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
	●DBポータル基本部分の連携システム開発	DB内容や規模を俯瞰可能な用語辞書に基づいたDB分類と、JST構築のDBポータル（WING）との間の連携システムを構築
●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	ﾌﾟﾛﾓｰﾀ領域DB「ｄｂQSNP」と確定ﾊﾌﾟﾛﾀｲﾌﾟ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ーとの連携システムの開発」

- **目的**
  - データを産生し、データを解析するメソッド名についてウェット系、ドライ系の両方から整理を進め、データ統合技術内での意味づけ関連づけを正確にするとともに、利用者にわかりやすいメソッド検索システムや辞書の提供を図る。
- **実施事項**
  1. ウェット及びドライ両方の実験手法名を収集し辞書構築を行った。
  2. バイオインフォマティクス解析メソッドオントロジーを利用した検索プログラムを作成した。
  3. ウェット系の実験手法を整理しデータや論文との連携を図るためのオントロジーについて検討した。
- **成果**
  1. ウェット系実験手法名の収集と辞書及びオントロジー作成。
     - ウェット系の実験手法名を英文、和文双方の書籍、文献から幅広く語彙を収集し辞書の素材構築を行った。
  2. ドライ系メソッドオントロジーを利用した検索プログラムを作成しWebリソースポータルに実装した。
     - 様々な名称がつけられているバイオインフォのメソッドを利用者の観点から分類し検索可能にした。

ウェット系の実験手法名辞書の構築

見出し	同義/類義語	英訳	(英) 同義/類義語	解説
トリチウムチミジン取込み実験	[3H]チミジン取込み実験	[3H]thymidine incorporation experiment		DNAの合成速度や合成量を調べたり、、3Hで放射標識したDNAを得る目的で、生物を[3H]チミジンで標識する実験。
エンハンサートラップ法	エンハンサートラップ, エンハンサートラップ実験	enhancer trap	enhancer-trap experiment	DNA組換え技術を利用し、活性の指標となる遺伝子とプロモーターの組み合わせがエンハンサーの近傍に位置したときに遺伝子活性が上昇することを利用し、エンハンサー領域を同定、単離する実験法。
紫外線照射	UV照射, UV照射実験, 紫外線照射実験	ultraviolet irradiation	Ultraviolet radiation	主として紫外線による損傷や影響を調べる目的で、物体や生物を紫外線にあてること。
トレーサー	トレーサー実験	tracer, orbital	tracer experiment	一般的には、ある事象の変化を追跡観測するための鍵となるもの。生物実験では、分子中の特定のアミノ酸等を放射線や蛍光で標識した特定の部位。
同位体トレーサー法	トレーサー実験	isotope tracer technique	tracer experiment	同位体を利用した実験技術で、特定の部位を同位体で標識した化合物を生物に投与し、同位体の存在を目印に、特定物質の代謝や合成経路、存在部位を調べる技術。

Webリソースポータルサイトの検索システム

「メソッドオントロジーとの連携システムの開発」実施項目（成果）

1. ウェットの実験手法名の辞書構築
   1) ウェット系の実験手法名収集について→英文、和文の実験書の索引から実験手法の語彙を収集した。 →Jabion用語辞書及び検索システムの利用
   2) ウェット系メソッドオントロジー構築について→収集した用語を分類する基本的なオントロジーを構築した。 →JST/BIRD（菅原）にて開発中のシステムと連携

2. バイオインフォマティクス解析メソッド索引辞書の構築
   1) ドライ系メソッド名について教科書・論文から約350名称を収集し辞書を作成した。
      →日本語バイオポータルプロジェクトにおけるタスクオントロジーの成果利用
   2) ドライ系メソッドオントロジーを利用した検索プログラムを作成しWebリソースポータルに実装した。

3. 構築した辞書オントロジーを利用して、論文から研究の目的と実験手法を抽出し、ライフサイエンスの研究モデルについて検討する。（今後の課題）

「遺伝子名揺らぎ吸収システムの開発」

- **目的**
  - 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（「遺伝子名」）と一般名称（「ファミリー名」）の辞書データ 構築を行う。
- 実施事項
  - 1．9種類の生物の遺伝子名およびファミリー名の基本辞書データの構築を行った。
  - 2．辞書データを管理するためのWebデータベースアプリケーションを実装した。
- **成果**
  - 遺伝子名辞書の作成
    - さまざまなデータベースで利用されている名称の収集と専門的キュレータによる編集によって、遺伝子が持つ多様な名称の関係を明示した辞書を実現した。
    - 遺伝子名の医学文献中で同定において、表記の揺らぎ（マルチタームワード）を吸収する名称の派生を実現した。
  - 辞書検索ユーザインタフェースの作成
    - 構築した大量の辞書データ（名称数約83万）をWebブラウザから簡便に検索可能なインタフェースを実現した。

