NII OPEN HOUSE 2021

研究用データセットの
シェアリング文化を創る！

[D]

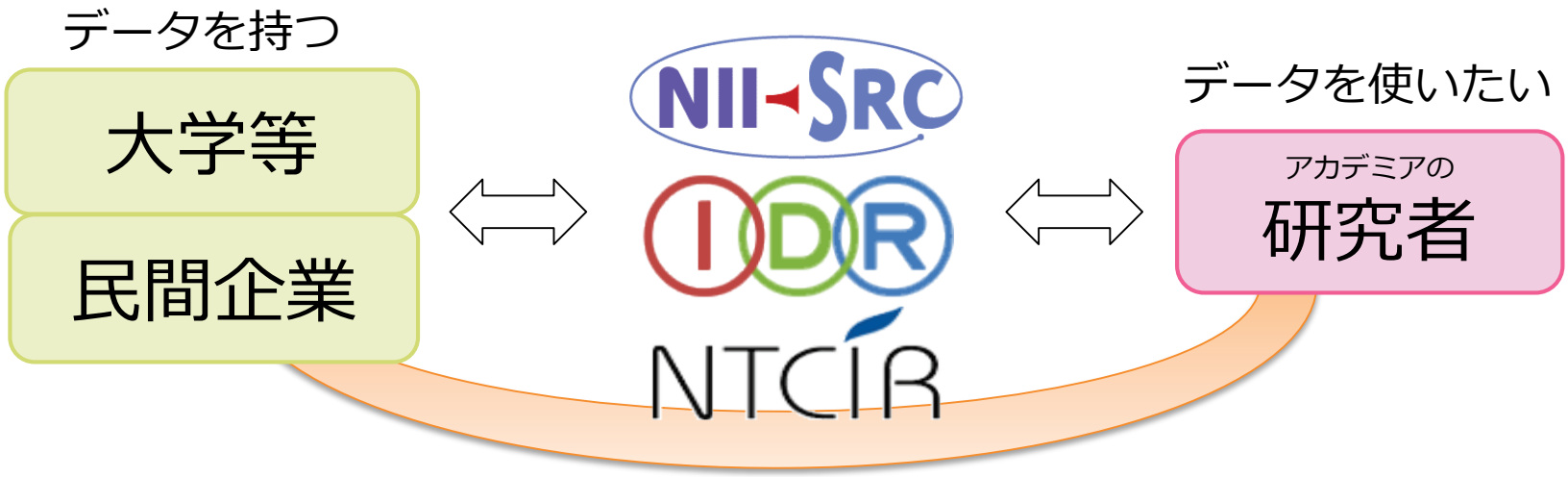
情報学データ資源の共同利用

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

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

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

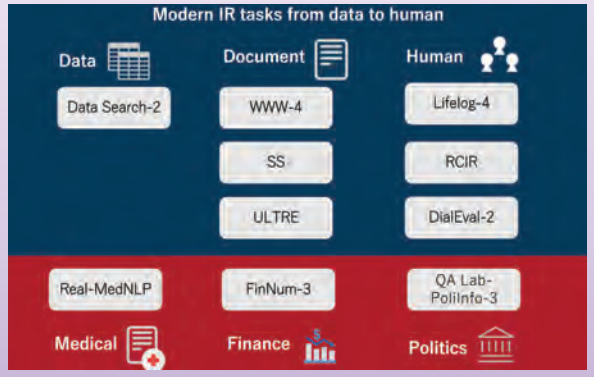
音声資源コンソーシアム

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

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…

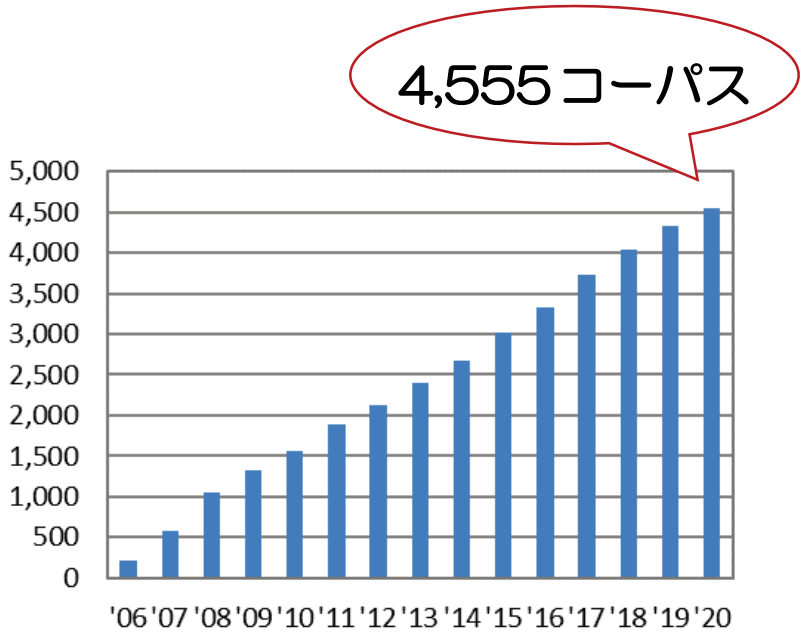
16. 音声研究用X線フィルムデータベース (X-Ray)
17. 特定領域研究「韻律と音声処理」日本語MULTEXT韻律コーパス (MULTEXT-J)
18. 中国語MULTEXTコーパス (MULTEXT-C)
19. 慶應義塾大学 研究用感情音声データベース (Keio-ESD)
20. 身体情報付き男・女・子どもの母音音声データベース (JVPD)
21. 東工大 多言語音声コーパス (TITML)
    a. インドネシア語 (TITML-IDN)
    b. アイスランド語 (TITML-ISL)
22. AWA長期間収録音声コーパス (AWA-LTR)
23. 南琉球新城方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）

