合研究所		○	
98	貯穀害虫・天敵図鑑	食品総合研究所		○	一部アノテ ーションに 利用可能
99	納豆研究文献データベース	食品総合研究所		○	
100	Fakara Plants(サヘルの植物資 源に関する情報)	国際農林水産業研究セン ター		○	
101	森林生物情報	森林総合研究所		○	
102	森林動態データベース	森林総合研究所		○	
103	スギゲノムデータベース	森林総合研究所	○		
104	東南アジア産材の木材特性デ ータベース	森林総合研究所		○	
105	日本産木材データベース	森林総合研究所		○	
106	木材標本庫データベース	森林総合研究所		○	
107	日本産木材識別データベース	森林総合研究所		○	
108	林業・林産関係国内文献データ ベース	森林総合研究所		○	
109	北海道樹木害虫図鑑	森林総合研究所・北海道支 所		○	

番号	データベース名	研究機関	ゲノム 視点 の DB	その他 の視点 の DB	備考
110	北海道で発生した森林昆虫に関する文献	森林総合研究所・北海道支所		○	
111	芽ばえ写真集	森林総合研究所・四国支所		○	
112	九州の主な樹木害虫	森林総合研究所・九州支所		○	
113	立田山の動物	森林総合研究所・九州支所		○	
114	立田山の森林	森林総合研究所・九州支所		○	
115	水生生物情報データベース	水産総合研究センター		○	
116	A-LINE Database	北海道区水産研究所、東北区水産研究所		○	
117	Off Tohoku Temperature Field	東北区水産研究所		○	
118	東北海区海況情報	東北区水産研究所		○	
119	リアルタイム海洋情報システム (種子島,串間,串本,片田)	中央水産研究所		○	
120	栄養塩クロロフィルデータベース	中央水産研究所		○	
121	水産海洋データベース	中央水産研究所		○	
122	日本海海洋環境部(日本海 NOAA 画像)	日本海区水産研究所		○	
123	大型クラゲ関連情報	日本海区水産研究所		○	
124	東シナ海・黄海のさかな	西海区水産研究所		○	
125	長崎近海水温速報	西海区水産研究所		○	

前述の事前調査を元に、調査項目を設定し、アンケート形式で情報収集を行うためのオンライン入力可能な Web インターフェースのデータベース調査項目蓄積システムを作成した(図 4.1)。本システムの調査項目は、入力者情報、データベース基本情報(DB 名、公開 URL、DB 管理者名等)、インターフェース情報(ユーザーインターフェース情報、アプリケーションインターフェース情報)、データ保存形式情報、データ項目情報及びその他の情報である。さらに本システムを利用して農林水産省配下研究機関の主要データベースの調査を実施し、分析結果の概要をまとめた(表 4.2)。

図 1 データベース調査項目蓄積システム

アンケート

生物統合DB構築のためのDB記録調査

入力者情報

氏名:

連絡先(E-mail):

データベース基本情報

DB名:

DB公開URL:

DB開発者(E-mail):

DB管理運営者:

データベースのインターフェース情報

ユーザーインターフェース

☐ HTML document ☐ BLAST search ☐ Clickable Map

☐ Keyword Search ☐ File Download

Other:

アプリケーションインターフェース

☐ DAS

Other:

データベースのデータ保存形式情報

データ保存形式

☐ Plain text ☐ Image data ☐ Blast DB ☐ Namazu index

☐ Oracle DB ☐ PostgreSQL DB ☐ MySQL DB ☐ Other Relational DB

2006年10月22日

RiceGAAS

入力者情報

氏名: 伊藤 浩司

連絡先: ikeda@abn.mech.uo.jp

DB基本情報

DB名: RiceGAAS

公開URL: http://ricegaas.dna.affrc.go.jp/

DB開発者: igor@dna.affrc.go.jp

管理運営: 伊藤 浩司

DBインターフェース情報

ユーザーインターフェース

HTML document: ☒ Clickable Map

Keyword Search: ☒ BLAST: File Download: ☒ BLAST: ☒ BLAST: Anonymous FTP: ☒ 不明

アプリケーションインターフェース

☐ DAS

データベースのデータ保存形式情報

データ保存形式

☐ Plain text ☐ Image data ☐ Blast DB ☐ Namazu index

☐ Oracle DB ☐ PostgreSQL DB ☐ MySQL DB ☐ Other Relational DB

DBデータ情報

保存形式	データ型	項目名	項目注釈	データ数
Sequence	BLAST/AC	Contig	-	4234
	Nucleotide	予測遺伝子 (cds)	207417	
	Nucleotide	予測遺伝子	207417	
	Protein	予測遺伝子 (翻訳)	207417	
Gene	GenePredictor	Gene Matrix	4234	
	Nucleotide	-	207417	
Protein	Protein	Protein	4234	
	Protein	予測遺伝子 (翻訳) より再翻訳	207417	
Protein	Protein	Protein	139630	
	Protein	Protein	4234	
Protein	Protein	Protein	207417	
	Protein	Protein	4234	
Protein	Protein	Protein	67767	
	Protein	Protein	67767	

入力画面

入力結果画面

図 4.1 データベース調査項目蓄積システム

表 4.2 ゲノム関連データベースの調査・分析結果概要

No.	データベース名称	DB 分類	DB インターフェース情報	DB 項目	データ数
1	日本産糸状菌類図鑑	総説型	HTML document	名称、学名、形態、分類等	-
2	Sugi Genome Database	プロジェクト型	HTML document	マップ、電気泳動画像、配列データ等	-
3	BombMap	プロジェクト型	HTML document Clickable Map	遺伝地図、物理地図、BAC 情報、プライマー情報	-
4	CROP SCIENCE DATABASE	総説型	HTML document	文献、作物学会誌	
5	イネ品種・特性データベース	知識モデル型	HTML document	水稻・陸稻の品種情報	11,587
6	Rice Mitochondrial Genome Information	プロジェクト型	HTML document	物理地図、ゲノム情報、Annotation 情報	108
7	PLACE	知識モデル型	HTML document	cis-acting regulatory DNA elements のモチーフ情報、文献から抽出	469
8	Rice PIPELINE	プログラム型	HTML document	ゲノム、蛋白質、FL-cDNA 等の検索	-
9	Rice Microarray Opening Site	プロジェクト型	HTML document	材料情報、主にアレイデータ	8,987
10	Rice Expression Database	プロジェクト型	HTML document	発現解析データ	-
11	Genomes TO Protein structures and functions	プログラム型	HTML document	配列 Annotation	-
12	タンパク質構造データベース	プログラム型	HTML document	蛋白質 1 次、2 次、3 次情報	357
13	Rice Proteome Database	プロジェクト型	HTML document Clickable Map	2次元画像	23
14	KAIKOBLAST	プログラム型	BLAST search	WGS contig, BACend Sequence 等	6,426,163