遺伝子名辞書

生物種等	遺伝子数	名称数
遺伝子ファミリ	12,110	27,923
ヒト	38,728	173,630
マウス	60,688	172,260
ラット	38,164	123,726
ゼブラフィッシュ	38,879	83,694
ショウジョウバエ	30,410	95,578
線虫	25,316	97,031
出芽酵母	6,190	33,030
分裂酵母	4,895	9,790
枯草菌	4,106	18,920
合計	259,486	835,582

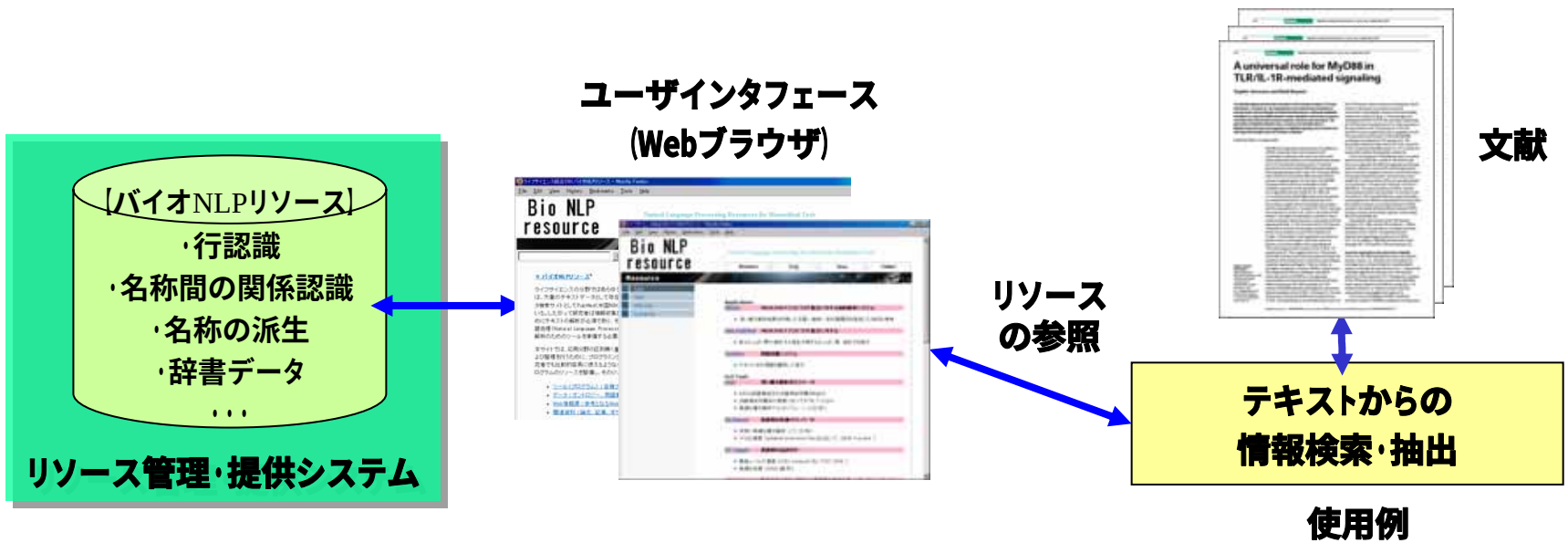
辞書検索ユーザインタフェース

「遺伝子名揺らぎ吸収システムの開発」実施項目（成果）

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→辞書データの収集とデータベース構築**
   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→Webブラウザで利用可能なユーザインタフェース作成**
      - **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→アルゴリズムを実装して、遺伝子名に適用**
   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）**
   - **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**

「バイオNLPリソースの整備」

- **目的**
  - 公開されているNLP(自然言語処理技術) 系のツール群を網羅的に収集してリソースとして整備･提供を行う。
- **実施事項**
  - 1. リソースの調査を行い、基本的なリソースの収集を行った。
  - 2. リソースを管理･提供するためのプログラムの開発を実施した。
- **成果**
  - バイオ系テキストの処理リソースの作成
    - 自然言語処理の知見･技術を有しなくても比較的容易に使える基本的なリソース群を実現した。
  - リソースへのアクセス用インタフェース作成
    - Webブラウザを利用して比較的簡単にリソースへのアクセスを実現した。



