



RIKEN MetaDatabase
<https://metadb.riken.jp>



RIKEN Open Life
Science Platform



理研オープンサイエンスポータル

オープンサイエンスを推進する 研究データの知識化開発研究

国立研究開発法人理化学研究所
情報統合本部 基盤研究開発部門
小林 紀郎



基盤研究開発部門

全理研研究センター・グループと連携し、以下の課題を行う

- **研究情報管理基盤**を構築・運用するための技術開発
- 研究手法のデジタル変革(DX)やデータ駆動型科学に必要とされる**研究支援**

研究センターの研究者と密に連携して研究データの集約と利活用促進を図り、理研のオープンサイエンスを推進する。

副部門長

部門長



井原 雅行



美濃 導彦



小林 紀郎

所内研究情報の集約・管理・共有基盤開発実施体制

データ知識化
開発ユニット



小林 紀郎

データ管理
システム
開発ユニット



實本 英之

分野毎の特性に合わせた研究データの集約・公開を推進するユニット群

生命科学データ
共有開発ユニット



大浪 修一

医科学データ
共有開発ユニット



清田 純

オープンサイエンスとは？

以下に掲げる目的を達成するための、情報通信技術を使った新たな科学研究の活動、理念、方向性等を総称してオープンサイエンスといたします

- **研究活動を加速。さらに研究の高度化、多様化に迅速に対応**
 - 国内外での共同研究やコンソーシアム等の連携を推進
 - ビッグデータ解析、データ中心科学の盛隆
 - データ、資料、プログラムの迅速な共有と共同編纂を促進
- **研究成果の社会還元を推進**
 - 学術論文の閲覧無料化（オープンアクセス）
 - 研究データの公開による研究成果の国民への還元
 - 産学間連携、学際的連携、研究者と公衆との対話
- **透明性、再現性のある研究成果の共有**
 - 研究データの公開による研究再現性の担保
 - 研究データや論文根拠データ公開による研究不正の抑止

オープンサイエンスの6原則 (<http://openscienceasap.org/open-science/>より)

オープンソース



ハード・ソフトウェアの再利用

オープンな方法論



研究・教育の手法開発

オープンアクセス



学術論文の閲覧無償化

オープンな査読



研究の質の担保

オープンデータ



研究データの公開

オープンな教材



教材の共有と再利用

単に研究データを公開すればよいのではなく、新たな価値創造につながる研究データの利活用戦略について検討する必要があります。理研はオープンデータの諸課題解決に取り組んでいます。

理研研究データポリシーの策定

- 研究者をメンバーとする委員会にて原案を策定
- 2020年3月に所として承認・公開

1. 目的

1. 国立研究開発法人理化学研究所（以下「研究所」という）は、研究データの取り扱いに関する基本方針を定めることにより、国と時代の要請に応えながら幅広い研究分野で先導的な研究を進めるとともに、**研究者の豊かな発想と成果を尊重しつつ、研究データの利活用促進を通じて研究開発成果の最大化及び社会への還元を図る。**

2. **研究データは、研究者の魂**といえるものであるが、公的資金で運営されている研究所で生成された**国民の知の共有財産**でもある。研究所は**研究データを国益に沿う形で最終的に原則公開**するという理念に基づいて、我が国および世界の科学技術の発展に貢献する。ただし、**個人のプライバシーの観点から保護が必要なデータ、研究契約等で公開に制限があるデータ等、特別な配慮が必要な研究データは公開の対象外とする。**

研究データ

研究所の研究活動を通じて取得、作成されたあらゆるデータ

利活用データ

論文発表等に伴い公開が義務付けられたデータ
利活用により科学技術の発展に貢献できるデータ

公開データ

所内外限定共有
データ

所内限定共有
データ

研究所が管理

- 原則リポジトリに登録
- 公開範囲や期間も登録
- アクセス権付与
- 必要な範囲で**メタデータ**を付与
- 保存期間は10年以上
- 二次利用許諾条件を明示