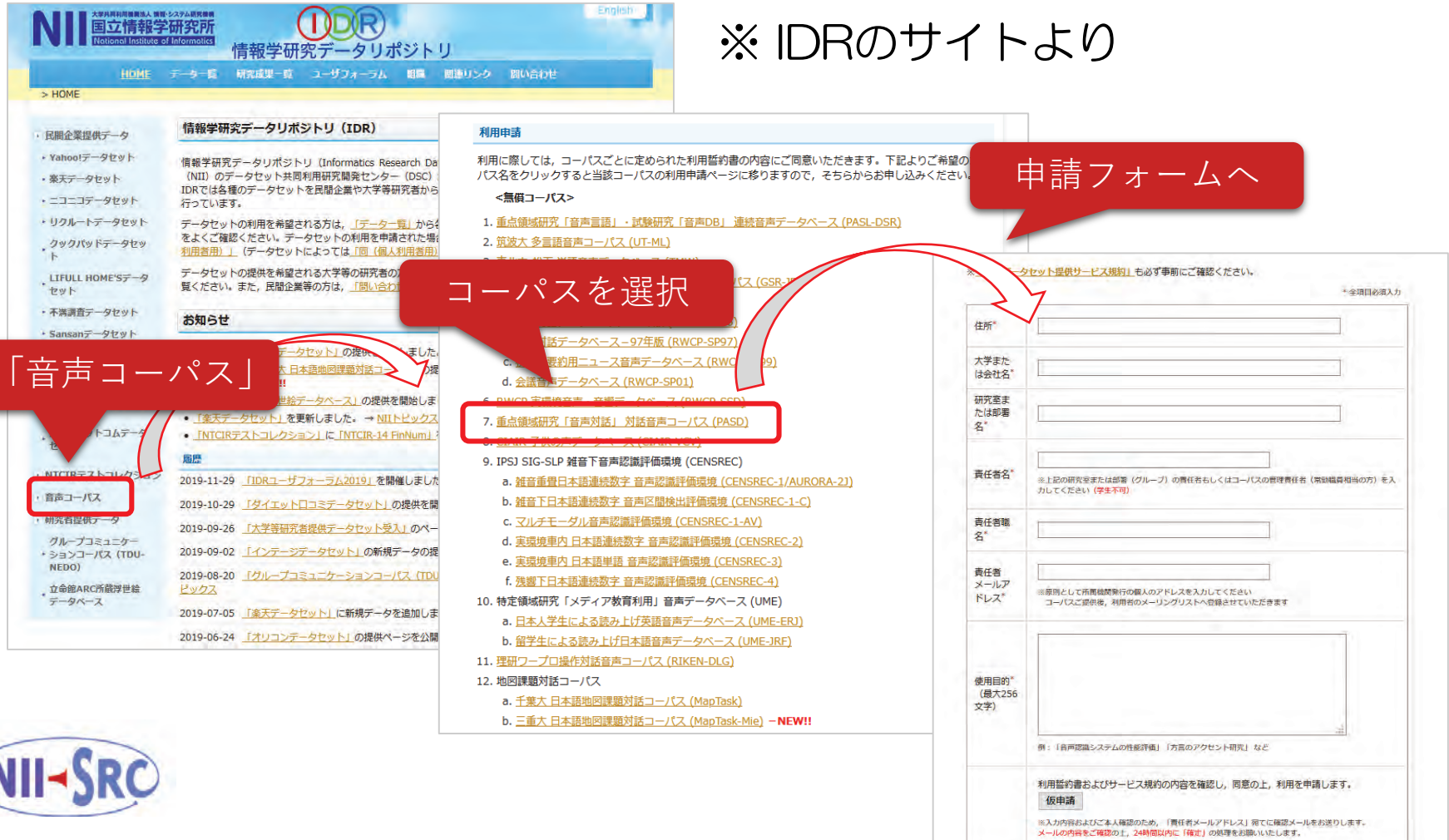
コーパス提供数の推移



異なり利用者数の推移

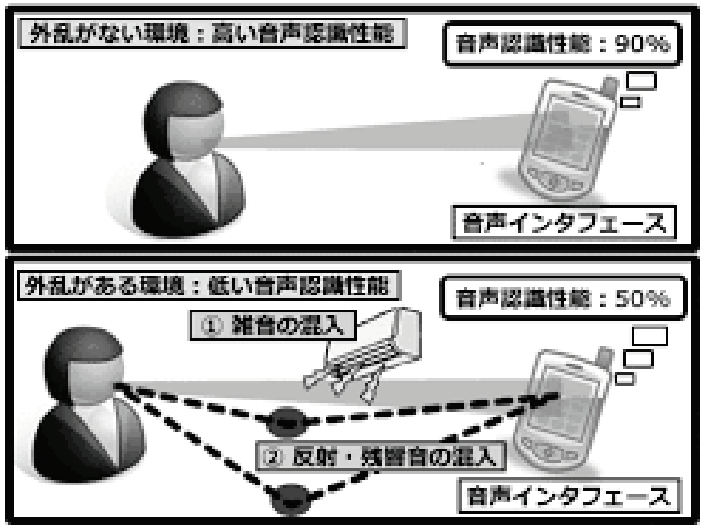
無償コーパスはオンライン配布中（2019.3～）

※ IDRのサイトより



音声コーパスを利用した研究の例（1）

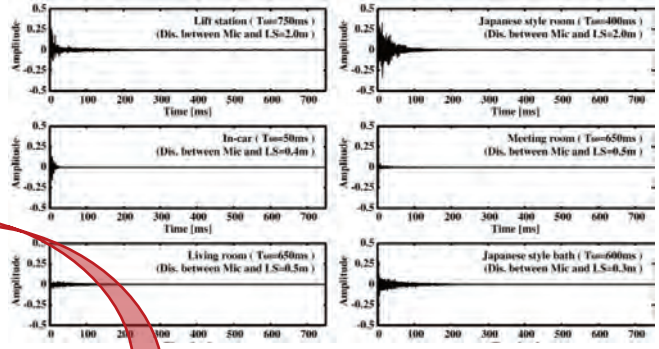
雑音や残響下での音声認識性能の劣化

