NBDCサイトビジット
2015.8.10
於理研バイオリソースセンター

生命と環境のフェノーム統合データベース 進捗報告

理研バイオリソースセンター
桝屋啓志

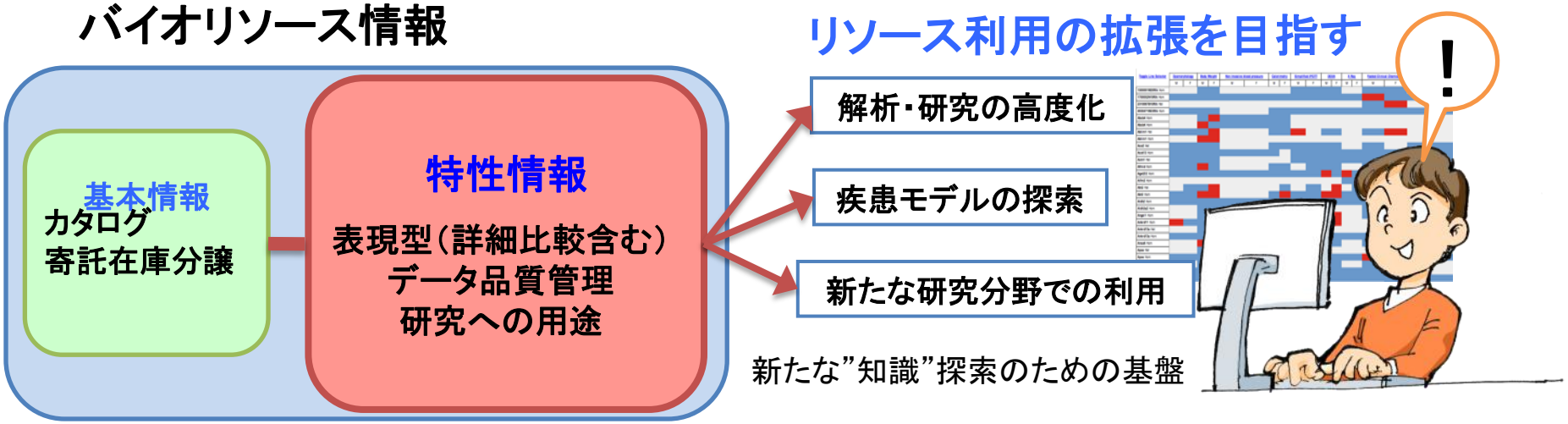
発表内容

- 本ユニットの紹介
- 終了時点における達成目標、通期研究開発計画
- H26年度実施項目、実施状況
- H27年度実施項目
- その他（通期計画に照らした懸念点、想定外の進展等） など

本ユニット（理研BRC・マウス表現型知識化研究開発ユニット）の紹介

理研BRC マウス表現型知識化研究開発ユニット

バイオリソース利用を高めるための情報基盤技術開発



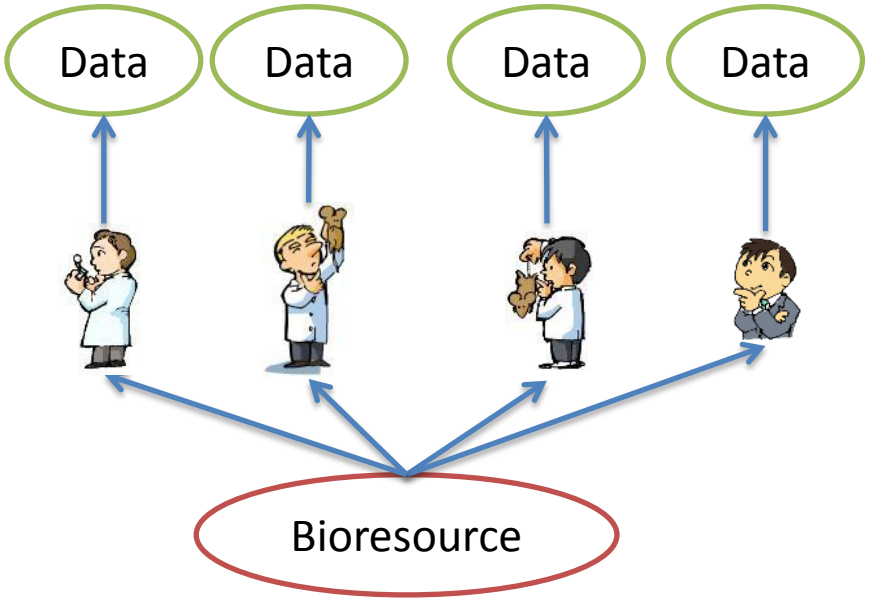
1. データを統合するための、理論的基礎の確立
2. リソース特性情報を統合するデータベースの開発
3. 新たなリソース利用に向けた、情報の国際連携