研究データを、戦略的に公開の範囲を決定して管理
具体的な分野の特性に合わせ、センター毎にガイドラインを定める

メタデータの種類

• カタログデータ

- データセットやプロジェクトを説明するメタデータ (Dublin Core等)
- データベースや研究者を検索

• 研究知識データ

- 実験、機器、材料などの詳細を記述したメタデータ
- 研究分野を問わず、知識の統合、共有、推論を目的とする
- データ解析時にパラメータとして参照

オープンサイエンスに適したFAIRデータとして共有・公開したい
研究DX化のために機械可読なメタデータを流通させたい

FAIR原則

オープンサイエンスに適したデータが満たすべき原則

Findable 発見可能

- F1. (メタ)データがグローバルに一意で永続的な識別子を有している。
- F2. データがメタデータによって十分に記述されている。
- F3. (メタ)データは、検索可能となるように管理・インデックスが作られている。
- F4. メタデータがデータの識別子を持ち、その所在を特定できる。

Accessible アクセス可能

- A1 (メタ)データに、標準化された通信プロトコルを経由し、識別子を用いてアクセスできる。
 - A1.1 プロトコルは公開されており、無料で、普遍的に実装可能である。
 - A1.2 プロトコルは、必要に応じて認証やアクセス権の付与が可能である。
- A2 データが利用できなくなっても、メタデータにはアクセス可能である。

Interoperable 相互運用可能

- I1. (メタ)データは、知識記述のために形式が定まり、アクセス可能であり、共有され、広く適用可能な言語で記述される。
- I2. (メタ)データの記述にはAIR原則に従う語彙を用いる。
- I3. (メタ)データは、他の(メタ)データを適切に参照する。

Re-usable 再利用可能

- R1. (メタ)データは、十分かつ正確な属性を持つ。
 - R1.1. (メタ)データは、明確でアクセス可能なデータ利用ライセンスとともに公開される。
 - R1.2. (メタ)データは、その来歴と関連づけられている。
 - R1.3. (メタ)データが分野毎のコミュニティの標準を満足する。

理研データベース電話帳

RIKEN Database Directory

List of databases

This is a catalog of databases operated in the RIKEN institute. You can filter this table by clicking icons of magnifying glass at columns.

Instance

Download SPARQL History Inquiry Show URI

1 - 20 of 123 1 2 3 4 5 6 7 20

List of databases	title	alternative	homepage	Metadata link	publisher	description	language	species	theme	dataType	reference
FANTOM4	Functional Annotation of the Mammalian Genome 4	FANTOM4	http://fantom.gsc.riken.jp/4/	http://metadb.riken.jp/metadb/db/SciNetS_ria187i	Center for Life Science Technologies Chief Scientist Laboratories RIKEN	RIKEN has designed to understanding how these components work together	eng jpn	Homo sapiens	DNA clones cell experimental procedures gene or	GeneExpression Image_Movie Interaction_Pathway Sequence	19377474 19377475 19377478 20211142
Systems and Structural Biology Center	Systems and Structural Biology Center	SSBC	http://metadb.riken.jp/metadb/db/SciNetS_ria46i	http://metadb.riken.jp/metadb/db/SciNetS_ria46i	Advanced Center for Computing and Communication RIKEN	Imported from SciNetS	eng		DNA clones experimental procedures protein		
Bacpedia	Bacpedia beta in Spring8-Center	Bacpedia	http://bacpedia.harima.riken.jp/bacteria/View/Frm/Contents/topview.aspx	http://metadb.riken.jp/metadb/db/SciNetS_rib220i	Advanced Center for Computing and Communication RIKEN	Available methods for the studies of protein 3-D structures in bacteria	eng		experimental procedures protein		23633602
RIKEN BRC Cell Bank	RIKEN BRC Cell Bank	Cell search system	http://www2.brc.riken.jp/lab/cell/search.php	http://metadb.riken.jp/metadb/db/rikenbrc_cell	BioResource Center RIKEN						
RIKEN BRC Mouse Catalog	RIKEN BRC Mouse Catalog	Animal Search System	http://www2.brc.riken.jp/lab/animal/search.php	http://metadb.riken.jp/metadb/db/rikenbrc_mouse	BioResource Center RIKEN						

DB名 URL +メタデータURI 説明

生物種, 対象, データ型, など.

