資料 3－3

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
●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（18年度）
     - 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（19年度）
     - 目的に関する記述を収集し抽出自動化を検討する（19年度）
   - 目的-メソッド名で与えられた文献の分類に関する検討（18年度、19年度）
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（18年度）
   - 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（18年度、19年度）
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（18年度）
   - 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（18年度）
   - 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（19年度）
   - ウェット-ドライ融合型の解析フロー抽出法を検討する（19年度）

「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

研究を目的と手法から分析
文献と分子データの新たな統合
と分類、利用を図る

1.

ウェットを中心とする研究

文献からの研究目的と実験メソッド名の抽出(18年度)

目的とメソッド名による文献の分類の検討(19年度)

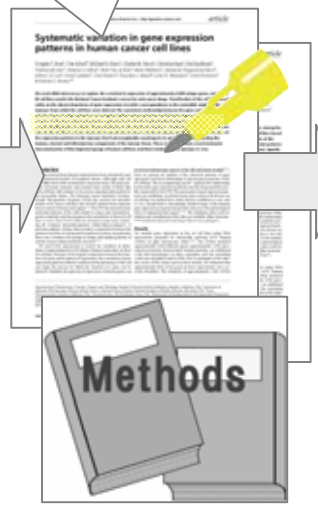
抽出自動化の検討(19年度)

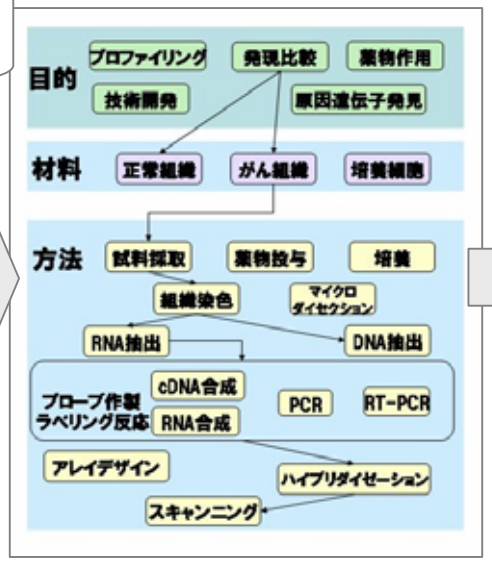
✓目的やメソッドによる文献解析

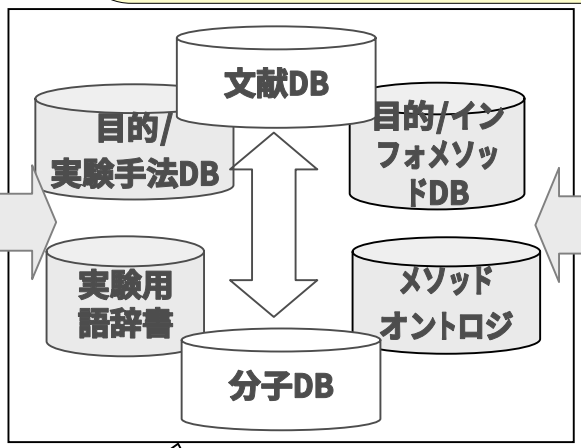
開発技術をウェットへ応用(19年度)

ウェット-ドライ融合型フロー抽出法の開発(19年度)

キュレーション支援システムと連携







✓実験研究者が目的や手法をもとにDBを検索利用できる環境

3.

バイオインフォマティクス解析

バイオインフォ解析メソッド名の収集と検索プログラム作成(18年度)

ポータル作製と連携

Webサービスやセマンティクwebの研究開発

文献に基づくバイオインフォ解析メソッドの抽出(NII/MRI/DDBJ)

メソッドオントロジの開発(BIRD/NII/MRI)

オントロジとウェブサービスを利用したバイオインフォマティクス解析環境の開発 (BIRD/DDBJ)

2.