No.	データベース名称	DB 分類	DB インターフェース情報	DB 項目	データ数
15	KAIKOGAAS	プログラム型	HTML document	WGS contig, 配列アノテーション等	7,788
16	SGP(Silkworm Genome Research Program)	プロジェクト型	HTML document	assembled WGS Contig	213,289
17	KAIKOcDNA	プロジェクト型	HTML document	EST	41,068
18	Swine Linkage Map Viewer	プログラム型	HTML document Clickable Map	連鎖地図情報	-
19	PEDE(Pig Expression Data Explorer)	キュレーション型	HTML document	EST 情報、SNP 情報、クラスター情報	261,047
20	KTSOS 害虫関連文献データベース検索システム	総説型	Keyword Search	雑誌名、内容キーワード	-
21	論文記事検索	総説型	Keyword Search	雑誌名、内容キーワード	-
22	IRGSP Physical Map	プロジェクト型	HTML document	物理地図情報	-
23	End-sequencing and in silico mapping of BAC clones from an indica rice variety, Kasalath.	プロジェクト型	HTML document	ゲノムマップ、物理地図	-
24	High-Density Rice Genetic Map	プロジェクト型	HTML document	遺伝地図情報	3,267
25	PCR-based genetic markers on rice chromosomes	プロジェクト型	File Download	プライマー情報	332
26	A YAC-Based Rice Transcript Map	プロジェクト型	File Download	物理地図、EST 情報	6,670
27	Mutant Panel Database (Tos17 mutant lines)	プロジェクト型	HTML document Clickable Map	変異体情報、画像	19,000
28	Knowledge-based Oryza Molecular biological Encyclopedia	プロジェクト型	HTML document	配列情報	32,127

No.	データベース名称	DB 分類	DB インターフェース情報	DB 項目	データ数
29	Rice Annotation Database	キュレーション型	HTML document Clickable Map RDBMS(Oracle)	PAC/BAC contig	1,922
30	Rice Annotation Project Database	キュレーション型	HTML document Clickable Map	ゲノム情報、アノテーション情報	40,041
31	貯穀害虫・天敵図鑑	総説型	HTML document	名称、学名、形態、画像、防除方法等	-
32	RiceGAAS	プログラム型	HTML document	配列情報、アノテーション情報	207,417
33	INE (INtegrated Rice Genome Explorer)	プロジェクト型	HTML document Clickable Map RDBMS(Oracle)	連鎖&物理地図情報、ゲノム情報	-
34	RGP (Rice cDNA Sequence Database)	プロジェクト型	HTML document	イネ EST 情報	61,361

a) データベースの分類と情報の種類

ゲノム視点のデータベースとそのアノテーション情報として利用可能なデータベースから 34 件に絞り込み、これらのデータについて、データベース分類、データベースインタフェース情報、データベース項目、データ数についてまとめた。データベース分類(本プロジェクト事務局による分類)結果では、「バンク型」0 件、「プロジェクト型」17 件、「プログラム型」8 件、「キュレーション型」2 件、「知識モデル型」2 件、「総説型」5 件に分類された。

・バンク型 (0 件)

研究者から投稿されたデータを集めたデータベース。データ形式は構造化されたテキストであり、DDBJ 塩基配列データベースが代表的な例である。今回収集した農林水産省傘下の研究機関が公開しているデータベースではこの型にあてはまるものは無い。

・プロジェクト型 (17 件)

様々な研究プロジェクトで計測されたデータを集めたデータベース。データ形式は各プロジェクト毎に決定されているため、リレーショナルデータベースシステムを採用しているものや構造化テキストを採用しているものなど多種多様である。今回収集した農林水産省傘下の研究機関が公開しているデータベースでは、プロジェクト型がその大多数を占める。

・プログラム型 (8 件)

定型的な機械処理により加工されたデータベース。データの質という意味では後述のキュレーション型に劣るが、データ構築やデータ更新を自動化できるためデータベース更新の効率に優れる。また、機械処理の機能をユーザーに提供し、RiceGAAS や KAIKOGAAS のようなユーザーが登録したデータを解析するというサービスも可能である。データ形式はプロジェクト型と同様に各データベース毎に決定されているため、リレーショナルデータベースシステムを採用しているものや構造化テキストを採用しているものなど様々である。INE (INtegrated Rice Genome Explorer)では、市販のリレーショナ

ルデータベースシステム(Oracle)が使用されている。今回収集した農林水産省傘下の研究機関が公開しているデータベースでは、プロジェクト型の次に件数が多い型である。

・キュレーション型 (2 件)

専門家がデータをチェックし注釈付けなどの作業を行いその結果をまとめた信頼性の高いデータベース。一般的にデータを前段で機械処理を行いその結果を専門家が評価、判断するという工程で作成されている。大量のデータをキュレーションするためには多数の専門家と多くの時間が必要であり、データ更新の頻度は低くなる。データ形式はプロジェクト型と同様に各データベース毎に決定されているため、リレーショナルデータベースシステムを採用しているものや構造化テキストを採用しているものなどがある。今回収集した農林水産省傘下の研究機関が公開しているデータベースでは、2 件(RAP-DB、RAD)のみがキュレーション型に相当する。RAD は、市販のリレーショナルデータベースシステム(Oracle)を使用している。

・知識モデル型 (2 件)

専門家が論文などの文献から有用な情報を抽出し、一括で利用可能にしたデータベース。抽出した情報から要素の関係などを考慮して知識モデルとして表現したパスウェイデータベースなどがその代表である。情報収集に専門家の多くの時間が必要となり、データ更新の頻度は低くめである。今回収集した農林水産省傘下の研究機関が公開しているデータベースでは、2 件(イネ・品種特性データベース、PLACE)が知識モデル型に相当する。

・総説型 (5 件)

専門家が論文などの文献を集めキーワードによる検索を可能としたデータベース。生物の形態で分類した図鑑もこの型に含まれる。情報収集には専門家の知識が必要であり、データベースの構築も機械処理が困難であるため、データ更新の頻度は低めである。図鑑においては画像を搭載したデータベースが多く見られる。今回収集した農林水産省傘下の研究機関が公開しているデータベースでは、5 件が総説型に相当しており、農林水産系のデータベースでは比較的多い型と言える。

データの種類から俯瞰すると、ゲノム情報(配列情報)、cDNA 情報、遺伝地図情報、物理地図配列情報関連データベース及びタンパク質(プロテオーム DB、2次元電気泳動情報)は充実している。また、形態・形質情報(表現型)に関する変異体データベース(レトロトランスポゾン Tos17 を利用した変異体データベース)及びイネ品種データベースも作成されている。ゲノム関連データベースや表現型データベースは、農業上重要な形態・形質関連遺伝子、例えばイネの開花、収量、耐病性、耐暑及び耐冷性関連遺伝子等の単離・機能解析に有効に利用可能であると考えられる。しかしながら、今後、イネの生命現象の解明や収量に関わる光合成機能の改変等の実用化研究を進展させるためには、代謝関連のデータベース(メタボローム研究領域)やホルモン関連データベース(ホルモノーム研究領域)の充実が必要であり、現段階では、これらの研究及びデータベースが欠落していると言える。

b) データ提供方法・利用方法

各データベースはいずれも Web を使用して公開されているが、使用言語は英語標記のみが 22 件、日本語のみが 6 件、英語&日本語が 6 件であった。研究プロジェクトで作成されたデータベースは、アカデミックな利用者(主に生命科学研究者)を対象としたデータベース構築が多く、日本の中学生や高校生、民間企業の研究者、また育種従事者等向けのデータベース開発、製作、公開となっていないものが多い。データベースの実質的な活用及び利用者の拡大を図るためには、日本語版と英語版を作成する事が望ましい。さらに、データダウンロード機能においては、一部のデータベースは、データベース作成に用いたデータセットのダウンロード機能を装備しているが、ほとんどのデータベースは装備されていない。このことは、利用者のデータベースあるいはデータの利用率向上を図るための重要な要因の1つであると考えられることから、今後情報のダウンロード機能装備が必要である。

