

研究用データセットの  
シェアリング文化を創る！

[D]

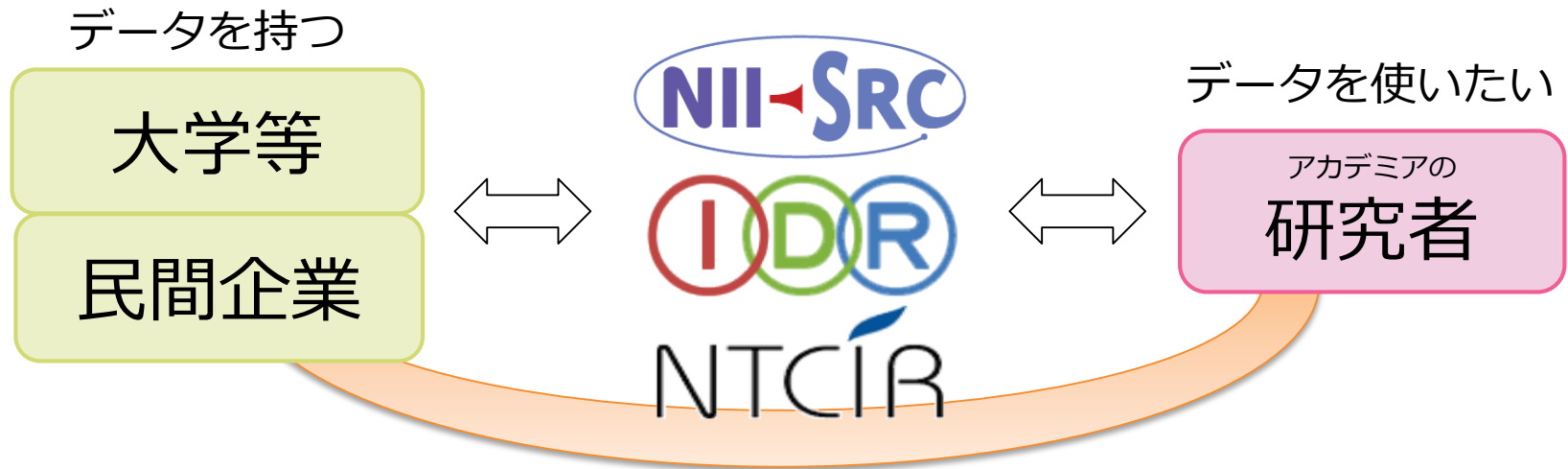
# 情報学データ資源の共同利用

データセット共同利用研究開発センター

大山敬三, 神門典子, 佐藤真一, 山岸順一,  
相澤彰子, 水野貴之, 菅原 朔, 大須賀智子



# センターの主な活動内容



産学界のデータを学術研究目的で大学等に提供

オープンデータにできないリアルなデータ（個人情報，著作権，…）を，適正な管理の下で，契約に基づき利用可能に

⇒ コミュニティを繋いで研究を活性化



# 各サービス／プロジェクトの概要



## 音声資源コンソーシアム

2006年～ 音声コーパスの提供窓口

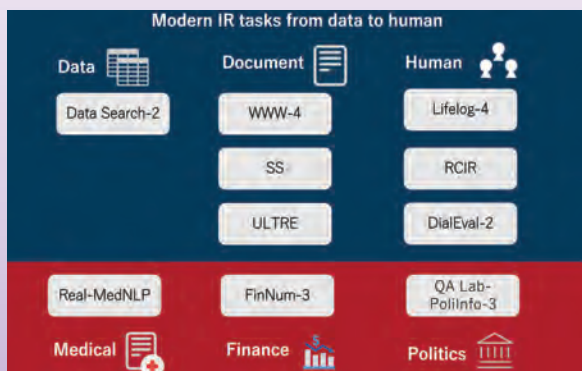


## 情報学研究データリポジトリ

2010年～ 企業データの提供窓口

2019年～ 研究者提供データの取り扱いも開始

## 情報アクセス研究のためのテストベッドとコミュニティ



### ・ NTCIRワークショップ

- 1年半サイクルでタスクを実施
- 共有テストコレクションを構築



### ・ NTCIRカンファレンス

- 各参加チームの成果を比較評価



### ・ テストコレクションの公開

# SRCで提供中の音声コーパス



音声資源コンソーシアム  
Speech Resources Consortium

NII-SRC  
Speech Resources Consortium English

ホーム SRCについて コーパスリスト 入手方法 資料室 FAQ リンク

ホーム コーパスリスト

## 音声コーパスリスト

→音声コーパス検索システム  
→音声コーパスを用いた研究成果一覧 (※DSC Reference Portalのサイトに飛びます)

### 無償コーパス

※当面は配布手数料(郵送料等)も無料とします

1. 重点領域研究「音声言語」・試験研究「音声DB」連続音声データベース (PASL-DSR)
2. 筑波大 多言語音声コーパス (UT-ML)
3. 東北大-松下 単語音声データベース (TMW)
4. 基盤研究(A)「日本語方言の地域差」方言音声コーパス (GSR-JD)
5. RWCP 音声データベース
  - a. 音声対話データベース - 96年版 (RWCP-SP96)
  - b. 音声対話データベース - 97年版 (RWCP-SP97)
  - c. 検索・要約用ニュース音声データベース (RWCP-SP99)
  - d. 会議音声データベース (RWCP-SP01)
6. RWCP 実環境音声・音響データベース (RWCP-SSD)
7. 重点領域研究「音声対話」対話音声コーパス (PASD)
8. CIAIR 子供の声データベース (CIAIR-VCV)
9. IPSJ SIG-SLP 雑音下音声認識評価環境 (CENSREC)
  - a. 雑音重畳日本語連続数字 音声認識評価環境 (CENSREC-1/AURORA-2J)
  - b. 雑音下日本語連続数字 音声区間検出評価環境 (CENSREC-1-C)
  - c. マルチモーダル音声認識評価環境 (CENSREC-1-AV)
  - d. 実環境車内 日本語連続数字 音声認識評価環境 (CENSREC-2)
  - e. 実環境車内 日本語単語 音声認識評価環境 (CENSREC-3)
  - f. 雑音下日本語連続数字 音声認識評価環境 (CENSREC-4)

## 45種類のコーパス

- 読み上げ/講演/対話/3人会話
- 単語/短文/長文
- 成人/乳幼児/高齢者, 非母語話者
- 方言, 多言語, 感情音声
- 雑音下, 残響下, などなど...