実験メソッド名の収集と辞書作製(18年度)

シソーラス/オントロジ作製 (19年度)

JST・BIRD・バイオ基幹情報資源の高準化と共用化
メソッド・オントロジーに関する研究と連携

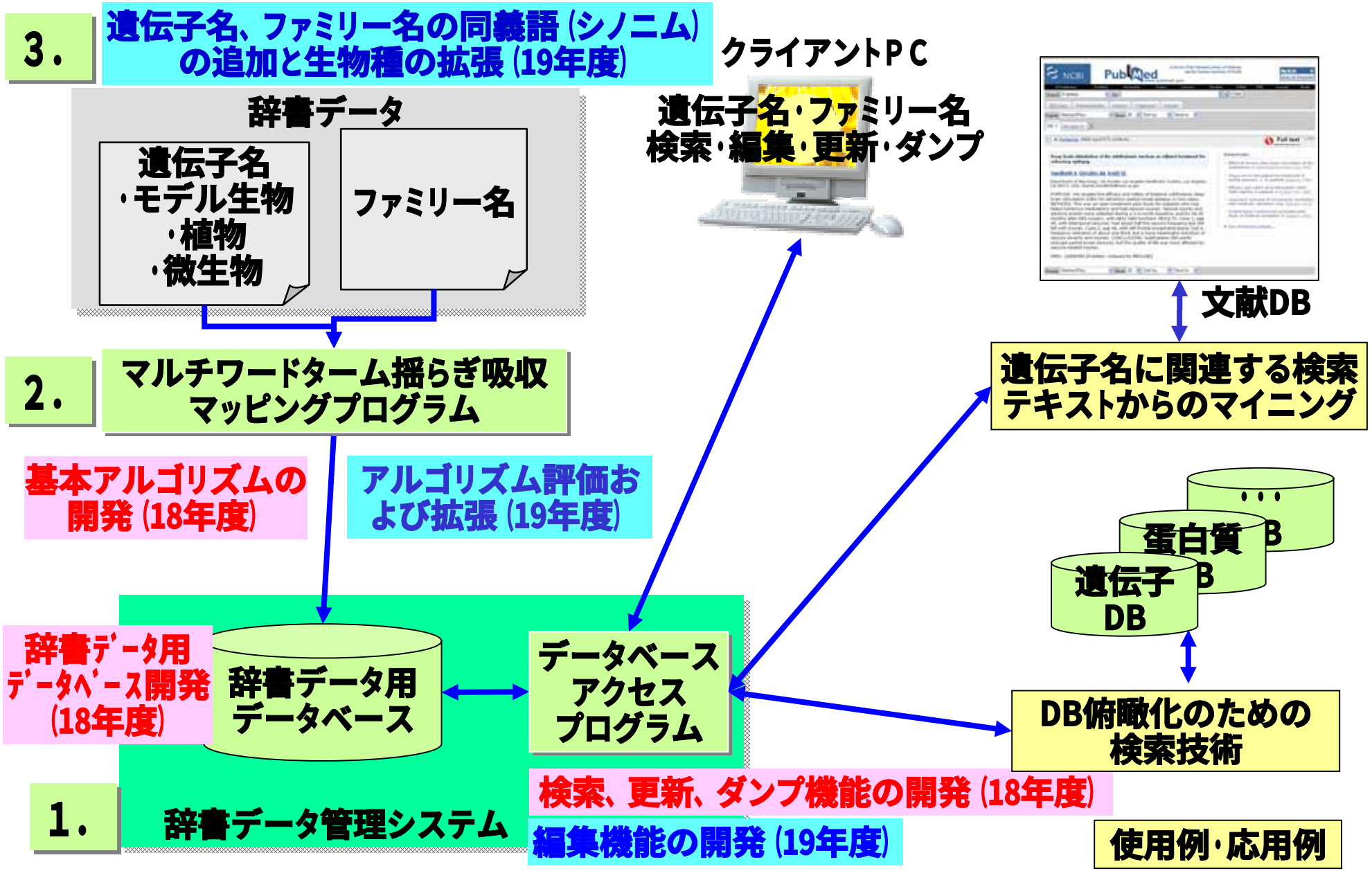
「遺伝子名揺らぎ吸収システムの開発」の実施項目（計画）

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**（18**年度**）
   - 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**（18**年度**）
   - ・**編集機能の開発**（19**年度**）

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**（18**年度**）
   - 2) **アルゴリズムの評価および拡張**（19**年度**）

3. **辞書データの拡張**（19**年度**）
   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

