



文献や研究成果の 検索基盤を利用した発見

大波純一 (国立情報学研究所オープンサイエンス基盤研究センター)

令和4年7月28日



自己紹介

■大学・大学院

生物学:動物の遺伝子解析

新卒後:IT企業で製薬企業向けシステム開発

生物系、情報工学系の技術と知識を習得。 オープンサイエンスの重要性を認識。 公共データベースの現場へ。



- ■2012年~2020年3月 科学技術振興機構バイオサイエンスデータベースセンター研究員:
 - ・バイオ系研究データの横断検索システムの開発
 - ・バイオインフォマティクスを用いたゲノム研究
 - ・オープンデータの推進活動

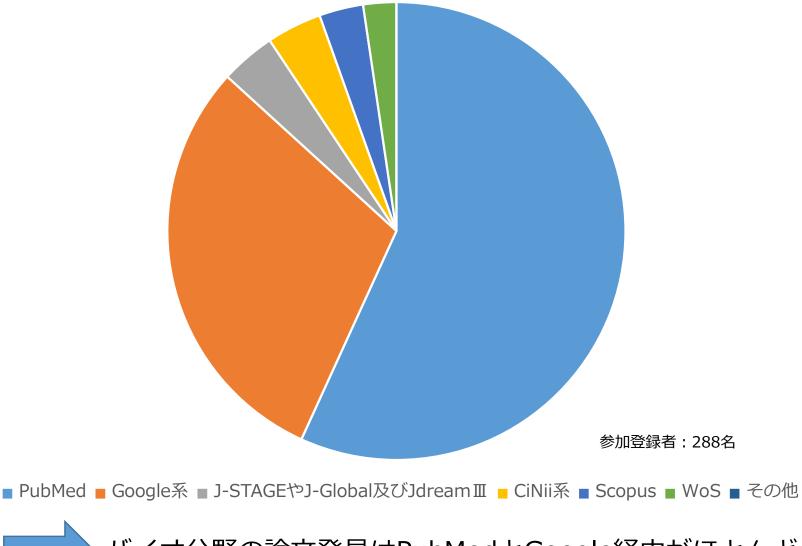


- ■2020年4月~(現在) 国立情報学研究所オープンサイエンス基盤研究センター:
 - ・新しい検索基盤、CiNii Research(サイニィリサーチ)担当。



今回のAJACSオンライン11アンケート結果

【講義1-1】あなたが最もよく使う文献検索サービスを教えて下さい。(複数回答可)







学術検索基盤の選択

Q:バイオの研究者は学術検索基盤を どのように使い分けるべきか?

	推奨する検索基盤		
バイオ分野	国際誌	論文	Pub Med.gov
全分野	国内誌	論文	JSTAGE
全分野	国際&国内誌	論文	Google Scholar
全分野	国内誌を中心として、関連する国際誌も	論文を中心と して関連する データ、書誌 も	CINII Research



本日の内容

- 1. CiNii Research開発の背景
- 2. 国立情報学研究所における開発
- 3. CiNii Research内部のデータ
- 4. CiNii Researchで研究データを発見する



1. CiNii Research開発の背景



学術情報を取り巻く近年の状況

課題:複雑化している学術情報をどう探すか?

学術情報の膨大化

※学術情報=学術論文・学術図書・学術データなど



例) イヌとオオカミの遺伝情報解析の論文

Supplementaryとして 12個体分のSRA Data

198.3GB

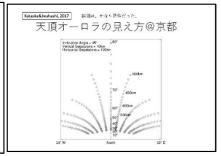
Axelsson E. et. al. "The genomic signature of dog domestication reveals adaptation to a starch-rich diet" Nature. 2013 Mar 21;495(7441):360-4. doi: 10.1038/nature11837. http://www.ncbi.nlm.nih.gov/pubmed/23354050

学際研究の重要性

(専門分野以外の情報へのアクセス)







片岡龍峰、オーロラと人間社会の過去・現在・未来 https://www.rois.ac.jp/sympo/2017/pdf/presen_aurora_180226.pdf

研究のグローバル化

(インターネット経由での海外からのアクセス)



国・地域ごとの学術情報発見

欧州: OpenAIRE



https://explore.openaire.eu/

英国: CORE



https://core.ac.uk/

米国: Semantic Scholar



https://www.semanticscholar.org/

韓国:dataon



https://dataon.kisti.re.kr/#section1

米国: PubMed



https://pubmed.ncbi.nlm.nih.gov/

独国:BASE



https://www.base-search.net/



分野ごとの情報発見

遺伝子配列: DDBJ



https://www.ddbj.nig.ac.jp/ddbj/index.html

※件数は2021年12月の公開統計情報

生命科学:データベース横断検索



https://dbsearch.biosciencedbc.jp/

※件数は2022年2月時点

物質材料学:MDR



https://mdr.nims.go.jp/ ※件数は2022年2月時点

人文社会科学: JDCat



https://jdcat.jsps.go.jp/ ※件数は2022年2月時点



企業が提供する学術発見

Google Scholar



http://scholar.google.com/

※件数は2014年の推計(http://arxiv.org/abs/1407.6239)

