ライフサイエンスデータベース統合推進事業
平成23年度キックオフ･ミーティング

基盤技術開発プログラム実施計画
概要説明

DBCLS

センター長　米澤　明憲

2011/06/10

発表内容

1. 研究計画実施体制について
   (スライド3~4)
2. DBCLSが提案する統合化について
   (スライド5~16)
3. 研究実施計画の概要
   (スライド17~37)

基盤技術開発プログラム実施体制：組織

長洲毅志　PO

研究代表者

ライフサイエンス統合データベースセンター
米澤明憲

研究開発題目：生命科学分野におけるデータベース統合化のための基盤技術開発

共同研究グループ

産総研・
生命情報工学研究センター
浅井潔

研究開発題目：解析プラットフォームによる統合利用環境の整備

共同研究グループ

京大・化学研究所
五斗進

研究開発題目：データ統合と新規分野データ活用のための基盤技術開発

基盤技術開発プログラム実施体制：メンバー

【DBCLS】

米澤明憲
大久保公策
岡本忍
川本祥子
金進東
坊農秀雅
箕輪真理
山口敦子
狩野芳伸
河野信

内藤雄樹
仲里猛留
山本泰智
飯田啓介
呉紅艶
大田達郎
小野浩雅
藤枝香
王悦
RA25名

研究協力者

中村春木
金城玲
片山俊明
川島秀一
小笠原理
岩崎渉

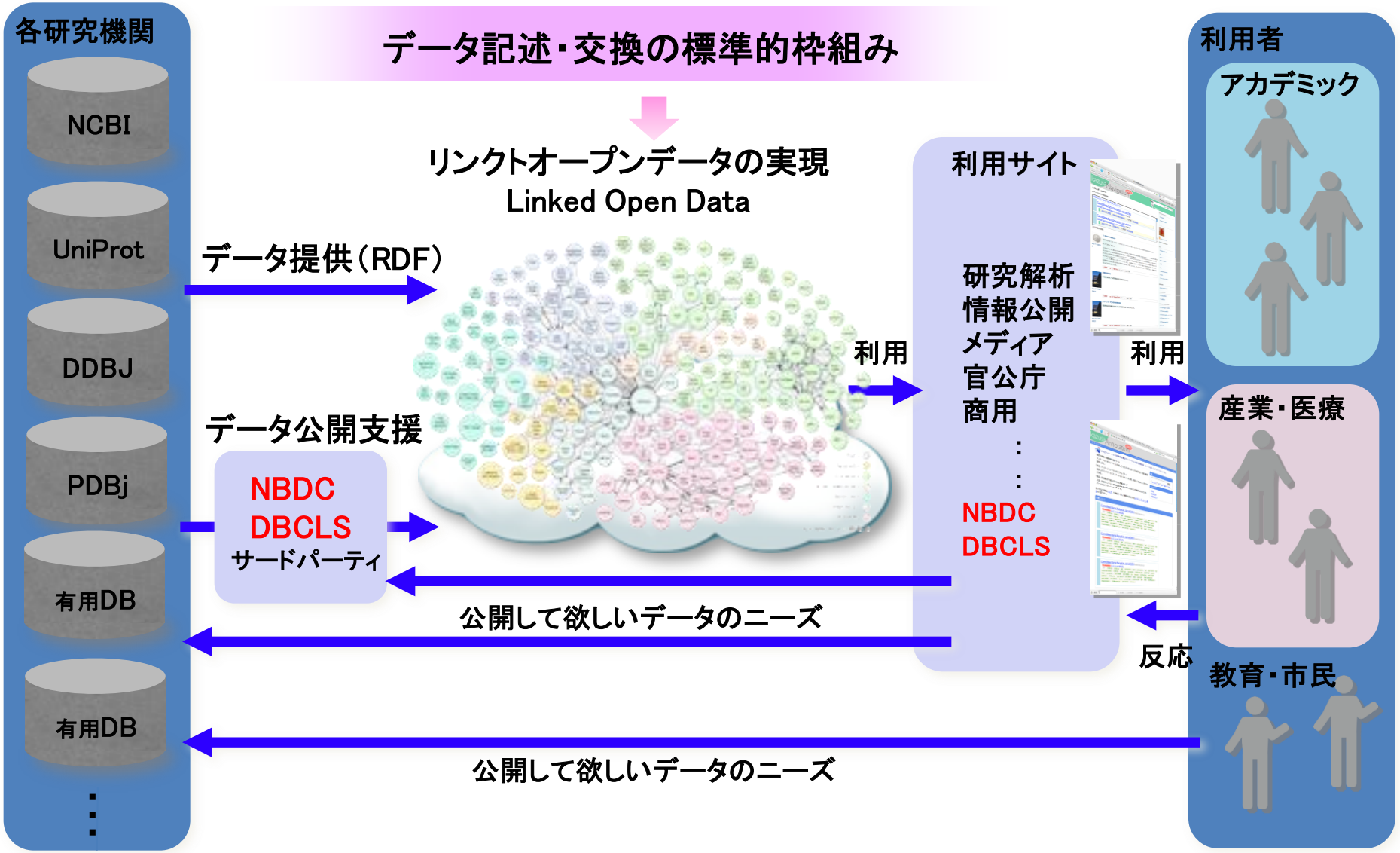
【CBRC】

浅井　潔
福井　一彦
田代　俊行

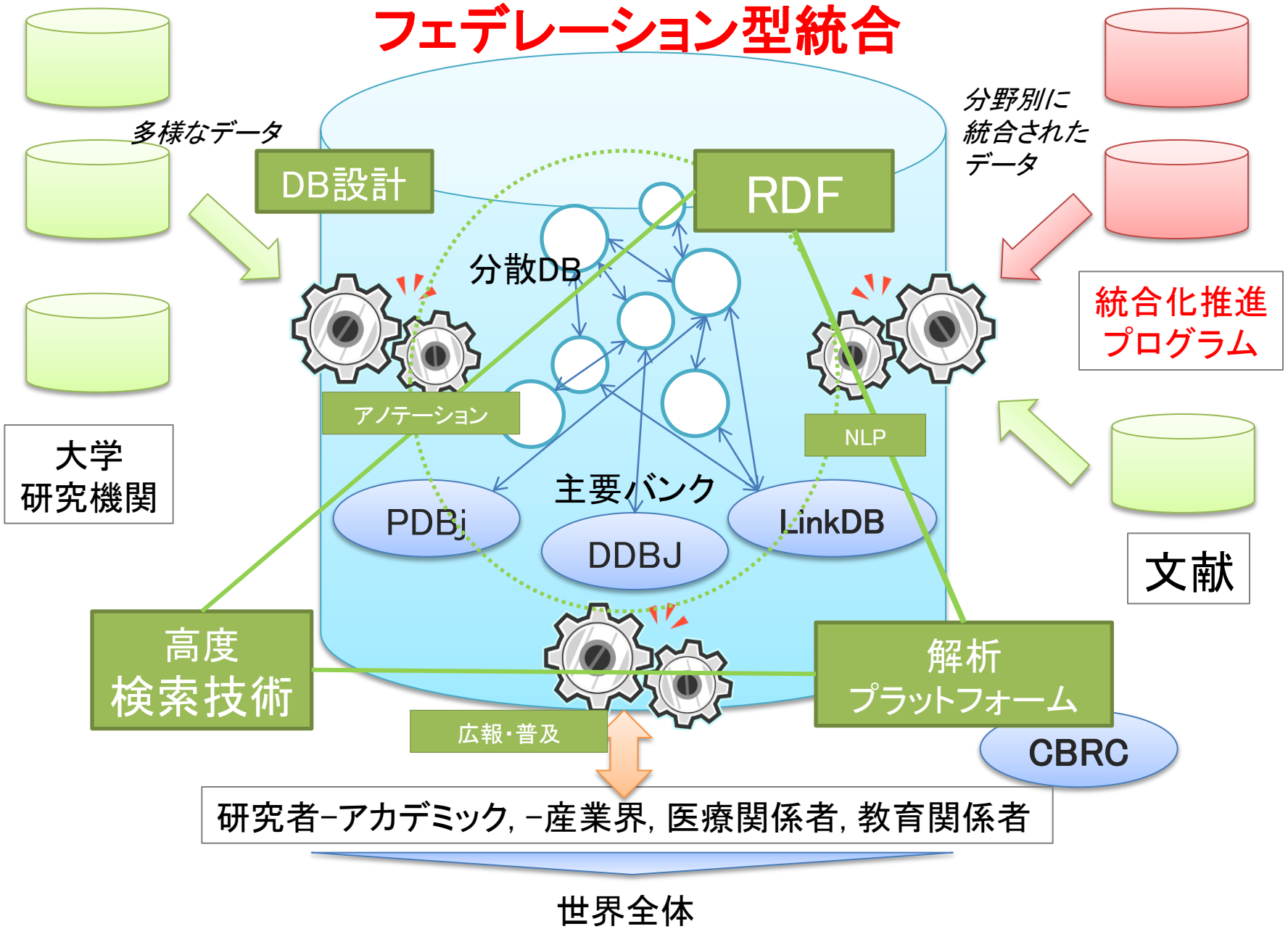
【京大】

五斗進
時松敏明
小寺正明
Diego Diez
中川善一
守屋勇樹

フェデレーション型データ統合による データ公開と利用のサイクル



DBCLSが提案する統合データベースの全体像



なぜフェデレーション型統合か

- 多種多様かつ大量な生物学データが世界中の様々な場所で日々作られている
- 目的手段がばらばらな解析ツールと解析プラットフォームが開発

1箇所で全てのデータを運用更新するスタイル(例：NCBI)ではダメ
- 人員面でも費用面でも非常に高コスト
- 限定的な種類のDBのみサポート可能

逆に，ばらばらに存在するままで，知識抽出などの処理を可能にするためには
- 統一的フォーマット(RDF)で記述し
- それを活用するための解析プラットフォームが必要

フェデレーション(連携)型統合とその狙い

DBCLSは,
RDFなど
セマンティックウェブ技術を持つ

＋

DBCLS は,
国内外のデータベース，解析ツール,
解析プラットフォームの開発者を集めて,構築してきた人的ネットワーク

DBCLSがフェデレーション型をとることで
NCBI（集中型）,EBI（分散型）とは
異なる連携型の国際的な拠点に

ライフサイエンス分野のDB統合化と活用に関わる
将来のあるべきインフラストラクチャー構築を狙う

統一的フォーマットの実現

セマンティックウェブ技術の活用

生物学者のDB利用のニーズは多様

多種多様なニーズに応えるために