(B) 日本におけるイネ品種育成方法の調査結果

日本におけるイネの育種は、1990年代から本格的に農林水産省の研究機関によって進められ、集団育種法(系統育種法)によって改良が行われてきた(図4.2)。この方法によって現在の主要銘柄であるコシヒカリをはじめとする多数の良食味品種が開発されている。しかしながら、集団育種法による選抜育種は、1) 育種期間が約10年と長期間を要すること、2) 選抜のために多品種の組み合わせ、膨大な系統数からの選抜が必要であり多大な労力を有すること、3) 系統選抜には、育種家の知識・経験及び選抜眼に依存する等の課題も含まれている(図4.3、4.4)。これらの課題を克服し、集団育種法に替わる新たな育種法を確立するためには、近年蓄積されているイネゲノム成果、DNAマーカーによる遺伝地図情報、物理地図情報及びゲノム配列情報等を有効に利用することが新たな品種開発の1つの方法であると考えられる。しかしながら、個々のデータベースは、独立して存在・機能しているが、育種家や育種現場で利用できる使い易いデータベースとなっていないのが現状である。独立に存在・機能している個々のデータベースを有機的に結合させ、育種家や育種現場に携わる研究者が利用可能なデータベースやブラウザを開発すれば育種現場においてゲノム研究で得られた情報を有効に活かしたゲノム育種(デザイン育種)が可能になると予想される(図4.5)。このような視点からもゲノム関連データベースの統合化は重要な位置づけにある。

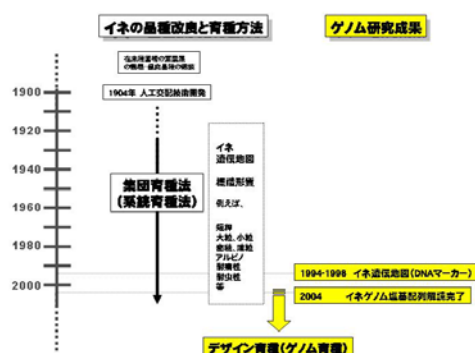


図 4.2 イネ育種の歴史

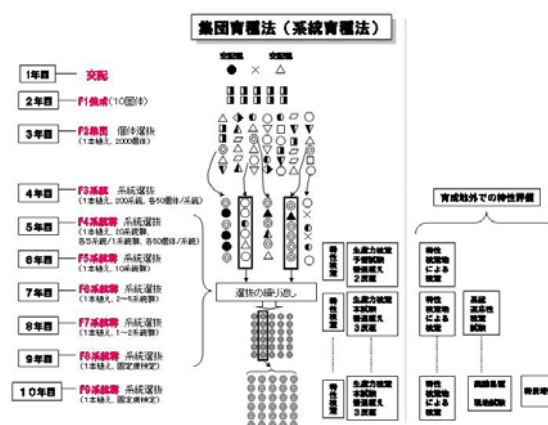


図 4.3 イネの育種法:集団育種法

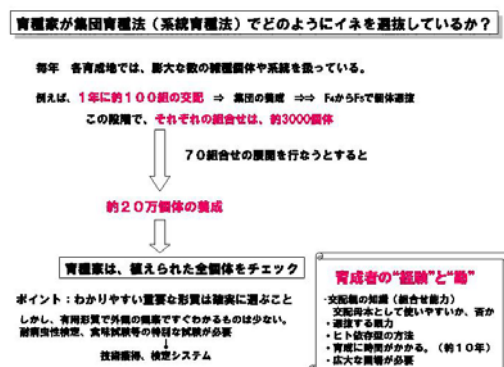


図 4.4 集団育種法の現状と課題

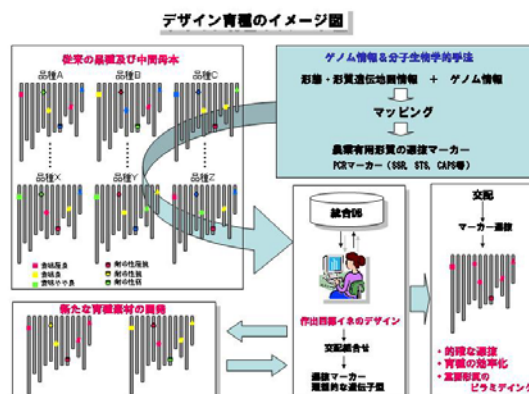


図 4.5 ゲノム育種の可能性

(C) データベース統合化のフィージビリティ・スタディ実施と今後の課題

農林水産省が保有しているデータベースについて、統合化のあり方や可能性(技術面)について検討し、データベースの種類・内容が最も充実している“イネ”

関連のデータベース統合化をフィージビリティ・スタディとして実施した。統合化の基本的考え方として、統合化したデータベースのユーザーをイネ育種家と生命科学研究者の両者を想定し、イネ統合データベースの骨格及び統合化データベースのイメージを図 4.6、4.7、4.8 に示した。

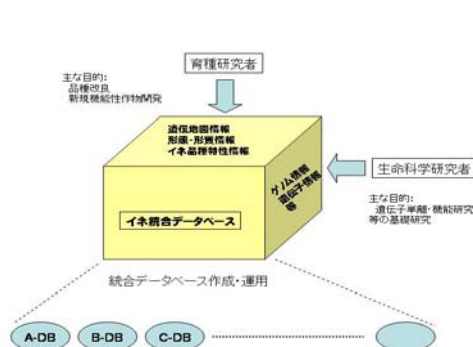


図 4.6 イネ統合データベースの利用者



図 4.7 イネ統合データベースの基本骨格

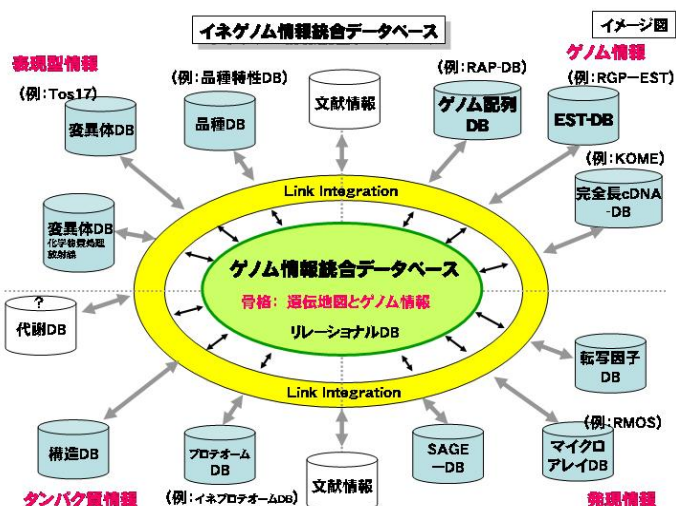


図 4.8 イネ統合化データベースのイメージ図

具体的には、収集・調査したデータベース34件から、先行して育種家向けに必要な遺伝地図情報及びマーカー情報(RFLP マーカー、STS マーカー、SSR マーカー情報)をもつデータベース及びイネゲノム配列情報をもつデータベースをリストアップし、フィージビリティ・スタディとしてイネ第1染色体上の各マーカーのゲノム上へのマッピングを行った。その結果、遺伝地図上のマーカーの配列順序とゲノム上にマップしたマーカーの配列順序に矛盾する箇所が認められた。大規模解プロジェクト型研究においては、扱うデータ量が膨大であり、“誤り”を皆無にすることは極めて難しいと思われるが、現存するデータベース、統合化データベースのいずれにおいても“データの信頼性”は重要であり、明確な“誤り”については修正する必要がある。同時に利用者による指摘や意見を受け入れる仕組みを設け、より信頼性の高いデータベース作成・維持のシステム作りが必要である。イネゲノム配列データ(RAP-DB Build4)上にマップされたマーカー情報(RFLP マーカー、STS マーカー、SSR マーカー情報)で矛盾の確認されたマーカーについては、随時オリジナル情報(写真及び遺伝子型情報)を確認し、修正を図った。