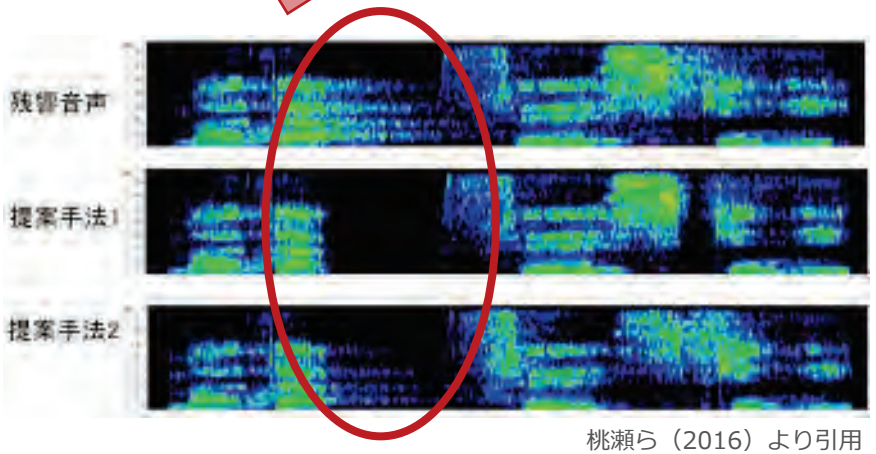


福森（2015）より引用

⇒ 事前の残響除去の研究

CENSREC-4：残響下音声認識評価環境

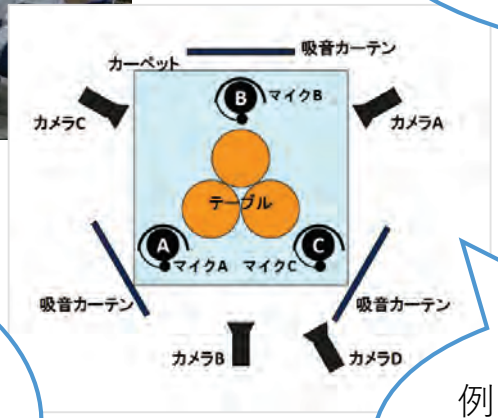




桃瀬ら（2016）より引用

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」

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」