あらゆるDBをDBの構造と独立に
標準的な形で再構築する必要性がある

➡ シンプルかつ標準的技術を用いた
再構築　＝＞　RDF化

RDFとは

データ記述・交換のための標準的枠組み

意味を機械可読な形で流通させることが可能

特徴

シンプル　**主語 述語 目的語** の三つ組で関係情報を表現

例:「UniProt ID O95819 の EC番号は 2.7.1.1」というデータ
→主語:　<http://www.uniprot.org/uniprot/O95819>
述語:　<enzyme>
目的語:　<http://purl.uniprot.org/enzyme/2.7.1.1>

クエリ言語が用意されている

SPARQLがデファクトスタンダードであり，いくつかの実装がすでに存在

IDとしてURIを利用

インターネット上で一意となる

RDF化による統合

例:「PDB のエントリ 3elj のタンパク質についてバイオアッセイの結果が欲しい」時

PDB

PDB ID	UniProtID
3elj	P45983
3hec	Q16539

⋮

PubChem BioAssay

AID	Name	Protein Accession
494405	Inihibitition of JNK1	P45983
490834	Inhibition of JNK1 at 10 uM	P45983

⋮

「PDB の UniProtID という項目も PubChem BioAssay の Protein Accession という項目も UniProt のIDを指している」という知識が必要

複数のDBを必要とするデータ取得は、それぞれのDBについて詳しい知識が必要

従来の解決方法

項目の対応を何らかの方法で明示的に示す

PDB

PDB ID	UniProtID
3elj	P45983
3hec	Q16539

PubChem BioAssay

AID	Name	Protein Accession
494405	Inihibitition of JNK1	P45983
490834	Inhibition of JNK1 at 10 uM	P45983

スキーマ名を揃える
対応表を用意する
など

RDF化による統合

PubChem BioAssay

AID	Name	Protein Accession
494405	Inihibition of JNK1	P45983
490834	Inhibition of JNK1 at 10 uM	P45983

<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=49083
<Name>
“Inihibition of JNK1“.

<http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=4908
<ProteinAccession>
<http://purl.uniprot.org/uniprot/P45983>