上記のようにイネゲノム配列情報及び DNA マーカー情報(育種における選抜マーカー)を整理し、最もデータが充実しているイネゲノム情報を利用して、育種家向け及び生命科学者向けのデータ統合ブラウザプロトタイプを開発中である(図 4.9)。本システムは、DNA マーカー情報をゲノム配列にマッピングした情報を利用し、その上に農業形質の遺伝子を配置し、文献情報も参照可能としている。データ統合ブラウザは、Ajax の技術を活用し、ゲノム上の領域を Google Maps のようにスクロール可能であり、任意の領域のゲノム配列も抽出可能とした。さらに、育種家向けとして DNA マーカーを配置したゲノムブラウザへのリンクを設け、生命科学者向けとして RAP-DB(Rice Annotation Project Data Base)のリンクを設けている。これらのリンクは、データ統合ブラウザで選択した領域で絞り込むことが可能である。

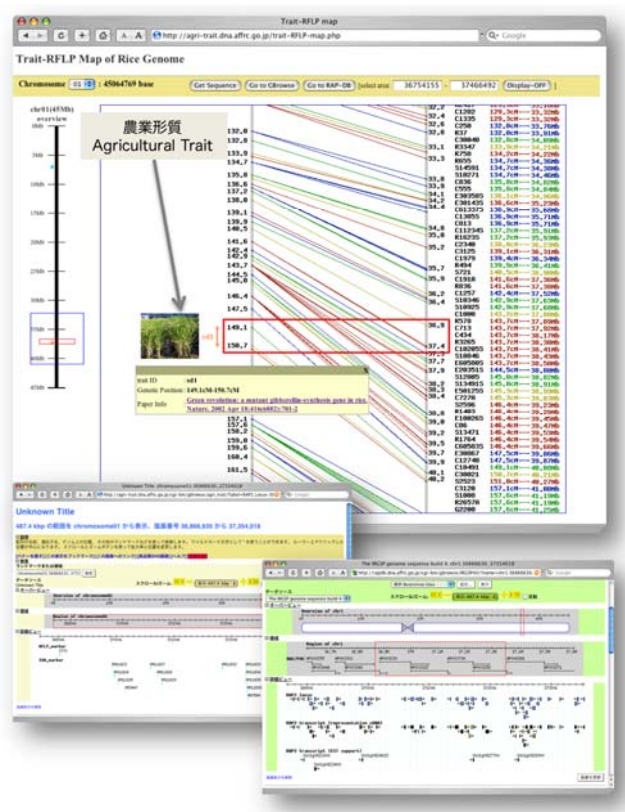


図 4.9 統合データベースブラウザプロトタイプ

また、データ統合ブラウザの開発と平行して、統合データベース検索用プラットフォームの開発も進めている。このシステムは、ユーザーが指定する入力キーワードをネットワーク上のリレーショナルデータベースに問い合わせる必要な情報を取得するイネゲノム用のシステムである。キーワード検索により、イネゲノムのアノテーション情報や変異体の情報等、データベース横断的に検索可能な検索用プラットフォームである(図 4.10)。



図 4.10 検索用プラットフォーム

今後の課題：

- ・ 農業上重要な形質に関する文献情報の収集・精査および質の高い形質情報の入力
イネの形質情報に関する論文等の文献情報は多数あり、時間と手間（人的労力による有用な情報整理）をかけて精査し、質の高い有用な情報のみを入力する必要がある。
- ・ 形質に関与する遺伝子・タンパク質、発現情報、変異体情報等との有機的な結合（リンク）
リンクを張るための共通項目の抽出（例えば、アクセッション番号）と用語辞書（同一の形質や部位を指す異なる表現方法の整理）の作成が必要である。
- ・ 有償の RDBMS (Oracle 等) からフリーの RDBMS (PostgreSQL) への情報の移植
- ・ 本調査で得た情報やデータベース統合化のフィージビリティ・スタディを踏まえて、今後データベース統合化は、農林水産生物ゲノム情報統合データベース作成（委託プロジェクト）の中で進める。

(D) データベース統合化の問題点について

農林水産省傘下研究機関で作成・公開しているデータベース群の調査・分析を実施し、それらの情報をもとにデータベース統合化のフィージビリティ・スタディを行った。

その結果、下記の問題点・課題があると考えられる。

- ・ データベース作成・構築は、主に政府資金のプロジェクト（大規模のデータ生産型プロジェクト）で行われているが、プロジェクト終了後に予算が確保（ハードウェア保守経費や情報更新）できないために、公開を継続できない場合がある。（制度的問題）
継続して公開する必要がある、事業的予算（安定した予算）の確保が必要と考える。
- ・ データベースの多くは、研究ジャーナル等に論文として投稿・記載されている場合が多いため、統合化にあたっては作成者の貢献を維持（評価）しつつ、統合化を図る仕組みが必要であると思われる。（制

度的問題)

- ・ 農林水産省傘下には、複数の独立行政法人研究機関が存在する。各独立行政法人の評価は個別に行われており、作成されているデータベースそのものも評価の対象になる。従って、独立行政法人毎に存在するデータベースの統合化においては、リンクを含む機能や新たなシステムを活用して情報の共有化を図る必要がある。(制度的問題)
- ・ 作成されているデータベースは、OS, RDBMS, アプリケーション、情報の種類等様々である。従って、データベース統合化にあたっては、データベース統合化を実施する専門集団の構築とデータの安全な保管管理及びデータの共有(公開・維持)を行う組織の構築が必要であると思われる。また、多額な公的資金を投入して得られたデータ及びデータベースの保管・管理については、危機管理の観点から、2重保存の必要性もあると考える。

5) 考察・今後の発展等

農林水産省傘下研究機関で作成・公開しているデータベース群の調査・分析を実施し、それらの情報をもとにデータベース統合化のフィージビリティ・スタディを行った。本調査研究を通して、得られた今後の課題・発展等について考察する。

- ・ データベース作成・構築は、主に政府資金のプロジェクト(大規模のデータ生産型プロジェクト)で行われているが、プロジェクト終了後に予算が確保(ハードウェア保守経費や情報更新)できないために、公開を継続できない場合がある。継続して公開する必要がある、事業的予算(安定した予算)の確保が必要と考える。データベースの多くは、研究ジャーナル等に論文として投稿・記載されている場合が多いため、統合化にあたっては作成者の貢献を維持(評価)しつつ、統合化を図る仕組みが必要であると思われる。農林水産省傘下には、複数の独立行政法人研究機関が存在する。各独立行政法人の評価は個別に行われており、作成されているデータベースそのものも評価の対象になる。従って、独立行政法人毎に存在するデータベースの統合化においては、リンクを含む機能や新たなシステムを活用して情報の共有化を図る必要がある。また、作成されているデータベースは、OS, RDBMS, アプリケーション、情報の種類等様々である。従って、データベース統合化にあたっては、データベース統合化を実施する専門集団の構築とデータの安全な保管管理及びデータの共有(公開・維持)を行う組織の構築が必要であると思われる。また、多額な公的資金を投入して得られたデータ及びデータベースの保管・管理については、危機管理の観点から、2重保存の必要性もあると考える。今後データベース統合化は、本調査で得た情報やデータベース統合化のフィージビリティ・スタディを踏まえて、農林水産生物ゲノム情報統合データベース作成の中で進めて行く予定である。

