

# 個別タスクの実施計画及び成果イメージ(案)

## 2. データベース統合化基盤技術開発

情報・システム研究機構 事務局

## 2. データベース統合化基盤技術開発

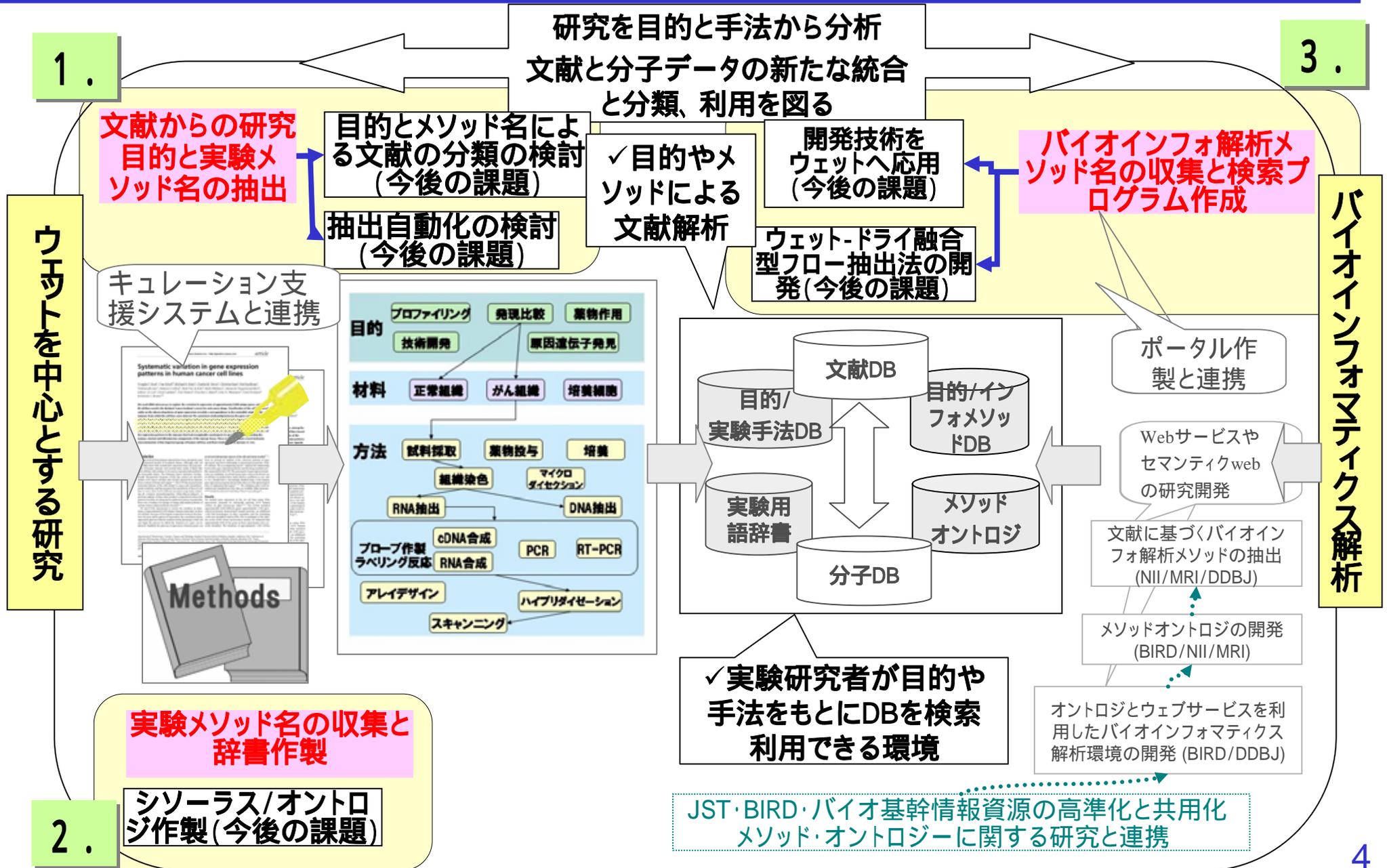
業務計画書項目	個別課題名	成果の概要
基盤知識表現 技術開発	メソッドオントロジとの連携システム開発	分子データをその生産動機から分類整理し利用に付するためのシステムおよび辞書
	遺伝子名揺らぎ吸収システムの開発	分子情報から医学文献までの遺伝子や蛋白質名称の正確な同定のための遺伝子固有名称と一般名称の管理システム、辞書
	バイオNLPリソースの整備	バイオNLP(自然言語処理)リソース(プログラム、データ)を公的サイトから収集し、リソースを管理しユーザーに提供するデータベースシステムを開発する
	解剖学用語表示システムの開発	分子レコードや文献臨床記録中の解剖学用語を検出し索引整理に利用するためのシステムおよび辞書
	バンクレコード俯瞰化のための索引技術開発	DDBJ、GEO、PDBj、dbSNPを一括整理検索するためのバンクレコードの総目録およびバンク内容の表現技術開発
癌研究知識表現技術開発	癌研究知識表現システム開発	癌研究分野での分子情報の統合整理サイト
多型知識表現技術開発	多型知識表現技術開発	プロモータ領域DB「dbQSNP」と確定ハプロタイプDB「D-HaploDB」の統合化に向けたXML化
キュレーター支援技術開発	キュレーション支援システムの開発	国内のキュレーションによるデータベース構築維持作業を援助するための論文解読情報抽出援助システム