保護された通信 | <https://www.w3.org/TR/hcls-dataset/>

W3C Interest Group Note

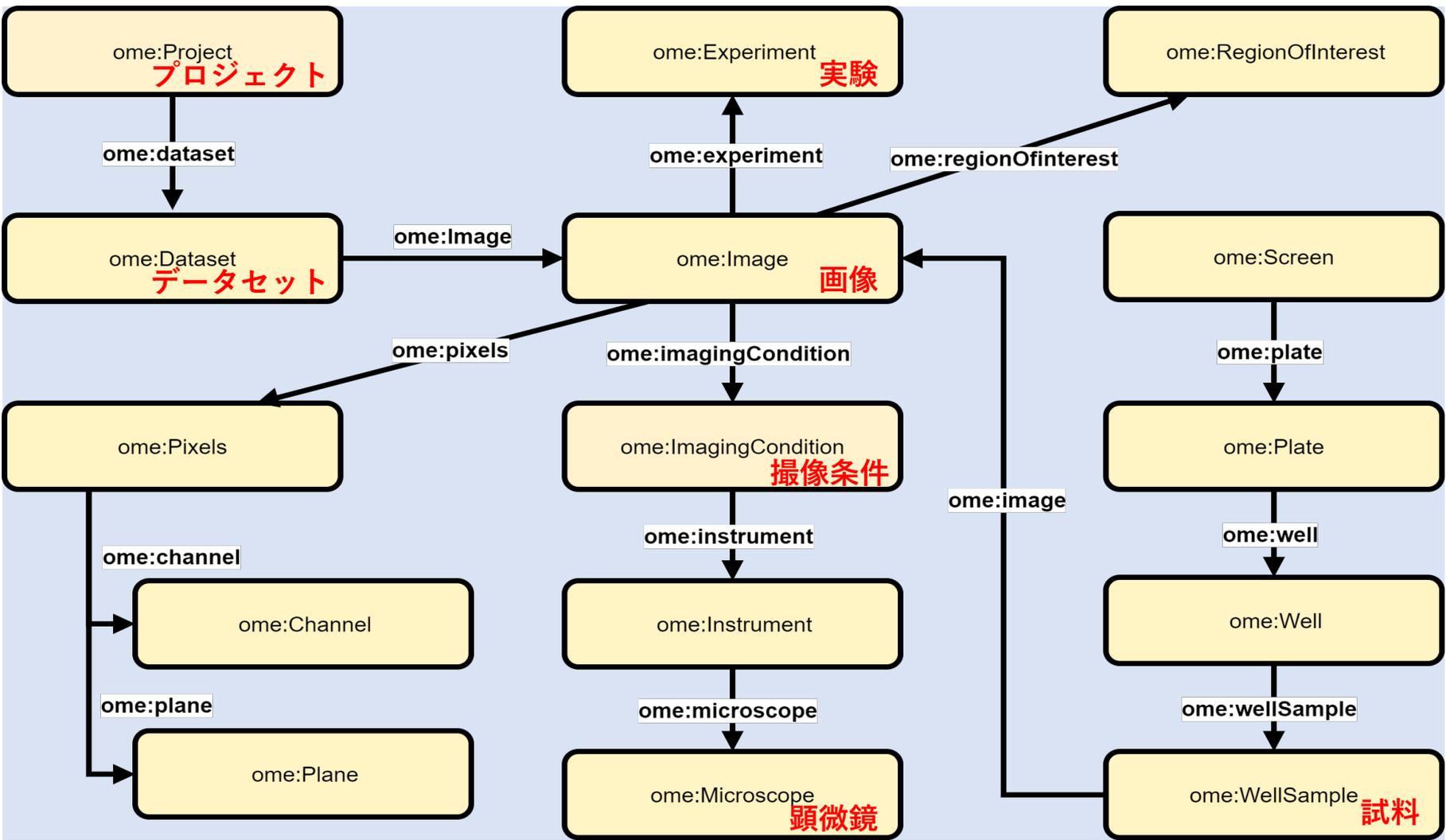
Dataset Descriptions: HCLS Community Profile

[W3C Interest Group Note 14 May 2015](#)