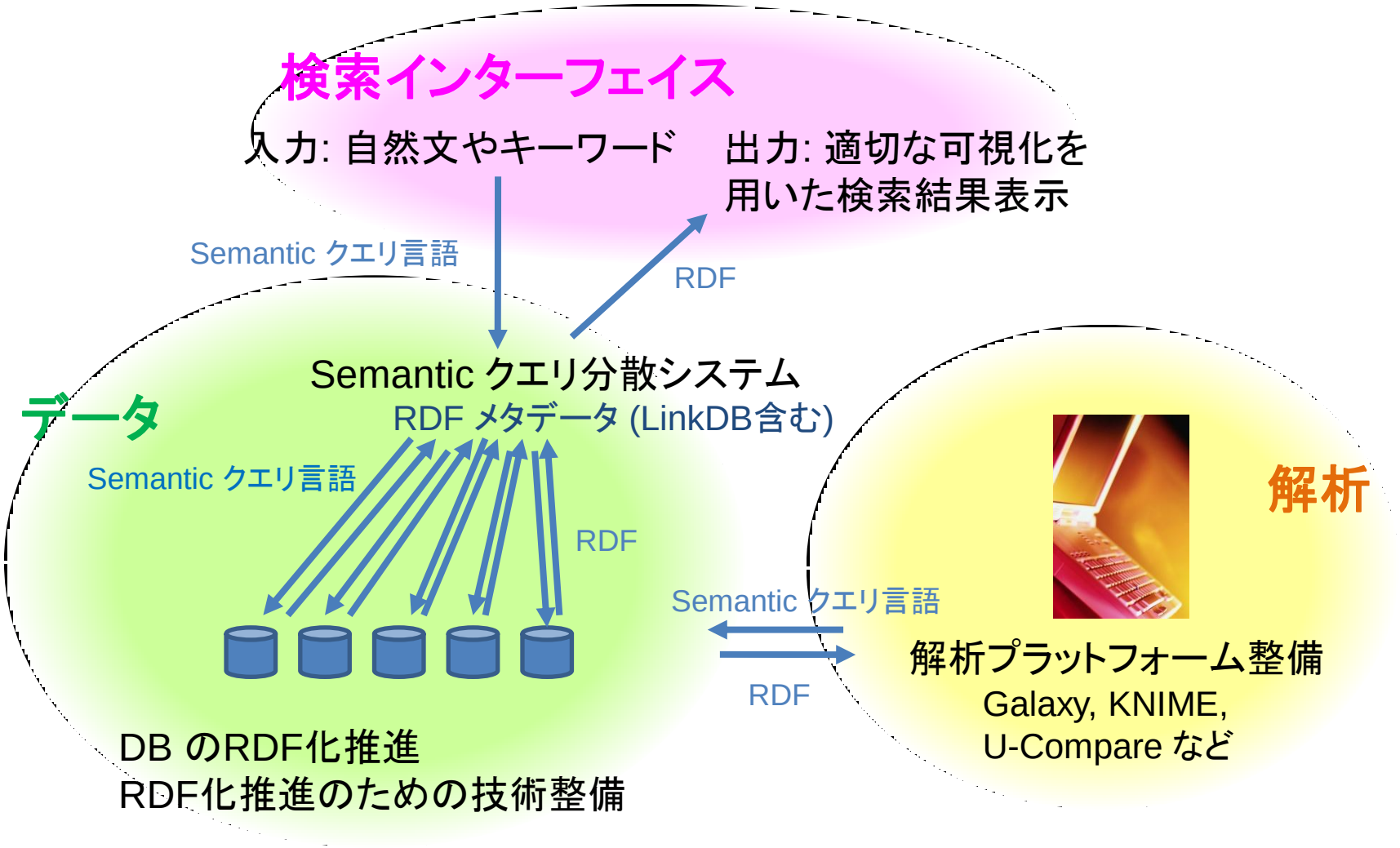
PDB

PDB ID	UniProtID
3elj	P45983
3hec	Q16539

<http://service.pdbj.org/mine/xpath/3ELJ>
<http://www.pdbj.org/schema/pdbx-v32.owl# link_to_uniprot>
<http://purl.uniprot.org/uniprot/P45983>

同じDBの同じエントリかどうか URL を見れば自明

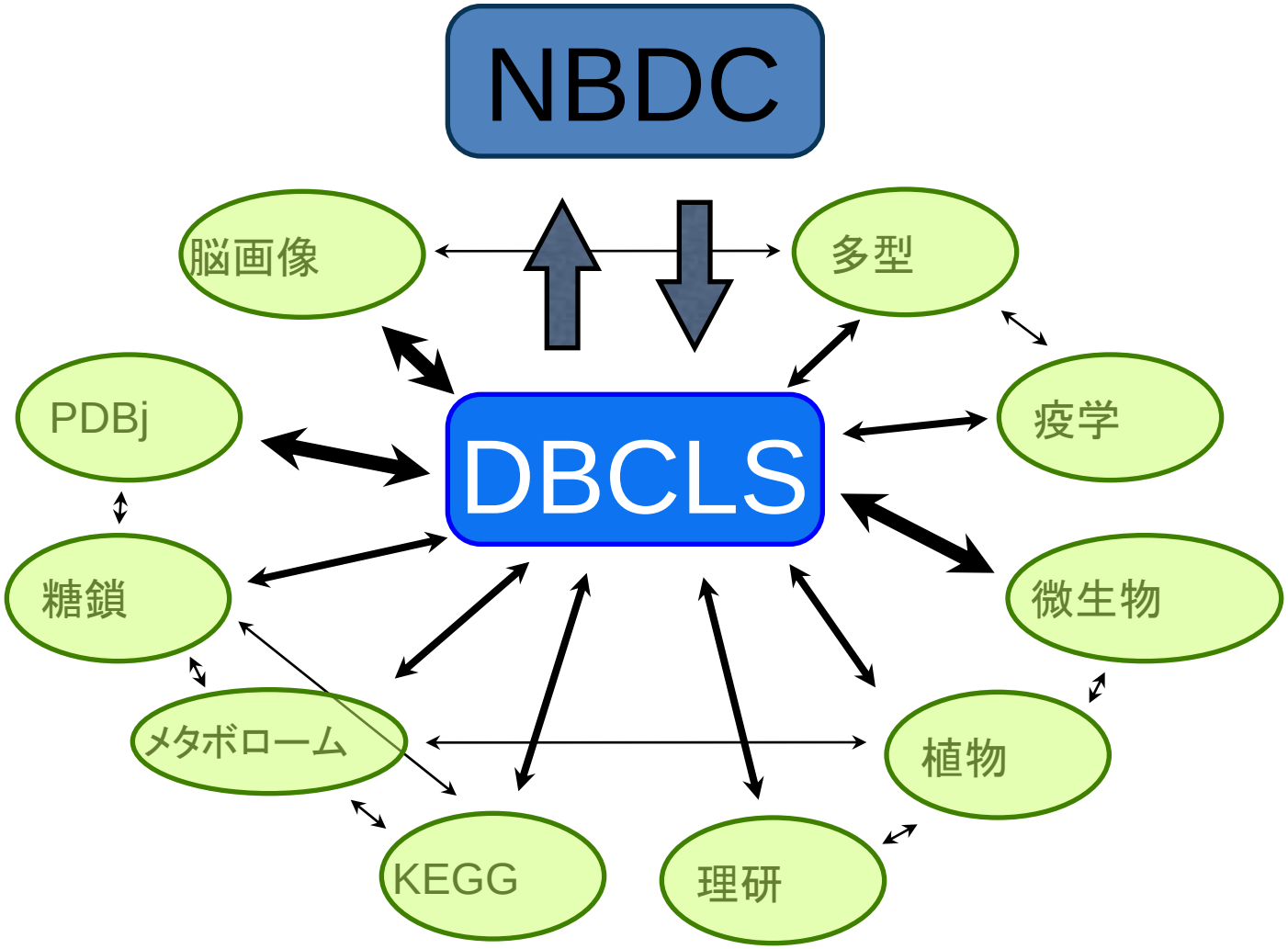
RDFによる統合利用環境の構築



国際連携-BioHackathonの開催

第1回  共催：CBRC	2008/2/11- 15 開催地: 東京 参加者 70名 （海外32名）	テーマ: ウェブサービスの標準化 (Towards interoperable web services in life science with Open Bio * libraries)
第2回  共催：OIST	2009/3/15- 21 開催地: 東京，沖縄 参加者 60名 (海外28名)	テーマ: 実用的な解析ワークフロー作成 (Integration of web services in bioinformatics applications – Verification through real world use cases –)
第3回  共催：CBRC	2010/2/8 – 12 開催地: 東京 参加者80名 (海外32名)	テーマ: セマンティックウェブ技術による生物学知識の解釈 (Interpretation of biological knowledge with Semantic Web technologies)
第4回 NBDC＆DBCLS&KyotoU BioHackathon	2011/8/21-26 開催地：京都 海外招聘者25名を予定	（仮）テーマ：LinkedData実現に向けての技術開発、オントロジー整備など