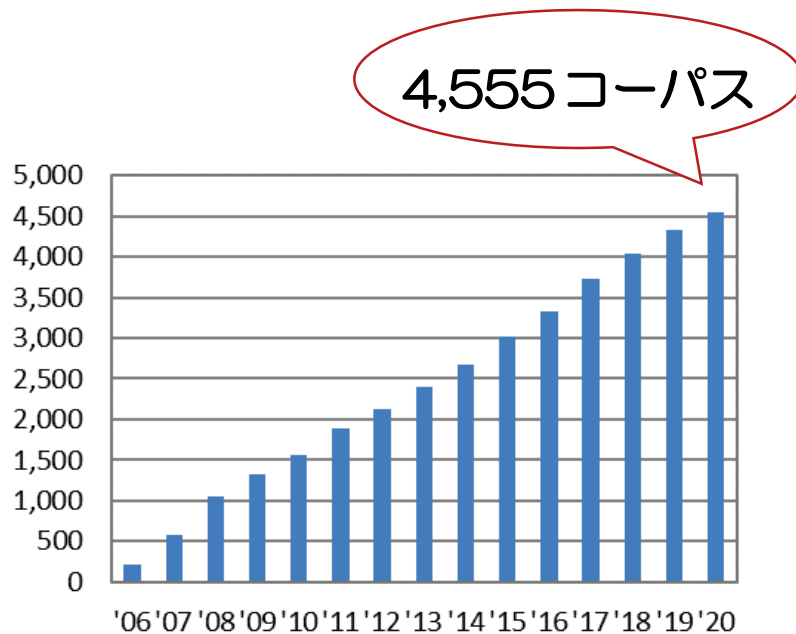
16. 音声研究用X線フィルムデータベース (X-Ray)
17. 特定領域研究「韻律と音声処理」日本語MULTEXT韻律コーパス (MULTEXT-J)
18. 中国語MULTEXTコーパス (MULTEXT-C)
19. 慶應義塾大学 研究用感情音声データベース (Keio-ESD)
20. 身体情報付き男・女・子どもの母音音声データベース (JVPD)
21. 東工大 多言語音声コーパス (TITML)
  - a. インドネシア語 (TITML-IDN)
  - b. アイスランド語 (TITML-ISL)
22. AWA長期間収録音声コーパス (AWA-LTR)
23. 南琉球新城方言音声データベース (Aragusuku)
24. 宮古大神島方言音声データベース (Oogami)
25. 感情評定値付きオンラインゲーム音声チャットコーパス (OGVC)
26. 千葉大学 3人会話コーパス (Chiba3Party)
27. 近畿大 児童の単語音声データベース (JWC)
28. 日本語単一話者オーディオブック・紙芝居朗読音声コーパス (J-KAC) **NEW**

### 有償コーパス

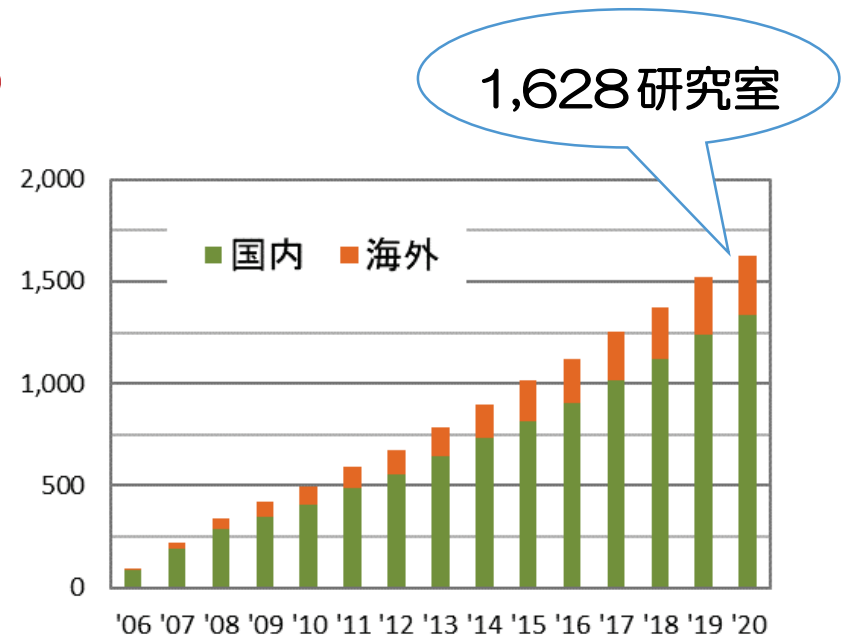
1. 日本音響学会 新聞記事読み上げ音声コーパス (JNAS)
2. 新聞記事読み上げ高齢者音声コーパス (S-JNAS)
3. 日本音響学会 研究用連続音声データベース (ASJ-JIPDEC)
4. NTT・東北大 親密度別単語了解度試験用音声データセット (FW03)
5. NTT・東北大 親密度別単語了解度試験用音声データセット2007 (FW07)
6. NTT 乳幼児音声データベース (INFANT)

# 音声コーパス提供実績：研究室単位（～2021.3）

## コーパス提供数の推移



## 異なり利用者数の推移



# 無償コーパスはオンライン配布中 (2019.3~)

※ IDRのサイトより

情報学研究データリポジトリ (IDR)