- 理研が公開しているデータベースのリスト
 - 日本標準ライフサイエンスデータベースカタログ “Integbio”
 - “W3C Database Description: HCLS Community Profile”
 - 汎用のデータセットメタデータ schema.org それぞれに準拠
- さらに、JPCOAR等の国内外の仕様や指針に順次対応していく

研究知識データの例 OME-OWL (OME-Core-201606)

光学顕微鏡撮像についてのデータスキーマ

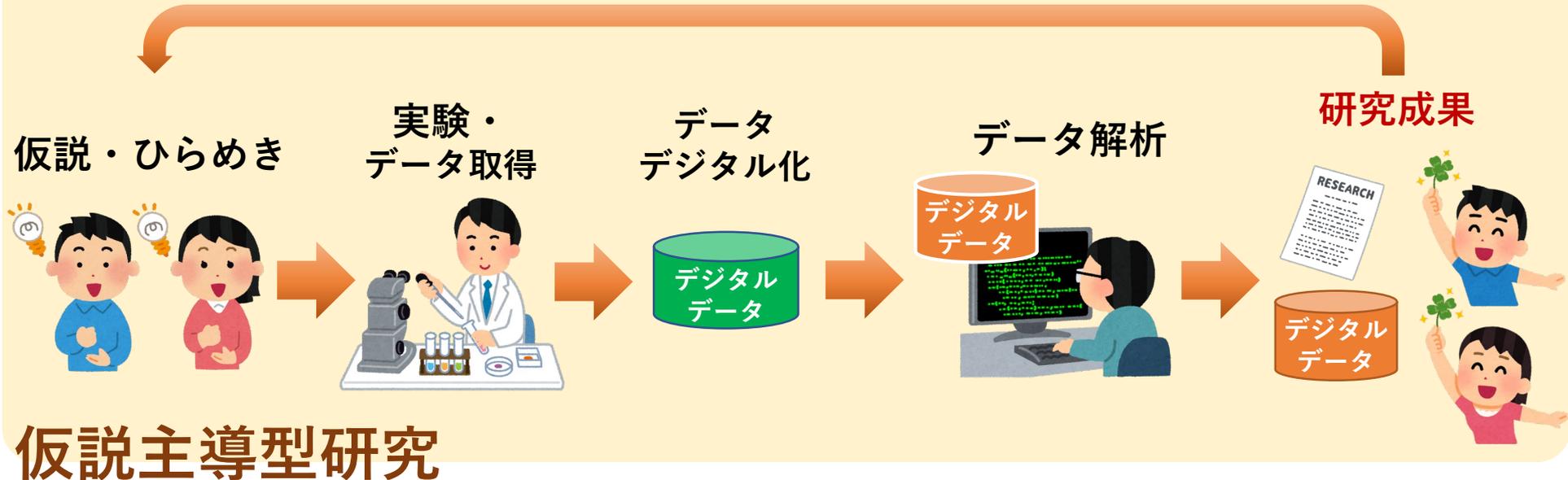


130の概念と 210のプロパティを定義

https://gitlab.com/openmicroscopy/incubator/ome-owl/-/tree/master/ontology/owl/ome_core