「バイオNLPリソースの整備」実施項目（成果）

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→基本的なリソースの構築**
      - **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→リソースの管理プログラムとWebブラウザを利用したユーザインタフェースの構築**
      - **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）**

「解剖学用語表示システムの開発」

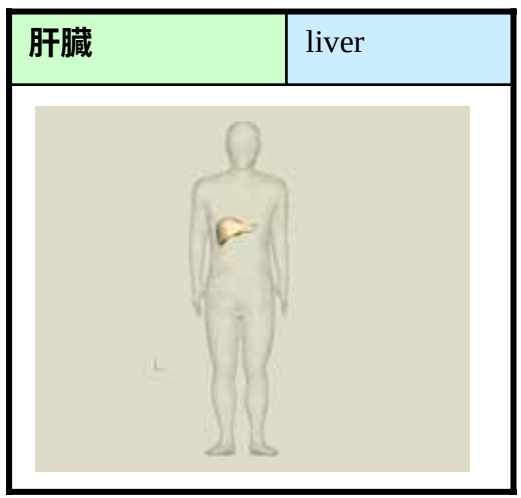
- **目的**
  - 解剖学用語という基盤的な概念(用語)の整理（解剖学用語辞書とオントロジーの構築）を行う。
- **実施事項**

  1．３Ｄ編集ソフト（FreeForm Modelling)を用いて、低解像度人体３ＤデータTAROの詳細化を行った。

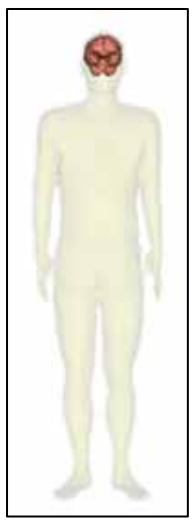
  2．解剖学３Ｄポリゴンマン辞書へのユーザデータマッピング機能を実装した。
- **成果**

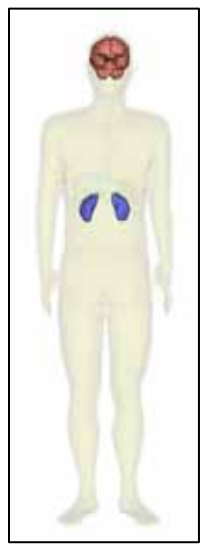
  1. 解剖学３Ｄポリゴンマン辞書の作成
     - 解剖学用語を３次元座標で定義することにより、概念を曖昧性なく表現。
  2. ３Ｄアナトモグラフィー API の作成
     - ユーザデータ（例：器官別の発現解析データ）の解剖学３Ｄポリゴンマン辞書へのマッピングによって、解剖学の観点からのデータ俯瞰を実現。
     - 由来の異なるデータの同時マッピングによって、それらのデータの統合的な比較観察を実現。