# 「メソッドオントロジとの連携システム開発」の実施項目(計画)

---

1. 研究目的と実験メソッド名による文献の分類法とシステムの確立
  - ・ 文献から研究目的と実験メソッド名を抽出する手法の調査
    - 遺伝子発現解析の500文献をテストデータとして予備調査を行う
    - 辞書構築と平行し実験メソッド名の抽出自動化を検討する(今後の課題)
    - 目的に関する記述を収集し抽出自動化を検討する(今後の課題)
  - ・ 目的-メソッド名で与えられた文献の分類に関する検討
2. ウェットの実験メソッド名辞書構築
  - ・ 実験書やMeSH等から遺伝子実験を中心に実験メソッド名を収集する
  - ・ 実験メソッド名の分類、構造化(シソーラス/オントロジ作成)を検討する
3. バイオインフォマティクス解析メソッド名の収集とメソッド名検索システムの作製(バイオインフォマティクスメソッドオントロジーとの連携)
  - ・ 論文を中心にメソッド名収集と整理を行う
  - ・ メソッド名検索システムを作製し、収集したメソッド名の表記揺れ等の評価を行う
  - ・ 開発した手法やシステムをウェットのメソッド名整理に応用する(今後の課題)
  - ・ ウェット-ドライ融合型の解析フロー抽出法を検討する(今後の課題)