Mendeley



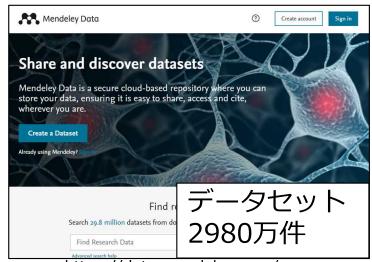
https://www.mendeley.com/ ※件数は2022年2月時点

Google Dataset Search



https://datasetsearch.research.google.com/ ※件数は2020年のKaggleデータより算出

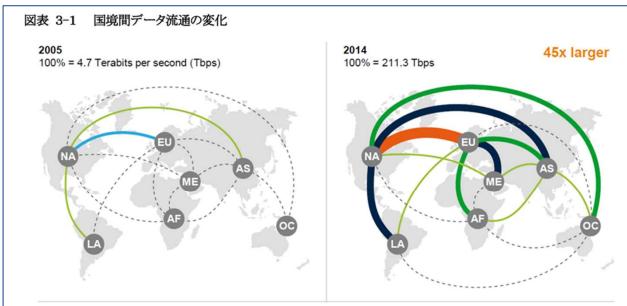
Mendeley Data



https://data.mendeley.com/ ※件数は2022年2月時点



データ流通量の増加と日本の立ち位置



──米国 ──イギリス ──ドイツ ──日本 個人のデータ提供に 対する理解度 個人の企業に パーソナルデータ 対する安心感 活用サービスに係る (情報加工等) 個人の利用意向 企業のデータ 利活用意欲 利活用意欲 (産業データ) (パーソナルデータ) (出典)パーソナルデータ提供等に係る消費者向け国際アンケート調査、 第4次産業革命への取組状況およびデータ流通・利活用に関する国際企業アンケート

NA:北米、EU:欧州、AS:アジア、ME:中東、AF:アフリカ、OC:オセアニア 黒色点線: <50Gbps、黒色実線:50 Gbps -100Gbps、緑色実線(細):100-500Gbps、青色実線(細):500-1000Gbps、緑色実線(太):1,000Gbps-5,000Gbps、青色実線(太):5,000 Gbps -20,000Gbps、オレンジ実線(太):20,000Gbps 以上

(出典)Mckinsey Global Institute DIGITAL GLOBALIZATION: THE NEW ERA OF GLOBAL FLOWS

ネット上の通信帯域の総量は 9年で45倍に。 米国・英国・独国と比較して、 日本のデータ利用・提供に関する 意欲は低レベル

安心・安全なデータ流通・利活用に関する調査研究の請負報告書 2017年3月 株式会社三菱総合研究所 https://www.soumu.go.jp/johotsusintokei/linkdata/h29_02_houkoku.pdf

データ流通・促進環境に関する国際比較



政策的な背景

第6期科学技術・イノベーション基本計画(2021年3月26日閣議決定) p.59-60より 【科学技術・イノベーション政策において目指す主要な<mark>数値目標</mark>】(主要指標)

機関リポジトリを有する全ての大学・大学共同利用機関法人・国立研究開発法人において、2025年までに、データポリシーの策定率が100%になる。

公募型の研究資金162の新規公募分において、2023 年度までに、データマネジメントプラン(DMP)及びこれと連動したメタデータの付与を行う仕組みの導入率が100%になる。



研究者の研究データ管理・利活用を促進するため、例えば、データ・キュレーター、<u>図書館職員</u>、URA、研究の第一線から退いたシニア人材、企業等において研究関連業務に携わってきた人材、自らの研究活動に資する場合にはポスドク等の参画や、<u>図書館のデジタル転換</u>等の取組について、<u>2022 年度までにその方向性を定める</u>。

【科技、文、関係府省】

「公的資金による研究データ管理・利活用(Percent Center for Open Science and Date Platform 関する基本的な考え方」の体制

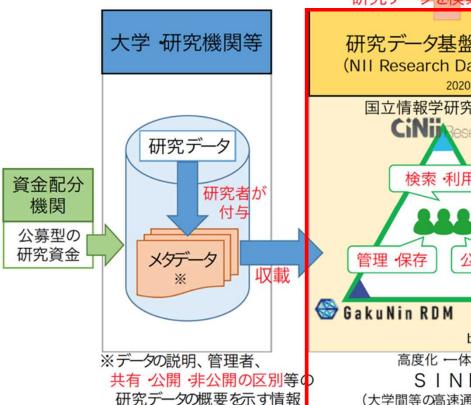
公的資金による研究データの管理・利活用に関する基本的な考え方について

研究データ基盤システムを中核としたデータ・プラットフォームの構築

- ○研究データの公開・共有を推進、産学官のユーザがデータを検索可能
- 〇ムーンショット型研究開発制度における試行(2020年度開始)、その後、次期SIPに導入
 - ➡ 全ての公募型の研究資金の新規公募分に導入(2023年度まで)

アカデミア、産業界等

研究データを検索・利活用







令和3年4月27日 内閣府, 統合イノベーション戦略推進会議 https://www8.cao.go.jp/cstp/tougosenryaku/9kai/siryo3-1.pdf