サイバー空間におけるバイオリソースの役割

生命科学の再現性を担保する最重要基盤

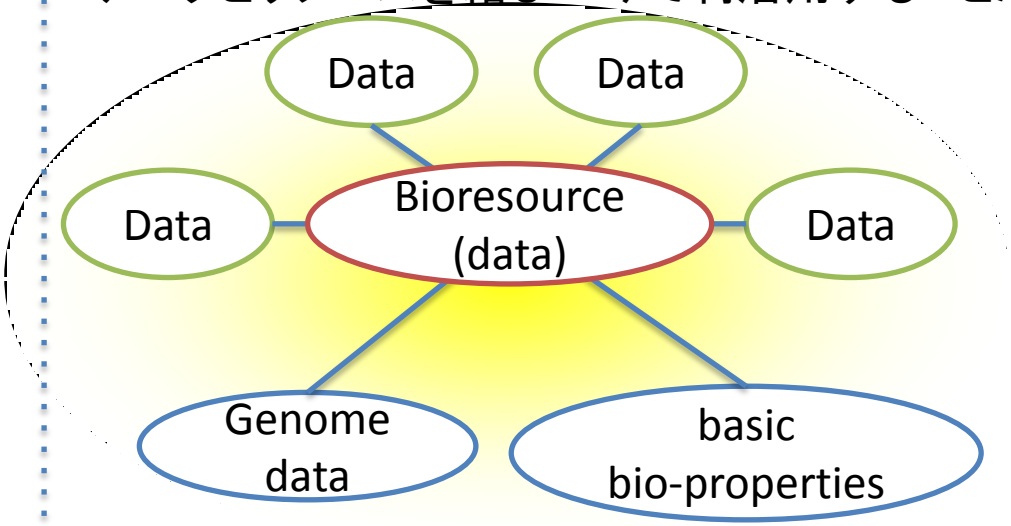
現実世界

バイオリソースの共通利用により
データ間の相互運用性が可能になる



サイバー空間

現実世界と同様に、
データとリソースを結びつけて利活用することが重要



標準的にデータを流通させる基盤が必要

- 標準化された固有ID
- リンク方法
- オープン性
- ソフトウェア横断性

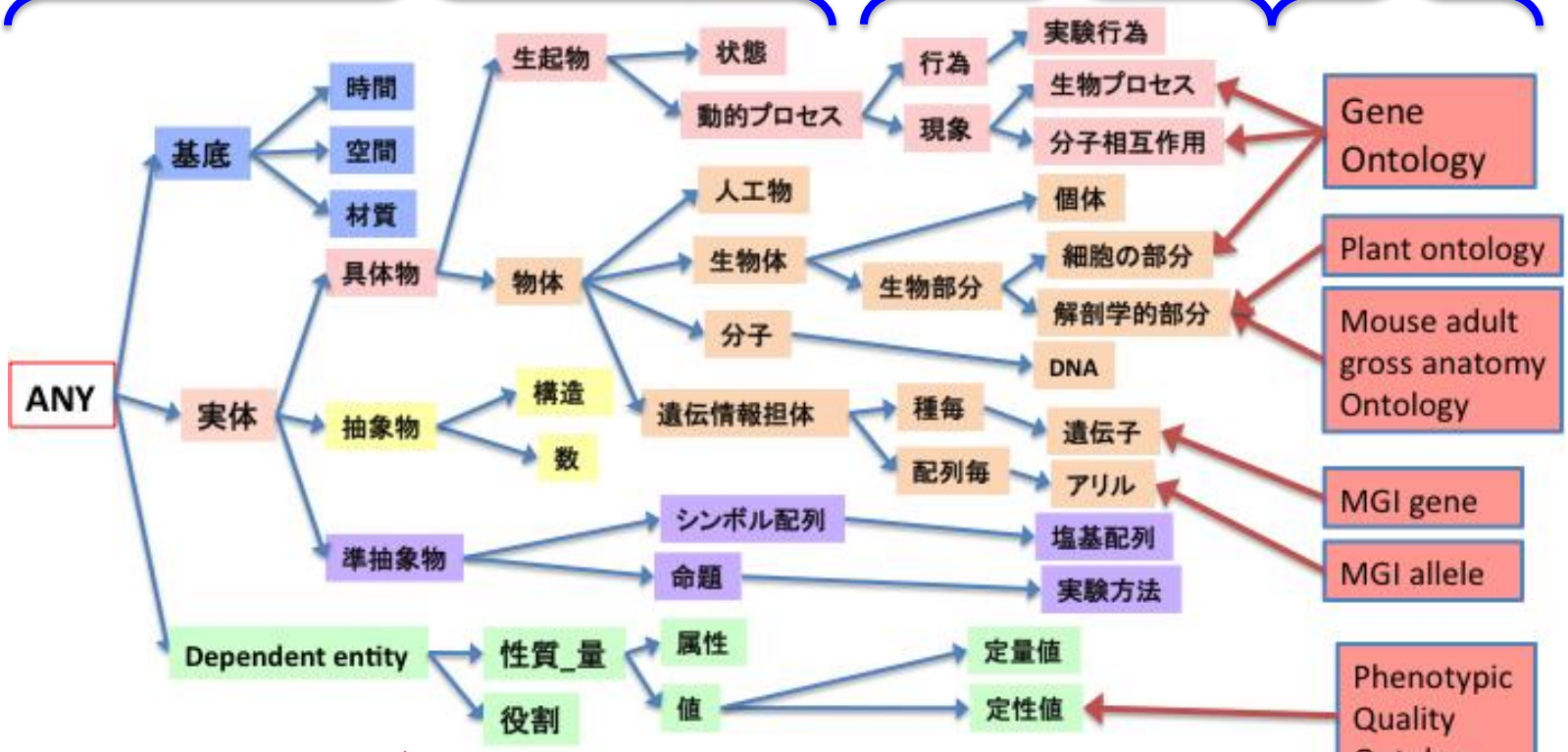
RDF

先行研究：上位オントロジーによる生命科学知識の統合化

上位オントロジー：
概念の普遍的な分類体系

生物学に登場する
概念：

生物分野のデータベース



- **統合データベースを設計するための指標**

Masuya et. al. NAR 2011 D861-870

桝屋・溝口 人工知能学会論文誌 2014 29: 311-327

J. Applied Ontology（投稿中）

本ユニットのアクティビティ

国内、国際コミュニティ

バイオリソースセンター

バイオリソース特性DB
の開発、国際連携

理研のデータの利便性を高める研究開発（理研情報基盤センターとの連携）

理研

バイオインフォマティクス検討委員会
情報基盤センター
RIKEN META DATABASE

オールジャパン・統合DBの作成
（表現型データの分野横断的共有：本課題）

All Japan

NBDC National Bioscience Database Center

DBCLS Database Center for Life Science

DDBJ DNA Data Bank of Japan

jCGGDB KEGG

SSBD Database

Microbe DB.JP

IMPC International Mouse Phenotyping Consortium

MGI

FIMRe

AMMRA

ANRRC

IRDiRC INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

難病情報センター Japan Intractable Diseases Information Center

orphanet

OBO open biomedical ontologies

BIO2RDF

EMBL-EBI RDF Platform

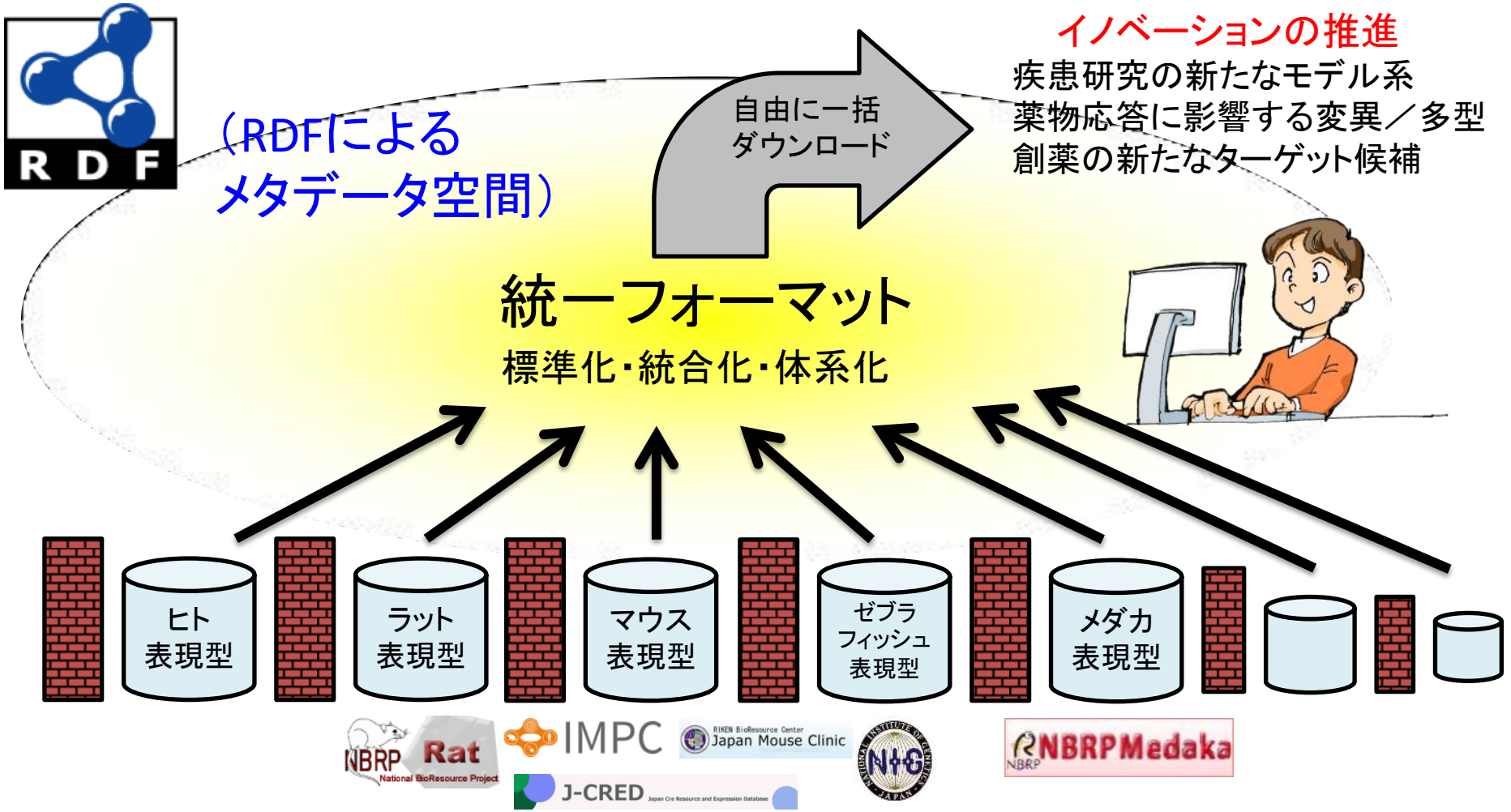
モデル生物研究

疾患研究

終了時点における達成目標、通期研究開発計画

目的