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」

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！

例：弁護士ドットコムデータセット ※2020.3提供開始

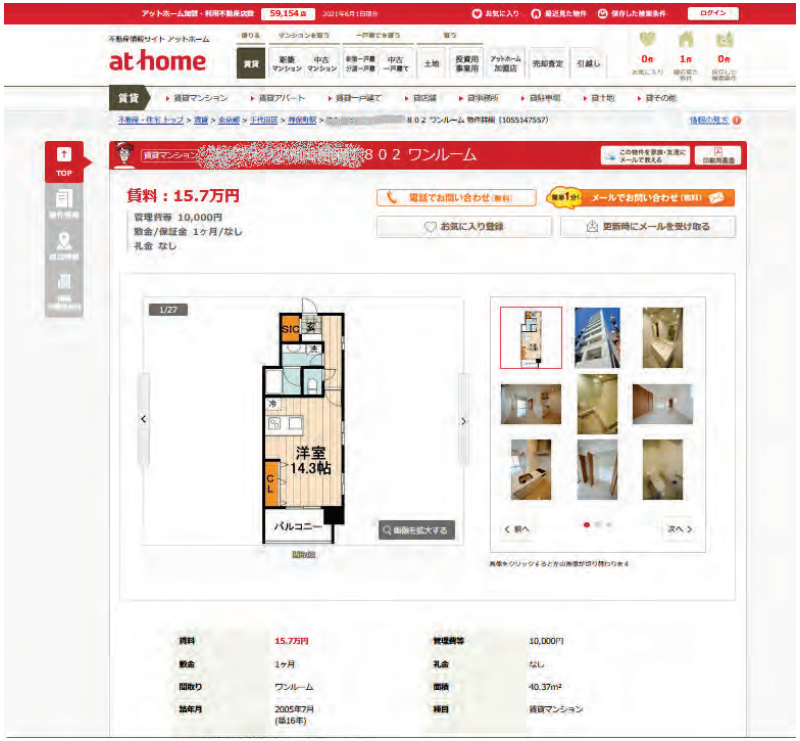
オンライン法律相談サービス「みんなの法律相談」に投稿された
約25万件の質問とそれに対する弁護士の回答のテキストデータ