6) 関連特許

該当なし

7) 研究成果の発表

(成果発表の概要)

1. 原著論文(査読付き)	該当なし
2. 上記論文以外による発表	該当なし
3. 口頭発表	該当なし
4. 特許出願	該当なし
5. 受賞件数	該当なし

(5) サブテーマ4 産業(農林水産業以外)データベースの統合にかかわる調査研究

(分担研究者名: 今西 規、所属機関名: 産業総合技術研究所)

1) 要旨

本サブテーマでは、産業(農林水産業以外)分野のデータベースの統合にかかわる調査研究を実施した。主な研究成果としては、まず、データベース基礎調査により経済産業省関連の各種データベースのカタログ作成が完了した。そして、産総研等の研究機関におけるデータベースの運用状況と、企業等におけるライフサイエンス分野のデータベースへの主なニーズを調査した。また、フィージビリティ・スタディを実施し、データベースの統合化や連携のためにはデータベース維持管理のための自動化システムが有効であることを見いだした。これらの結果をふまえて、データベースの客観的な評価法や、今後考えられる統合データベースの方式などについて考察を行った。

2) 目標と目標に対する結果

本サブテーマでは、産業(農林水産業以外)分野のデータベースの統合にかかわる調査研究として、「①データベース基礎調査」、「②サブテーマ研究検討委員会の開催」、「③データベース統合に関するフィージビリティ・スタディの実施」、を行う。具体的には、産業技術総合研究所をはじめとする経済産業省配下の機関にて構築されたライフサイエンス分野のデータベースの調査を行い、そのカタログ作成を進め、データベース便覧にまとめる。また、検討委員会やインタビュー形式での調査を行うことにより、創薬をはじめとする産業界で必要とされる公共データベースに期待される役割や満たすべき要件を明らかにする。また、フィージビリティ・スタディとして H-InvDB と GeMDBJ 等の試験的統合化を試み、データベース間の連携・統合化のために必要な各種要件を明らかにする。これらの活動を通じて、経済産業省におけるバイオデータベース関連の制度設計やロードマップ作成に資する試案を検討する。目標に対する結果は、詳細を下記項目4)に記載した。

3) 研究方法／調査方法

本サブテーマの調査方法は、「①データベース基礎調査」では NEDO や産業総合技術研究所、製品評価機構のホームページや各種の報告書類を使い、一部に担当者に対するインタビューを使いながら行った。「②サブテーマ研究検討委員会の開催」は、主に産業界のデータベースへのニーズを明らかにするために、企業で研究開発を行っている研究者等へのインタビュー等を行った。「③データベース統合に関するフィージビリティ・スタディ」としては、独立に運営されている複数のデータベースの連携の最適な方法を模索するために、実際のデータベースの相互リンクを作成し、さらにその自動管理システムの設計と試作を行った。このほか、各種のデータベースの評価基準に関するデータ収集とその分析は、文献データベース等を利用して実施した。

4) 研究結果

(A) 経産省ドメインのライフサイエンス分野データベースの基礎調査

新エネルギー・産業技術総合開発機構(NEDO)、産総研、製品評価技術基盤機構(NITE)にて公開されているライフサイエンス分野データベースの基礎調査を行った。結果はデータベース便覧として電子媒体で保管するとともに、ホームページ上で公開した

(http://staff.aist.go.jp/t.imanishi/DB_binran2007v4.xls)。

a) 新エネルギー・産業技術総合開発機構の成果報告書調査

新エネルギー・産業技術総合開発機構(NEDO)ホームページ上で公開されている成果報告書の調査を行った。「データベース」をキーワードとした NEDO 成果報告書全文の検索と、分野ごとの報告書のマニュアルでの洗い出しによって、最終的にバイオ産業に関連があると思われる成果報告書 533 件の存在を確認した。これらの成果報告書は全文を閲覧可能であるため一件ずつ報告書を調査したところ、データベース構築を行ったという記述のある成果報告書が 14 件あった。このうち一般に公開されていて現在利用可能なデータベースはわずか5件であった。

b) 産業技術総合研究所管轄下のデータベース調査

産業技術総合研究所(産総研)ホームページ上で公開されている研究情報公開データベースよりバイオ関連データベースの調査を行った。その結果、バイオ産業に関連のある 9 つデータベースの存在を確認した(微細藻類遺伝子工学、脳画像、転写産物データベース、古細菌情報コレクション、糸状菌ゲノム・転写制御、Digital Brain Atlas at neuron level、古細菌遺伝子ネットワークデータベース、微生物データベースシステム、創薬関連の膜たんぱく質データベース)。また、産総研・生命情報科学研究センター(CBRC) ホームページ上のデータベースについても調査を行い、公開されている 4 つデータベースの存在を確認した(ASTRA、GENIUS II、PDB-REPRDB、EzCatDB)。以上のデータベースの概要をまとめ、分類を行った。

c) 製品評価技術基盤機構管轄下のデータベース調査

製品評価技術基盤機構(NITE)ホームページ上で公開されている研究情報公開データベースよりバイオ関連データベースの調査を行った。NITE のホームページより、公開されている 6 つデータベースの存在を確認した(DOGAN、NBRC Culture、ヒト cDNA クローン検索、人間特性データベース、基本動態計測手法の開発、生体障害データベース)。以上のデータベースの概要をまとめ、分類を行った。

(B) インタビューによる各機関のデータベース運用状況の調査

研究成果報告書およびインタビュー形式での聞き取り調査に基づき、経済産業省傘下の主要な研究機関(産業技術総合研究所、新エネルギー・産業技術総合開発機構(NEDO)、製品評価技術基盤機構(NITE))を対象として、生命科学分野の公開データベースの運用状況を調査した。その結果、主に以下のことが明らかになった。

a) 産総研(*1)とNITE(*2)には、研究成果のデータベース化と維持のための持続性のあるシステムが存在する。

(*1)産総研のRIO-DBは、産総研の研究成果を公開するための枠組みを提供している。一定の基準による評価のもと、研究の過程で蓄積されたデータのデータベース化(形式の統一化、網羅性を補うための計測やデータの電子化等)を実施できる(全体で1.2億円/年、91種類のデータベース[バイオ系約10件]が対象)。また、イノベーション推進室にて、イノベーション創出のための基盤形成に資するデータベースの管理・運営を支援している。

(*2)NITEでは、中期目標にて、微生物のゲノム配列決定からアノテーション(注釈づけ)、データベース構築、公開、維持・更新までを一貫して位置づけ、実施している。

b) NEDO プロジェクトにおいては、データベースの構築がプロジェクトの基本計画に位置づけられているか否かにかかわらず、プロジェクトの成果物は成果報告書のみである。なお、研究成果の利用・普及については「その実現に努めること」としており、プロジェクトで構築されたデータベースの維持・更新は、実施者に委ねられている。