フェノタイプと関連データの情報共有と標準化を、研究分野の垣根を超えて実現する



終了時点における達成目標

	第1期（終了）	第2期
主要な統合対象データ	マウス系統：約5000系統 培養細胞：約3600株 微生物株：約15000株 （理研バイオリソースセンター） シロイヌナズナ変異株 ：3700 件環境資源科学科学研究センター）	【提供データ拡充】提供するフェノーム情報の質と量の拡大 分子表現型： マウスCreドライバー系統　マーカー遺伝子の発現情報：約50系統　（熊本大） ゼブラフィッシュのジーントラップ系統　マーカー遺伝子の発現情報：約500系統　（遺伝研） 単純表現型： マウス網羅的表現型解析データ：約100系統（理研） ラット網羅的表現型解析データ：約1000系統（京大） メダカ表現型：約500系統（基生研） 複合表現型： 疾患オントロジー：約5000疾患（リンク） 国際的な標準病名ICD10：約12000疾患（リンク） NBDC各グループと連携、RDF技術に基づく相互利用を実現
データ収集方法	データベースより直接収集（バイオリソースセンター等）	【データ収集技術】生物学者向け表現型データ入力S開発 データベースからの直接収集に加え、一般ユーザー（ゲノム編集コンソーシアム）からの直接入力システムの開発
可視化	つながり検索、お勧めリソース	【データ活用技術】モデル生物表現型と疾患の関連づけと可視化 疾患研究者が最適な研究モデル系を探せるシステム（SPARQLによるインターフェース開発）

第1期データの利活用性改善

- 第1期に作成したデータについて
  - 第1期に作成したデータに関しては、RDFとして流通させやすい形式に修正する
  - 第2期におけるスキーマ拡張に合わせて、一部のデータの修正を行う。

- データ公開基盤に関して
  - 第1期データの公開基盤だった「サイネス」に関しては、運用コスト、ネイティブRDFとして外部機関との連携のしやすさ等の面から、別の公開基盤を検討する
    （例：TogoDB(DBCLS)、あるいは理研の新たな公開基盤）

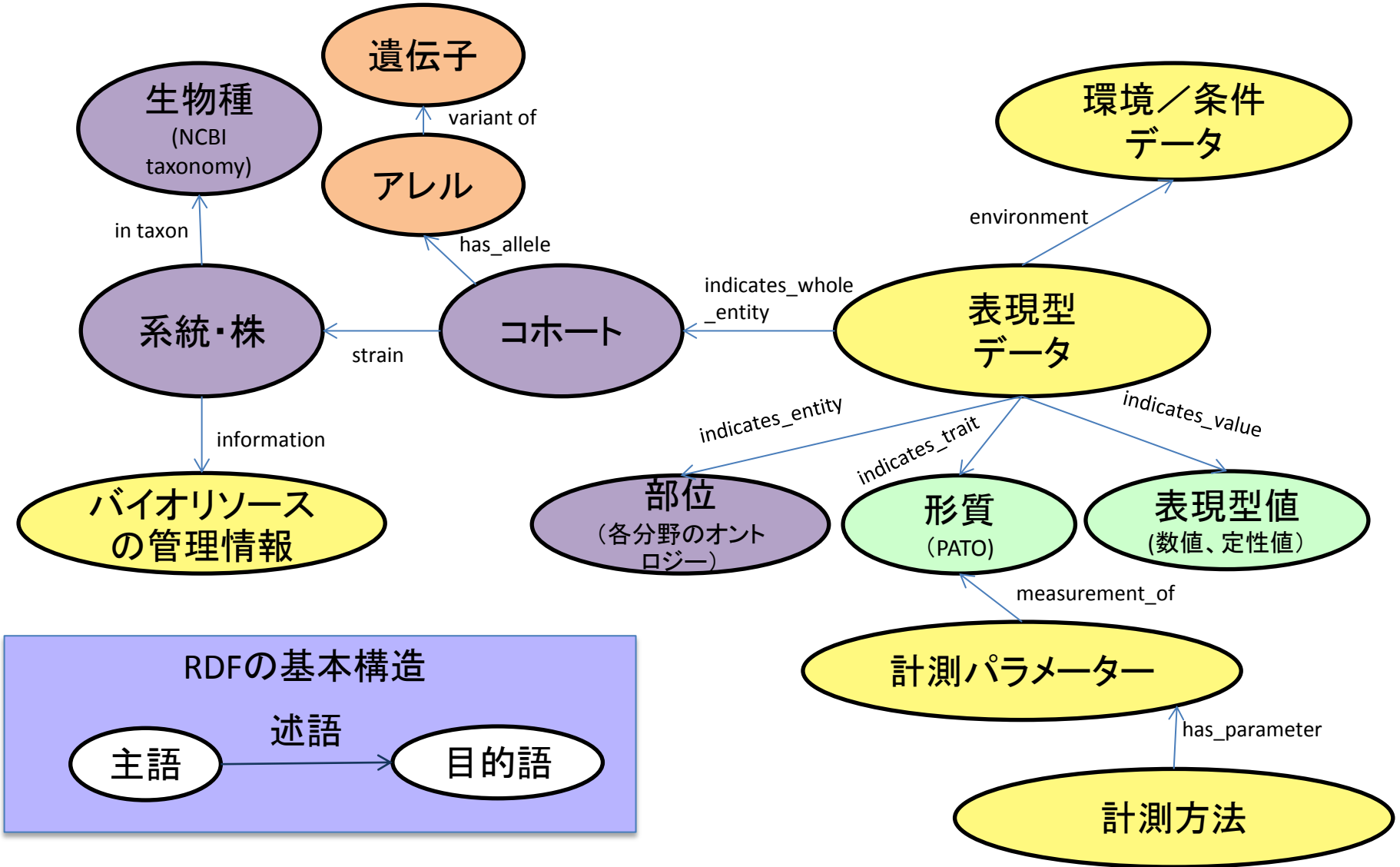
通期研究開発計画

研究開発項目	H26年度	H27年度	H28年度
1. 提供するフェノーム情報の質と量の拡大 ・データ整備ワークフローの確立と、オントロジーの選定 ・データのRDF化と公開 ・分子データとの連結	←→	←→	←→
2. 生物学者向けの表現型データ入力システム開発 ・オントロジーの収集、仕様調査、設計 ・プロトタイプ開発 ・アプリ改変と公開	←→	←→	←→
3. モデル生物表現型と疾患との関連づけと可視化 ・基本設計とオントロジーの収集と検証 ・Webアプリケーションの設計 ・Webアプリ開発とその公開	←→	←→	←→