仮説主導型研究の研究DX化

結果の検証・仮説の再検討



仮説主導型研究の研究DX化

DX化で研究加速・研究領域拡大: 研究者への強力な後押し

結果の検証・仮説の再検討

仮説
ひらめき

実験・
デジタル
データ取得

データ管理基盤上で
収集・統合

統合データ
解析

(想定外を含む)
研究成果最大化
イノベーション

統合

デジタル
データ

公共・共有
デジタル
データ

オープン
データ

デジタル
データ

デジタル
データ

オープン
データ

検索・取得・参照

BIG DATA

研究成果データ公開

共有データ
膨大な公共データ

研究DX化

部門の役割

- ・オープンサイエンスに適したデータ収集・統合・利活用推進
- ・データ管理基盤の構築

高度化したAIによる解析や探索
実験の自動化・遠隔化
仮説の自動生成

データ駆動型研究

解析結果解釈
実験による検証

分野専門家



情報学的アプローチ
知の探索



部門の役割

- ・オープンサイエンスに適したデータ収集・統合・利活用推進
- ・データ管理基盤の構築



統合データ
解析



(想定外を含む)
研究成果最大化
イノベーション



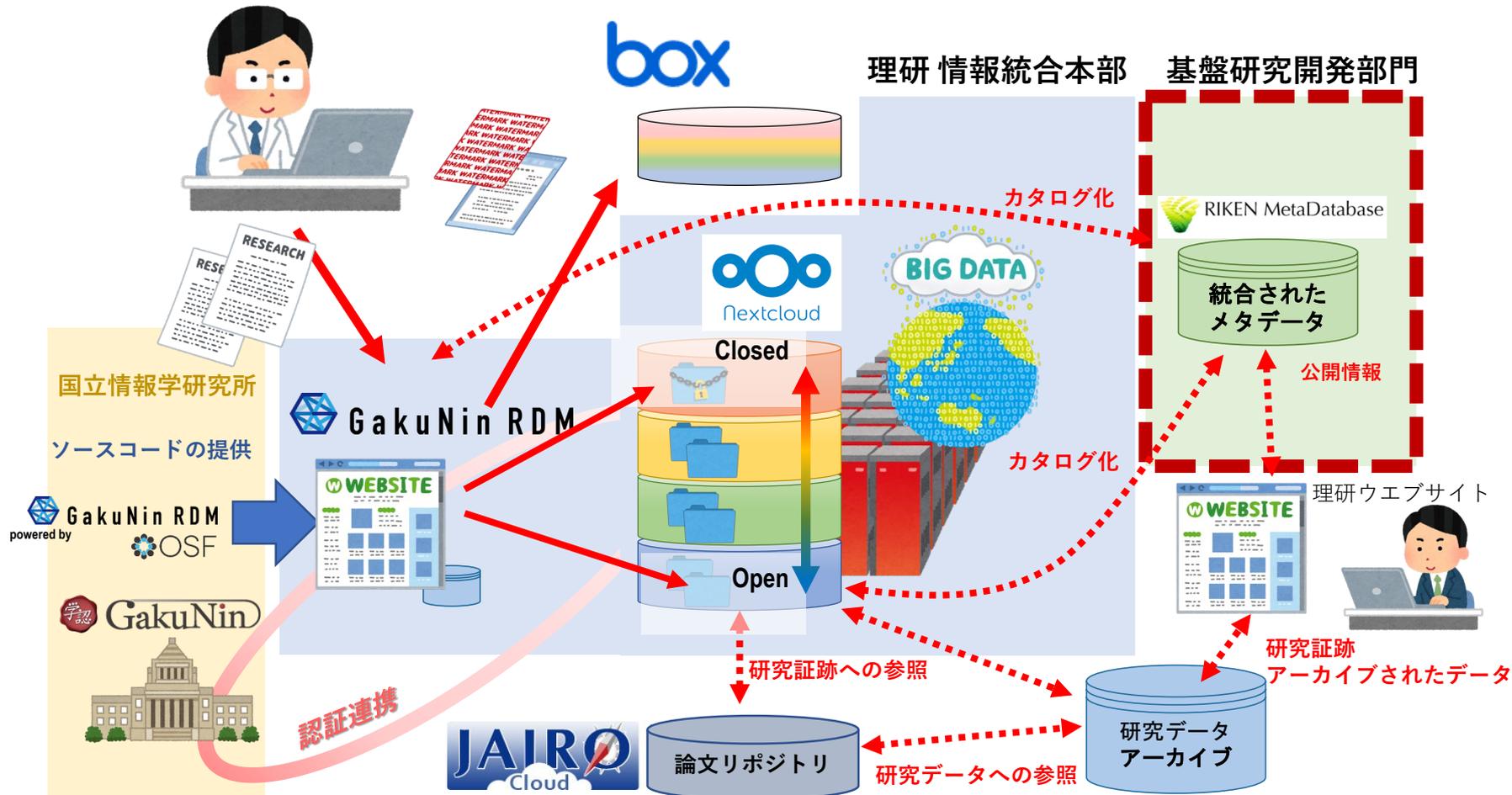
研究成果データ公開

共有データ
膨大な公共データ

研究DX化

高度化したAIによる解析や探索
埋もれた知識の抽出
専門家への「気づき」を提供

理研研究情報管理基盤 (理研オープンサイエンス基盤)