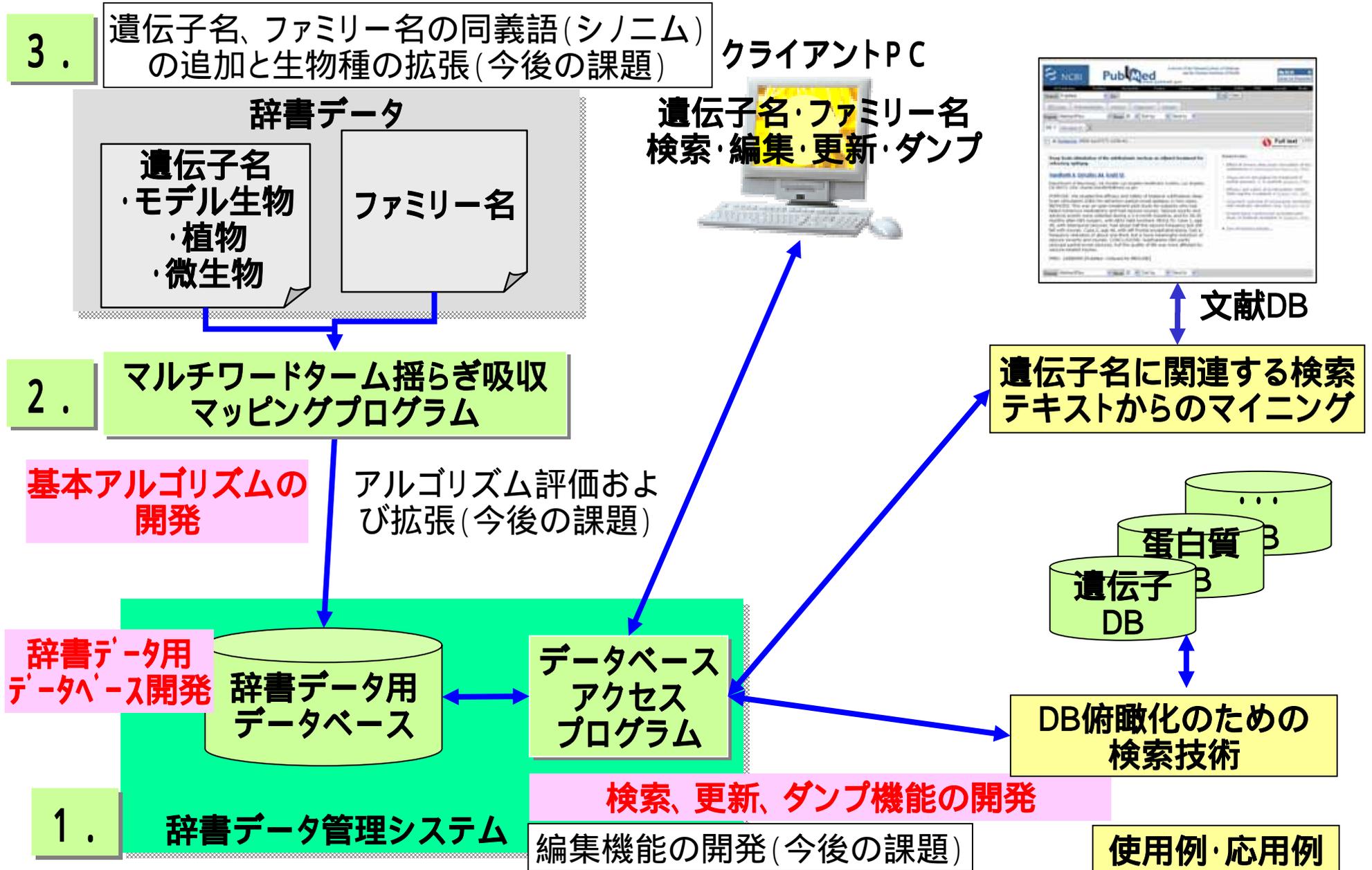
# 「メソッドオントロジとの連携システム開発」の成果イメージ



# 「遺伝子名揺らぎ吸収システムの開発」の実施項目(計画)

1. 辞書データ管理システムの構築
  - 1) 辞書データ用データベースの開発
  - 2) データベースアクセスプログラムの開発
    - ・検索、更新、ダンプ機能の開発
    - ・編集機能の開発(今後の課題)
  
2. マルチワードターム揺らぎ吸収マッピングプログラムの開発
  - 1) 基本アルゴリズムの開発
  - 2) アルゴリズムの評価および拡張(今後の課題)
  
3. 辞書データの拡張(今後の課題)
  1. 遺伝子名、ファミリー名の同義語(シノニム)の追加と生物種の拡張

# 「遺伝子名揺らぎ吸収システムの開発」の成果イメージ



# 「バイオNLPリソースの整備」の実施項目(計画)

---

## 1. バイオNLPリソースの収集・開発・整備

### 1) リソース(プログラム、データ)の収集および開発

- ・公的サイトからオープンソースで使用できるリソースの収集
- ・収集できないプログラムの新規開発およびデータの作成

### 2) リソースの拡張および修正(今後の課題)

## 2. リソース管理・提供システムの開発

### 1) 基本機能の開発

- ・バイオNLPリソース用データベース開発
- ・データベースへのアクセスプログラム(検索・更新・追加・削除・ダウンロード)

### 2) ユーザインタフェース(Webサイト)の評価および拡張(今後の課題)

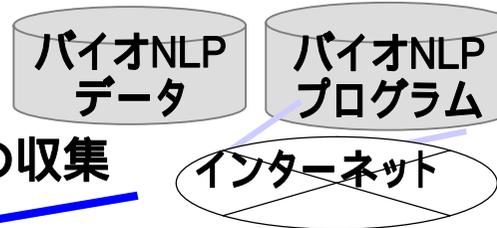
# 「バイオNLPリソースの整備」の成果イメージ

1.