H26年度実施項目、実施状況

RDFスキーマの見直し

生命の解析データを表現する汎用RDFスキーマ設計



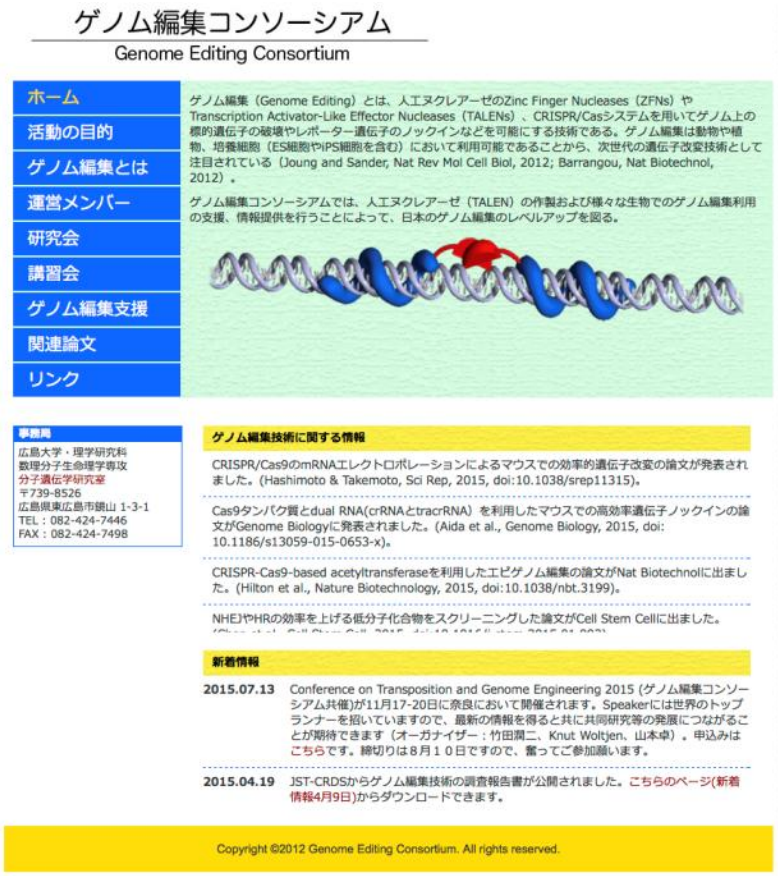
H26年度実施状況：提供データ拡充

収集対象	目標数	データ取得数	備考
マウスCreドライバー系統（熊本大・理研BRC)	50系統	130系統	RDFスキーマ作成 テストデータ作成
ゼブラフィッシュ ジーントラップ系統（遺伝研）	500系統	(500系統）	データスキーマ協議中 （テストデータ作成）
マウス網羅的表現型データ(理研BRC)	100系統	約50コホート （21系統）	RDFスキーマ作成済 国際プロジェクト（IMPC）とのデータ連携を協議
コンソミックマウス網羅的表現型データ（遺伝研）		60コホート (30系統）	RDFスキーマ作成
糖鎖グループマウス表現型		40系統	RDFスキーマ作成
ラット網羅的表現型データ（京大）	1000系統	216コホート (172系統）	RDFスキーマ作成 テストデータ作成
ラット表現型データ（RGD)		約1000系統	RGDよりインポート
メダカ表現型（基生研）	500系統	15系統	RDFスキーマ作成 テストデータ作成 基生研と連携で、オントロジーを用いた表現型アノテーションを開始
BRCマウス系統	（第1期課題）	5,722系統	第1期からのRDFスキーマの見直し テストデータ作成
BRC細胞株	（第1期課題）	3,775株	第1期からのRDFスキーマの見直し テストデータ作成
JCM微生物株	（第1期課題）	14,752株	第1期からのRDFスキーマの見直し 黒川Gとのデータ連携準備 テストデータ作成

H26年度実施状況：データ収集技術

生物学者向け表現型データ入力システム仕様確定

ゲノム編集コミュニティより、編集の結果得られた生物の表現型情報を収集するソフトウェア



ゲノム編集コンソーシアム
代表：広島大学・山本卓

生命科学としての重要性：

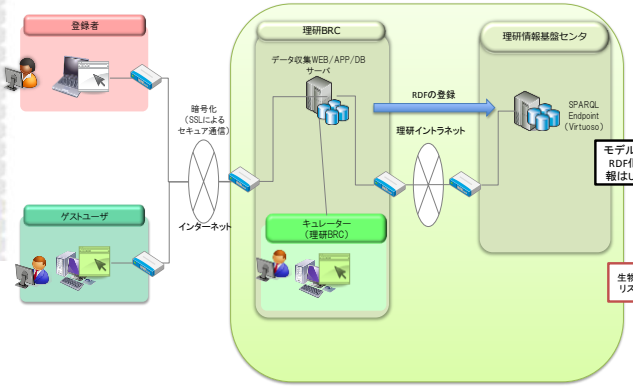
様々な生物種で遺伝子改変系統が爆発的に増加すると予測され、改変内容、結果として得られた表現型の情報共有し、重複努力の防止、再現性確保が課題