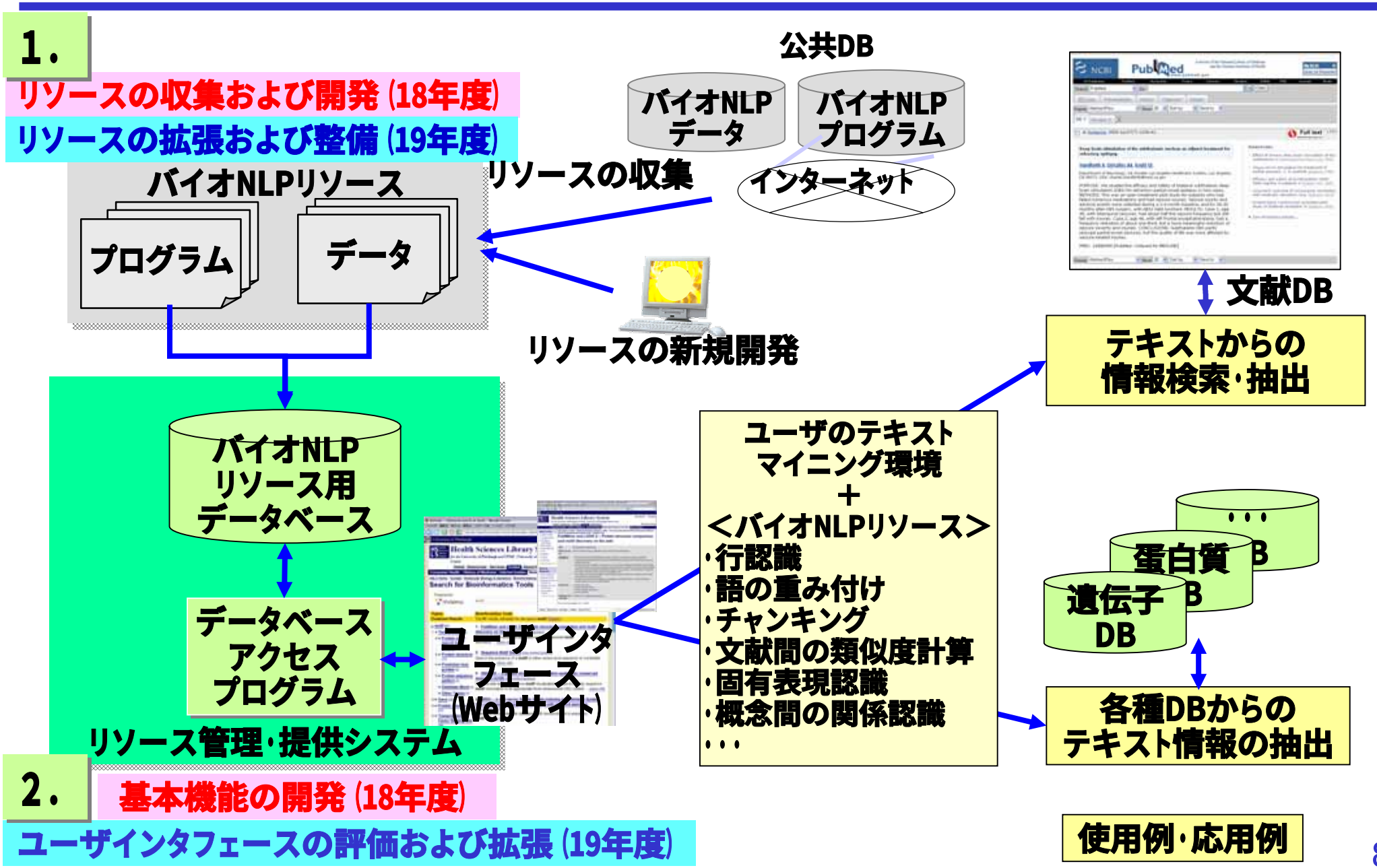


「バイオNLPリソースの整備」の実施項目（計画）

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（18年度）
      ・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（19年度）

2. リソース管理・提供システムの開発
   1) 基本機能の開発（18年度）
      ・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（19年度）