音声コーパス

コーパスを選択

申請フォームへ

- 利用申請
- 利用に際しては、コーパスごとに定められた利用規約の内容にご同意いただけます。下記よりご希望のパス名をクリックすると当該コーパスの利用申請ページに移りますので、そちらからお申し込みください。
- <無償コーパス>
1. 重点領域研究「音声言語」・試験研究「音声DB」連続音声データベース (PASL-DSR)
  2. 筑波大 多言語音声コーパス (UT-ML)
  3. 筑波大 多言語音声コーパス (UT-ML)
  4. 筑波大 多言語音声コーパス (UT-ML)
  5. 筑波大 多言語音声コーパス (UT-ML)
  6. RWCP 実環境音声・音声データベース (RWCP-CSS)
  7. 重点領域研究「音声対話」対話音声コーパス (PASD)
  8. 重点領域研究「音声対話」対話音声コーパス (PASD)
  9. IPSJ SIG-SLP 雑音下音声認識評価環境 (CENSREC)
    - a. 雑音重畳日本語連続数字 音声認識評価環境 (CENSREC-1/AURORA-2)
    - b. 雑音下日本語連続数字 音声区間検出評価環境 (CENSREC-1-C)
    - c. マルチモーダル音声認識評価環境 (CENSREC-1-AV)
    - d. 実環境車内 日本語連続数字 音声認識評価環境 (CENSREC-2)
    - e. 実環境車内 日本語単語 音声認識評価環境 (CENSREC-3)
    - f. 雑音下日本語連続数字 音声認識評価環境 (CENSREC-4)
  10. 特定領域研究「メディア教育利用」音声データベース (UME)
    - a. 日本人学生による読み上げ英語音声データベース (UME-ERJ)
    - b. 留学生による読み上げ日本語音声データベース (UME-JRF)
  11. 理研ウェブ操作対話音声コーパス (RIKEN-DLG)
  12. 地図課題対話コーパス
    - a. 千葉大 日本語地図課題対話コーパス (MapTask)
    - b. 三重大 日本語地図課題対話コーパス (MapTask-Mie) - NEW!!

※ 各コーパス提供サービス規約も必ず事前にご確認ください。

※ 全項目必須入力

住所\*

大学または会社名\*

研究室または部署名\*

責任者名\*

※ 上記の研究または部署 (グループ) の責任者もしくはコーパスの管理責任者 (補助職員相当の方) を入力してください (学生不可)

責任者職名\*

責任者メールアドレス\*

※ 原語として所属機関発行の個人アドレスを入力してください  
コーパスご提供後、利用者のメールアドレスへ返信させていただきます

使用目的\* (最大256文字)

例: 「音声認識システムの性能評価」「方言のアクセント研究」など

利用規約およびサービス規約の内容を確認し、同意の上、利用を申請します。

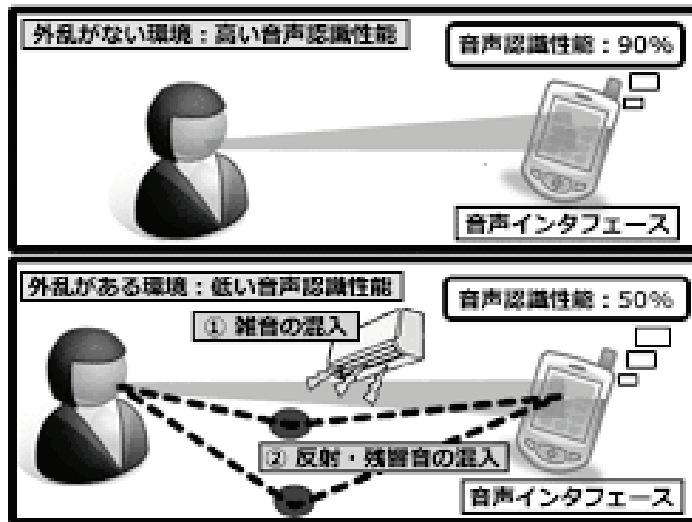
仮申請

※ 入力内容およびご本人確認のため、「責任者メールアドレス」宛てに確認メールをお送りします。メールの内容をご確認の上、24時間以内に「確定」の処理をお願いします。



# 音声コーパスを利用した研究の例（1）

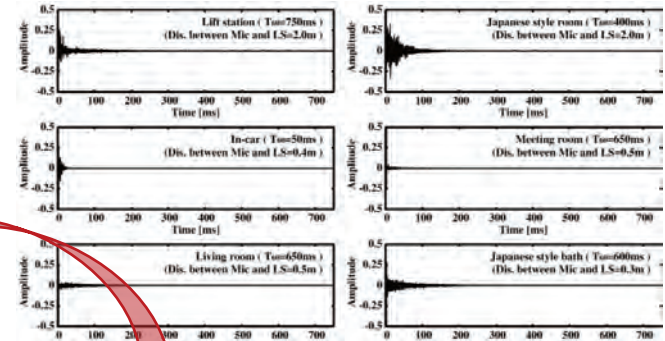
雑音や残響下での音声認識性能の劣化



福森（2015）より引用

⇒ 事前の残響除去の研究

CENSREC-4：残響下音声認識評価環境



桃瀬ら（2016）より引用

# 音声コーパスを利用した研究の例 (2)

## Chiba3Party : 千葉大学3人会話コーパス

### 音声情報処理

例：石本ら (2020)  
「三人会話の次話者  
選択に関わる言語  
・音響特徴の分析」

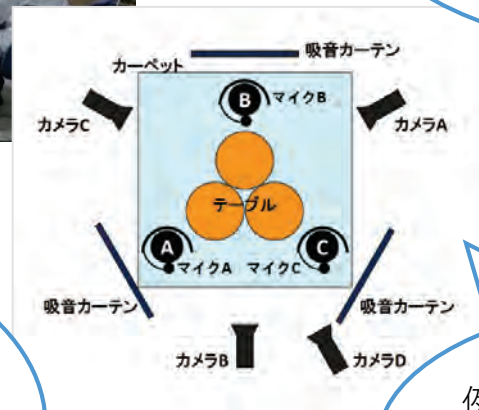


### 会話分析

例：中川 (2015)  
「日本語話し言葉における  
情報構造と語順」  
例：松村 (2016)  
「対面会話における相槌の  
機能の決定要因について」

### 対話システムへの応用

例：木村ら (2015)  
「多人数会話における発話  
タイミング検出のための  
ムーブの推定」



### 日本語教育

例：太田ら (2019)  
「コーパスにもとづく  
学習文法項目の選定  
とレベル設定」



# IDRで提供中の民間企業のデータセット

- Yahoo! データセット
- 楽天データセット
- ニコニコデータセット
- リクルートデータセット
- クックパッドデータセット
- LIFULL HOME'Sデータセット
- 不満調査データセット
- Sansanデータセット
- インテージデータセット
- オリコンデータセット
- ダイエットロコミデータセット
- 弁護士ドットコムデータセット
- アットホームデータセット ※2020/12 新規提供開始!