解剖学３Ｄポリゴンマン辞書



３Ｄアナトモグラフィー



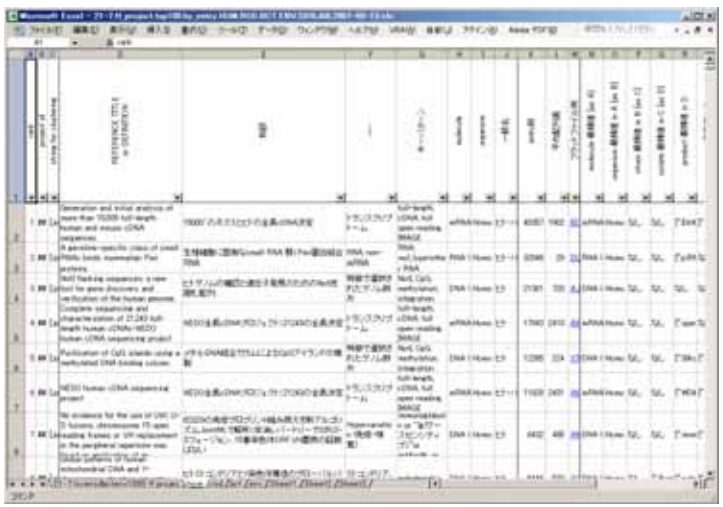


「解剖学用語表示システムの開発」実施項目（成果）

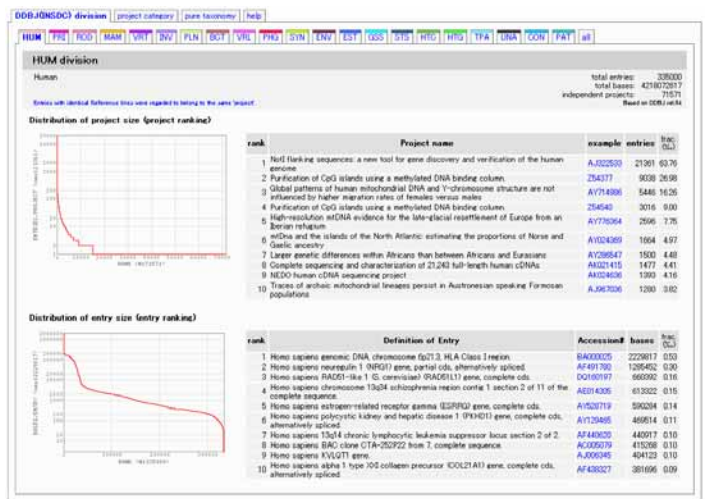
1. 解剖学ボクセル辞書作成環境の構築
   1) 画像データ（CTやMRIデータ、Visible humanのカラー写真）から臓器を抽出するためのアルゴリズム、ツールの調査、検討
      → 既存ツールの利用による半自動方式の採用→低解像度ボクセルデータTAROの利用
      → 自動化の検討（今後の課題）
   2) 3Dボクセルデータの編集機能の整備
      → 既存ツールの採用 →FreeFormModellingによる3D辞書構築作業の確立
2. 解剖学ボクセル辞書構築
   1) 主要臓器の位置、輪郭、サイズの再現→解剖学3Dポリゴンマン辞書の作成
   2) 各臓器の詳細構造の再現（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 →3Dアナトモグラフィー API開発

「バンクレコード俯瞰化のための索引技術開発」

- **目的**
  - バンクの内容につき広く我が国の研究者に知っていただき、我が国のより多くの研究者にバンクからデータを引き出し、自在に利用していただくこと、および、データバンクを通じた分子レベルの生物学研究の目次的俯瞰の提供
- **実施事項**
  1. データバンク目次用索引辞書構築システムの開発
  2. 索引辞書を適用し、DDBJとGEOを対象にしたデータバンク目次データベースおよび公開システムを構築
- **成果**
  1. 索引辞書構築システム
     - データバンクレコードの目的別分類を可能とする索引辞書を構築
  2. データバンク目次公開システム
     - INSDCとGEOについて総合データ目次を提供
     - 目次項目毎のデータのダウンロードが可能



索引情報データベース



公開システム

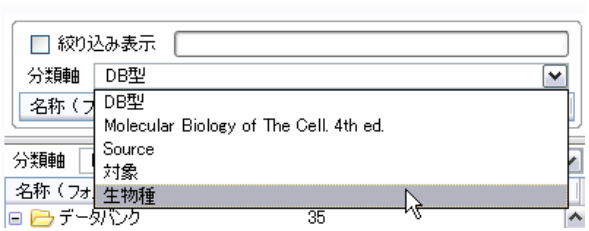
「バンクレコード俯瞰化のための索引技術開発」実施項目（成果）

1. **研究プロジェクト索引構築**
   1) **引用文献情報抽出、レコードクラスタリング**
      →配列(DDBJ)、蛋白質構造(PDBj)、発現(GEO, CIBEX)のレコードクラスタリング
      多型(dbSNP)のデータバンクのレコードクラスタリング(今後の課題)
   2) **研究プロジェクト索引辞書構築　→索引辞書構築システムを開発**
      索引辞書構築システムを開発
      試行評価結果のフィードバックと再構築、KWICによる結果表現、生物種区分活用
   3) **データバンクレコード索引付け　→索引情報データベースを構築**
      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
      「データ目次」としての利用を想定
   4) **公開システム →DDBJとGEOを対象にした索引情報データベース公開システムを構築**
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 →生物種系統の階層情報を解析**