データ産出時から公開まで すべての研究データ管理プロセス をサポート
オープン・アンド・クローズド戦略 に基づいたデータの公開、限定共有

セマンティックウェブ技術で知識化を実現

メタデータをウェブ上で流通させる技術

メタデータとは、データを説明するデータ

Resource Description Framework (RDF)

機械可読なメタデータをウェブ上で記述する標準技術

オントロジー

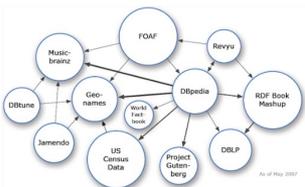
ある領域の概念と、概念間の関係の集合として知識を記述

Linked Open Data (LOD) や Knowledge Graph (KG) とも呼ばれる

生命科学はセマンティックウェブの利用が最も進んでいる分野

LODクラウド = グローバル知識

生命科学LOD

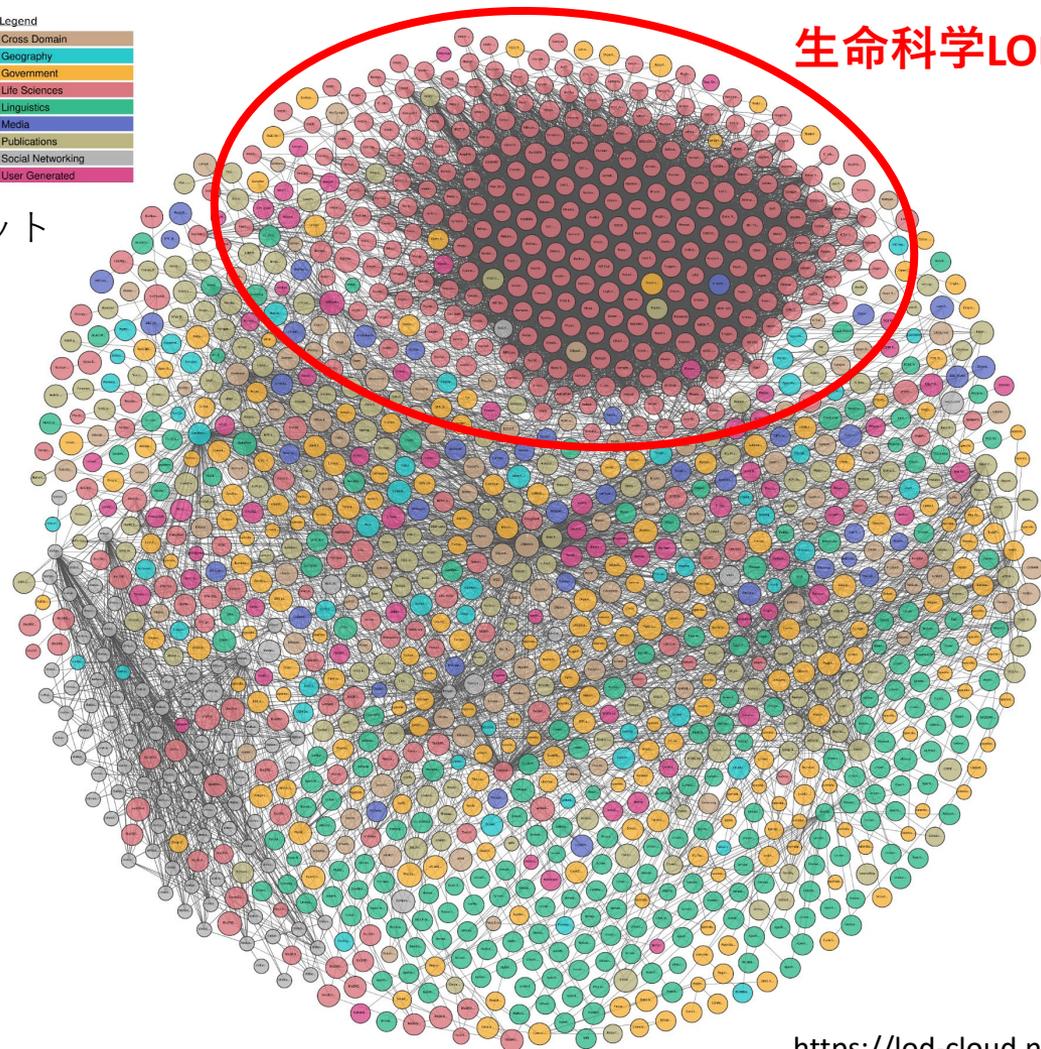


12 データセット
(2007年5月)