## リソースの収集および開発

リソースの拡張および整備(今後の課題)

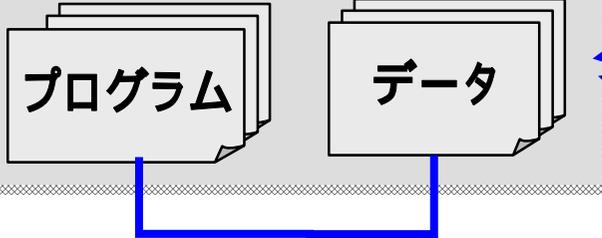
公共DB



文献DB

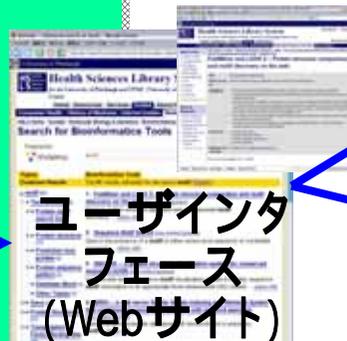
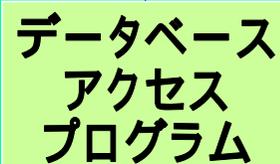
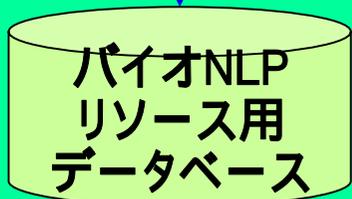
テキストからの  
情報検索・抽出

バイオNLPリソース

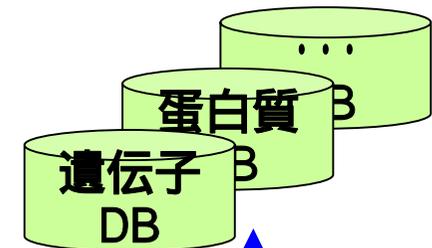
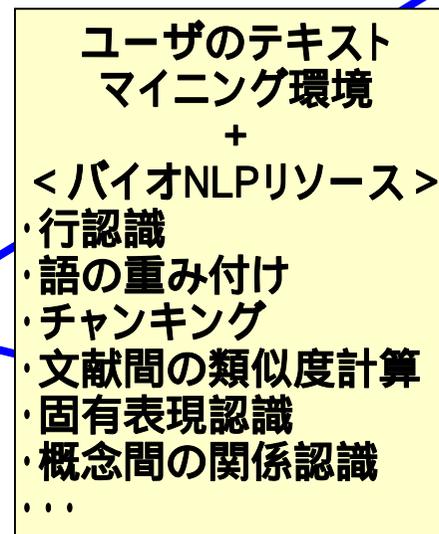


リソースの収集

リソースの新規開発



リソース管理・提供システム



各種DBからの  
テキスト情報の抽出

2.

## 基本機能の開発

ユーザインタフェースの評価および拡張(今後の課題)

使用例・応用例

# 「解剖学用語表示システムの開発」の実施項目(計画)

1. 解剖学ボクセル辞書作成環境の構築
  - 1) 画像データ(CTやMRIデータ、Visible humanのカラー写真)から臓器を抽出するためのアルゴリズム、ツールの調査、検討
    - 既存ツールの利用による半自動方式の採用
    - 自動化の検討(今後の課題)
  - 2) 3Dボクセルデータの編集機能の整備
    - 既存ツールの採用
2. 解剖学ボクセル辞書構築
  - 1) 主要臓器の位置、輪郭、サイズの再現
  - 2) 各臓器の詳細構造の再現(今後の課題)
3. 解剖学用語間の関係抽出による知識の構造化の検討(今後の課題)
  - ・ 応用例; 解剖学用語に関連する検索、テキストからのマイニングシステムの構築
4. 解剖学ボクセル辞書を用いた、解剖学用語の可視化の検討(今後の課題)
  - ・ 応用例; 発現データの人体モデルへのマッピング

# 「解剖学用語表示システムの開発」の成果イメージ

画像情報 (MRI、カラー画像、低解像度ボクセルモデル)