統合化推進プログラムとの連携



各グループにヒアリングしてニーズを把握

（優先してRDF化するDB，整備するオントロジー・辞書のフィードバック）

公募要領から

優先的に取り組む課題

1. 先端的なプログラミング技術によるインターネットを活用した高度検索技術開発を行うこと

2. 国内の基盤的データベースおよび本事業で構築される分野別統合データベースのRDF 化を実現するための, 標準フォーマット, オントロジーの提供, RDF 化の支援を行い, RDF コンテンツを公開すること

3. 一連の作業を自動化するための仕組みを構築するとともに, 高度なインターネット技術, データベース技術を応用した統合利用環境を整備すること

4. 一貫性のあるポリシーのもとで実用的なオントロジー, 辞書, コーパス, 標準化技術を開発すること

5. 個人ゲノム等大規模データを活用する技術を開発すること

公募要領から
全体の進捗に合わせ進める課題

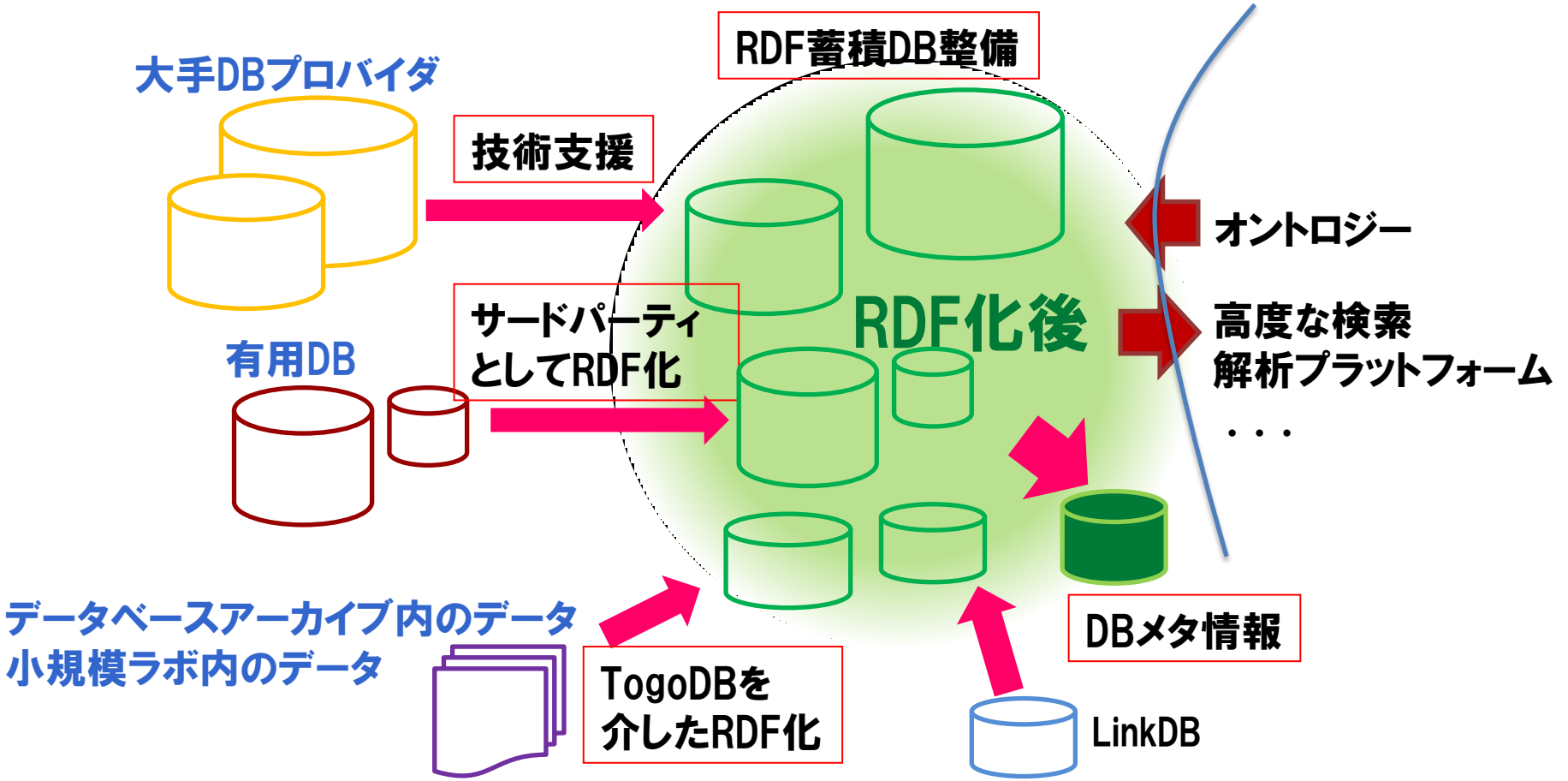
6. コーパス構築を含んだ文献，画像等コンテンツ活用技術を開発し多様な検索に対応すること
7. 論文解読技術や文献管理システムをベースに論文作成や管理，アノテーション作成を支援する技術を開発し多様な検索に対応すること
8. 医療用画像等に関わる画像データの利用技術（管理・検索・標準化・定量的解析等）を開発すること
9. 統合データベース利用のためのコンテンツ作成（データベースやツールの使い方に係わるコンテンツや，チュートリアルの動画コンテンツ等）とアノテーション支援を行うこと
10. 統合検索システムとアーカイブシステムの有用性の高い高機能化技術を開発すること
11. バイオサイエンスデータベースセンターでの，個人ゲノム等の個人情報に関わるセキュリティ，公開範囲等の検討を踏まえ，適正な内容で公開可能となるよう技術開発を行うこと
12. 脳画像，単一セル計測など，新規分野データの統合化に関わる要素技術を検討し，将来のデータベース化に必要なツール開発等をタイムリーに行うこと

7つの課題に再編

1. データベースのRDF による統合化
   →DBCLS, 京大
2. 解析プラットフォームによる統合利用環境の整備
   →CBRC, DBCLS, 京大
3. インターネットを活用した高度検索技術の開発
   →DBCLS, 京大
4. RDF化に資するオントロジー, 辞書,コーパス整備,
   標準化技術開発→DBCLS, 京大
5. 大規模データの利用技術開発
   →DBCLS, 京大
6. 情報統合化・知識発見のためのキュレーション支援
   →DBCLS
7. 統合データベースに関わるコンテンツの作成 , 整備
   →DBCLS