2. 国立情報学研究所における開発



国立情報学研究所が提供する学術コンテンツ

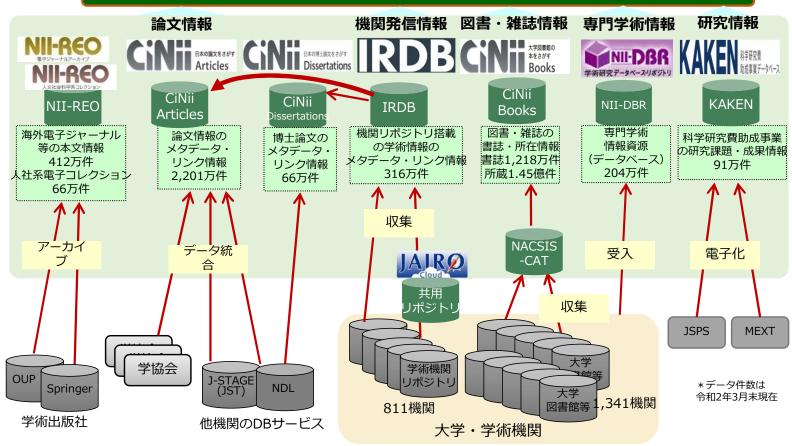
NIIの中で既に学術情報発見のためのサービスが多数存在



学術コンテンツ基盤:学術コンテンツの確保と提供

提供する党体で

NIIの提供する学術コンテンツ





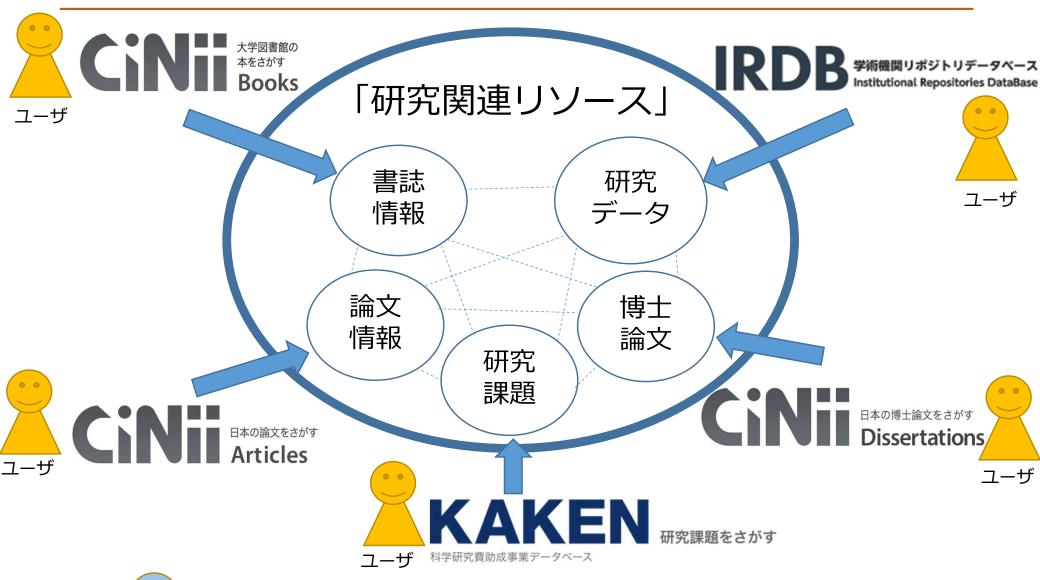
CiNii Research開発の背景

NIIの次世代サービスとして、学術情報発見のために、 「何ができて」「NIIは何をしたいか」を 徐々に明確にしていった。

- 2017年度
 - 学術情報発見のための次世代CiNiiとして計画始動
 - 「グラフデータベースによる知識基盤」CiNii Research(仮)
 - オープンサイエンス検索基盤、NII統合検索基盤としての役割
 - グラフDB、インフラ、フロントエンドのプロトタイプ構築
- 2018年度
 - ORCID認証、学認連携のプロトタイプ
 - データソース(CiA、CiB、CiD)取り込み
- 2019年度
 - 名寄せ高度化、外部機関DBとのリンク
- 2020年度
 - プレ版環境リリース
- 2021年度
 - 本番リリース