(C) データベース評価法に関する調査

検討会議を数回開催し、データベース評価のための尺度に関する調査を行った。

学術論文を評価するための尺度としては、論文の引用回数を用いることが多い。同様に、データベースを評価するための尺度として、そのデータベースに関する学術論文の引用回数を用いることが考えられる。しかし、データベースの用途は必ずしも学術分野に限らないため、論文での引用回数では適切な評価とは言えないかもしれない。しかも、論文が別の論文に引用されるまでには発表から十分に時間が経過していなくてはならず(2年程度は必要と思われる)、データベースの公開直後に引用回数で評価を行うことは困難である。

そこで、データベースを評価するための別の評価として、データベースのアクセス数を使うことが考えられる。実際、経済産業省モデル事業「ゲノム情報統合プロジェクト」(平成17-19年度)では、ヒト遺伝子の統合データベースである H-Invitational Database の評価のために、アクセス数を用いている。しかし、アクセス数にもさまざまな問題がある。例えば、Googleなどの検索サービス提供会社は「ロボット」と呼ばれる自動ソフトを使ってデータベースに頻繁にアクセスする場合があります、正確な利用者数の集計を困難にしている。

こうした状況から、データベースの評価のために最も適切な尺度を明らかにすることを目標に、考えられる数種類の尺度を調査した。ここでは、論文の引用回数、データベースの被リンク数、テキスト内でのデータベース名の言及回数について調査した。そして、互いに相関が高いかどうかを調べた。

主な発見としては、データベースの被リンク数が論文の引用回数とやや高い正の相関を示した。Googleにより「類似フィルターあり」、「外部リンク URL 数(ホスト重複除去)」の条件で数えた被リンク数が、トムソン社の Web of Science での論文引用回数との間で 0.56 という相関係数を示し、最も高い相関であった。「ホスト重複除去」とは、同一のウェブサイトからの複数のリンクを重複して数えないことである。なお、この調査には43の国内外のデータベースを用いた。一方、アクセス数についてはごく限られた計測データしか入手できなかったために詳細な検討ができなかったが、被リンク数や論文の引用回数が多いデータベースはアクセス数も多い傾向があるものと推測された(数件の産総研の公開データベースでの情報に基づく)。

(D) データベース統合化に関するフィージビリティ・スタディとしての「リンク自動管理システム」開発

異なる機関で独立に運営されているデータベースの間での密な連携を実現するためには、データの対応関係を調べて相互リンクを設置することが効果的である。しかし、多数のデータベースの間での相互リンクを維持管理するには、データベースの組み合わせが非常に多くなるため、膨大なコストがかかる(図 5.1)。そこで、相互リンクの管理を完全に自動化するための「リンク自動管理システム」を設計・試作し、GeMDBJ と H-InvDB の統合化を試みた。本システムの詳細は以下の通りである。

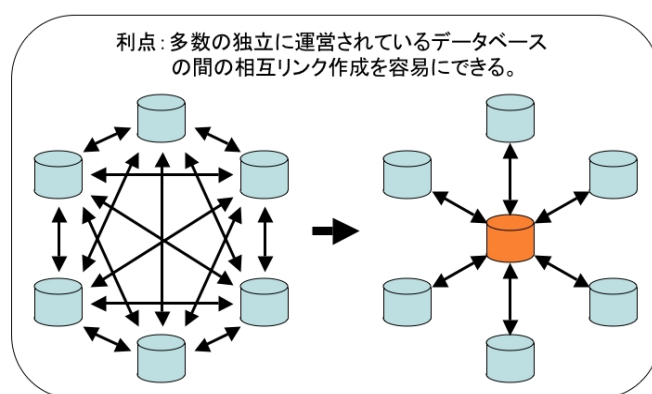


図 5.1 相互リンクの管理を容易にする効果のイメージ図

「リンク自動管理システム」は、DDBJ/EMBL/GenBank のアクセッション番号などを「共通 ID」と定義し、各々のデータベースが独自に管理する ID と共通 ID の対応関係を、各データベースから毎日ダウンロードしてデータ対応表を作成する。これを共通 ID 管理データベースと呼ぶ。

共通 ID 管理データベースが持つ情報を参照することにより、あるデータベース中のデータと別のデータベース中のデータの間の対応関係を抽出し、該当するデータを参照するための URL を返す cgi を作成した。この cgi をリンク自動転送システムと呼ぶ(図 5.2)。このシステムが提供する cgi へのリンクを設定することにより、外部データベースへのリンクが容易に設定できる。例えば、H-InvDB から GeMDBJ へのリンクを設定する場合には、

http://biodb.jp/hfs.cgi?db1=H-InvDB&type=HIX_ID&db2=GemDBJ&id=HIX0000794

などのように cgi を呼び出す。するとこの cgi が GemDBJ 中の該当するデータの URL を自動的に返し、その URL が開く。ここで重要なことは、リンク先データベース中での対応するデータの ID を知らなくても、リンクが設置できることである。共通 ID 管理データベースの内容は自動的・定期的に更新されるため、データの更新にともなう外部データベースへのリンクの再設定の人手での作業が不要になる。また、対応がなくなかったデータはその旨が表示されるため、「リンク切れ」により利用者へストレスを与えることも発生しなくなると期待される。以上のシステムを設計し、ソフトウェア部分は外注にて作成した。本システムはテストサーバ上で期待通りに動作している。

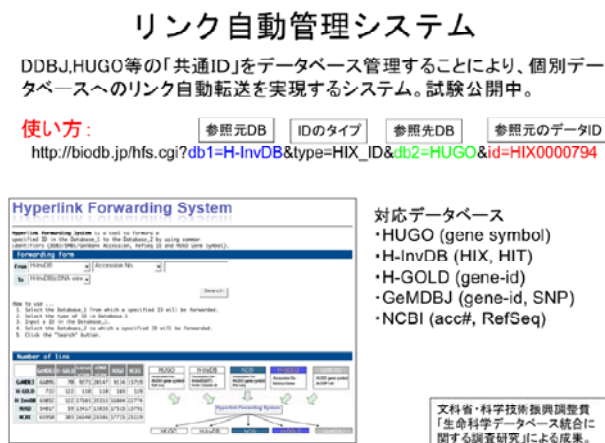


図 5.2 リンク自動管理システムの概要

(E) 産業分野でのライフサイエンス分野データベースに関するニーズ調査

製薬、食品、分析機器などの産業分野を対象に、インタビューの実施や各種統計資料の調査によって、ライフサイエンス分野の公共データベースに関するニーズ調査を行った。目的は、産業分野で必要とされるライフサイエンス関連データベースの洗い出しと、現実および潜在的ニーズの調査である。2006 年から2008 年にかけてこれらの産業分野の数社を訪問したり、各種公開資料を使った調査を行った。

まず、製薬業界の一般的な状況に関して、社団法人バイオ産業情報化コンソーシアム(JBIC)の佐藤清事務局長から次のようなコメントをいただいた。『製薬業界では「情報」の取得のために巨額の投資をしており、特に海外の市販データベースを多数採用している。活用するデータベースの種類としては、文献データベースはもちろん、それ以外にも化合物、パスイエイ情報、臨床試験情報、特許など非常に多岐にわたる。したがって、産業界が本当に必要とする公共データベースの種類を特定することは、非常に難しい。一方、ゲノム情報の活用は限定的でしかない。また、外資系の製薬会社は海外の本社で各種「情報」へのアクセスが可能であるため、日本の公共データベースには大きな期待はしていない可能性がある。』