- 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



https://www.athome.co.jp/ より引用

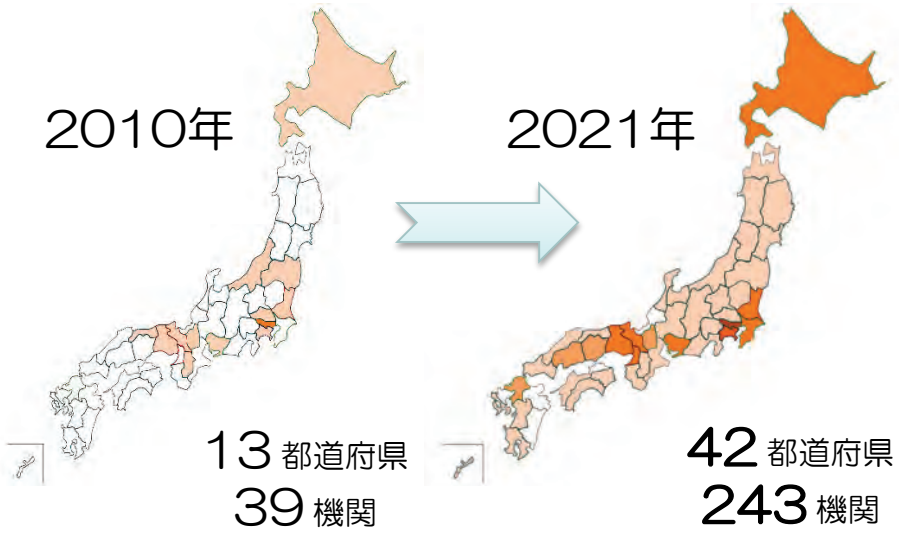
- 賃料，価格
- 物件概要
  - 面積
  - 間取り
  - 構造
  - 築年
- 立地
  - 所在地（丁目）
  - 最寄り沿線・駅
  - 徒歩分
- 諸設備