既存の知識基盤環境

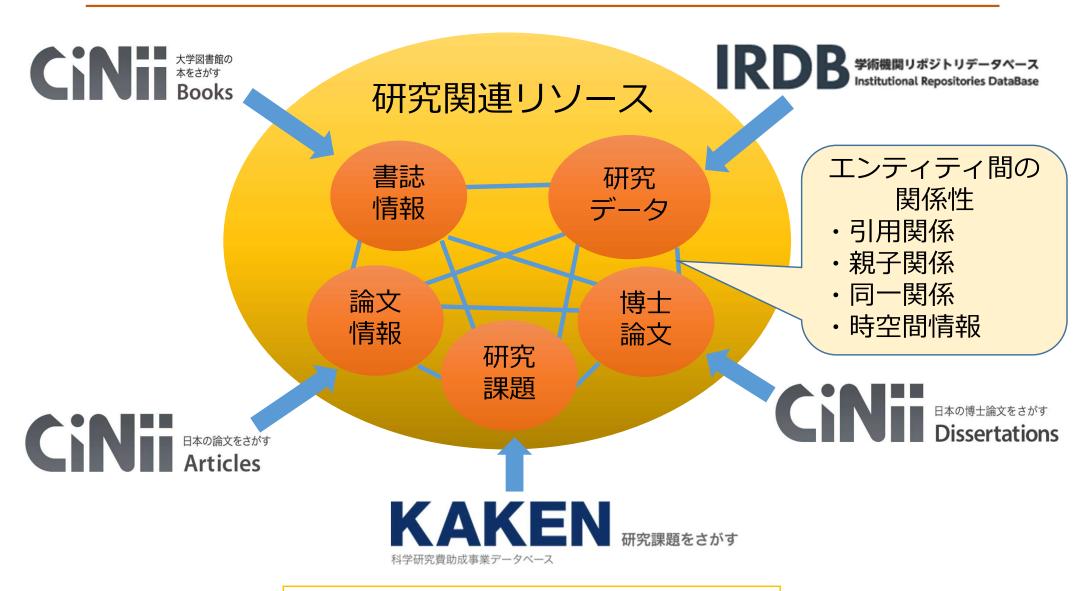




- ・別々の基盤から、関連データへの部分へのアクセスのみ
- ・基盤を跨いだ情報間の関係性が見えづらい・・・



研究関連リソースへのアクセスと関係性

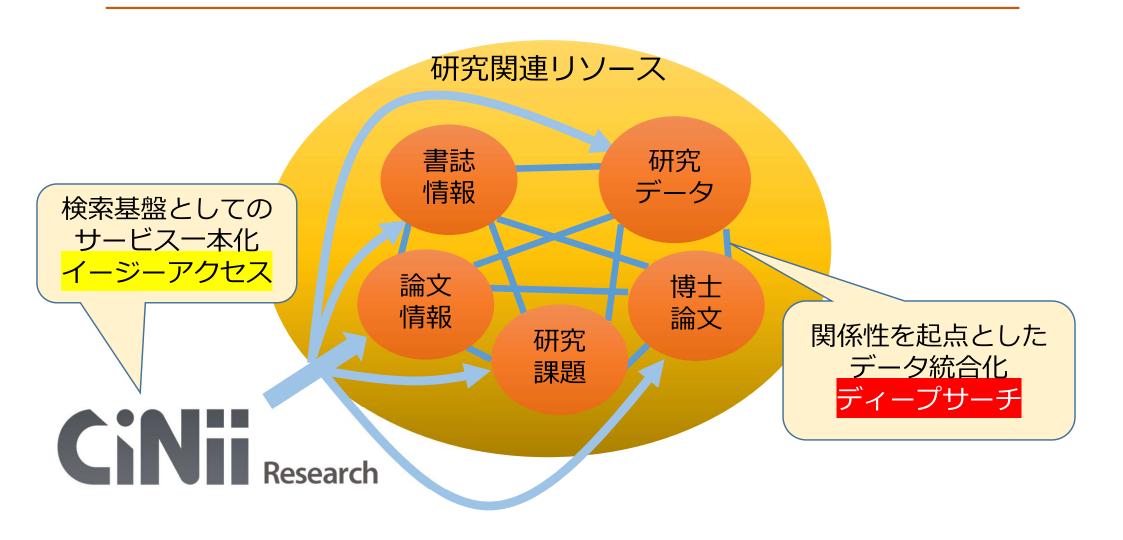




知識発見基盤としての一本化 エンティティの関係性を介した統合化



CiNii Researchが解決する研究リソース検索





2つの要件を満たす、新しい検索基盤として、 CiNii Researchを開発



CiNii Researchの基本動作

• 基本動作

- 2000万件超の研究データ、論文、書誌情報、プロジェクトデータの
 シンプルな横断検索(=イージーアクセス)
- 単体ページにおける、関連データ・関連文献・関連プロジェクト・関係するID等の表示(=ディープサーチ)



・想定ユーザと利用法

- 研究者 → 論文検索で研究データやプロジェクトデータも取得
- ・図書館員 → 書誌や論文などの研究成果を包括的に管理
- 調査担当者 → プロジェクトや組織ごとの研究成果を集計



研究データ基盤: NII Research Data Cloud

2017年から開発開始 ⇒ 2021年から運用開始

ISNI

外部ID発行サービス

- データ収集装置や解析用計算機とも 連携
- 研究遂行中の研究データなどを共同 研究者間やラボ内で共有・管理
- 組織が提供するストレージに接続し た利用が可能

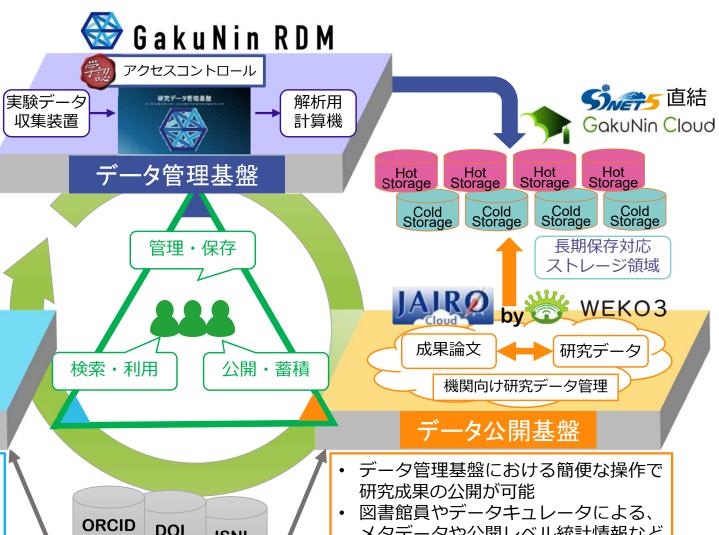
分野別

リポジトリ

Research

海外の 研究データ

公開基盤



データ検索基盤

メタデータ集約・管理

知識ベースの構築

- 機関リポジトリ+分野別リポジトリや データリポジトリとも連携
- 研究者や機関、研究プロジェクトの情 報と関連付けた知識ベースを形成
- 研究者による発見プロセスをサポート

メタデータや公開レベル統計情報など の管理機能の提供



3. CiNii Research内部のデータ



学術検索基盤 CINII Research

2021年4月より正式公開された、国立情報学研究所の新サービス一つの検索画面から多様な学術情報へ、イージーアクセス



山中伸弥先生に、人生とiPS細胞について聞いてみた



平尾誠二と山中伸弥「最後の一年」

山中, 伸弥, 平尾, 誠二, 平尾, 惠子 講談社 2017

🚺 新型コロナウイルス感染症(COVID-19)が心理臨床業務に与えた影響について

...このデータは、新型コロナウイルス感染症(COVID-19)が心理臨床業務に与えた影響について実 施されたwebアンケートの結果である。実施時期は、緊急事態宣言が出された直後の2020年5月17日から6月16日までで、有効回答者は465名である。...