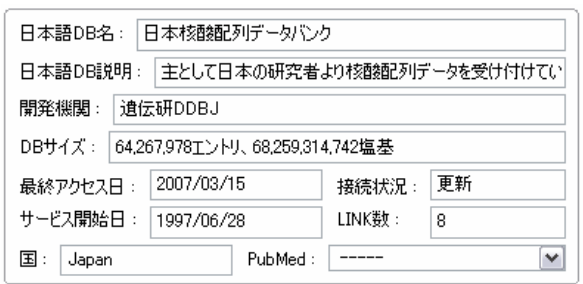
「DBポータル基本部分の連携システム開発」

- **目的**
  - DBポータル連携のためのデータベース及び利用システムを開発する。
- **実施事項**
  1. ＤＢポータル連携データベースの開発を行った。
  2. ＤＢポータル連携データ利用システムの開発を行った。
- **成果**
  1. ＤＢポータル連携データベースの作成
     - ＤＢポータル連携データを提供するデータベースを作成した。
  2. ＤＢポータル連携データ利用システムの作成
     - ユーザデータ（例：器官別の発現解析データ）の解剖学３Ｄポリゴンマン辞書へのマッピングによって、解剖学の観点からのデータ俯瞰を実現。
     - 由来の異なるデータの同時マッピングによって、それらのデータの統合的な比較観察を実現。

連携システム分類軸



連携システム提供情報



ＤＢポータル連携データ利用システム

「DBポータル基本部分の連携システム開発」実施項目（成果）

- データベース連携システム
  - 内閣府調査の成果とJST-WINGを連携する→対象データによる連携データベースの作成を実施
    - （連携対象）　→　成果一覧からWINGproへリンク
    - ・内閣府調査：網羅的に収集したライフサイエンスデータベースとその多次元的な分類
    - ・JST-WING：データベース解説と分類情報
    - 両データベースに対するアクセスインターフェースを開発
    - データベースの情報を取得し、融合して利用可能とする　→　WINGproとして構築
  - 両データベースの更新への対応（今後の課題）

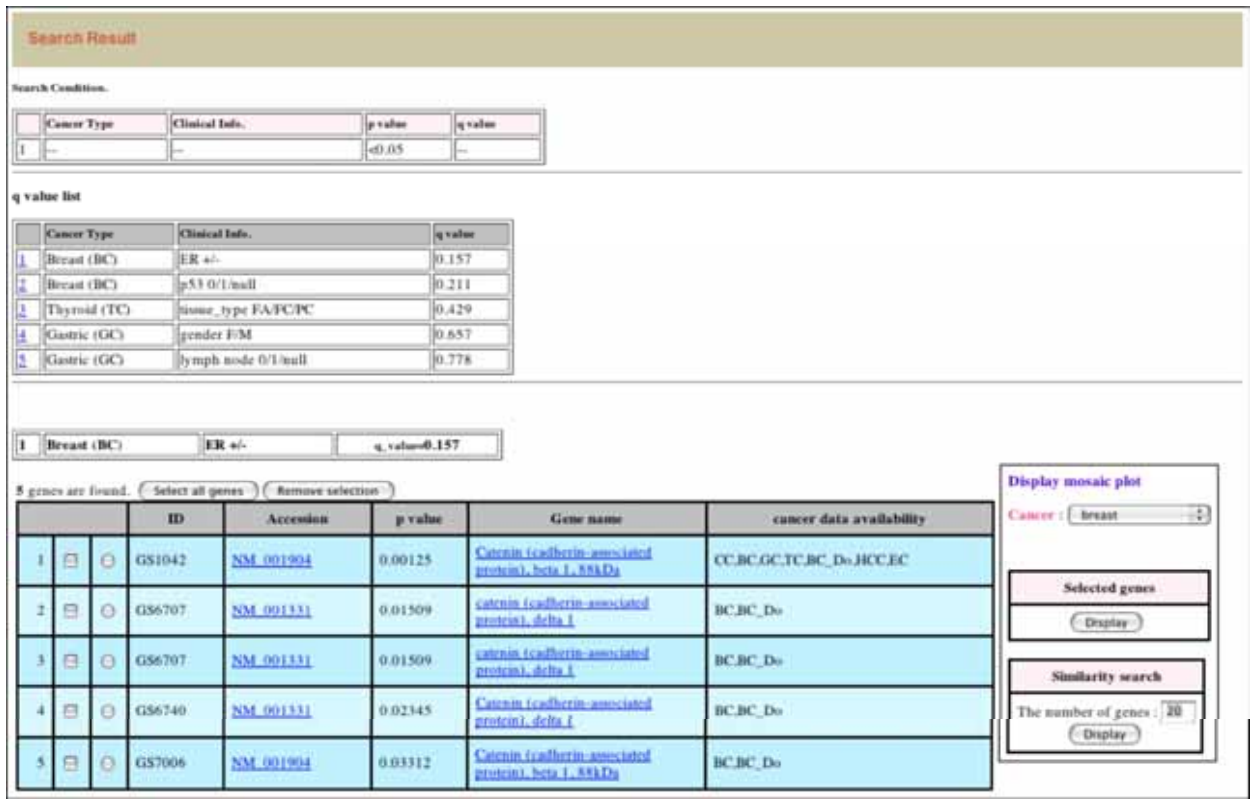
- データベース連携ナビゲーションシステム　→融合データに対する分類軸の整備
  - 連携した両データベースの分類軸を利用したナビゲーションシステムを開発
  - →　研究者の興味や研究内容に対して有用な分類軸の提供→　ディレクトリで一覧表示
  - →　分類軸の組み合わせとキーワードによる対象データベースの絞込み機能を提供