（1）データベースのRDFによる統合化

セマンティックウェブ技術を活用した
統一的フォーマットの実現

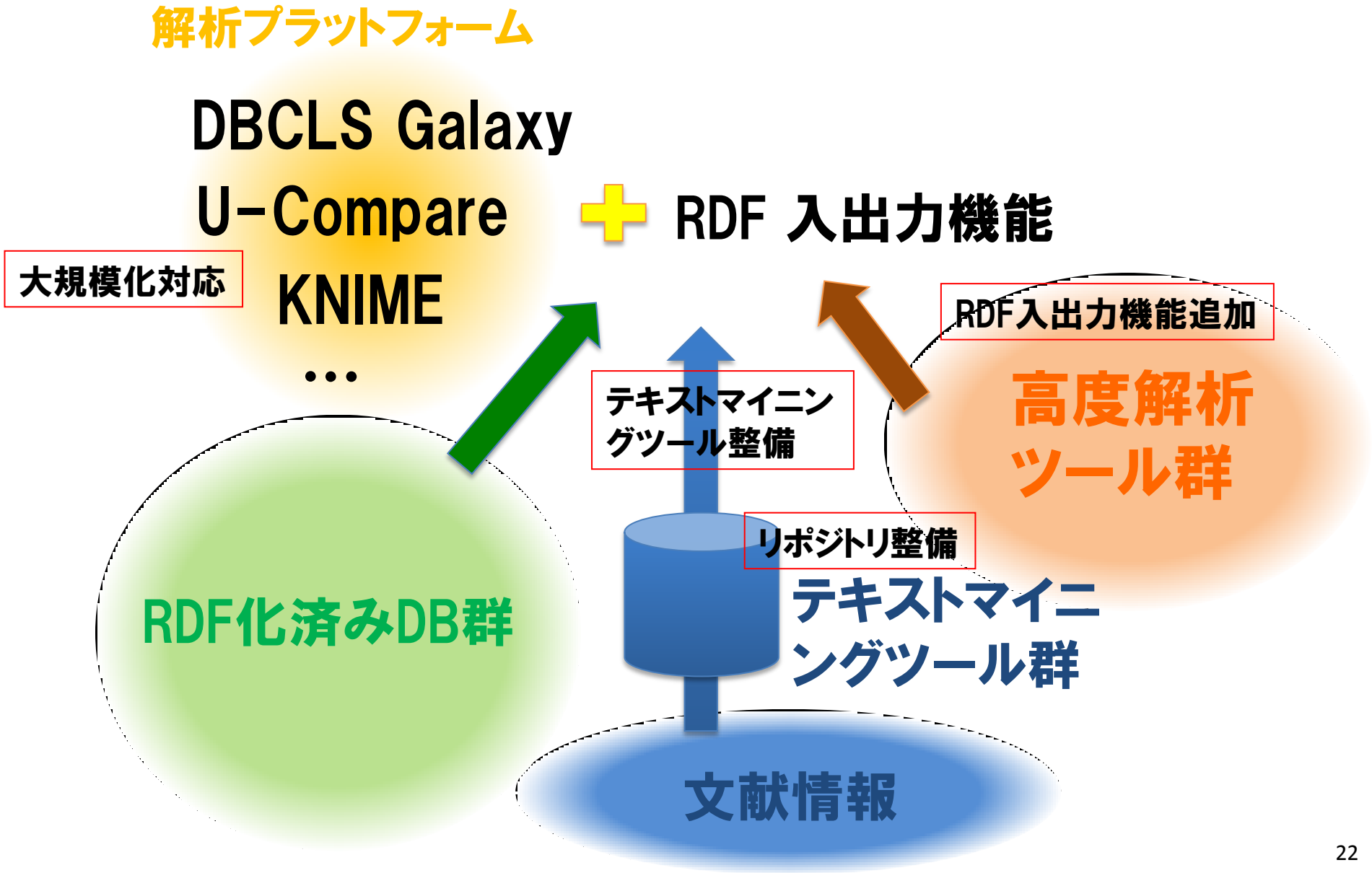


(1) データベースのRDFによる統合化

今年度研究開発内容

1. RDF化ガイドライン整備に着手 技術支援
2. RDF化すべきDBの優先順序調査 サードパーティとしてRDF化
3. TogoDB RDF化機能プロトタイプ開発 TogoDBを介したRDF化
4. DBメタデータRDF蓄積仕様検討 RDFメタ情報蓄積
5. 既存のRDFストアの調査・比較検討 RDF蓄積DB整備
6. LinkDB の RDF化着手 (京大化研に技術支援)

(2) 解析プラットフォームによる統合利用環境の整備



(2) 解析プラットフォームによる統合利用環境の整備

今年度研究開発内容

リポジトリ整備

1. テキストマイニングツール群の整備とリポジトリ化

2. 1.を DBCLS Galaxy に反映

テキストマイニングツール整備

3. U-Compare ワークフロー分散処理化

大規模化対応

4. ツールの RDF 入出力機能追加
（CBRCに技術支援）

RDF入出力機能追加

(3)インターネットを活用した高度検索技術の開発
既開発サービスの機能向上

本課題の役割：RDF化による統合を目標に，NBDCと協力し既開発サービスの機能向上，高度検索技術の開発を行う

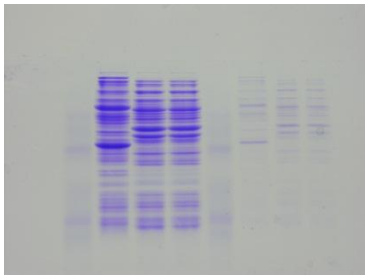
- DBカタログ
  - (H23)カタログデータのRDF化，RDF化チームへのデータ提供
  - (H23-H24)セマンティック検索の実現
- 横断検索
  - (H23)NBDCサーバへの実装と簡単な機能改良
  - (H24-H25)RDF化データへの対応
- アーカイブ
  - (H23)検索機能の利便性向上
  - (H24-H25)分野別統合化から求められる機能の追加

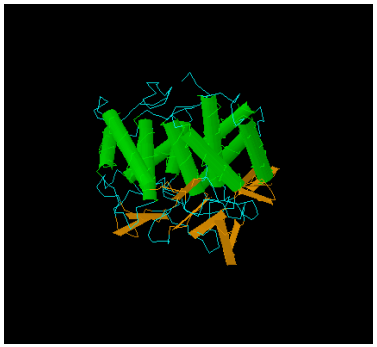
有用分野のプロトタイピング

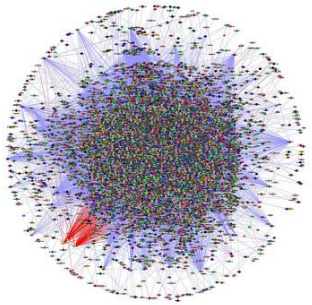
基盤的技術（主に実施項目1と4）を活用したユースケースを提示
DB利用者とDB開発者・実装者の橋渡し