1,301 データセット
(2021年5月)

- Legend
- Cross Domain
 - Geography
 - Government
 - Life Sciences
 - Linguistics
 - Media
 - Publications
 - Social Networking
 - User Generated



BioPortal
(生命科学オントロジーリポジトリ)

オントロジー
概念(クラス) **934**
13,395,455

<https://bioportal.bioontology.org/>
 (2021年11月8日現在)

BioPortal Statistics	
Ontologies	934
Classes	13,395,455
Properties	36,286
Mappings	55,648,584

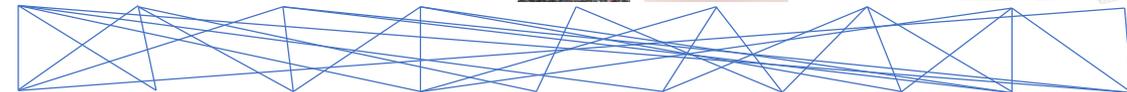
<https://lod-cloud.net>

生命科学データは大規模で互いに深く連携している

理研メタデータベース:

多様な生命科学データの解析、統合、公開を推進

<https://metadb.riken.jp>



理研はこれまで研究成果を個別のデータベースを構築して公開してきた



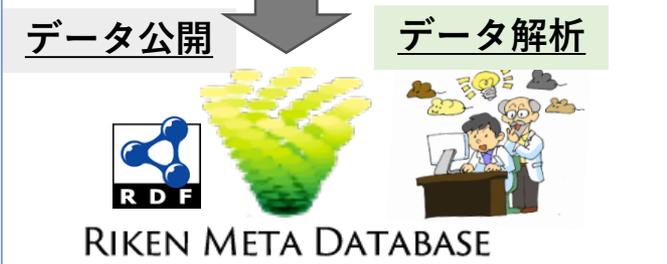
オントロジーを使ってデータ作成、再編、統合



RIKEN META DATABASE

- 世界標準のセマンティックウェブに基づいたデータ管理、公開基盤
- 機械可読な知識を提供

ウェブ上で様々な研究分野のデータを統合



30の公共オントロジーによるデータ記述
64の理研DBと72の外部DBを統合(計 136DB)



クラス (概念) 数	3,902
インスタンス数	35,877,237
RDFトリプル数	188,975,025
プロパティ数	1,904
月間ユーザ数	6,315
月間アクセス数	1,024,244

データベースを跨ぐ相互参照数

なるべく他のデータベースとリンクするよう、人手でメタデータの構造を定義している。

データベース名	相互参照している	
	データベース数	オントロジー数
Metadata of BRC mouse resources and phenotypes	13	14
Metadata of BRC cell resources	13	10
NIG Mouse Phenotype Database Metadata	13	10
Metadata of NBRP Rat	12	13
Metadata of Functional Glycomics with KO mice database (beta)	12	10
RIKEN Database Directory	12	5
bioresource_schema	10	12
NBRP Medaka Phenotype Metadata	8	10
Metadata of JCM resources	8	9
Metadata of quantitative data and datasets of microscopy images provided from SSBD database	8	9