- データベース情報表示インターフェース　→融合データ検索・表示インターフェースを作成
  - ナビゲーションシステムにより選択された一連のデータベースの情報を表示
  - →　日本語解説、オリジナルデータベースへのアクセスリンク、維持管理状況、
  - データサイズ、関連データベース間の比較、データベースの利用状況等
  - →　日本語解説、アクセスリンク、維持管理状況　提供済み
- 連携ポータルデータベースの公開と更新
  - ブックマーク自動更新機能（今後の課題）

「癌研究知識表現技術開発」

- **目的**
  - 実験的なデータベース作成等を通じて癌の分子データと臨床情報の統合、表現を想定ユーザーにわかりやすい形で実現する。
- **実施事項**
  1. 癌遺伝子発現臨床情報データベース の機能拡張。
  2. 臨床情報の整理（成人病センター乳腺内分泌外科症例）。
- **成果**
  - CGED (Cancer Gene Expression Database)の機能拡張。
  - 従来の機能に加えて臨床情報から遺伝子を検索する機能を追加。

追加機能画面→
(例)
転移のある癌とない癌で発現の異なる遺伝子を検索する



「癌研究知識表現技術開発」実施項目（成果）

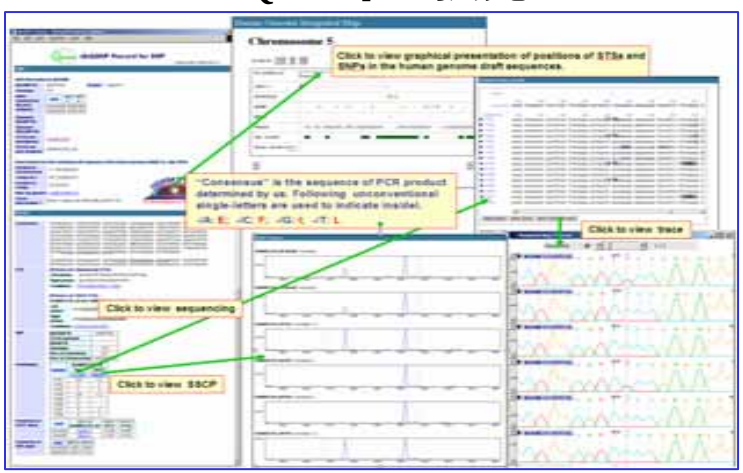
1. **実験的データベースの開発**
   1) **癌遺伝子発現臨床情報データベース**(Cancer Gene Expression Database, CGED)
      → **臨床情報から遺伝子を選択する機能を付加　実装済み**
      → **神経膠腫など新規データのアップロード　乳癌（抗癌剤耐性研究）、胃癌、甲状腺癌はアップロード済み、神経膠腫などは次年度以降の予定**
   2) **肺癌分子情報臨床情報統合データベース**
      → **遺伝子発現以外のゲノム情報を付加した新しいデータベース　今後の課題**

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　乳癌は終了、他は今後の課題**
   2) **臨床医のニーズにあった知識表現技術のデザイン　デザインは終了、次年度以降データベースとしてとして構築**