以上のことから、製薬業界が期待する公共データベースは、既存の情報を整備したものなどではなく、新しい、オリジナリティのあるデータベースでなくてはならない。』

次に、ホームページ等の公開資料を使って、各種の産業分野でのライフサイエンス関連データベースの活用例を探し、具体的な事例の収集を行った。化学、医療機器・分析機器、食品の3分野の国内大手企業の一覧から、特に「データベース」と関連が深いと思われる企業を11社選択した。売上高で比較すると、これらの会社は大手製薬会社と同程度かそれ以上の規模であることがわかった。次に上記11社のホームページ上の情報を「データベース」というキーワードによって全文検索し、ライフサイエンス関連データベースを利用あるいは開発した事例を探した。その結果、5件の事例がみつかった。例えば、富士フイルムでは医療用の画像処理サービスを実施しているが、正常および病態での画像データベースを使っている。また、テルモや島津製作所では医療機器や診断機器を使った検査・診断を行うためのデータベースが使われている。食品分野でも、機能性食品の開発等のためにデータベースが活用されている。このように、化学、医療機器、食品の分野でのライフサイエンスデータベースの利用や開発の事例が確認できた。

次に、分析機器や食品分野の状況を具体的かつ詳細に調査するため、企業を訪問してインタビューを実施した。ここでは、「企業における研究開発や製品開発の中で、どのようなデータベースを使っているか。また、将来どのようなデータベースがあれば研究や製品開発に有用であるか」等を調査することを目的とした。

味の素ライフサイエンス研究所の岩槻健氏、木村英一郎氏、戸塚一彦氏、四方菜穂子氏、尾上弘美氏、西尾陽介氏からは、食品の製品開発や基礎研究におけるライフサイエンス関連データベースの活用状況などのお話を伺った。要約すると、味の素では「特保」を含めて健康に貢献できるさまざまな食品の開発をめざしているが、研究テーマとしては味覚に関する研究や、消化管における遺伝子発現解析などを実施している。研究者が最もよく使うデータベースは遺伝子発現・配列データ・文献であり、主に NCBI の GenBank、OMIM、UniGene、PubMed、GEO などを用いている。が、これらのデータベースを十分に使いこなしているという実感はない。他にどのような有用なデータベースがあるのかよくわからない。そこで、ミシュランのように評価をつけたデータベースのカatalogを提供するサイトが欲しい。今後欲しいデータベースは、分子生物学者が利用しやすい臨床データベース、医者が使いやすいオミックス・データベース、病気のデータベース、薬のデータベースなどである。

タカラバイオ・ドラゴンジェノミクスセンターの北川正成氏、鈴木徹氏、水野政彦氏らからは、ゲノム解析を中心とするバイオテクノロジー分野の各種の分析受託、解析受託サービスの概要や、データベース等への期待について伺った。要約すると、ドラゴンジェノミクスでは、NCBI から RefSeq、EntrezGene などの配列データを定期的にダウンロードして使っているほか、UniProt のタンパク質配列データ、SNP データ、KEGG のパスウェイ情報、機能性 RNA 情報なども活用している。また、データベースと同様に重要なニーズとして、各種のデータ解析ソフトがある。民間企業で解析を行う際は、ソフトウェアのライセンスの扱いがポイントとなる。現状では海外のソフトを多く使っているが、日本製の解析ソフトの充実を望む。

以上の結果から、データベースは企業における研究開発現場で広く使われており、データベースに関するニーズは多様であることがわかる。また、膨大な知識と測定データの両方を整理して、わかりやすく提供するデータベースが求められていると言える。

5) 考察・今後の発展等

各種データベースの調査の過程で気づいた現行制度の問題点を指摘する。これらは将来データベースの統合化を実現する上でも考慮されるべき点であろう。

(A) データベース基礎調査について

NEDOのバイオテクノロジー・医療技術分野のプロジェクトの中には、基本計画の中にデータベース構築が位置づけられている場合と、プロジェクトの実施過程で必要となりデータベースが構築される場合があった。

研究内容の重要性にかかわらず、他の研究者や産業化にも有用な情報が生産される可能性がある。公的資金による研究データの公開と維持・管理する仕組み(事業)があれば、プロジェクトの成果の汎用化につながると考えられる。

(B) データベース評価法に関する考察

データベースの評価尺度調査の結果、データベースの被リンク数が論文の引用回数とやや高い正の相関を示したことは、興味深い。仮に、学术论文の評価と同様に、データベースをそのデータベースに関する論文の引用回数で評価することを是とするのであれば、被リンク数を用いて代用することも可となる。また、データベースのアクセス数もおそらくこれらの2つの尺度との間で正の相関を示すであろう(推測である)から、以上3種類の尺度は、いずれもデータベース評価のために用いることができると考えて良さそうである。

これとは別に、被リンク数はデータベースの重要性だけでなく公開からの経過年数にも依存するという説がある(大久保教授による平成17年度の本調査研究により、古くから維持されている「コドン使用頻度データベース」へのリンク数が調査対象の国産データベースの中で最も多い被リンク数を示したことが明らかにされている)。もしそうであれば、単純な被リンク数や引用回数ではなく、その数をデータベースの経過年数で割った数字によって評価を行うことが、より適切であるかもしれない。ちょうど、学術雑誌の評価にインパクトファクターを用いるのと同様である。このような単位時間あたりの被リンク数の是非については、継続して調査を行う必要があると考えている。

被リンク数が少ないデータベースについては、概して、データベースのポータルサイト(データベースのデータベースのようなもの)を作成すれば、状況を改善できる可能性があるだろう。

(C) リンク自動管理システムの発展

リンク自動管理システムが実現した「リンク維持の完全自動化」のアイデアは、「統合データベース」の維持管理に費やされるコストの大幅な削減につながる可能性を秘めている。今後の追加機能のアイデアとしては、ID対応表のダウンロード機能、ID更新ルール(追加、削除、融合、分裂)のルール策定、対応DBの追加などが考えられる。この技術開発をさらに進展させることにより、将来は自動管理された統合データベースが実現できるかもしれない。

(D) 将来の統合データベースのありかた

これまでの調査研究の成果をふまえて、将来の統合データベースのありかたについて、2つの可能性を提案する。

a) メタ検索エンジン型統合データベースの活用

今後、バイオ分野で増加が見込まれるデータベースを統合的に管理するためには、メタ検索エンジンを用いたバーチャルな統合化が有益な方法の一つと考えられる。メタ検索とは、入力されたキーワードを複数のデータベースに対して横断的に検索し結果を返す機能である。まず、メタ検索エンジン機能を持つ海外のサイト Bioinformatic Harvester <http://harvester.embl.de/> (図 5.3)がある。