・疾病に関連する分野
・プロテオミクス分野

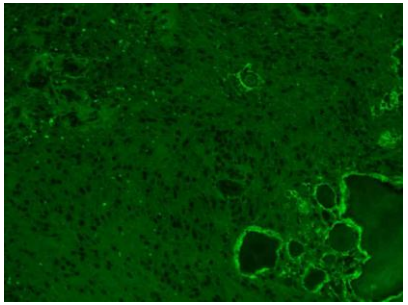
多様なデータ:
配列、物理化学的性質、立体構造、PPI、翻訳後修飾、発現情報







http://www.flickr.com/photos/andytrop/5234332602/



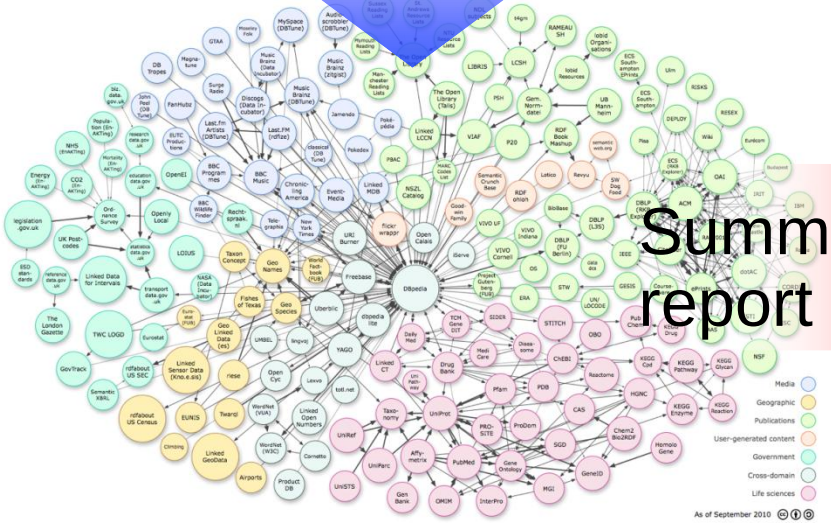
ニーズの調査・情報の適切な表示法の検討（H23~H24）
プロトタイプの実装（H23後半~H25）

(3)インターネットを活用した高度検索技術の開発

検索結果の情報提示法と可視化の開発

Query

Visualization

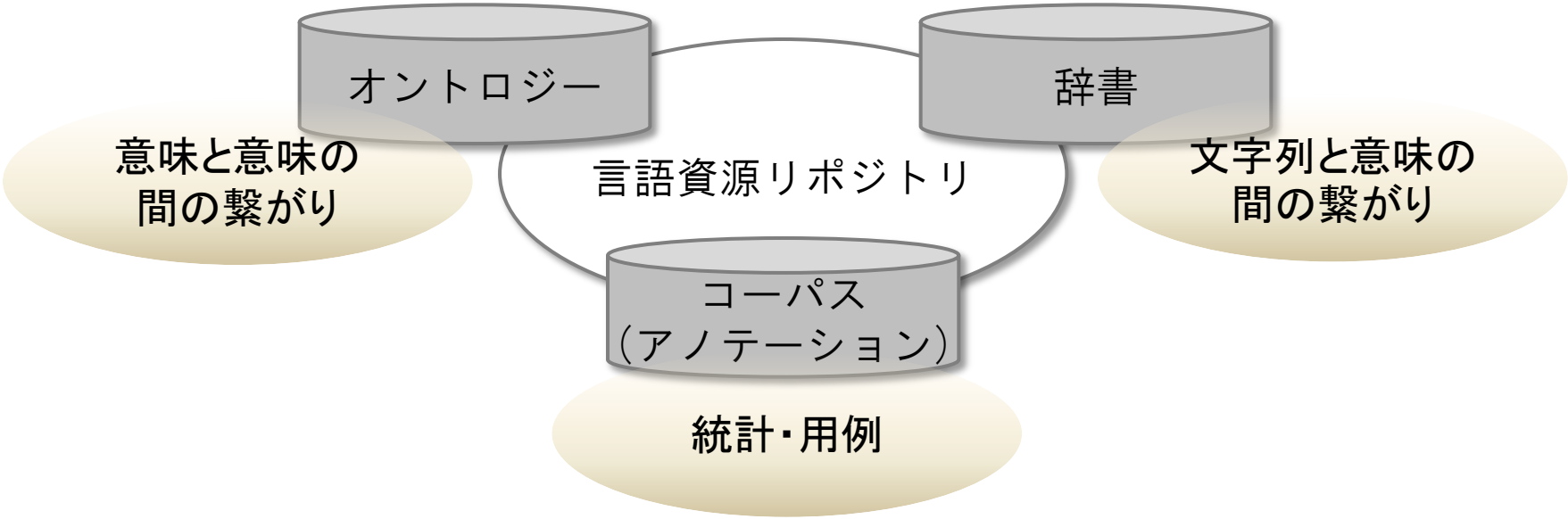


Linked DBs

Summarized report

(4) RDF化に資する オントロジー, 辞書, コーパスの整備, 標準化

- 効果的なRDF化の為には概念の標準化が必要
  - オントロジー及び辞書の整備
- オントロジー・辞書の開発、そして文献データの統合の為にはコーパス・アノテーションの確保が必要
  - コーパス・アノテーションリポジトリの構築



(4) RDF化に資する オントロジー, 辞書, コーパスの整備, 標準化

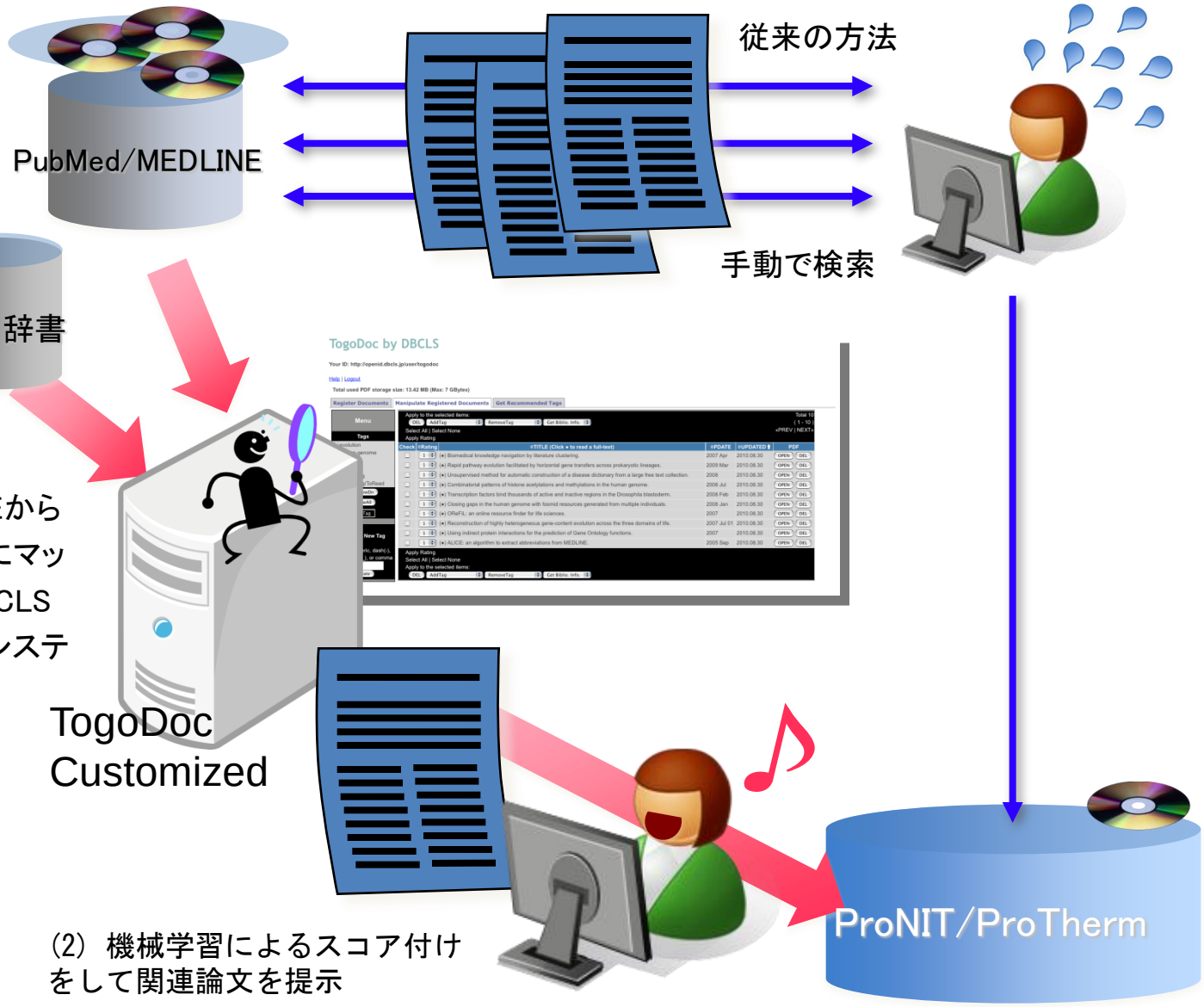
- 標準フォーマットの開発(H23)
  - オントロジーや辞書、コーパスとアノテーションが相互参照できるように統合的フォーマットを開発
- トップレベルオントロジーの開発(H23)
  - 様々なオントロジーの統合に向けトップレベルオントロジーを開発
  - OBO FoundryのBasic Formal Ontology(BFO)との互換性を保つ
- 自然言語処理、テキストマイニングツールの整備(H23-25)
  - 辞書、オントロジー, コーパス開発に必要なツールの整備
  - 既存ツールのドメインへの特化
- 反応パターンの整備（京大グループ）(H23)
  - 代謝パスウェイに登場する酵素反応について, 反応パターンを定義