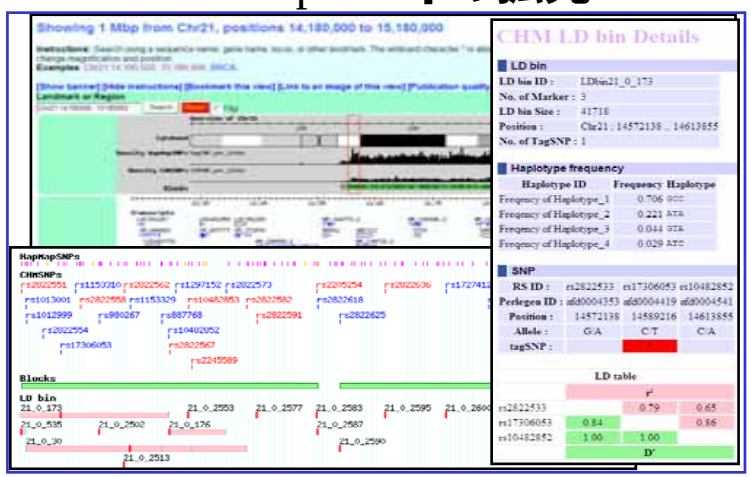
「多型知識表現技術開発」

- 目的
  - 日本人ゲノム多型情報を高度化し、医療情報との統合のためのデータポータビリティーを図る。
- 実施事項
  1. 疾患の主要な遺伝的要因である遺伝子発現調節領域多型の公開データベース「dbQSNP」を拡充する。
  2. 要因遺伝子探索に必須な全ゲノム確定ハプロタイプ構造の公開データベース「D-HaploDB」を拡充する。
  3. 他のデータベースとの多型データ相互利用促進のため上記2データベースの標準言語（XML）化を行う。
- 成果
  1. 「dbQSNP」の拡充
     - 種々の疾患要因候補遺伝子領域にあるSNP情報を収集し、同データベースに記述されたSNPを 1.0×10^4 個超とした。
  2. 「D-HaploDB」の拡充
     - 既存の280k個のSNP に加えて500K個のSNPによる確定ハプロタイプ情報を収集・追加し、同データベースを高度に精細化した。
  3. 「dbQSNP」及び「D-HaploDB」のXML化
     - 上記2データベースのPML（XMLを多型情報記述に特化させた標準言語）版を構築した。

「dbQSNP」の拡充



「D-HaploDB」の拡充



両DBのPML化

「多型知識表現技術開発」実施項目（成果）

1. 遺伝子発現制御ゲノム領域多型データベース「dbQSNP」の拡充

   自己免疫疾患、がんへの関与が疑われる約100個の遺伝子のゲノム領域にあるSNP配列及びその正常日本人でのアレル頻度を直接配列決定及び定量SSCP解析により決定し、上記データベースを拡充した。現在約1.0 X 10^4個のSNP情報が記載されている。

   → 「dbQSNP」での検索対象領域におけるSNP密度は、最大とされる現在のHapMap計画多型データベースでの密度の約2倍

   → 更なる拡充の検討（今後の課題）

2. 日本人確定ハプロタイプデータベース「D-HaploDB」の拡充

   新たに100個の胞状奇胎についてAffymetrix社アレイチップを用いた5 x 10^5個（既存データの約2倍）のSNPタイピングを行い、ゲノムワイド連鎖不平衡地図を飛躍的に高精度化した。