など

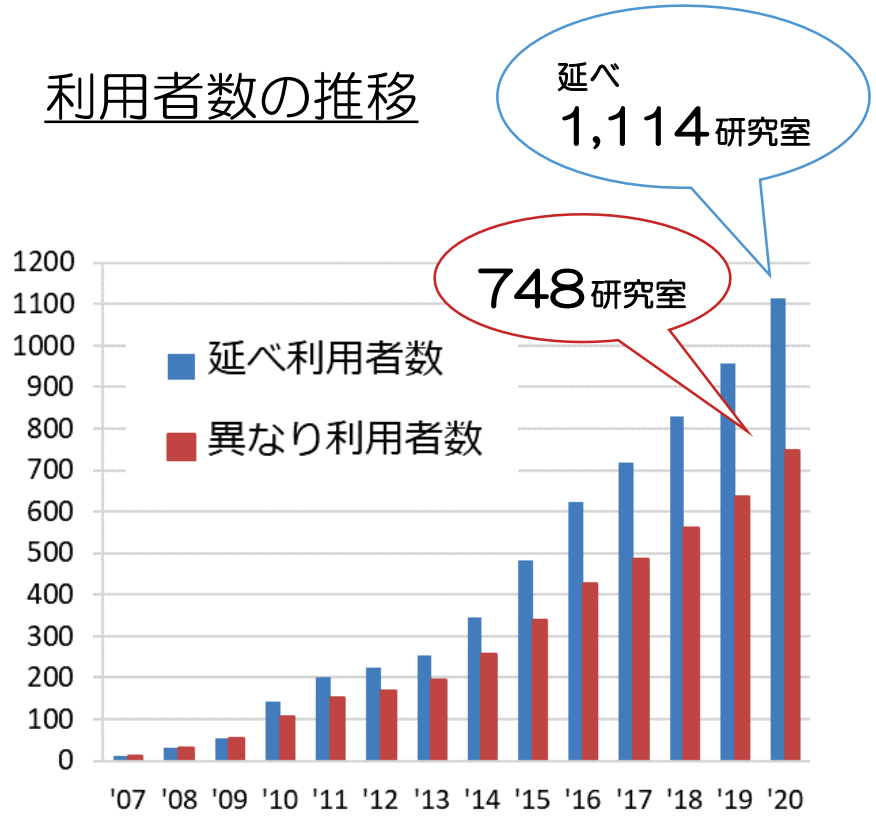
企業データ提供実績：研究室単位（~2021.3）

データ提供先

（ニコニコ，Sansanデータセット除く）

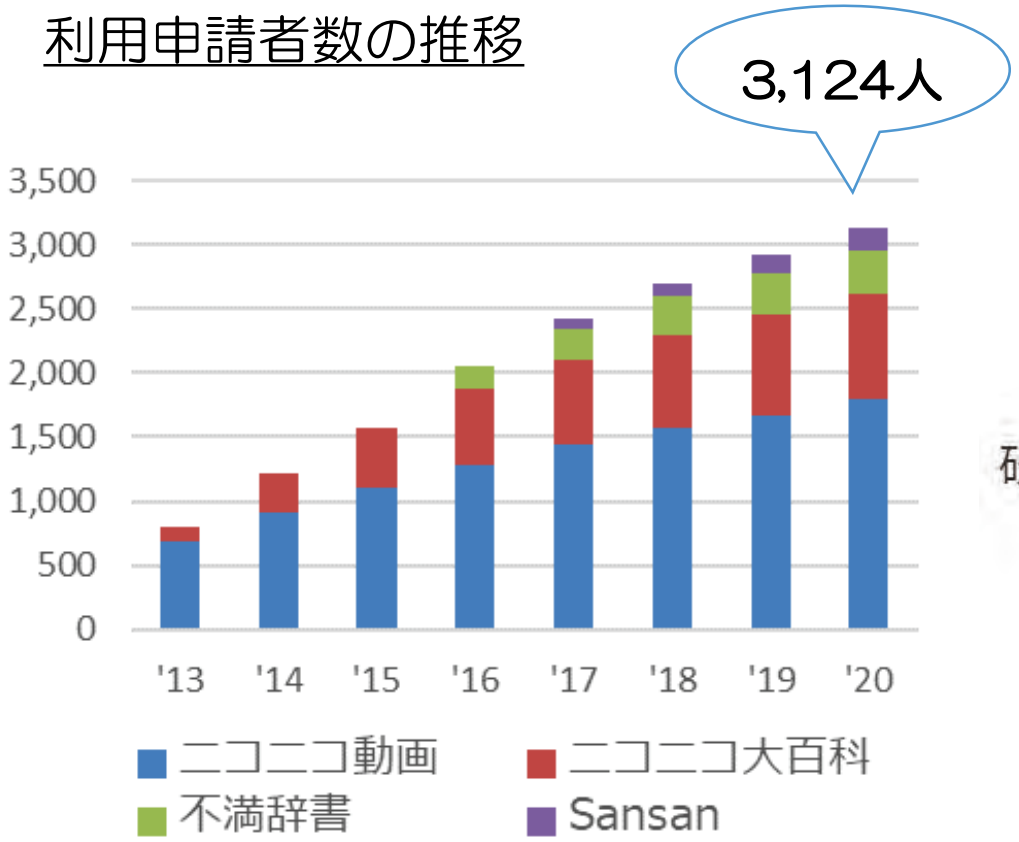


利用者数の推移

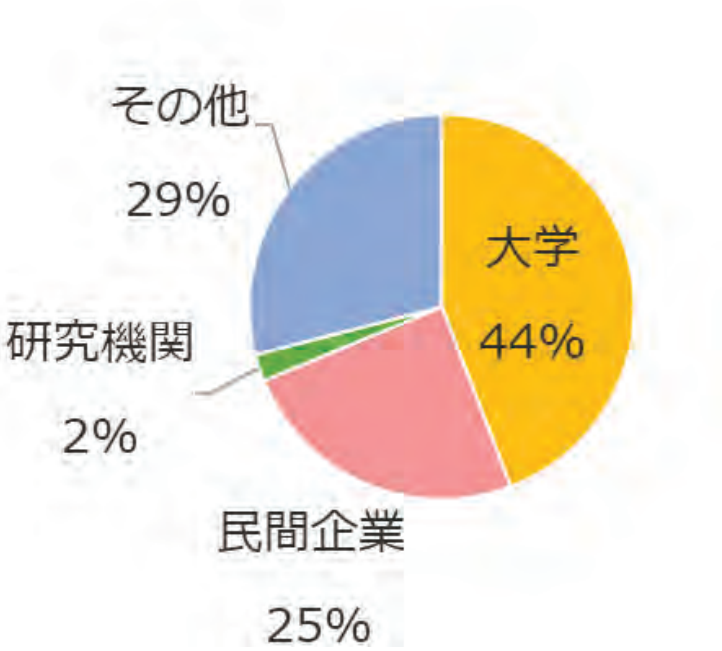