The screenshot shows the homepage of the National Institute of Informatics (NII) Information Research Data Repository (IDR). The page is in Japanese and features a navigation menu at the top with links for HOME, データ一覧 (Data List), 研究成果一覧 (Research Results), ユーザーフォーラム (User Forum), 連絡 (Contact), 関連リンク (Related Links), and お問い合わせ (Contact Us). A red box highlights the '民間企業提供データ' (Data provided by private companies) section, which lists various data sets including Yahoo!, Rakuten, NicoNico, Recruit, Cookpad, LIFULL HOME'S, and others. Below the list, there is a 'お知らせ' (Notice) section with several news items, including the start of audio course provision and updates to the NII database. The footer of the page includes the year 2021 and the month 03.

# 例：弁護士ドットコムデータセット ※2020.3提供開始

オンライン法律相談サービス「みんなの法律相談」に投稿された約25万件の質問とそれに対する弁護士の回答のテキストデータ

- 質問のカテゴリ
- 「弁護士が同意」の人数
- 質問者が付与した「ありがとう」「ベストアンサー」



<https://www.bengo4.com/> のトップページより引用

# 例：アットホームデータセット ※2020.12提供開始！

住宅情報サイト「アットホーム」に2015年～2019年に掲載された全国の賃貸物件・売買物件のデータ

The screenshot shows a rental listing on the At Home website. The main heading is "賃貸マンション 802 ワンルーム". The rental price is listed as "賃料: 15.7万円". Below this, there are details for management fees, deposit, and lease terms. A floor plan is shown with a highlighted room labeled "洋室 14.3帖" and a balcony. To the right, there are several small photos of the apartment's interior. At the bottom, a table summarizes the key details of the listing.

賃料	15.7万円	管理費等	10,000円
敷金	1ヶ月	礼金	なし
間取り	ワンルーム	面積	40.37㎡
築年月	2005年7月 (築16年)	種類	賃貸マンション

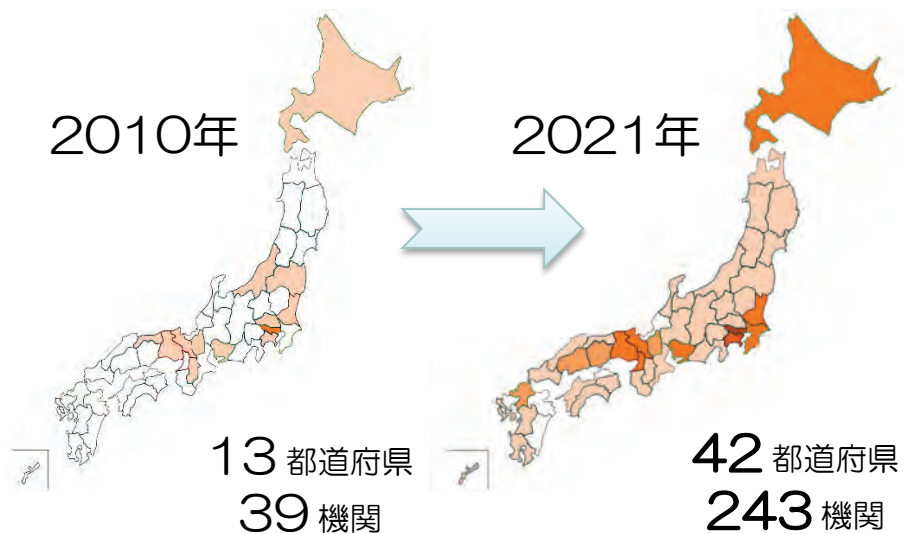
- 賃料, 価格
- 物件概要
  - 面積
  - 間取り
  - 構造
  - 築年
- 立地
  - 所在地 (丁目)
  - 最寄り沿線・駅
  - 徒歩分
- 諸設備