COVID-19/SARS-CoV-2関連のプレプリントに関する分散表現データセット— 2020年05月17日版—

等 小柴, 和弘 林 2020-06

データ本体への アクセスリンク

Putative mechanism of hypotensive action of platelet-activating factor in dogs

山中, 伸弥 大阪市立大学 博士 (医学) 甲第1162号 1993-03-24 博士号情報 NDLデジタルコレクション DOI Web Site ほか1件

研究プロジェク

■ BRG1によるクロマチン構造変化を介した細胞初期化促進メカニズムの解明 京都大学 2011 - 2013 (科研費)

勇 再プログラム化によるヒト体細胞幹細胞化のための基盤技術開発 多田 高 京都大学 2007 - 2010 (科研費)

ファンディング情報



CiNii Researchに含まれるデータ件数

3000万件以上の研究データ、プロジェクト、論文、書誌情報を搭載

2021年5月時点のCiNii Research検索エントリ数

論文	9,258,151
研究データ	191,988
本	12,737,134
博士論文	673,510
その他成果物	6,552,789
プロジェクト	967,919
計	30,381,491

これらのデータは、NIIが以前から公開していた学術サービス、



等のデータをそのまま搭載した・・・わけではありません。

JPCOARスキーマをベースとした 検索フィールド定義の設定



J P C O A R

オープンアクセスリポジトリ推進協会

JPCOARスキーマガイドライン JPCOAR Schema Guidelines スキーマ説明 改訂履歴 FAQ 問い合わせ 関連サイト

ホーム

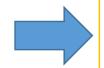
JPCOARスキーマは、オープンアクセスリポジトリ推進協会(JPCOAR)が策定した新しいメタデータ規格です。日本の機関リポジトリのメタデータの国際的な相互運用性を向上させ、日本の学術的成果の円滑な流通を図ることを目的としています。

このウェブサイトでは、JPCOARスキーマの説明(ガイドライン)やFAQ(よくあるご質問)といった、JPCOARスキーマに関する各種情報を提供しています。

XMLスキーマ定義と項目一覧 (Excel、PDF) は、https://github.com/JPCOAR/schema ☑ をご参照ください。

https://schema.irdb.nii.ac.jp/





メタデータの標準化 国際的な相互運用性の向上



ID付与状況

データソース種別	データ種別	全件	DOI	NAID	NCID	CRID	
CIB	Book	12310091	1888	0	12310091	0	
CID	Dissertation	669422	188589	669422	0	0	
CROSSREF	Article	2916310	2916310	2102753	0	0	
CROSSREF	その他成果物	10017294	9790061	259233	55	0 - 1 //	
DATACITE	Dataset	1	1	0	0	合計約	1900万件
DATACITE	その他成果物	1	1	0	0		
DBPEDIA	Article	41	0	41	0	0)	DOI
DBPEDIA	Book	1	0	0	1	、夕庆	中国中に
IDR	Article	1272	195	349	0	一石司	せ同定に
IDR	Book	2	0	0	2	7	利用
IDR	Dataset	140	28	1	0	4	בדוני.
IDR	Dissertation	5	0	2	0		
IRDB	Article	1907943	409584	1621731	0	11966	
IRDB	その他成果物	682	19	2	0	0	
JALC	Article	4927670	4863062	42080	0	0	
JALC	Dataset	175924	175924	O	0	0	
JALC	その他成果物	53448	1996	38	0	0	
KAKEN	Article	3363897	810620	328126	0	1315049	
KAKEN	Book	375175	11	0	43594	59143	
NINJAL	Article	297	0	297	0	0	
NINJAL	Book	39	0	0	39	0	
NINJAL	Dataset	4	0	0	0	0	
NINJAL	Dissertation	8	0	8	0	0	
NINJAL	その他成果物	397	51	0	0	0	
RUDA	Dataset	61	0	0	0	0	
SSJDA	Article	480	0	480	0	0	
SSJDA	Book	112	0	0	112	0	
SSJDA	Dataset	1363	0	0	0	0	
SSJDA	その他成果物	594	0	0	0	0	

※ 2021-02-04 時点

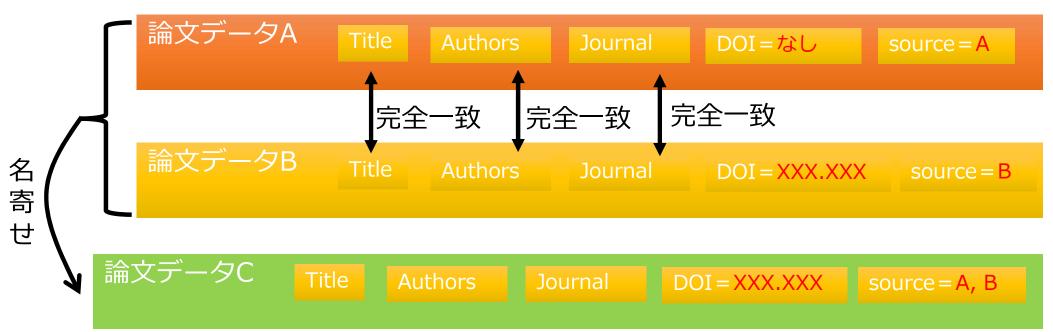


データの名寄せの効果

■IDを持たない論文にDOI を付与

CiNii Researchで収集したDOIを持たない論文と 持つ論文のマッチングを行い、同一論文と判定されたCiR の論文に対してDOI を付与

マッチングアルゴリズムの一例)



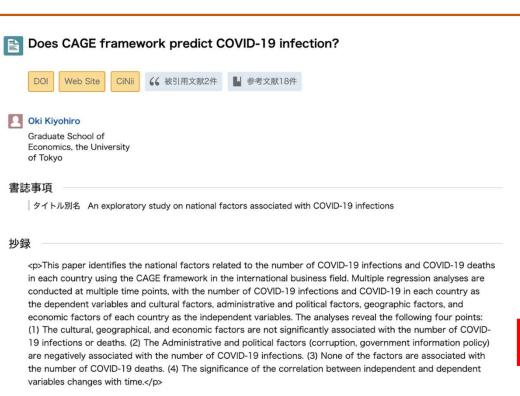
※他のメタデータはAとBの両方を含める。但しISSNや出版年が異なる場合は名寄せしない