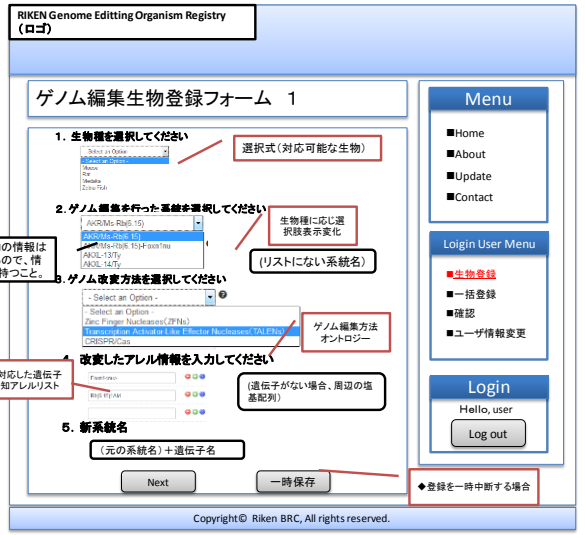
コンソーシアムとしての必要性：

技術および活動アピールとして、データベースが有用

システムの特徴／目標

オントロジー語彙を利用して、ばらつきの少ないデータを容易に入力できること。



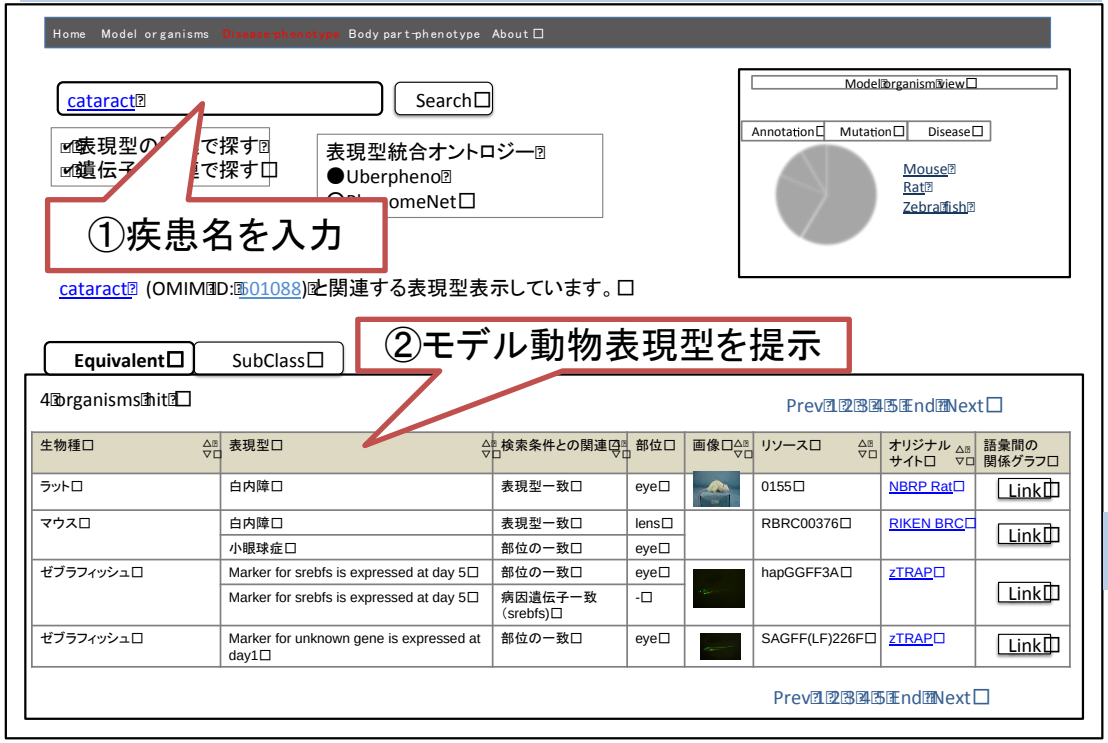


H26年度実施状況：データ活用技術

モデル生物表現型と疾患の関連づけシステム仕様確定

各種生物の表現型データを、特にヒト疾患との関連に注目して可視化する

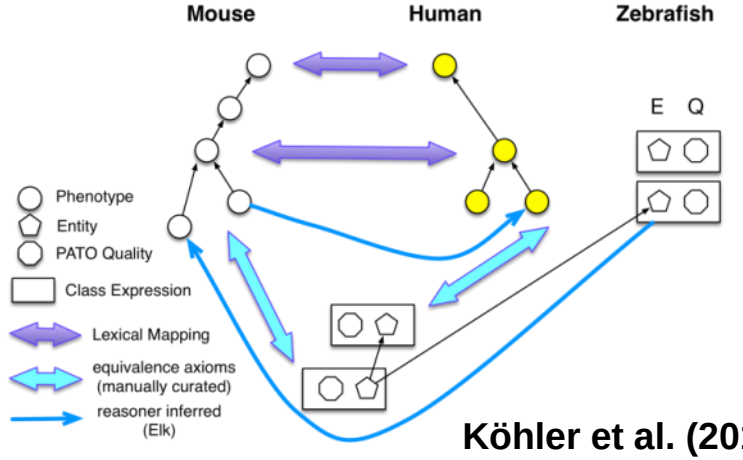
画面設計（一部：白内障の研究モデル動物を探索）



生物種	表現型	検索条件との関連	部位	画像	リソース	オリジナルサイト	語彙間の関係グラフ
ラット	白内障	表現型一致	eye		0155	NBRP Rat	Link
マウス	白内障	表現型一致	lens		RBRC00376	RIKEN BRC	Link
	小眼球症	部位の一致	eye				
ゼブラフィッシュ	Marker for srebfs is expressed at day 5	部位の一致	eye		hapGGFF3A	zTRAP	Link
	Marker for srebfs is expressed at day 5	病因遺伝子一致（srebfs）	-				
ゼブラフィッシュ	Marker for unknown gene is expressed at day1	部位の一致	eye		SAGFF(LF)226F	zTRAP	Link

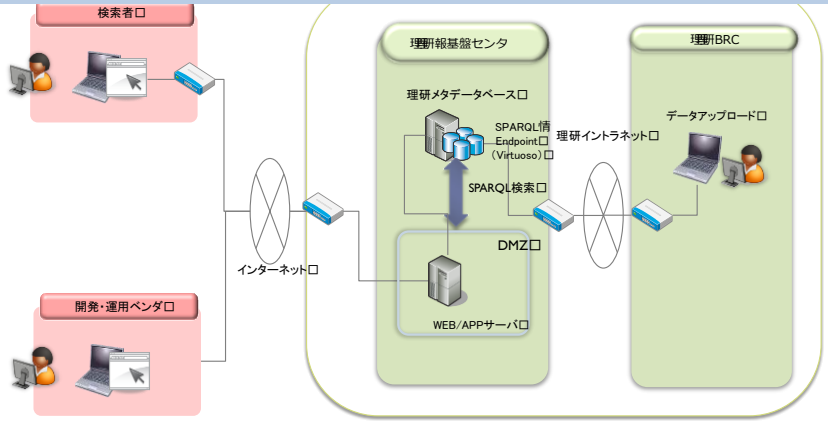
Prev 1 2 3 4 5 End Next

内部設計：オントロジー推論結果をRDF化



Köhler et al. (2014)

全体構成：RDFレポジトリを利用したWebアプリ



H27年度実施項目

進捗状況まとめ

- 【データ収集技術】生物学者向け表現型データ入力システム開発
  - 発注準備中
- 【データ活用技術】モデル生物表現型と疾患の関連づけと可視化
  - 開発中
- 【提供データ拡充】提供するフェノーム情報の質と量の拡大
  - データ公開基盤の選定
  - 各データのRDF化および公開
  - ポータルサイト作成
  - 国際／国内連携の拡大
  - アクセス数

データ公開基盤の選定
理研メタデータベース：http://metadb.riken.jp

理研情報基盤センターが開発、2015年4月より運用

メタデータの世界標準技術であるセマンティックウエブやRDFの枠組みに基づいてライフ系データの公開を支援するデータベース基盤整備

- スプレッドシートからの比較的容易なRDF化
- 表形式、リレーショナルDB的なRDF可視化
- オンントロジーツリー表示
- データダウンロード（RDF, スプレッドシート）
- SPARQLエンドポイント
- DBカタログ、Federated-query（分散型のデータ統合）実現等で、NBDC、DBCLSと連携