データベース間連携が密であり、統合の効果が大きい

メタデータ構造定義時に、自動的に参照可能なデータを検索、推論する支援があるとよい

理研オープンライフサイエンスプラットフォーム

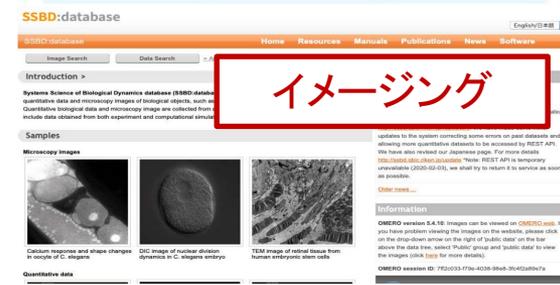
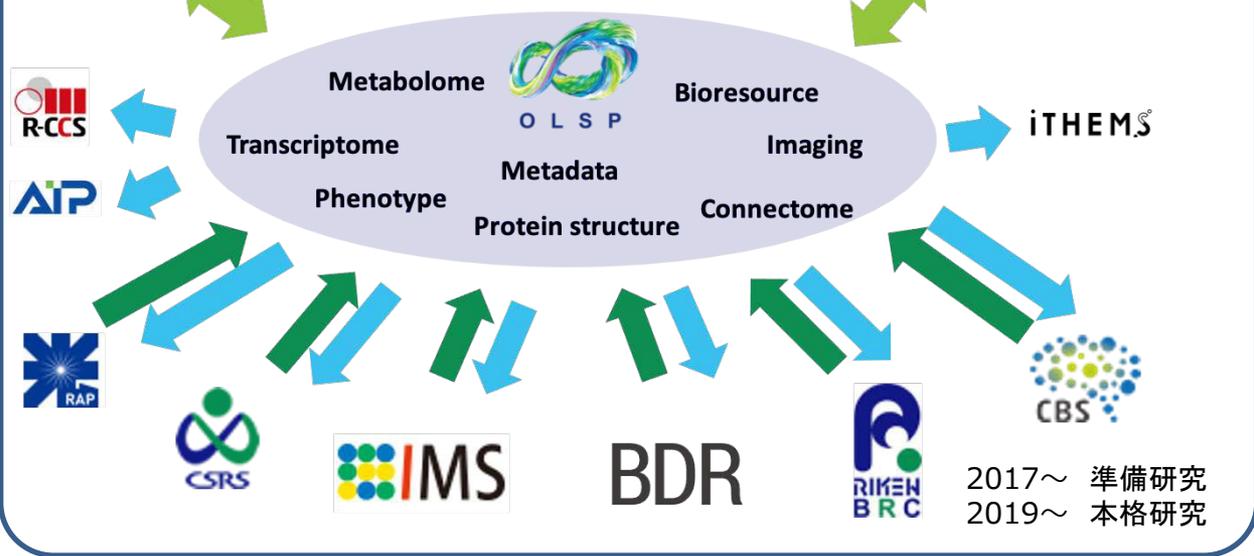
Open Life Science Platform (OLSP)

- 理研が生成する最先端の生命科学データを統合し、世界と共有
- オープンサイエンスに必要な技術や仕組みを創出
- 世界でオープンサイエンスを先導

理研の生命科学を世界に公開
データの再利用を推進

大学
研究所
企業

他分野のオープン
サイエンス基盤



2020年9月 生命科学データ共有開発ユニット
としての活動を開始

生命科学データ共有開発ユニットの活動

最重点目標: 世界においてオープンライフサイエンスを理研が牽引

- 各拠点の公開データベースの統合
 - **メタデータ**レベルでの連携・統合
 - **オントロジー**の指針作り、コンサルタント
- 生命科学分野データベースの国際連携の促進
 - 既存の国際連携のさらなる加速
 - 新たな計測機器への対応
- 各拠点の未公開データの活用
 - **メタデータ**をデータ産出時に取得する仕組みの整備
- オープンサイエンスによる成果創出の促進
 - 研究センターとの研究面での連携強化
 - 公募プロジェクトによるユニット(OLSP)活動の普及
- 人材育成
 - データサイエンティストの育成

研究知識化のカギ メタデータ

メタデータ(特に研究知識データ)の実用化研究は発展途上

- オントロジー(語彙)選択やスキーマ設計の難しさ
 - ドメイン研究者が知識を持っている
 - 情報研究者がドメイン研究者と共同でメタデータスキーマを設計
 - データの利活用を考慮した設計
- メタデータ作成のコストが高い
 - 計測機器がメタデータを作成
 - データ産出時にメタデータを作成する研究フローを作る
 - メタデータの質担保にはデータの意味が分かるキュレータの存在が大きい

解決に向けて

- より多くのメタデータを作成し、設計パターン集を作成
 - メタデータ設計、知識統合を加速
- メタデータの標準化を推進
 - よく知られたスキーマを再利用
 - 国際コンソーシアム等に参加し自ら策定する