など

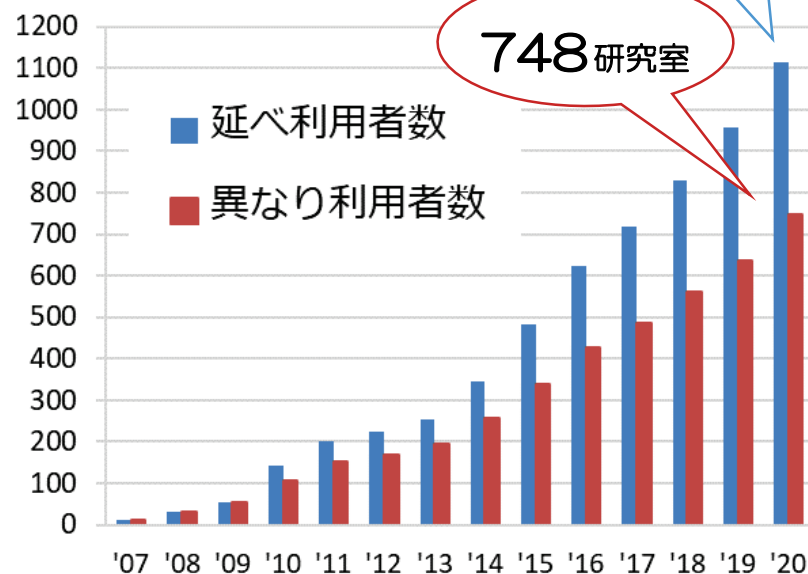
# 企業データ提供実績：研究室単位（～2021.3）

## データ提供先

（ニコニコ， Sansanデータセット除く）

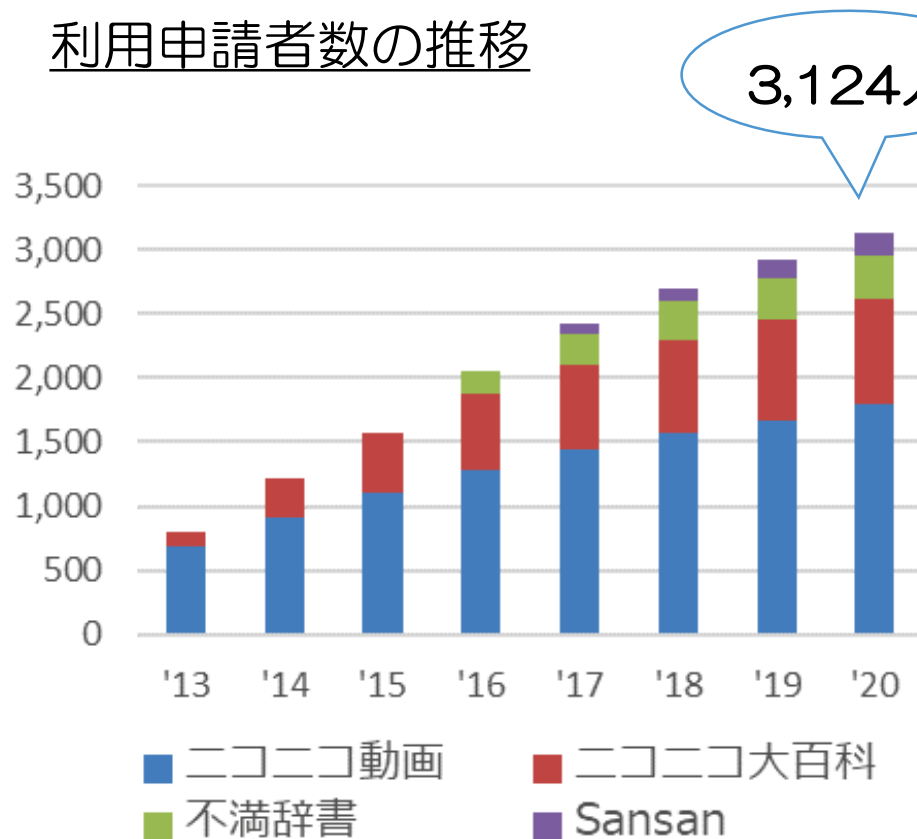


## 利用者数の推移

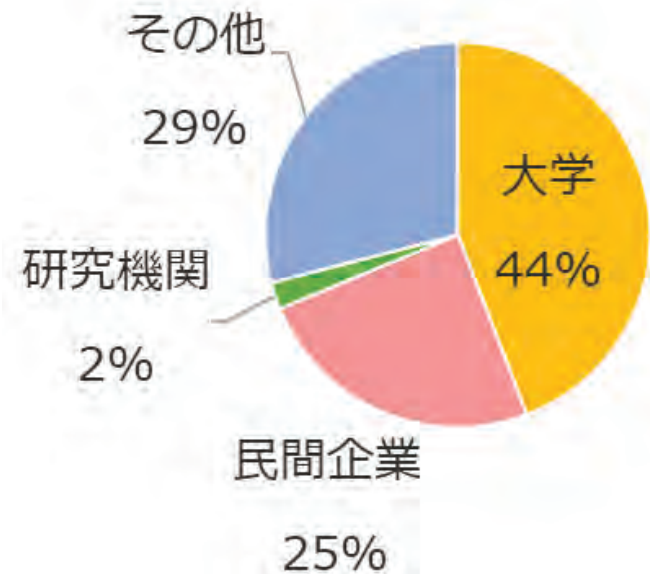


# 企業データ提供実績：個人単位（～2021.3）

## 利用申請者数の推移



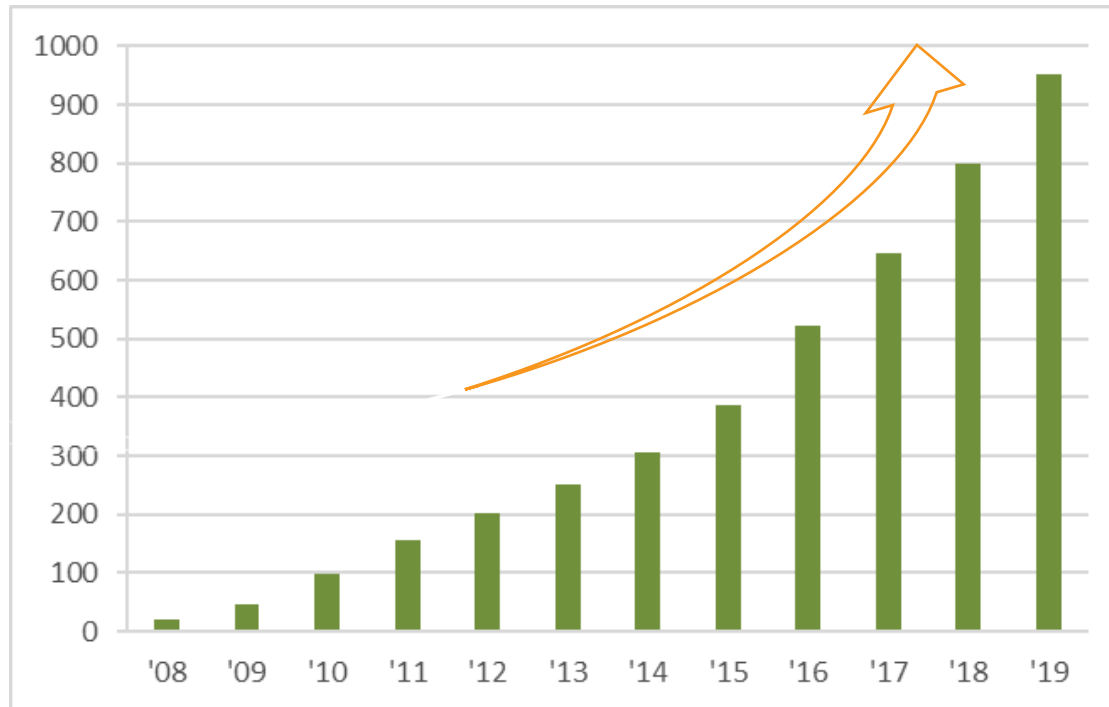
## 利用者の所属



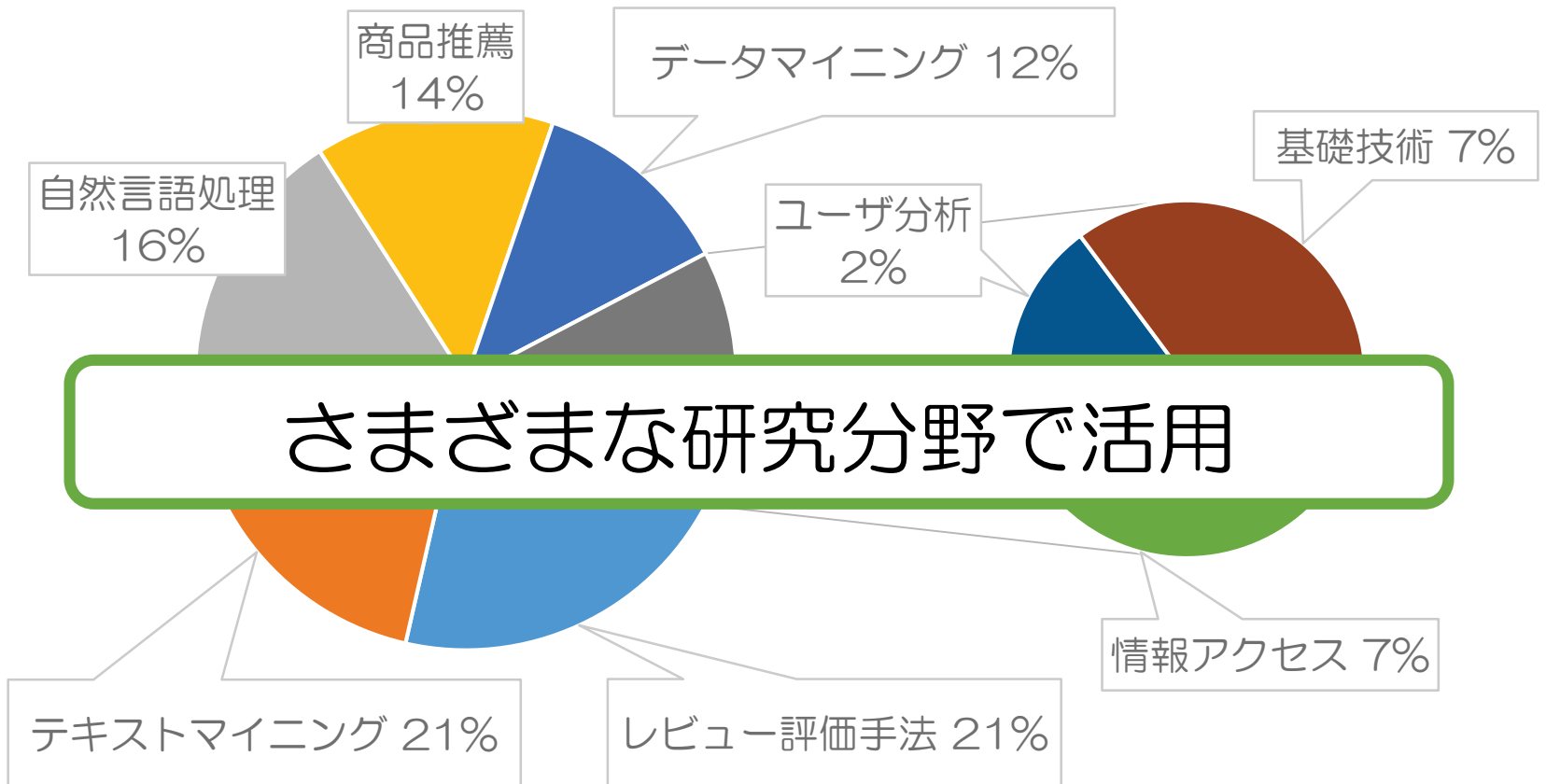
# 企業データを利用した研究発表（～2019年度）

## 研究成果発表数の推移

※毎年度の利用報告ベース



# 楽天市場データを利用した研究の例



# データを利用した論文リストは公開しています

「研究発表」

研究発表

リクルートデータセット [9件]

ランダムフォレストを用いたカスタマーレビューの半自動分類

小暮 枝梨子, 宮藤 文祐, 石津 昌平  
インテリジェントシステムシンポジウム2017予稿集, 154-157 (2017)

定数分析を用いた自作自産のロコモ判定モデルの構築と評価

谷 芽蒔, 藤原 智也, 波多野 昌治  
IDRユーザフォーラム2017, (2017)

レビュー投稿者の満足度を考慮した評判情報の可視化と店舗検索への応用

日高 加奈, 豊田 哲也, 大原 剛二  
IDRユーザフォーラム2017, (2017)

価値観モデルを適用した協調フィルタリングによるハイブリッド型推薦手法



<https://dsc.repo.nii.ac.jp/>



# イベント「IDRユーザフォーラム」

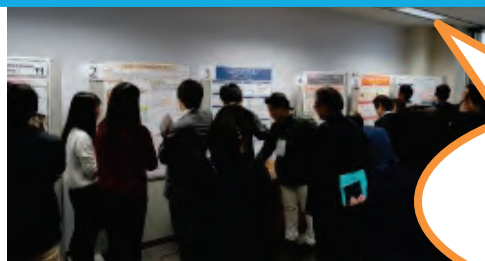


データの提供者と利用者との交流の場

## IDRユーザフォーラム2021

2021年11月下旬開催予定

<https://www.nii.ac.jp/dsc/idr/userforum/>



昨年度より  
日本DB学会  
データチャレンジ  
も併催！



# 研究者等提供データセットの受入を開始

- 受入要項を公開

- 流れ：

受入依頼書の提出



NIIで審査・手続き



提供同意書を提出

The screenshot shows the NII IDR website interface. At the top, there is a navigation bar with 'HOME', 'データ集', '研究成果一覧', 'ユーザフォーラム', '組織', '関連リンク', and 'お問い合わせ'. Below this, a sidebar lists various data sets, with '研究者提供データ' (Researcher-provided data) highlighted in a red box. The main content area is titled '大学等研究者提供データセット受入' (Acceptance of data sets provided by universities and researchers), also highlighted in a red box. It includes a date '2019/09/26 作成' and a description of the service. Below the description, there are sections for '受入条件' (Acceptance conditions), '受入手続き' (Acceptance procedure), and a list of steps for submission. The first step is highlighted in a red box.