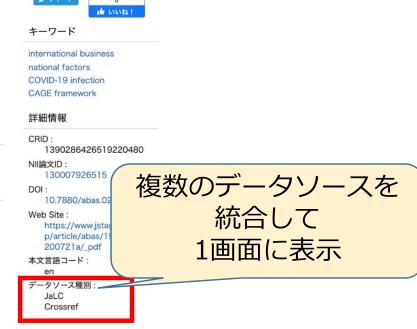


データの合理的な表示 各エンティティのメタデータリッチ化



CiNii Researchでヒットする論文の例







Annals of Business Administrative Science

Annals of Business Administrative Science 19 (5), 175-192, 2020 特定非営利活動法人 グローバルビジネスリサーチセンター

被引用文献 (2)

Reconsideration of value stream mapping and cross-functional integration in the digitalization of operations

Fukuzawa Mitsuhiro Annals of Business Administrative Science 19 (6), 263-276, 2020 -ln extant studies regarding value stream mapping (VSM) and cross-functional integration in Western journals, there is a deviation from the essence of lean production and flow management, which aim

DOI Web Site CiNii 参考文献20件

被引用情報



CiNii Researchでヒットする研究データ



● 研究データあり ● 参考文献1件

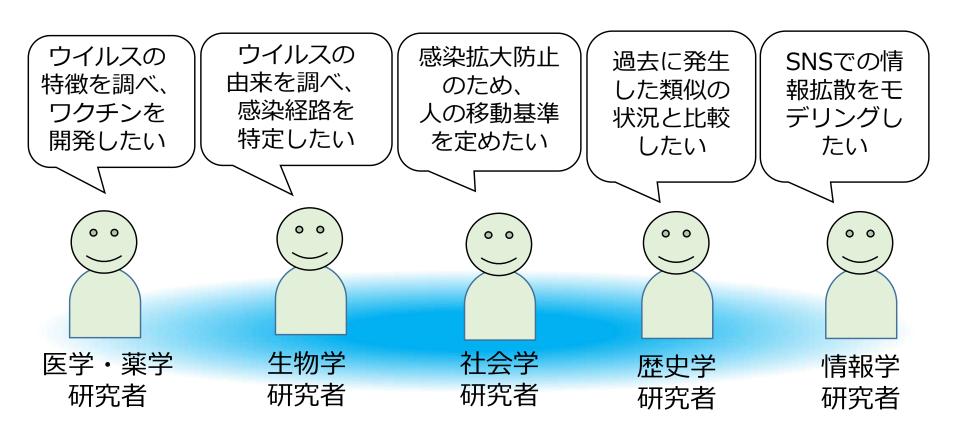


4. CiNii Researchで 研究データを発見する



COVID-19に関する研究

コロナ禍を収束させるため、多様な分野の研究者が 各立場から状況を分析したり、解決策を模索している。



これらの研究には、背景情報や先行する研究データなど、 踏まえておくべき前提情報が必要



2018年度以前のコロナウイルス研究

COVID-19を引き起こした新型コロナウイルス(SARS-CoV-2)に関係する「コロナウイルス」研究は、どの程度実施されてきたか?

科学研究費助成事業データベースKAKENで、研究課題名にコロナウイルス関連の文字列が含まれる、2018年度までの研究プロジェクト数を検索。

コロナウイルス **49**課題



coronavirus 21課題



SARS: 29課題



MERS: 5課題





2021年度「コロナウイルス」をタイトルに含む課題は78課題

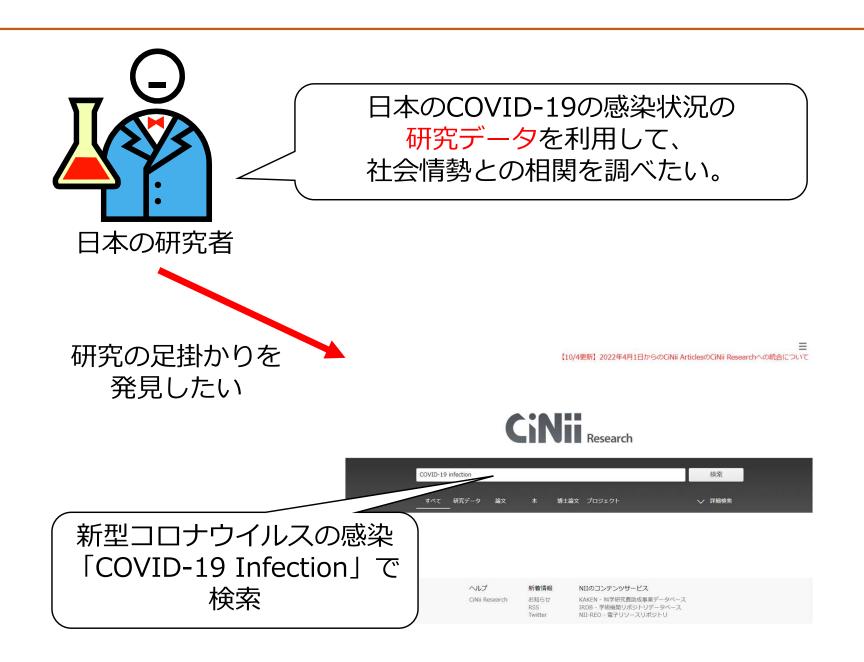
(2021/11/30時点。下記検索も同日)