(5) 大規模データの利用技術開発

1. メタ情報による大規模ゲノム配列データの整理・再利用促進技術開発
   1)ダイジェスト版の構築・開発・維持
   2)RNA配列を中心にしたDB利用技術開発
2. 遺伝子発現リファレンスデータセット整備
   1)RefExの構築維持管理
   2)SRAからのデータ抽出技術開発
3. 医療用画像データの利用技術開発

(6) 情報統合化・知識発見のためのキュレーション支援

キュレーション支援システムの開発

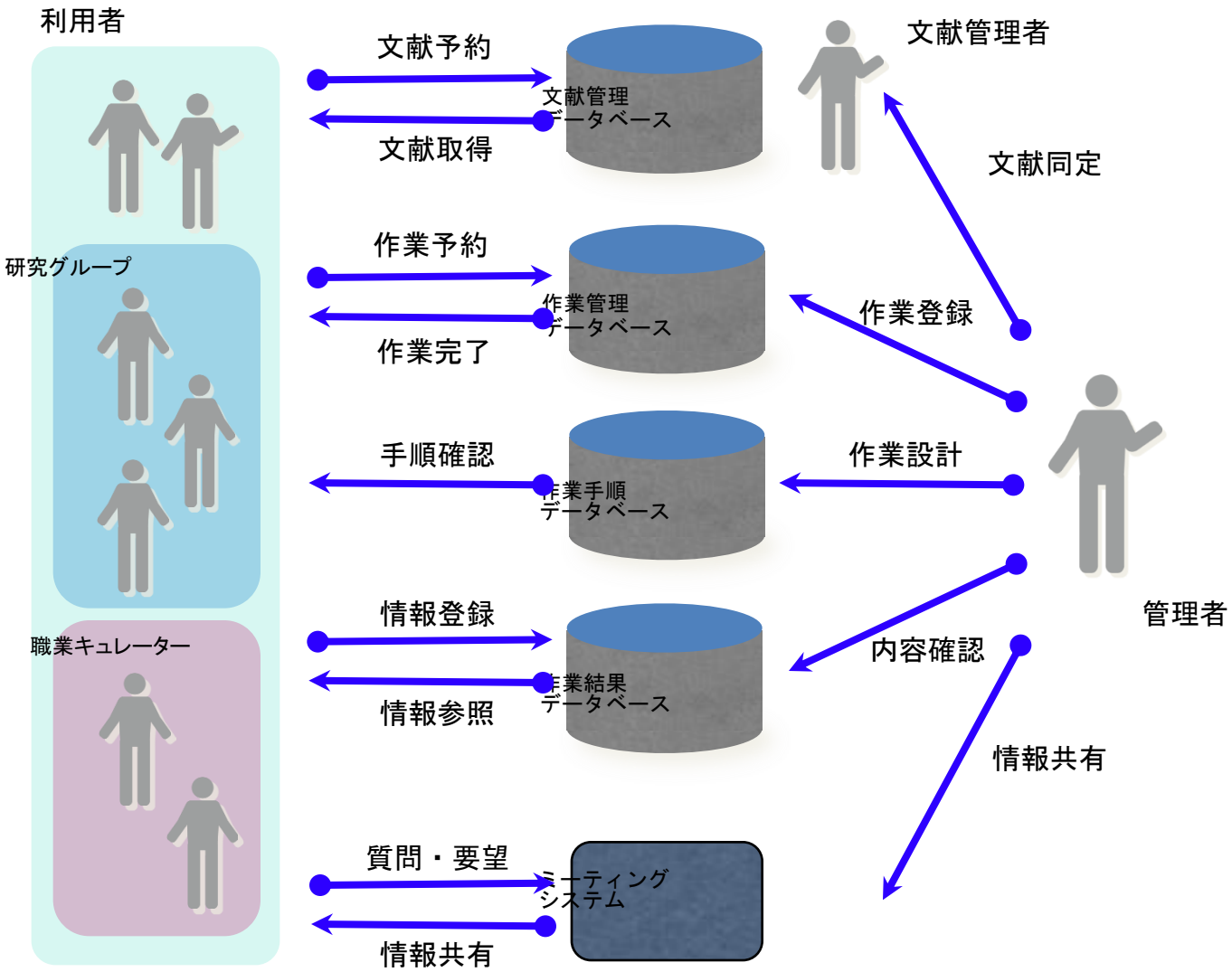


(1) PubMed/MEDLINEから特定のキーワード群にマッチする文献情報をDBCLS内部で構築した検索システムを利用して取得

(2) 機械学習によるスコア付けをして関連論文を提示

(6) 情報統合化・知識発見のためのキュレーション支援

協働キュレーション作業運用技術の整備



（7）統合データベースに関わるコンテンツの作成、整備

1. 動画によるチュートリアル作成

1) 更新動画の維持管理システム整備

2) 音声付与システム開発

2. 良質な日本語コンテンツの作成、整備

1) 日本語レビューのオンラインジャーナルの立ち上げ

3. 生物画像コンテンツの作成、整備

1) ヒト3Dマップ（個体レベル）の整備

2) 細胞、分子レベルの画像コンテンツの検討

産総研グループ

解析プラットフォームの基盤技術開発、利用環境の整備及びを主としてタンパク質情報や機能性RNAに対して実施する。解析結果の統合化された可視化

ワークフロー統合利用環境の構築・整備

- 国内・外の既存解析ツールをノード化し解析プラットフォームを利用したワークフローの拡張及び整備
- CBRC独自に開発中である解析ツールのノード化

プラットフォームとDB連携

- 解析ツールのSemantic Web対応
- プラットフォーム上からDBCLSが進めるRDF化されたDBとの連携機能開発

解析・検索結果の可視化

- 解析結果や検索結果の可視化ツール開発
- 解析プラットフォームとレポート作成機能を連携し容易に解析結果をまとめる機能を開発
- プラットフォーム上からDBと連携可能なノード及び解析結果とDBから取得した情報を融合させた表示機能の開発