**大学等研究者提供データセット受入**

2019/09/26 作成

情報学研究データリポジトリ (IDR) では、大学等の研究者（作成者）が研究用に作成した各種データセットを他の研究者（利用者）へ提供するための支援サービスを行っています。

本サービスでは、著作権や個人情報保護など様々な理由によりオープンデータとしての公開が難しいデータセットでも、作成者と協力して利用者の範囲、利用目的や利用方法などの条件を定め、IDRを通じて利用許諾契約を締結した利用者にデータセットを提供することによって、できるだけ作成者の負担を軽減しつつ、多くの研究者がデータセットを利用できるようにします。

**受入条件**

学術研究の対象として高い価値を有し、相当数の利用者が見込まれるデータセットが、本サービスの受入の対象となります。本サービスはアーカイブではありませんので、保存を目的とした受入はしません。

法令上または研究倫理上の観点から、IDRを通じた提供に必要な条件を満たさない場合や、IDRの能力を超える設備や取扱いが要求される場合は、受入をできません。

詳細については下記「書類一覧」の「研究者等提供データセット受入要項」をご覧ください。また、不明な点がございましたら、お気軽にお問い合わせください。

**受入手続き**

1. 所属機関の契約担当者に、下記「書類一覧」の「研究者等提供データセット受入要項」第3条の「提供者」について確認してください。  
※ 提供者は原則として研究代表者が所属する法人とします。ただし、例外として、当該法人によって適切であると判断された場合のみ、個人またはグループを提供者とすることができます。この場合、そのような判断の証となる資料の提出をお願いすることがあります。  
※ データセット作成時とは所属法人が異なる場合や、作成したグループが複数の法人にまたがる場合などは、予めご相談ください。
2. 「研究者等提供データセット受入要項」の内容に同意の上、下記「書類一覧」の「データセット受入依頼書」に必要事項を記入し、添付資料とともに、電子メールの添付ファイルで下記「問い合わせ窓口 (IDR事務局)」に提出してください。  
なお、様式は提供者に応じて「法人用」または「個人・グループ用」のいずれかを使用してください。
3. 提出いただいた「データセット受入依頼書」に基づき審査を行います。審査の過程でデータセット受入依頼書の修正や追加資料の提出をお願いすることがありますので対応をお願いします。

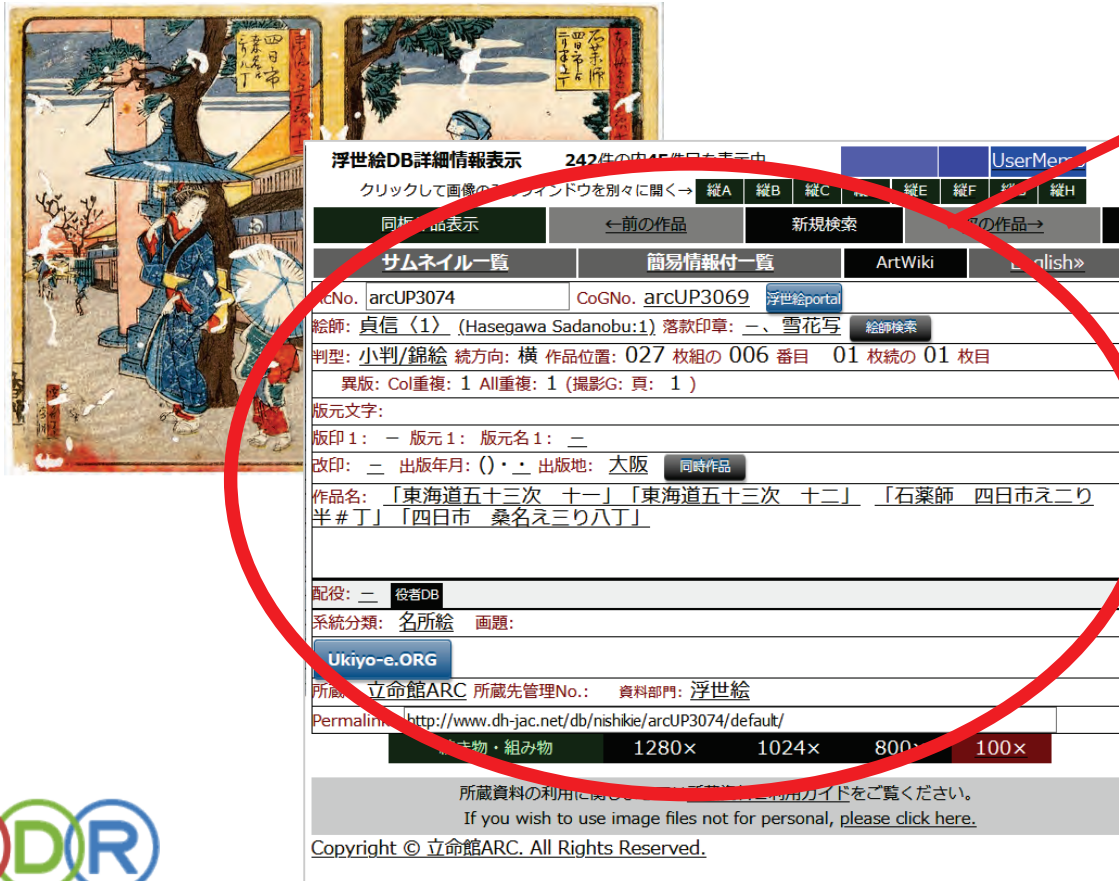


# 研究者等提供データセットの例（1）

## 立命館ARC所蔵浮世絵データベース

※2020.3提供開始

浮世絵の書誌情報  
約11,000件  
(画像データURL付き)



The screenshot shows a detailed record for a Ukiyo-e print. The title is 「東海道五十三次 十二」 「石薬師 四日市え二り半 #丁」 「四日市 桑名え三り八丁」. The artist is 貞信 (Hasegawa Sadanobu). The record includes fields for accession number (arcUP3074), collection number (arcUP3069), and various metadata such as dimensions (1280x1024x800x100) and a copyright notice for 立命館ARC.