解剖用語  
(ヒトについてはLSDの2,500語が目標)

発現データの人体モデルへのマッピング



文献DB

既存ツールの利用

ボクセル辞書作成環境

自動化の検討  
(今後の課題)

可視化 (今後の課題)

解剖学用語に関連する検索  
テキストからのマイニング



用語間の関係抽出  
(今後の課題)

応用例

座標解剖辞書

用語	意味
眼	E3 G3
頭	E2 F2 G2 H2 D3 E3 F3 G3 H3 I3 E4 F4 G4 H4 E5 F5 G5 H5
右耳	D3
左耳	I3
口	E5 F5 G5
首	F6 G6
右腕	C7 O8 O9 O10 O11 D7
左腕	J7 J8 J9 J10 J11 I7
心臓	G8
肝臓	E10 F10 E11
右脚	E13 E14 E15 E16 D16
左脚	H13 H14 H15 H16 I16
体幹	E7 F7 G7 H7 E8 F8 G8 H8 E9 F9 G9 H9 E10 F10 G10 H10 E11 F11 G11 H11 E12 F12 G12 H12

ボクセル辞書 : 解剖学用語を臓器の3次元座標で定義

主要臓器の位置、輪郭、サイズの再現

各臓器の詳細構造の再現 (今後の課題)

# 「バンクレコード俯瞰化のための索引技術開発」の実施項目(計画)

---

## 1. 研究プロジェクト索引構築

### 1) データバンクレコードからの引用文献情報抽出、及びレコードのクラスタリングによる研究プロジェクトの同定

配列(DDBJ)、蛋白質構造(PDBj)、多型(dbSNP)、発現(GEO, CIBEX)の4種のデータバンクの各レコードから抽出した引用文献をもとに、各レコードのクラスタリングを実施する。得られたクラスターを研究プロジェクトと呼ぶ。

### 2) 研究プロジェクト索引辞書構築システムの開発

各研究プロジェクトに対して、引用された文献を評価し、索引として利用できるキーワードを付与する。試行評価結果のフィードバックと再構築、キーワードのマーキングによる結果表現を可能とするシステムを開発する。生物種区分を活用。