→公開される論文や研究データも 爆発的に増加

https://kaken.nii.ac.jp/ja/search/?qa=コロナウイルス&fcdt=2021%2C2021

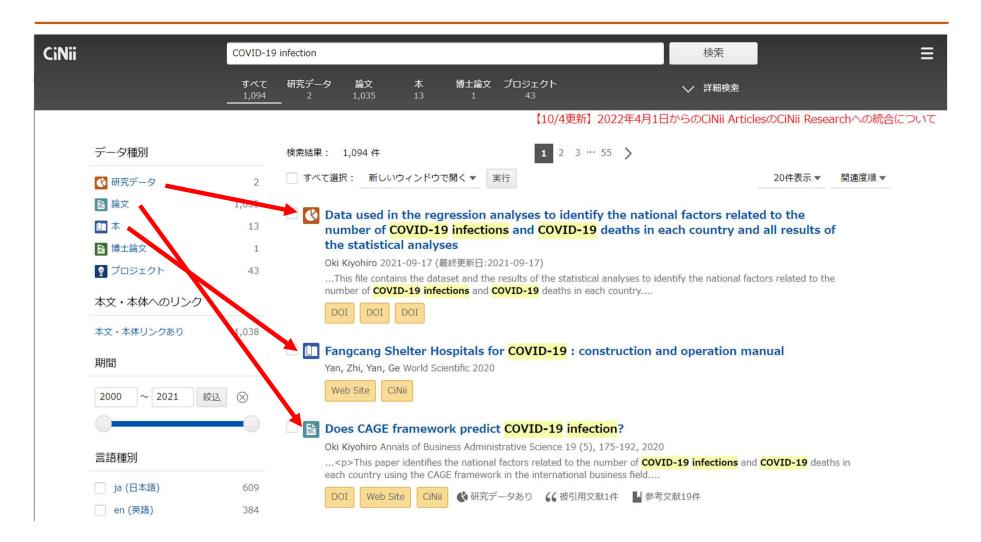


(ケース例)COVID-19の研究データを探す





COVID-19の感染研究の成果情報

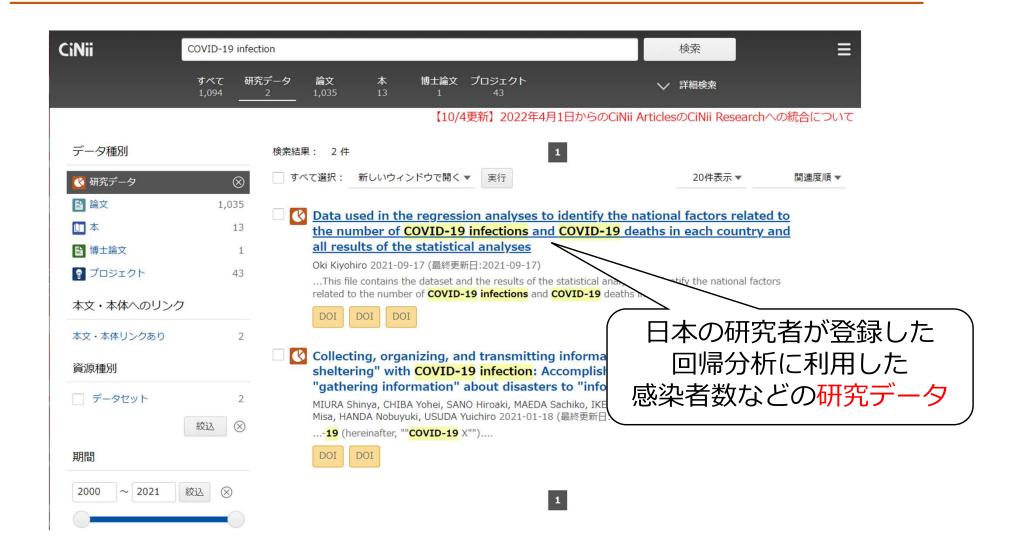


COVID-19の研究に関する情報が、1000件以上ヒット。 →研究データで絞り込み

https://cir.nii.ac.jp/ja/all?q=COVID-19



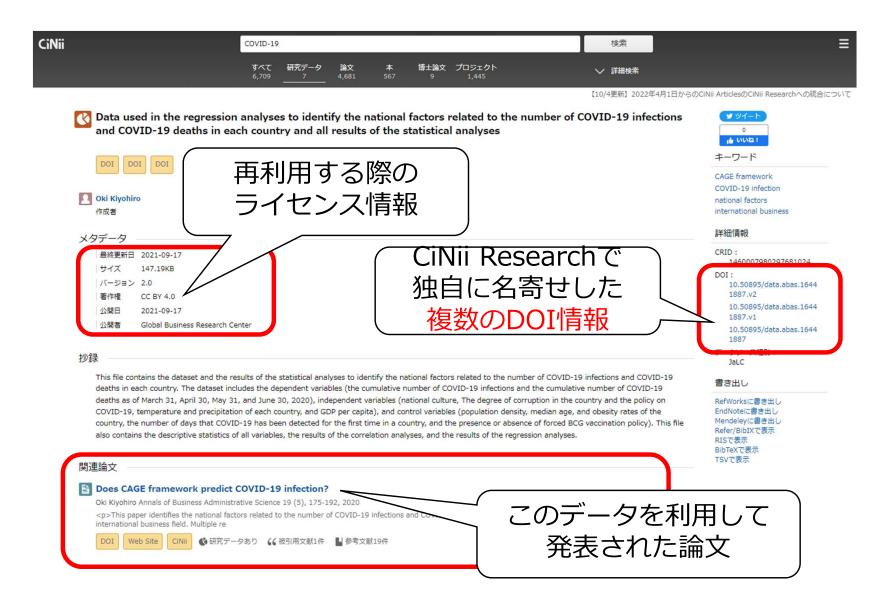
COVID-19の感染に関する研究データ



https://cir.nii.ac.jp/ja/data?q=COVID-19&sortorder=4



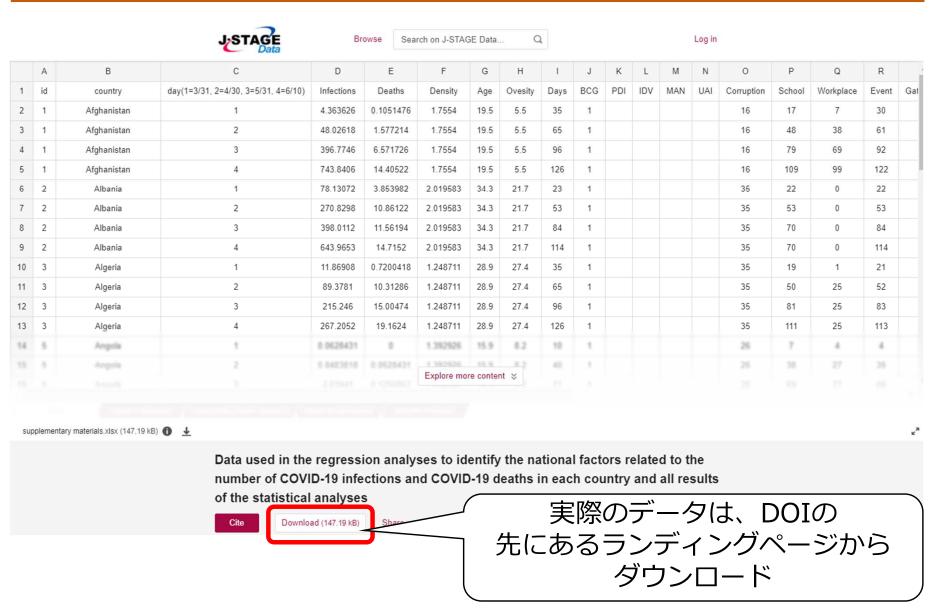
研究データ単体ページ



https://cir.nii.ac.jp/ja/crid/1460007980297681024



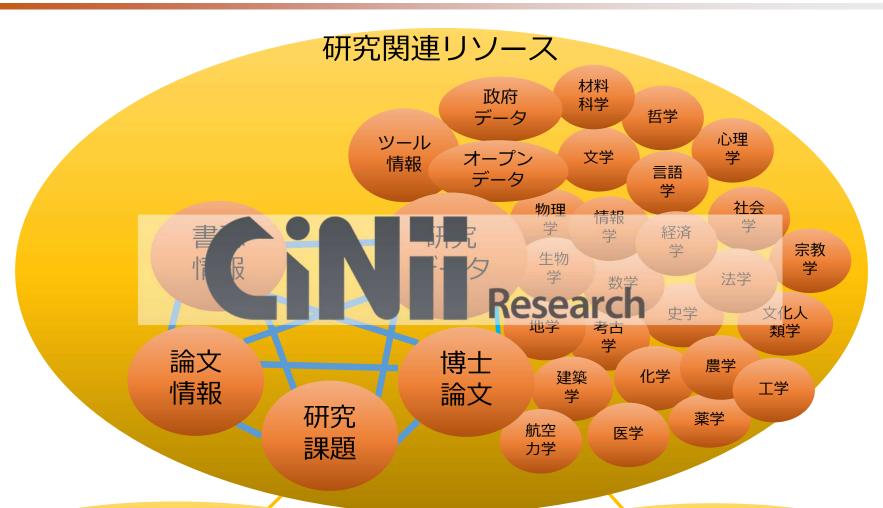
DOIから研究データ本体へ