arcNo.	arcUP3074	CoGNo.	arcUP3069
絵師	貞信〈1〉 (Hasegawa Sadanobu:1)	落款印章	一、雪花写
判型	小判/錦絵	続方向	横
作品位置	027 枚組の 006 番目	01 枚統の 01 枚目	
異版	Col重複: 1 All重複: 1 (撮影G: 頁: 1)		
版元文字			
版元1	-	版元名1	-
改印	二	出版年月	( ) . ( ) . ( )
出版地	太阪		
作品名	「東海道五十三次 十二」 「石薬師 四日市え二り半 #丁」 「四日市 桑名え三り八丁」		
配役	二	役者DB	
系統分類	名所絵	画題	
所蔵	立命館ARC	所蔵先管理No.	
資料部門	浮世絵		
Permalink	http://www.dh-jac.net/db/nishkie/arcUP3074/default/		
寸法	1280x	1024x	800x
単位			100x

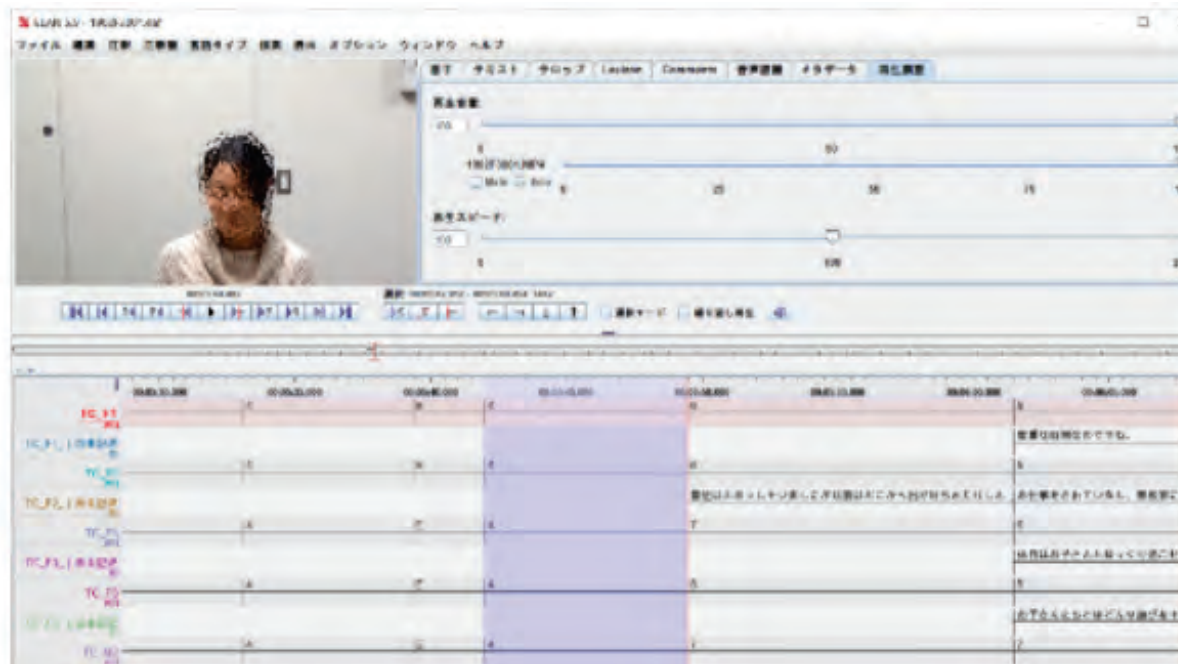
立命館大学アート・リサーチセンターの  
サイトより引用

# 研究者等提供データセットの例（2）

- 大阪大学 マルチモーダル対話コーパス（Hazumi）

- 対話エージェントと人との対話の映像とKinect等のデータ  
（1名あたり15分程度，20代～70代の計89名分）

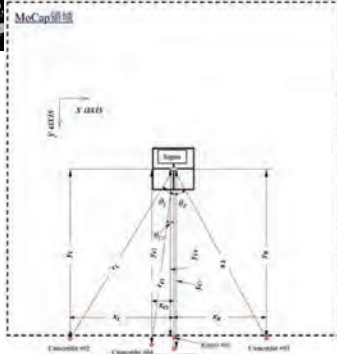
※2020.8提供開始



# 研究者等提供データセットの例（3）

- 工学院大学 多用途型日本手話言語データベース (KoSign)
  - 6,000超の単語（+数対話）の高精細カメラによる映像データと光学式モーションキャプチャによる3次元動作データ等

※2021.6提供開始！



3DCG画像の  
描画ツール付き

# NTCIRテストコレクション



## 情報アクセス研究のためのテストベッドとコミュニティ

NTCIRテストコレクション

NTCIRテストコレクション

名称	タスクデータ (1)		文書データ	
	説明	入手方法	説明	入手方法
ACLIA	NTCIR-7 NTCIR-8	オンライン申請		
CLIR	NTCIR-3 NTCIR-4 NTCIR-5 NTCIR-6	オンライン申請	(2)	
CLQA	NTCIR-5 NTCIR-6	オンライン申請	(2)	
CQA	NTCIR-8	(3)	「Yahoo!知恵袋」データ	覚書 → 申請方法
Crosslink	NTCIR-9 & NTCIR-10	オンライン申請	(2)	
FinNum	NTCIR-14 * データ更新: 2020/01/17	オープンアクセス	(2)	
INTENT / IMine	NTCIR-9 NTCIR-10 NTCIR-11 (IMine) NTCIR-12 (IMine)	オンライン申請	(2)	
Lifelog	NTCIR-12	オンライン申請	(2)	
MATH / MathIR	NTCIR-10 NTCIR-11 NTCIR-12 MathIR	オンライン申請	(2)	

構築したテストコレクションをIDRより提供中



# ぜひご利用ください！

- データセットを研究にご利用になりたい方
- データセットの提供をお考えの研究者の方

サイト：<https://www.nii.ac.jp/dsc/idr/>



- データセットの提供をお考えの企業の方
- メール：[idr@nii.ac.jp](mailto:idr@nii.ac.jp)