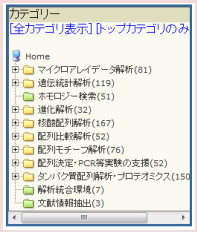
産総研グループ

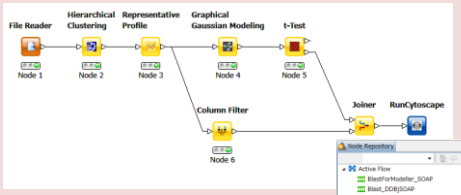
基盤技術開発プログラム

ワークフロー統合 利用環境の構築・整備

分散している既存解析ツールの統合化

CBRC独自解析ツールのノード化





効率性・柔軟性を考慮したプラットフォーム

コンテンツ作成

利用者のためのチュートリアル

プラットフォームとDB連携

RDF化

解析プラットフォームとDBに格納されたデータ連携による高度な知的基盤

Semantic Web Serviceとの相互運用性

・解析ツールのRDF化
・KNIMEからのDB利用

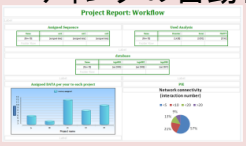
解析・検索結果の可視化

表示機能

解析結果の可視化及びDB連携ノードやその検索結果を統合した可視化ツール開発

レポーティング作成機能

レポーティングの自動化



DBCLS Database Center for Life Science

セマンティックWEB
RDF/OWL/SPARQL

統合DB

連携

京都大学グループ

（1）DBGET/LinkDBシステムの統合利用環境への応用

（2）メタゲノム・メタメタボローム等新規分野データ活用技術の開発

（3）反応オントロジーの整備

DBGET/LinkDB統合データベース検索システム（170 DBを統合）

カテゴリ	bget	bfind	blink	DB数
1. KEGGデータベース(DBGET版)	yes	yes	yes	22
2. その他のDBGETデータベース	yes	yes	yes	19
3. Web上の検索可能データベース	no	yes	yes	18
4. Web上のリンクのみのデータベース	no	no	yes	110
5. PubMedデータベース	yes	no	yes	1

LinkDB：6億以上のリンク情報
- 順引きリンク：データベース中に記述されているリンク
- 逆引きリンク：他のデータベースから参照されているリンク
- 等価リンク：データベース間で同じ化合物・遺伝子の関係を定義したリンク

blink: データベース間のリンク情報

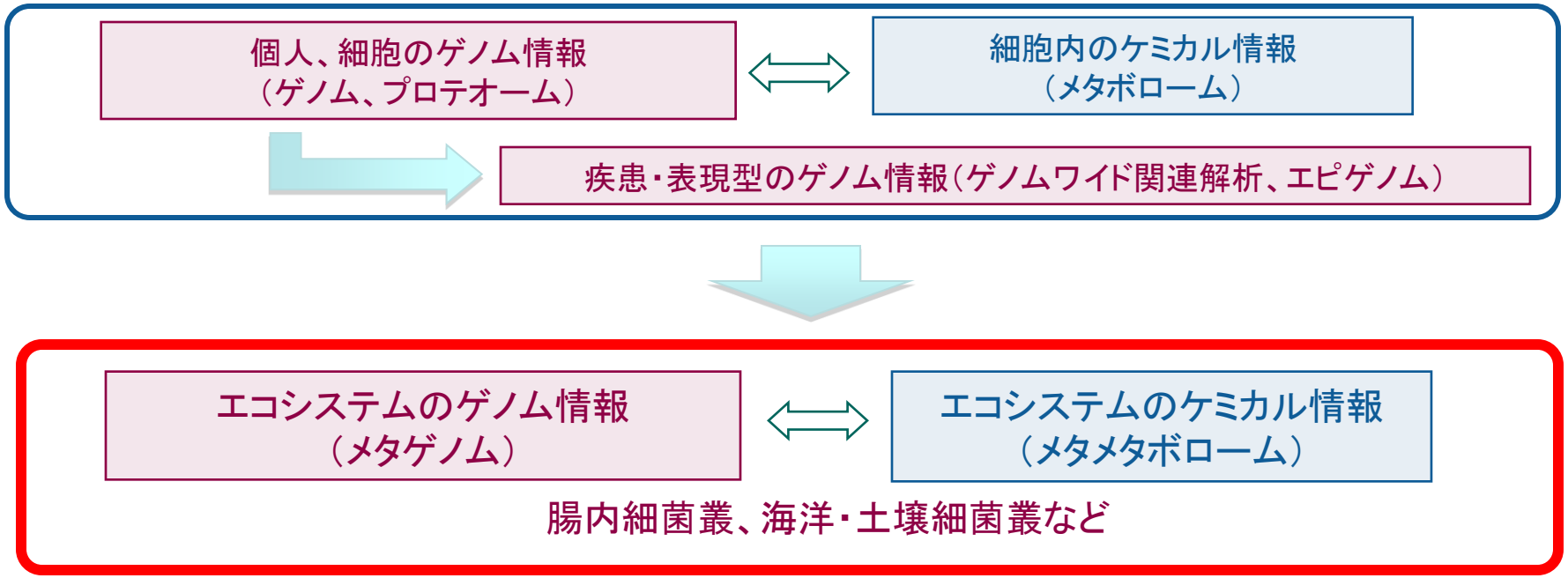
bget: エントリ取得、bfind: キーワード検索

- LinkDBのRDF化
  - 化合物データベースの等価リンクから始めて他のデータベースに拡張する
- RDF化したLinkDBの等価リンク情報を横断検索に応用する
  - ヒットしたエントリのうち同じ化合物・遺伝子の情報をまとめるなど
- 共通で使われているIDなどを利用した等価リンクの自動更新化を検討
- ゲノムネット計算サービスのAPI化
  - ホモロジー検索、モチーフ検索との統合

（1）DBGET/LinkDBシステムの統合利用環境への応用

（2）メタゲノム・メタメタボローム等新規分野データ活用技術の開発

（3）反応オントロジーの整備



- メタゲノムデータ、メタメタゲノムデータ利用技術開発
  - データベース化支援技術：機能アノテーション、パスウェイ再構築
- メタゲノムとメタメタボロームデータの統合化に関わる要素技術開発
  - 反応データによるゲノムとメタボロームの関連付け技術
  - 新規パスウェイ予測のための技術
- 反応オントロジーの整備
  - 遺伝子と反応タイプの関連付けによる反応分類システムの開発

統合化のステップと研究計画