### 3) データバンクレコードの索引付け

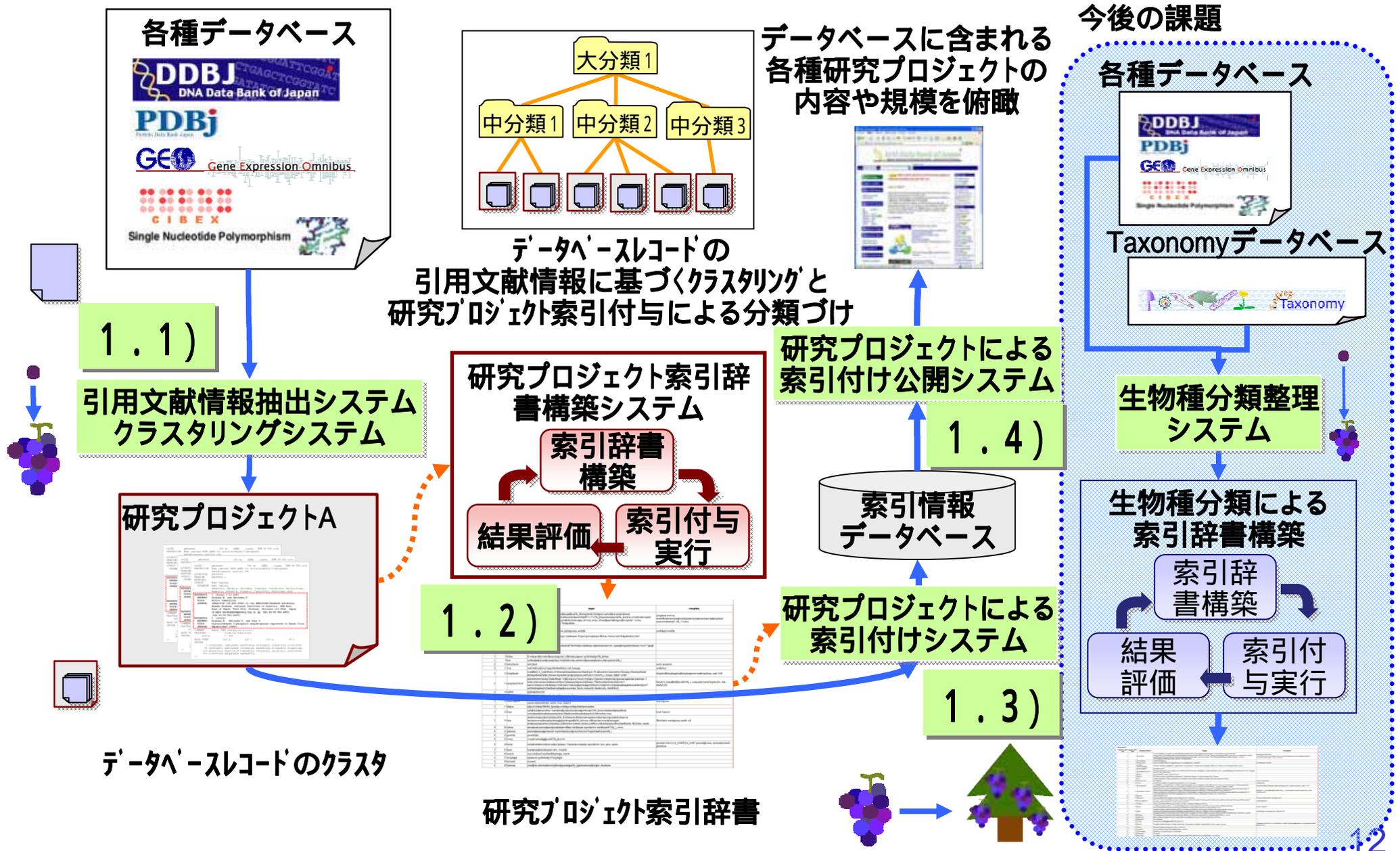
2) で得られた索引辞書を用いて研究プロジェクトに索引付けを行い、索引情報データベースを構築する。「データ目次」としての利用を想定

### 4) 公開システムの構築

## 2. 生物種分類による索引構築(今後の課題)

### 1) Taxonomyデータベース 問題解析に着手

# 「バンクレコード俯瞰化のための索引技術開発」の成果イメージ



# 「癌研究知識表現技術開発」の実施項目(計画)

---

## 1. 実験的データベースの開発

### 1) 癌遺伝子発現臨床情報データベース(Cancer Gene Expression Database, CGED)

臨床情報から遺伝子を選択する機能を付加

神経膠腫など新規データのアップロード(今後の課題)

### 2) 肺癌分子情報臨床情報統合データベース

遺伝子発現以外のゲノム情報を付加した新しいデータベースプロトタイプ(今後の課題)

## 2. 臨床情報の収集とニーズ調査

### 1) 成人病センター内の各種固形癌の臨床情報収集

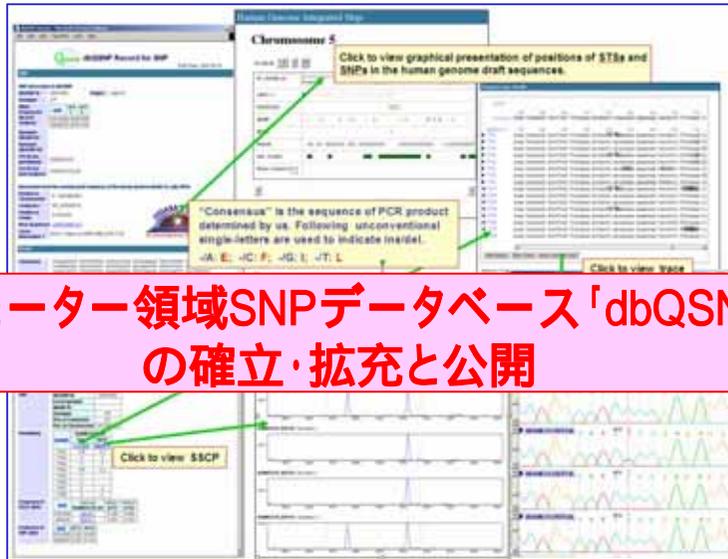
### 2) 臨床医のニーズにあった知識表現技術のデザイン(今後の課題)