   → 疾患遺伝子関連解析での「D-HaploDB」の有効性が飛躍的に向上

   → 更なる拡充の検討（今後の課題）

3. 「dbQSNP」及び「D-HaploDB」の標準化（データポータビリティの強化）

   データベース記述標準言語XMLのゲノム多型記述のための機能拡張版であるPMLを採用し、上記二個のデータベースのPML版を構築した。

   → 他のデータベースとの多型データ相互利用促進手段の確立

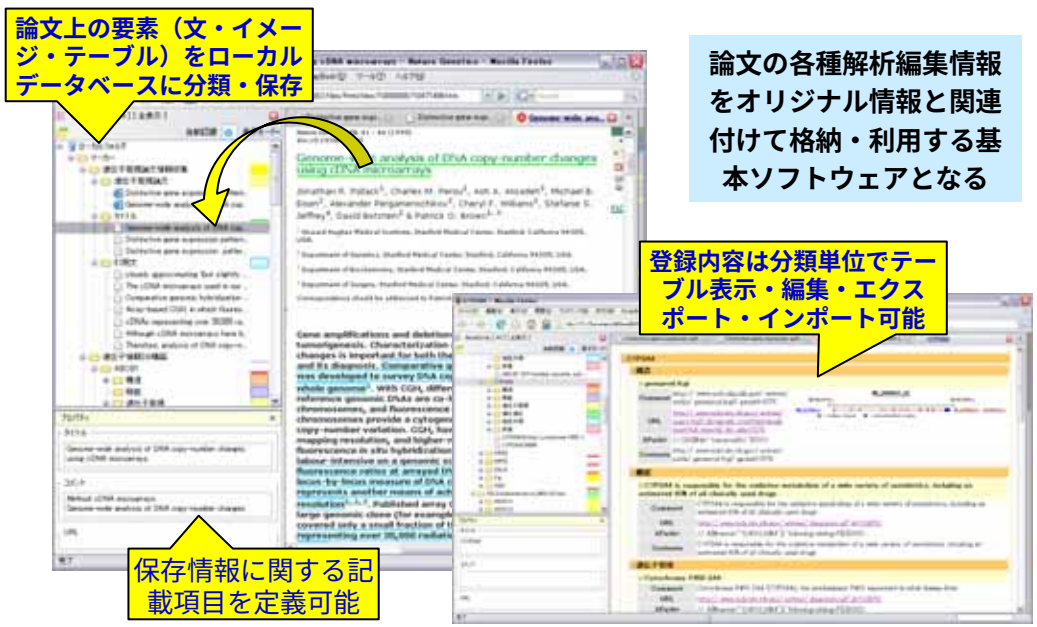
   → 他の多型表記標準言語GVML（現在開発途上）による記述（今後の課題）

4. 国内に散在する日本人ゲノム多型情報データベースとのデータ相互利用基盤の確立と医療情報との統合化（今後の課題）

「キュレーション支援システムの開発」

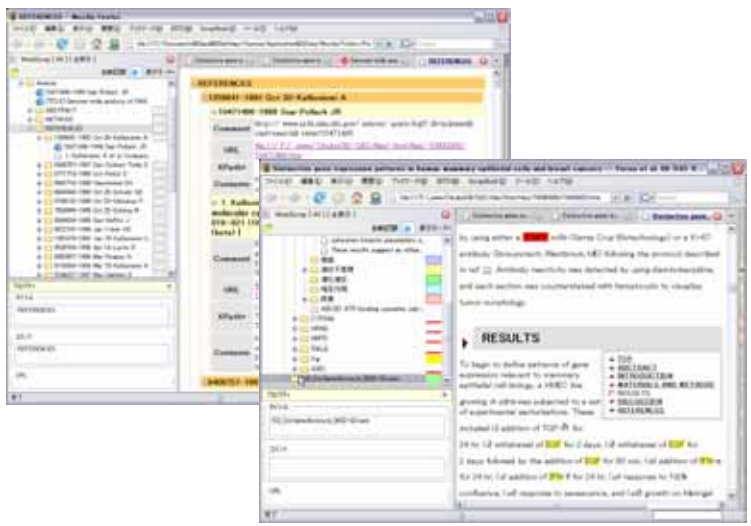
- **目的**
  - 論文を対象としたデータベース構築作業を支援するソフトウェアの開発を行う。
- **実施事項**
  1. 論文情報解析･編集ソフトウェアのベースシステムの開発を行った。
  2. 論文情報解析･編集用各種解析モジュールの開発を行った。
- **成果**
  1. 論文要素情報スクラップソフトウェアの作成
     - 電子化ドキュメント上のオブジェクトをクライアントサイドで分類･蓄積するシステムを完成。
  2. 論文情報解析･編集用各種モジュールを作成
     - 辞書マッピング、ナビゲーション、論文構成認識、引用論文情報収集を行う基本モジュールを作成。
     - 各種モジュールは、WiredScrapの追加モジュールとして開発。

論文要素スクラップソフトウェア



論文情報解析・編集用モジュール

論文構造解析・引用論文情報収集モジュール



辞書マッピングモジュール

「キュレーション支援システムの開発」実施項目（成果）

- 論文構造解析・編集クライアントの開発
  - 1) 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 →論文構成認識モジュール作成
  - 2) 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 →辞書マッピング・ナビゲーションモジュール等作成
  - 3) 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→引用論文モジュール作成
  - 4) 論文構成要素の編集・分類・表示を行う（今後の課題）
  - 5) 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）

  論文情報解析・編集クライアントのベースシステムを完成
  → 「WiredScrap」実用レベルの論文（電子化ドキュメント）要素スクラップシステム

- 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

- 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