https://jstagedata.jst.go.jp/articles/dataset/Supplementary_materials/16441887/2



CiNii Researchの将来イメージ



海外研究データ基盤





















JDCat (人社データカタログ)

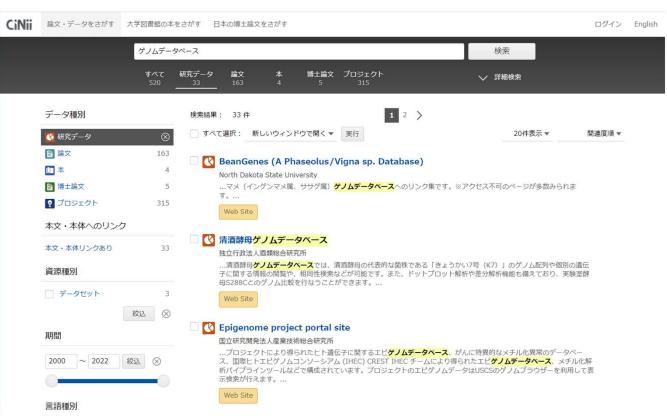


6月より検索対象にNBDCの Integbioデータベースカタログを追加しました

- 生命科学系データベースを一覧から探す -Integbio データベースカタログ



DOIやJ-GLOBAL IDを利用した成果 物や研究者情報の 名寄せを実施。



NBDCが作成したライフサイエンス分野の 研究データカタログを統合し、検索結果に表示。



まとめ

- 学術情報を取り巻く状況から「大規模研究データ」への需要と、日本の政策的な「学術データ利用」への期待から、研究データ基盤システムNII Research Data CloudとCiNii Researchが誕生した。
- CiNii Researchの内部には、論文や書誌情報のみならず、研究データや3000万件以上の多様な学術情報が統合されている。
- 今後は海外の学術基盤や分野別データ基盤とも連携し、さらに新たな知の循環へ貢献する。



ご清聴ありがとうございました!