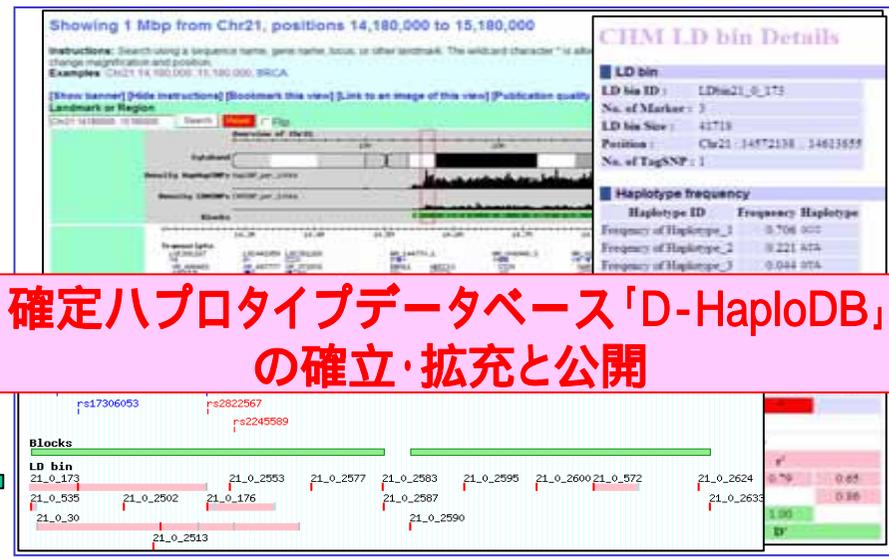
# 「多型知識表現技術開発」の実施項目(計画)

1. プロモーター領域SNPデータベース「dbQSNP」の確立・拡充と公開
  - 1) 波形トレースデータ(SSCP解析やシーケンシング解析)からSNP情報を確定するためのアルゴリズム、ツールの調査、検討
    - 既開発ツールの利用によるデータの拡充
    - 情報処理プロセスの簡易化、自動化の推進
  - 2) 公開データベースのXML化
    - 種々の多型関連XML表現の比較検討
    - 「dbQSNP」のPML化
2. 確定ハプロタイプデータベース「D-HaploDB」の確立・拡充と公開
  - 1) 胞状奇胎解析DNAアレイデータからのSNP判定アルゴリズムの検討
  - 2) 拡大DNAアレイデータによる「D-HaploDB」の拡充
  - 3) 「D-HaploDB」のPML化
3. 遺伝性疾患関連外部データベースのデータポータビリティの検討(今後の課題)
  - ・ 応用例; 遺伝性疾患関連外部データベース検索、テキストからの情報抽出システムの構築
4. 疾患とSNP/ハプロタイプの関連付けと統合データベース化(今後の課題)
  - ・ 応用例; SNP/ハプロタイプによる疾患感受性判定