「バイオNLPリソースの整備」の成果イメージ

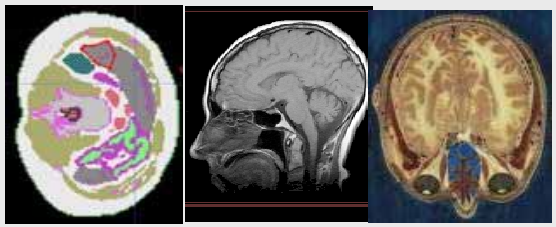


「解剖学用語表示システムの開発」の実施項目（計画）

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（18年度）
      → 自動化の検討（19年度）
   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（18年度、19年度）
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（18年度）
   2) 各臓器の詳細構造の再現（19年度）
3. 解剖学用語間の関係抽出による知識の構造化の検討（19年度）
   - 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（19年度）
   - 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)

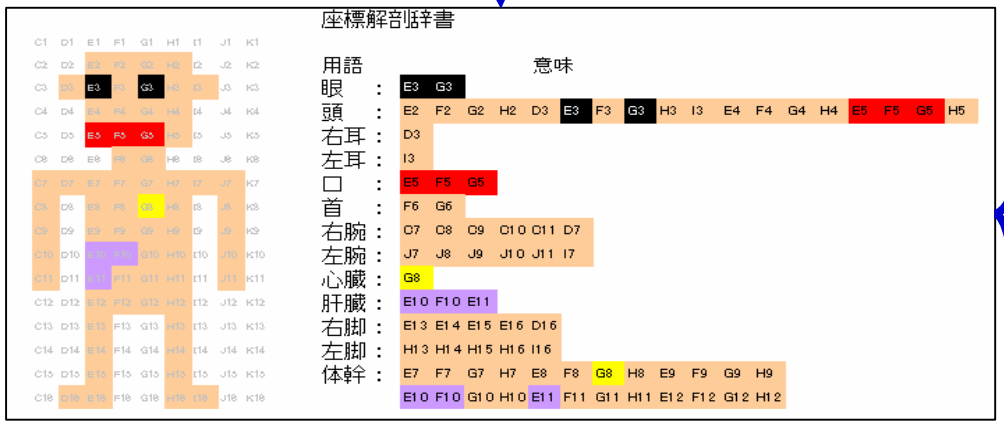


解剖用語
(ヒトについてはLSDの2,500語が目標)

既存ツールの利用
(18年度)

ボクセル辞書作成環境

自動化の検討
(19年度)



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 (18年度)

各臓器の詳細構造の再現　(19年度)

発現データの人体モデルへのマッピング

可視化 (19年度)

文献DB

解剖学用語に関連する検索
テキストからのマイニング

構造化された
解剖学用語

用語間の関係抽出
(19年度)

応用例

「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（18年度）
   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（19年度）
   1) Taxonomyデータベース 問題解析に着手（18年度）

「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

18年度

各種データベース

DDBJ DNA Data Bank of Japan

PDBj

GEO Gene Expression Omnibus

CIBEX Single Nucleotide Polymorphism

1．1)

引用文献情報抽出システム
クラスタリングシステム

研究プロジェクトA

データベースレコードのクラスタ

大分類1

中分類1　中分類2　中分類3

データベースに含まれる
各種研究プロジェクトの
内容や規模を俯瞰

データベースレコードの
引用文献情報に基づくクラスタリングと
研究プロジェクト索引付与による分類づけ

研究プロジェクト索引辞
書構築システム

索引辞書
構築

結果評価

索引付与
実行

1．2)

研究プロジェクト索引辞書

研究プロジェクトによる
索引付けシステム

1．3)

研究プロジェクトによる
索引付け公開システム

1．4)