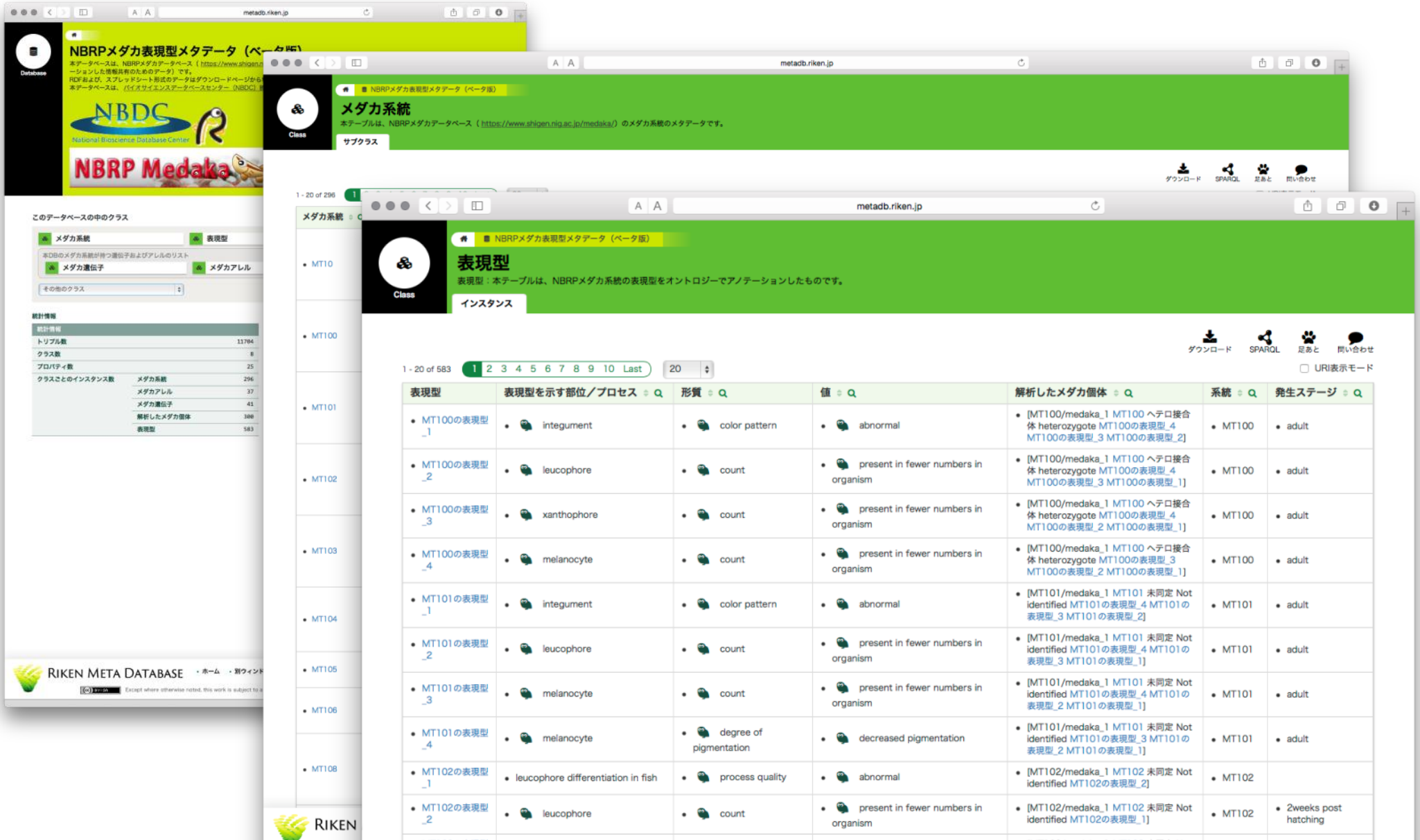
- 生命と環境のフェノーム統合データベース（本課題）の公開基盤として採用
  - 当ユニットが開発に協力
  - バイオリソースセンターが理研データのRDF化（活用しやすいデータ作成）を先導する役割
  - データ更新等、細かな要望をお願いしやすい
- 第1期において、サイネスより公開されていたデータも、ネイティブなRDFに変換し、公開

H27年度実施状況：提供データ拡充

収集対象	目標数	データ取得数	進捗	公開
マウスCreドライバー系統（熊本大・理研BRC)	50系統	130系統	テストデータ作成	
ゼブラフィッシュ ジーントラップ系統（遺伝研）	500系統	(500系統）	データスキーマ協議中 （ゲノムデータ記述方法の検討）	
マウス網羅的表現型データ(理研BRC)	100系統	約1800系統	IMPCデータインポート RDFテストデータ作成中	
コンソミックマウス網羅的表現型データ（遺伝研）		60コホート (30系統）	テストデータ作成中	
糖鎖グループマウス表現型		40系統	テストデータ作成中	
マウスENU誘発突然変異		70系統3000変異	テストデータ作成中 NGSデータスキーマ策定中	
ラット網羅的表現型データ（京大）	1000系統	216コホート (172系統）	テストデータ作成	
ラット表現型データ（RGD)		約1000系統	公開準備中（オントロジーとして）	
メダカ表現型（基生研）	500系統	約300系統	RDFデータ公開 基生研と連携で、オントロジーを用いた表現型アノテーションを続行	□
厚労省指定難病リスト	-	約500語彙	テストオントロジー作成 DBCLSと連携	
BRCマウス系統	（第1期課題）	5,722系統	RDFデータ公開 一部データ見直し中	□
BRC細胞株	（第1期課題）	3,775株	RDFデータ公開 一部データ見直し中	□
JCM微生物株	（第1期課題）	14,752株	MCCVによるRDFデータ公開 MicrobeDB.jp, DBCLSと連携	□