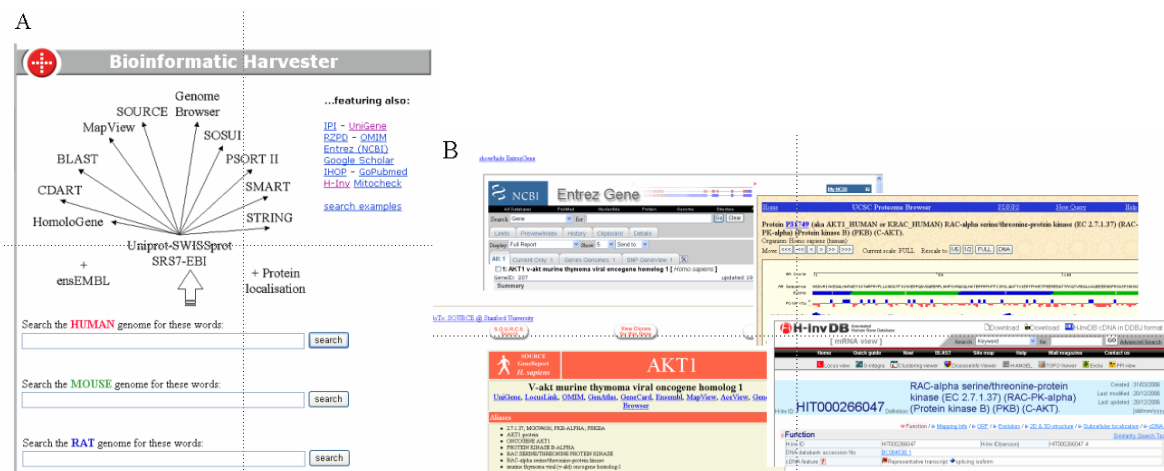


図 5.3 Bioinformatic Harvester

A: 検索キーワード入力画面。B: 検索結果。Human のデータベースで遺伝子 AKT を検索した結果を示す。Entrez Gene (NCBI)、SOURCE (Stanford Univ.)、UCSC Proteome Browser、H-InvDB の結果を示している。これらの結果が同一ウィンドウで閲覧できることが特徴である。

このようなメタ検索エンジンによる統合化は、実質的にポータルサイトの機能も兼ねていると考えられる。また、入力と閲覧されたページのログを残しておけばアクセス数をカウントすることも可能である。さらに、ポータルサイトがメタ検索エンジン機能を有していれば、ポータルサイト自体の利用者も多くなると予想されるので、各データベースの広報活動の一助となるばかりでなく、アクセス数や利用者数をカウントすることも可能となる。よって今後ポータルサイトを構築する場合、メタ検索機能は一つの有益な選択肢と考えられる。ただし、このような単純な構造のポータルサイトでは玉石混交な統合データベースになる危険性もある。以下に述べるアイデアとの併用をすれば、これに対応可能だと考えられる。

b) 登録型のデータベース・ポータルサイトの可能性

登録型のデータベース・ポータルサイトは、各研究プロジェクトで構築されたデータベースの情報をポータルサイトが規定した方法で登録し、一般公開するものである。ただし、この場合データベースそのものをポータルサイト内のサーバで管理するのではなく、各データベースの概要のみを管理し、そこから各データベースサイトへのリンクを貼るという形態をとる。もう一つの重要な機能として、公的資金研究によるデータベースの公開・運営・維持(各データベースを評価し、それに基づいて予算的な支援を行う)を実施する基盤としての機能が考えられる。

本サブテーマの課題による成果の継続性・発展性に関しては、「経済産業省統合データベースプロジェクト」がH20に開始される予定であり、これまでの調査研究で得られた知見を参考にしつつ、経済産業省関連のデータベース整備が進められてゆく見通しである。

6) 関連特許

該当なし

7) 研究成果の発表

(成果発表の概要)

1. 原著論文(査読付き)	該当なし
2. 上記論文以外による発表	該当なし
3. 口頭発表	招待講演:2回
4. 特許出願	該当なし
5. 受賞件数	該当なし

1. 原著論文(査読付き):該当なし。

2. 上記論文以外による発表:該当なし。

3. 口頭発表

招待講演

- 1) 今西規:「ヒト遺伝子と転写産物の統合データベース H-InvDB」、一橋記念講堂、「情報とシステム 2007、利用者のためのライフサイエンスデータベース-その現状と将来-」、2007 年 3 月 1 日
- 2) 今西規:「ヒト遺伝子の統合データベース開発とデータマイニング」、国立遺伝学研究所、研究会「生物情報資源の相互運用性」、2008 年 3 月 25 日

4. 特許出願:該当なし。

5. 受賞件数:該当なし。

IV. 実施期間終了後における取組みの継続性・発展性

科学社会における課題の整理とあるべき統合の形の描出は完成した。したがって実行に移すために、機構と管理(Institutional and management,)、予算と経済、法と政策、文化行動、と技術の5つの側面からそれぞれの分野の専門家グループと利用者となる研究者の代表およびデータ生産プロジェクトを遂行する代表者を加え、相互に十分な理解を図りながら、研究社会や市民社会に対し議論の内容を透明に公開しつつ検討を加え誰もが納得できる計画を策定するべきである。

V. 自己評価

1. 目標達成度

多角的検討によって必要な課題は抽出整理が出来たと評価する。しかしながら機構と管理面、予算と経済面、法と政策面や文化と行動面については一部著作権、知財専門の弁護士に相談する機会を設けたが一般に研究代表者の机上の勉強によるところが大きく、それぞれの専門家からの助言や提案を仰ぐ必要があったと反省される。一方で課題に忠実に上記の助言提言を十分に行うことの出来る人材が我が国に存在するかどうかは疑問である。

2. 情報発信

十分に行うことができた。報告書のみならず基礎調査データおよび運営会議記録の全てをweb siteの形で構造化記録し、情報システム機構統合データベースセンターより一般公開している。従って研究事業の透明性は極めて高く保たれたと考える。情報発信については我が国の科学政策を考える事業であるために国外向けの広報活動は全く行っていない。一方で国内向けには数回の公演やシンポジウム、新聞記事などを通じて十分に説明に勤め理解を呼びかけた。また調査期間中に個別に100名を超える様々な背景の研究者と面談し全ての機会に本課題の意義とそのときまでの調査内容について説明することで公知を図った。

3. 研究計画・実施体制

研究計画では技術的課題およびフィージビリティ・スタディに人的にも時間的にも不必要なウェイトを置いたことを反省すべきである。一方でそれ以外の機構と管理(Institutional and management), 予算と経済、法と政策、文化行動、の4側面に対する調査研究は事務局と遺伝研のみで行ったために、程度が不十分な面があり、また研究分担者との理解の程度の差を生む結果と共に、研究運営委員会において委員との現状認識等の差を生む結果となり、合意に達するに至らなかった。

実施体制も同様に課題のウェイトの見誤りによるもので反省すべき点が多い。

データベースを専門にする研究者に分担をお願いしたことはチーム全体としての専門領域を限定する結果となり反省すべきである。また、関係府省の担当官に運営委員会にご参加いただいたが、現状認識等のずれから問題意識を共有することから困難があり、運営委員会において十分な合意形成はできず、研究代表者側から見た場合に、期待した情報の提供や助言等の問題解析や計画立案に建設的に役立つ情報が得られることは少なかった。

4. 実施期間終了後における取組みの継続性・発展性

科学社会における課題の整理とあるべき統合の形の描出は研究運営委員会において十分な合意形成はできなかったが、研究者代表者としての今後の検討に資するものとして留意すべき事項を提示した。今後、機構と管理(Institutional and management), 予算と経済、法と政策、文化行動、と技術の5つの側

面から それぞれの分野の専門家グループと利用者となる研究者の代表およびデータ生産プロジェクトを遂行する代表者を加え、相互に十分な理解を図りながら、研究社会や市民社会に対し議論の内容を透明に公開しつつ検討を加え誰もが納得できる計画を策定するべきである。