索引情報
データベース

19年度

各種データベース

DDBJ　PDBj　GEO Gene Expression Omnibus　CIBEX

Taxonomyデータベース

Taxonomy

生物種分類整理
システム

生物種分類による
索引辞書構築

索引辞
書構築

結果
評価

索引付
与実行

「癌研究知識表現技術開発」の実施項目（計画）

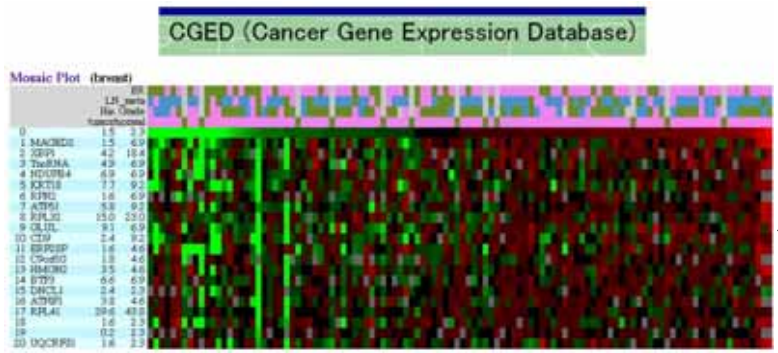
1. 実験的データベースの開発
   1) 癌遺伝子発現臨床情報データベース(Cancer Gene Expression Database, CGED)
      → 臨床情報から遺伝子を選択する機能を付加（18年度）
      → 神経膠腫など新規データのアップロード（19年度）
   2) 肺癌分子情報臨床情報統合データベース
      → 遺伝子発現以外のゲノム情報を付加した新しいデータベースプロトタイプ（19年度）

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（18年度、19年度）
   2) 臨床医のニーズにあった知識表現技術のデザイン（19年度）

「癌研究知識表現技術開発」の成果イメージ

実験的データベースの開発

CGED (Cancer Gene Expression Database)



臨床情報の収集とニーズ調査

患者ID	SEX	最終予後	overall survival (OAS)	OAS打ち切り	年齢	Tobacco	CEA	根治性	Size	T	N	M	ステージ分類	組織型
4	1	原病死	414	非打ち切り	71	200	9.1	0	32	2	12	0	2B	Ad
70	2	原病死	860	非打ち切り	69	0	<1.0	0	32	2	11	0	2B	Ad
110	2	原病死	758	非打ち切り	70	200	1.8	0	32	2	12	0	2B	Ad
5	2	生存中	2876	打ち切り	53	0		0	60	2	12	0	2B	Ad
7	1	原病死	1337	非打ち切り	64	1720	22.0	1	22	1	21	0	3A	Ad
8	2	生存中	2398	打ち切り	53	0		0	40	2	2	0	3A	Ad
14	2	原病死	2094	非打ち切り	58	25	2.2	0	70	4	21	0	3B	Ad
15	2	生存中	2316	打ち切り	65	0	3.2	0	30	2	21	0	3A	Ad
17	2	生存中	2001	打ち切り	57	0	17.3	0	42	2	21	0	3A	Ad
18	2	生存中	1936	打ち切り	59	0	4.5	0	18	1	0	0	1A	Ad
20	2	生存中	1839	打ち切り	53	0	<1.0	0	32	2	0	0	1B	Ad
21	1	原病死	404	非打ち切り	59	1050	10.7	0	11	4	12	0	3B	Non
25	1	生存中	1056	打ち切り	49	0	1.2	0	18	2	12	0	2B	Ad
26	2	原病死	793	非打ち切り	67	0	22.8	1	32	4	21	0	3B	Ad
2	1	原病死	1148	非打ち切り	54	0	1.2	0	20	1	21	0	3A	Ad
3	2	生存中	972	打ち切り	70	0	28.7	0	28	1	21	0	3A	Ad
4	2	生存中	2205	打ち切り	48	0	1.1	0	20	1	0	0	1A	Ad

肺癌分子情報臨床情報統合データベース

検索機能付加
臨床情報→遺伝子
(18年度)

プロトタイプ完成
(19年度)

肺癌臨床情報
＋分子情報
データファイル
(18年度)

癌臨床情報データベースの基本デザイン
(19年度)