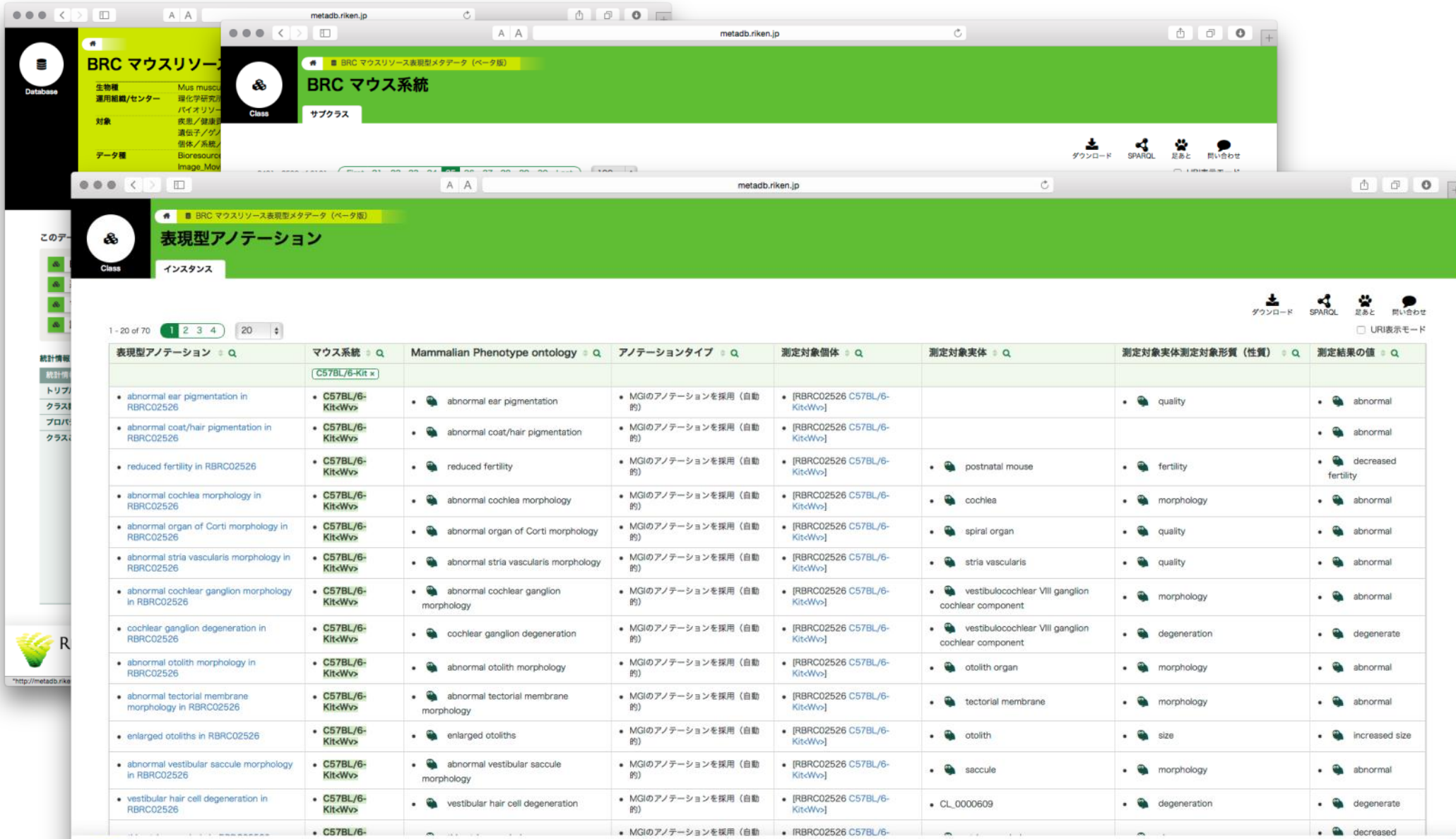
メダカDB

http://metadb.riken.jp/metadb/db/NBRP_medaka



マウスDB

http://metadb.riken.jp/metadb/db/rikenbrc_mouse

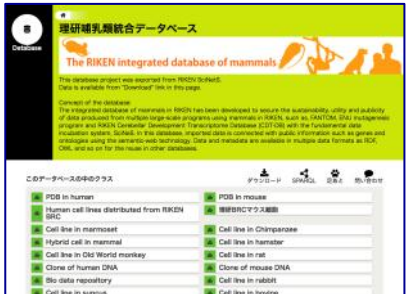


表現型アノテーション	マウス系統	Mammalian Phenotype ontology	アノテーションタイプ	測定対象個体	測定対象実体	測定対象実体測定対象形質（性質）	測定結果の値
	C57BL/6-Kit ×						
abnormal ear pigmentation in RBRC02526	C57BL/6-Kit<Wv>	abnormal ear pigmentation	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]		quality	abnormal
abnormal coat/hair pigmentation in RBRC02526	C57BL/6-Kit<Wv>	abnormal coat/hair pigmentation	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]			abnormal
reduced fertility in RBRC02526	C57BL/6-Kit<Wv>	reduced fertility	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	postnatal mouse	fertility	decreased fertility
abnormal cochlea morphology in RBRC02526	C57BL/6-Kit<Wv>	abnormal cochlea morphology	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	cochlea	morphology	abnormal
abnormal organ of Corti morphology in RBRC02526	C57BL/6-Kit<Wv>	abnormal organ of Corti morphology	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	spiral organ	quality	abnormal
abnormal stria vascularis morphology in RBRC02526	C57BL/6-Kit<Wv>	abnormal stria vascularis morphology	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	stria vascularis	quality	abnormal
abnormal cochlear ganglion morphology in RBRC02526	C57BL/6-Kit<Wv>	abnormal cochlear ganglion morphology	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	vestibulocochlear VIII ganglion cochlear component	morphology	abnormal
cochlear ganglion degeneration in RBRC02526	C57BL/6-Kit<Wv>	cochlear ganglion degeneration	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	vestibulocochlear VIII ganglion cochlear component	degeneration	degenerate
abnormal otolith morphology in RBRC02526	C57BL/6-Kit<Wv>	abnormal otolith morphology	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	otolith organ	morphology	abnormal
abnormal tectorial membrane morphology in RBRC02526	C57BL/6-Kit<Wv>	abnormal tectorial membrane morphology	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	tectorial membrane	morphology	abnormal
enlarged otoliths in RBRC02526	C57BL/6-Kit<Wv>	enlarged otoliths	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	otolith	size	increased size
abnormal vestibular saccule morphology in RBRC02526	C57BL/6-Kit<Wv>	abnormal vestibular saccule morphology	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	saccule	morphology	abnormal
vestibular hair cell degeneration in RBRC02526	C57BL/6-Kit<Wv>	vestibular hair cell degeneration	MGIのアノテーションを採用（自動的）	[RBRC02526 C57BL/6-Kit<Wv>]	CL_0000609	degeneration	degenerate
	C57BL/6-		MGIのアノテーションを採用（自動	[RBRC02526 C57BL/6-			decreased

ポータルサイト：J-phenome

http://jphenome.info

各データベースへのリンク



モデル動物／細胞分野での国際連携

国際マウス表現型解析コンソーシアム
（マウス全遺伝子の網羅的表現型解析）
12カ国18研究機関

IMPCのデータをRDF化を理研でトライすることを合意済み。
全データを定期的インポートの体制確立（現在約1800系統）。
テストデータ作成中

マウスゲノムデータベース
米国ジャクソン研究所

BIO2RDFのデータに更新の問題があるため、
一部データを理研でRDF化することを合意済み
テストデータ作成中。

細胞株オントロジーコンソーシアム
3カ国8機関

細胞株のRDFスキーマ（バージョン1）を策定済み
RDFデータ公開済み

アジアマウス変異体リソースコンソーシアム
8カ国10研究機関

アジアマウスリソースの統合DB化を検討中

アジアリソースセンターネットワーク
15カ国60研究機関

RDFによるリソース情報統合の有用性に関して
NBDCでの経験を含め、レポート化を提案

国内（分野横断的）連携

NBDC統合化推進プログラム

分野横断的に共通利用できるRDFスキーマを策定
測定値（数値、単位）、バイオサンプル、文献等

厚労省指定難病の病名約500語彙をオントロジー化
国際希少疾患コンソーシアムのオントロジー（ORDO）とのリンクを付すことで相互運用性を確保する。国内の各種疾患語彙についても、国際的にも相互運用性の向上を目指す。
人的ネットワーク構築の段階。

NGSメタデータのRDFスキーマを作成中
マウス、微生物データとリンク予定

理研QBiC・発生動態研究チーム

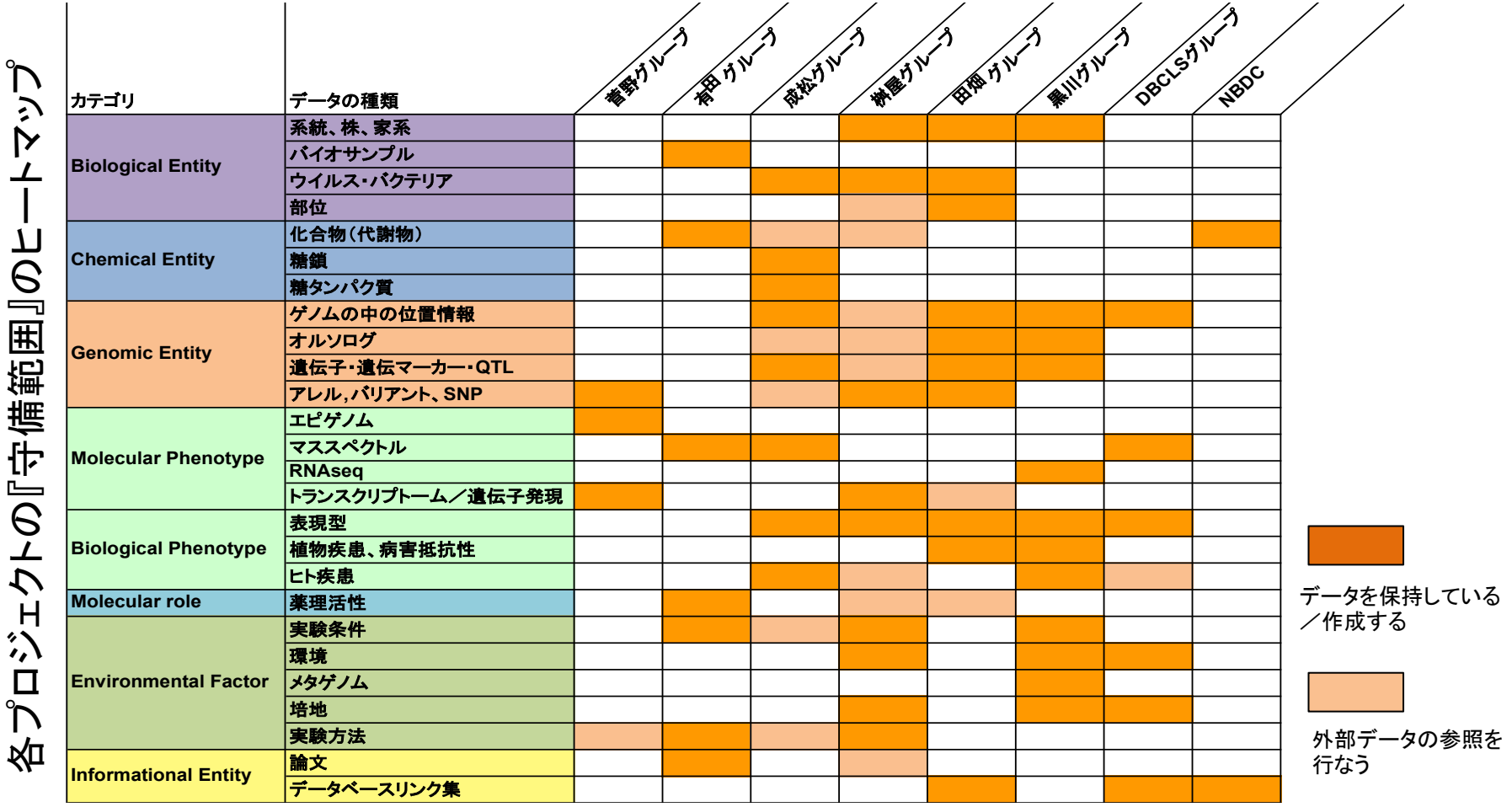
画像RDFメタデータの策定中。生命系で用いられる画像DB「OMERO」との連携へ