# 「多型知識表現技術開発」の成果イメージ



プロモーター領域SNPデータベース「dbQSNP」  
の確立・拡充と公開

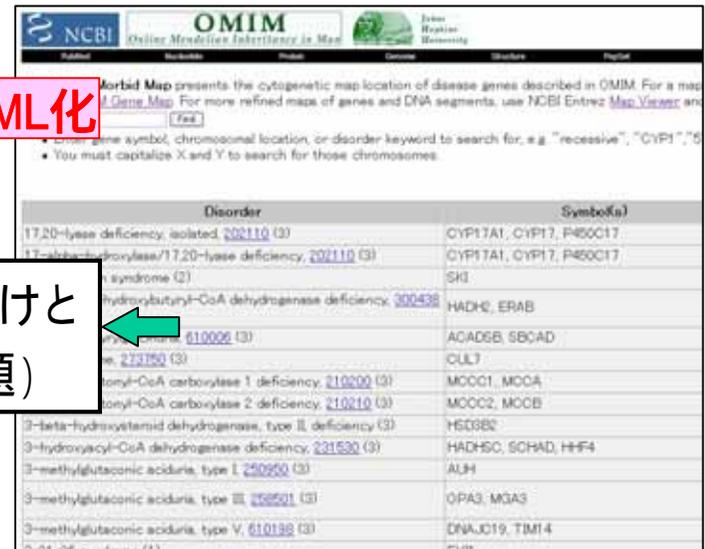
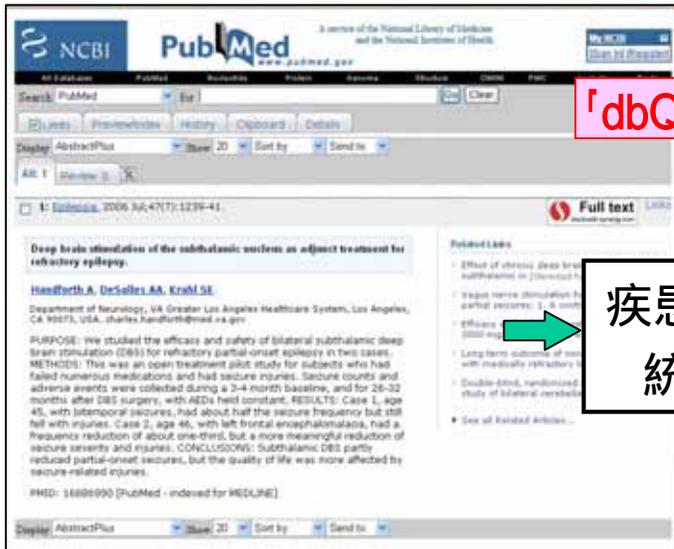


確定ハプロタイプデータベース「D-HaploDB」  
の確立・拡充と公開

種々の多型関連XML表現の  
比較検討

「dbQSNP」および「D-HaploDB」のPML化

疾患とSNP/ハプロタイプの関連付けと  
統合データベース化(今後の課題)



# 「キュレーション支援システムの開発」の実施項目(計画)

1. 論文構造解析・編集クライアントの開発
  - 1) 論文操作基本機能の開発
    - i) 学術論文データ(HTML)の構成(セクション、センテンス、フレーズ)を認識・編集操作する機能を開発
    - ii) 学術論文データへの辞書のマッピング、要素情報マップ表示・ナビゲーション、構成情報マップ作成の各機能を開発
    - iii) 学術論文データ中の引用論文を引用センテンスと共に認識し、対象センテンスの編集、PubMed対応情報の取得、引用論文の情報取得を行う機能を開発
  - 2) 論文構成・アノテーションモデルの適用 (今後の課題)
    - 1) 論文構成要素の編集・分類・表示を行う (今後の課題)
    - 2) 論文構造基本モデルの定義による論文構造の俯瞰 (今後の課題)
2. 論文構造解析・編集情報集約システムの開発 (今後の課題)
  - 1) 論文構造アノテーションデータの収集・編集システム
  - 2) 辞書、マップ、モデル等の共有システム
3. 論文構造解析・編集情報利用システムの検討 (今後の課題)
  - ・ 応用例; 遺伝子発現データベース、測定対象、測定手法に関連する論文情報データベースの構築

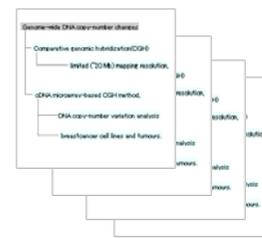
# 「キュレーション支援システムの開発」の成果イメージ

学術論文(全文: HTML形式)



・アノテーション対象抽出・分類項目  
・専門用語辞書  
・論文構成モデル

発現データベース関連論文のアノテーション他



i) 構成情報作成編集

新規開発

論文構造に基づく  
内容マップ

論文(全文)アノ  
テーション提供

iii) 引用論文取得表示

アノテーション実施(今後の課題)

ii) 辞書のマッピング

ii) 構成マップ編集

ii) ナビゲーション

アノテーションの  
集約と編集

構造化  
文献DB  
アノテ  
ションDB

1. 論文構造解析・編集クライアントの開発

1) 論文操作基本機能の開発

2. 論文構造解析・編集情報集約システムの開発(今後の課題)

3. 論文構造解析・編集情報利用システムの検討(今後の課題)

2) 論文構成・アノテーションモデルの適用(今後の課題)