臨床医、癌研究者のニーズに答える知識表現技術の開発

「多型知識表現技術開発」の実施項目（計画）

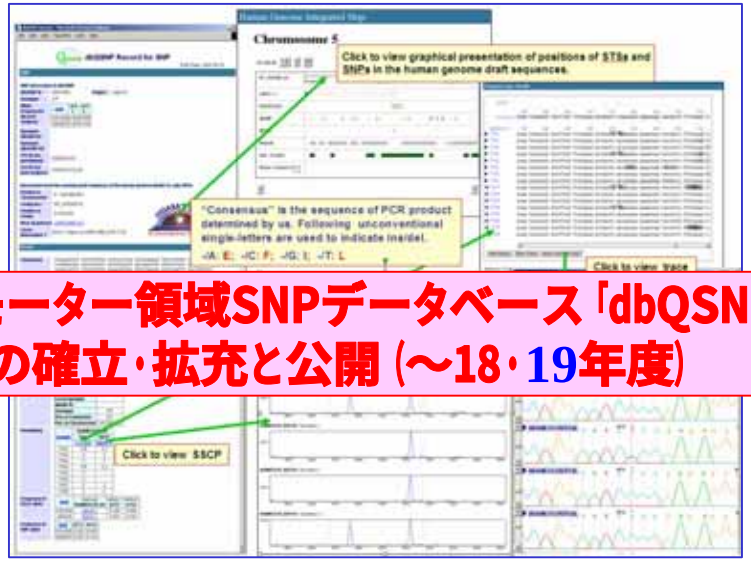
1. プロモーター領域SNPデータベース「dbQSNP」の確立・拡充と公開
   1) 波形トレースデータ（SSCP解析やシークエンシング解析）からSNP情報を確定するためのアルゴリズム、ツールの調査、検討
      → 既開発ツールの利用によるデータの拡充（～18年度）
      → 情報処理プロセスの簡易化、自動化の推進（18・19年度）
   2) 公開データベースのXML化
      → 種々の多型関連XML表現の比較検討（18年度）
      → 「dbQSNP」のPML化とその公開（18・19年度）

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（18年度）
   2) 拡大DNAアレイデータによる「D-HaploDB」の拡充（18年度）
   3) 「D-HaploDB」のPML化とその公開（18・19年度）

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ーの検討（19年度）
   - 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（19年度）
   - 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



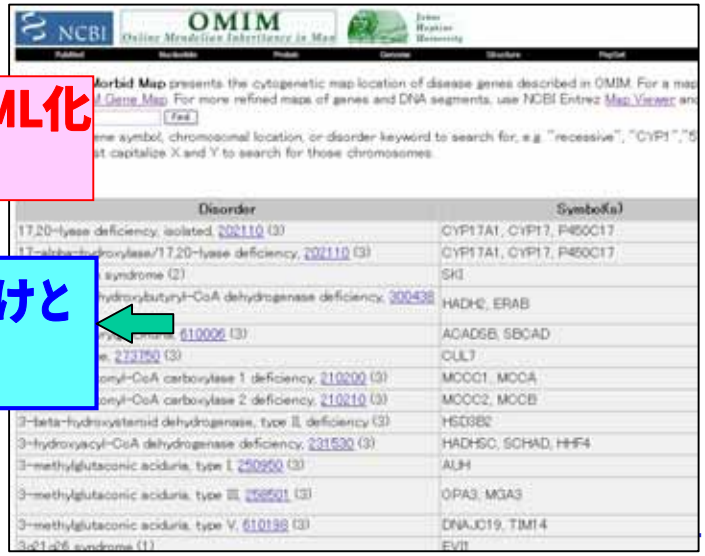
プロモーター領域SNPデータベース「dbQSNP」の確立・拡充と公開（～18・19年度）

確定ハプロタイプデータベース「D-HaploDB」の確立・拡充と公開（～18 ・19年度）

種々の多型関連XML表現の比較検討（18年度）

「dbQSNP」および「D-HaploDB」のPML化（18・19年度）

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（19年度）



「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（18年度）
      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（19年度）
      1) 論文構成要素の編集・分類・表示を行う（19年度）
      2) 論文構造基本モデルの定義による論文構造の俯瞰（19年度）

2. 論文構造解析・編集情報集約システムの開発（19年度）
   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

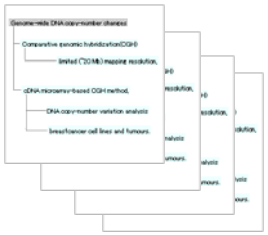
3. 論文構造解析・編集情報利用システムの検討（19年度）
   - 応用例； 遺伝子発現データベース、測定対象、測定手法に関連する論文情報データベースの構築

「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



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

i) 構成情報作成編集　新規開発（18年度）

iii) 引用論文取得表示

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

アノテーション実施（19年度）

アノテーションの集約と編集

構造化文献DB

アノテーション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（18年度）

2）論文構成・アノテーションモデルの適用（19年度）

2．論文構造解析・編集情報集約システムの開発（19年度）

3．論文構造解析・編集情報利用システムの検討（19年度）

応用例