微生物リソースに関して、MicrobeDB.jp・DBCLSの提案するRDFスキーマ／オントロジーでデータを作成。本プロジェクトで作成したデータが、そのまま流通し利活用されるネットワークができつつある。
さらに、国際標準化に向け、アジア（アジアリソースセンターネットワーク：ANRRC）、世界（微生物世界データセンター：WDCM）を通じて、働きかけを開始。

統合化推進プログラム（TPP）連携

統合に向けたRDF化支援、データの相互利用へ向けた調整と連携（毎月のSPARQLthonにて調整）

各プロジェクトの『守備範囲』のヒートマップ



桝屋Gは、すべてのデータ種について連携先があり、継続して統合化を行っていく必要がある

次年度計画

- 【提供データ拡充】提供するフェノーム情報の質と量の拡大
  各データのゲノム情報や、NBDC内各プロジェクトとの連結を行い、そのデータをNBDCあるいは、理研のRDFレポジトリから公開する。
  - ゲノムデータ
  - 糖鎖データ
  - パスウェイデータ

- 【データ収集技術】生物学者向けの表現型データ入力システム開発
  今年度開発したプロトタイプアプリケーションの改変と公開、データ収集を開始する。

- 【データ活用技術】モデル生物表現型と疾患との関連づけと可視化
  今年度開発したプロトタイプアプリケーションに基づき、さらなるアプリ開発とその公開を行う

その他（通期計画に照らした懸念点、想定外の進展等） など

- RDF化において、共通語彙／オントロジーの利用は、データの利便性を左右する重要な要素。供与先でのデータ整備（＝アノテーション）が必要になる場合が多い。
  ＝＞キュレーション作業の予算措置を求める声
  - メダカ:
    我々が勧めるスキーマ／オントロジーでのアノテーション作業を実施。
    （国際的にデータ連携を進めたいというインセンティブが働いた）
  - ゼブラフィッシュ:
    より利活用されるためには、さらに細かい粒度の表現型データが必要だが、予算がなく作業できない。
    一方で、コミュニティに役立つデータ統合様式を真剣に求めている。

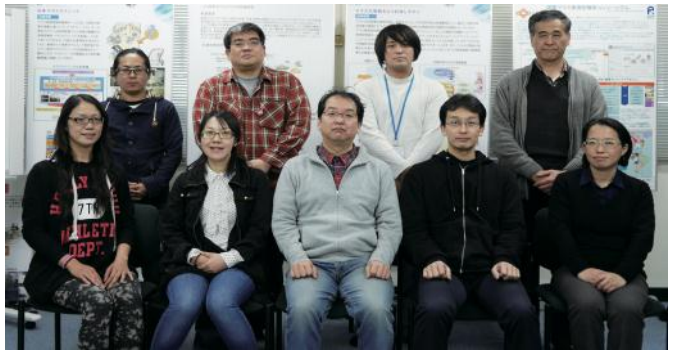
- ゲノム編集データ入力システム:
  一般生物学者個人のデータ登録に対するインセンティブをどれだけ生み出せるか

- 想定外の進展:
  - 微生物情報の国際化
  - 疾患情報をキーとした連携
  - NGSデータとの連結

データ統合と利活用、今後の課題

- モデル動物：疾患研究への利用に向けて
  - 国内開発のモデル動物に向け、さらなる拡充
  - 疾患および、ヒトバリエーション解析データとのさらなる連携、統合状態の可視化
  - パスウェイデータとの連携
- 細胞株：急速に拡大するiPS細胞ニーズへの対応、データ利活用を通した疾患研究の発展
  - 今後拡張される疾患特異的iPS細胞バンク（5000人15,000株）の情報利活用基盤
    - データ種：　核型、未分化性、分化能、分化後細胞、全ゲノムデータ、コントロールとしての健常人由来iPS
    - 公開データ（匿名化済み）の試料提供者が特定されないための、データ利用のレギュレーション
    - ヒトバリエーション、疾患情報、動物モデルとの連携
- 植物フェノーム x 環境（シロイヌナズナ）
- 情報技術面：ありふれた技術の利用拡大と、ビッグデータ利用の最先端研究
  - 世界標準APIであるRDFの特徴を生かし、データ可視化プログラムの共用をさらに進める。（プログラミングの敷居をさらに低める）
  - 効率的なFederated queryの実装、高負荷への対応
  - 巨大RDFデータ（グラフデータ）を用いた大規模関連解析
- 人材育成：個々のデータの確認／バルクで俯瞰の双方ができるセンスとスキル
  - アノテーター（データ作成）：データ生産が行われるところに必要？　あるいは集中化？
  - 開発者（データ可視化）

メンバー・謝辞



理研バイオリソースセンター
マウス表現型知識化研究開発ユニット

斎藤実香子
大島和也
高山英紀
大城望
高月照江
田中信彦
桝屋啓志

理研バイオリソースセンター
吉木淳
中村幸夫
大熊盛也
小幡裕一

理研情報基盤センター
戀津魁
小林紀郎

基礎生物学研究所
バイオリソース研究室
金子裕代
成瀬清

国立遺伝学研究所
哺乳動物遺伝研究室
高田豊行

国立遺伝学研究所
初期発生研究部門
川上浩一

京都大学院医学研究科
庫本高志

理研QBiC・発生動態研究チーム
遠里由佳子
京田耕司
大浪修一

DBCLS
川島秀一
片山 俊明
山本泰智

大阪大学・産業科学研究所
古崎晃司

国立遺伝学研究所
大量遺伝情報研究室
藤澤貴智

NBDC統合化推進プログラムの皆様