企業データ提供実績：個人単位（～2021.3）

利用申請者数の推移



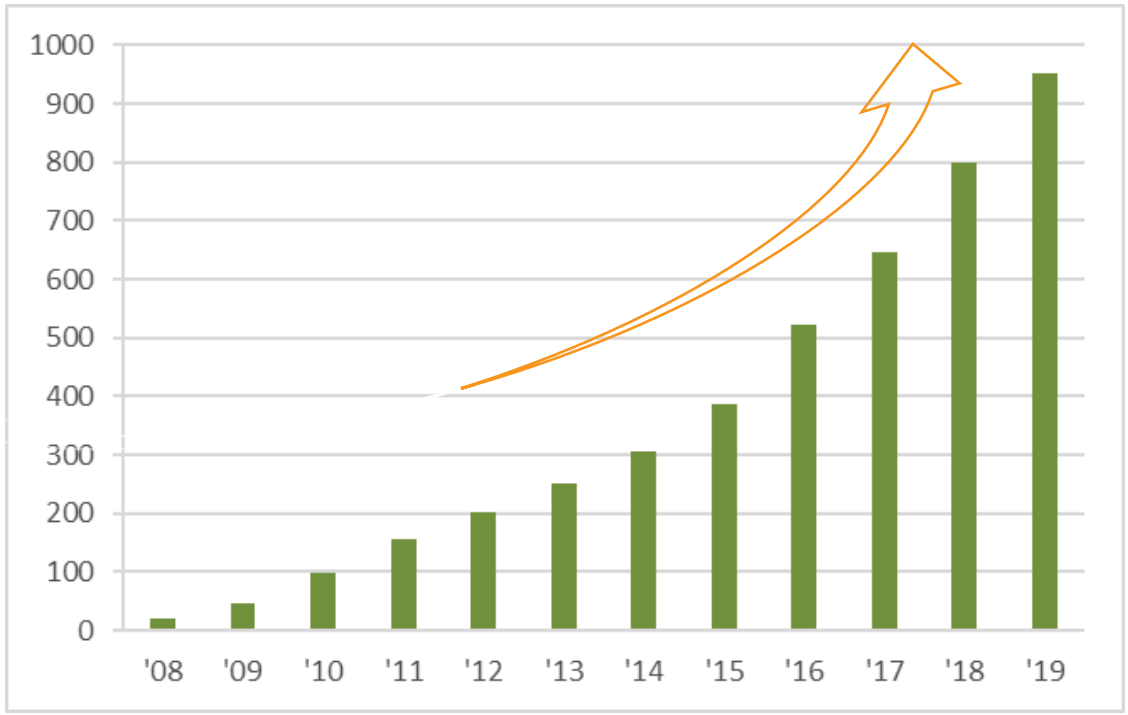
利用者の所属



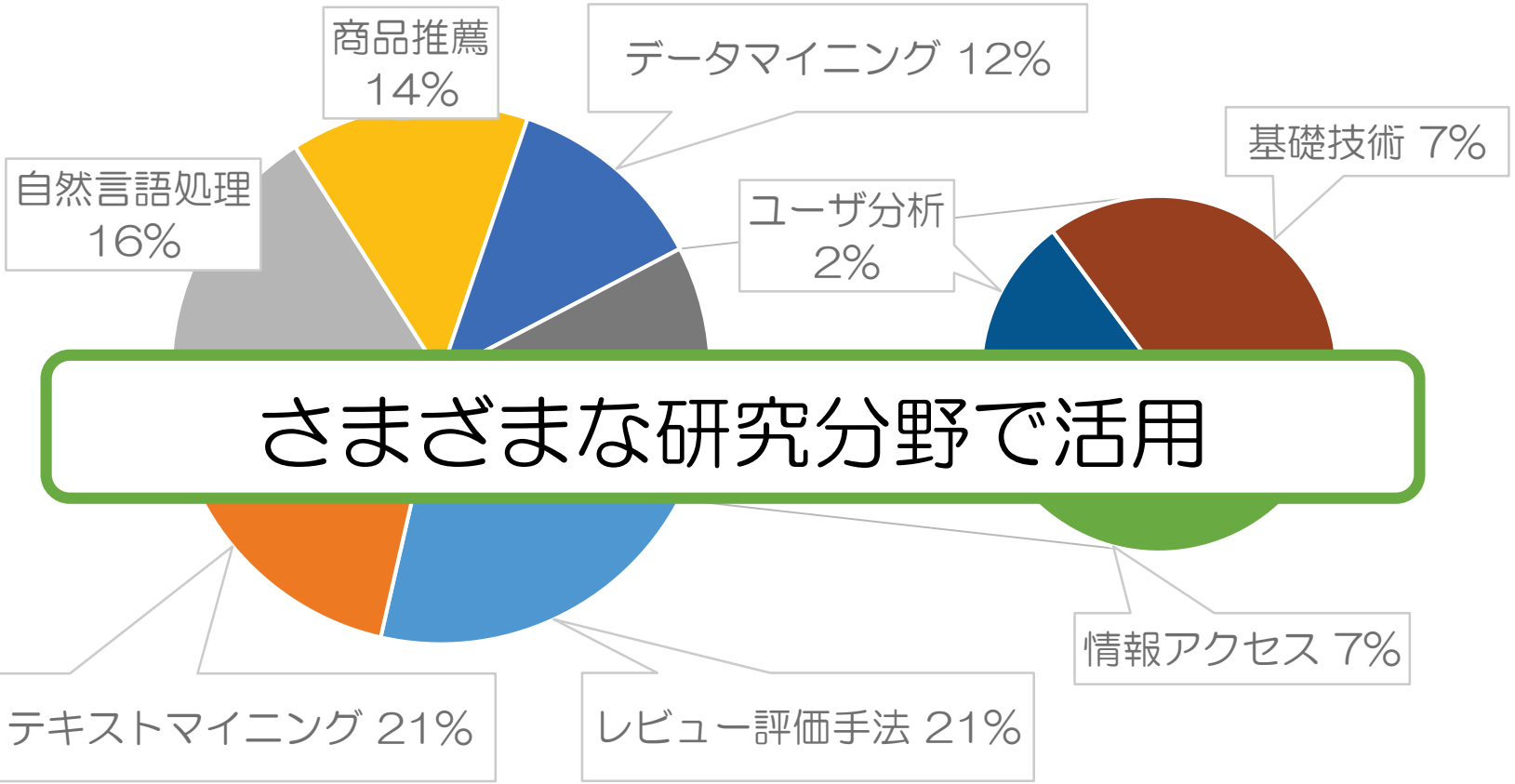
企業データを利用した研究発表（~2019年度）

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

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

イベント「IDRユーザフォーラム」

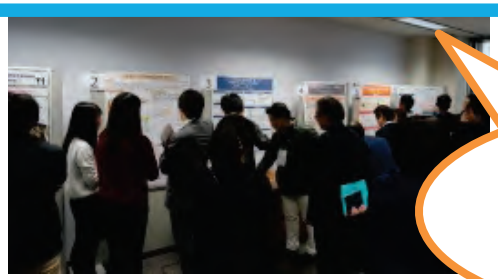
データを持つ
産学界
⇔ IDR ⇔
データを使いたい
研究者

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！

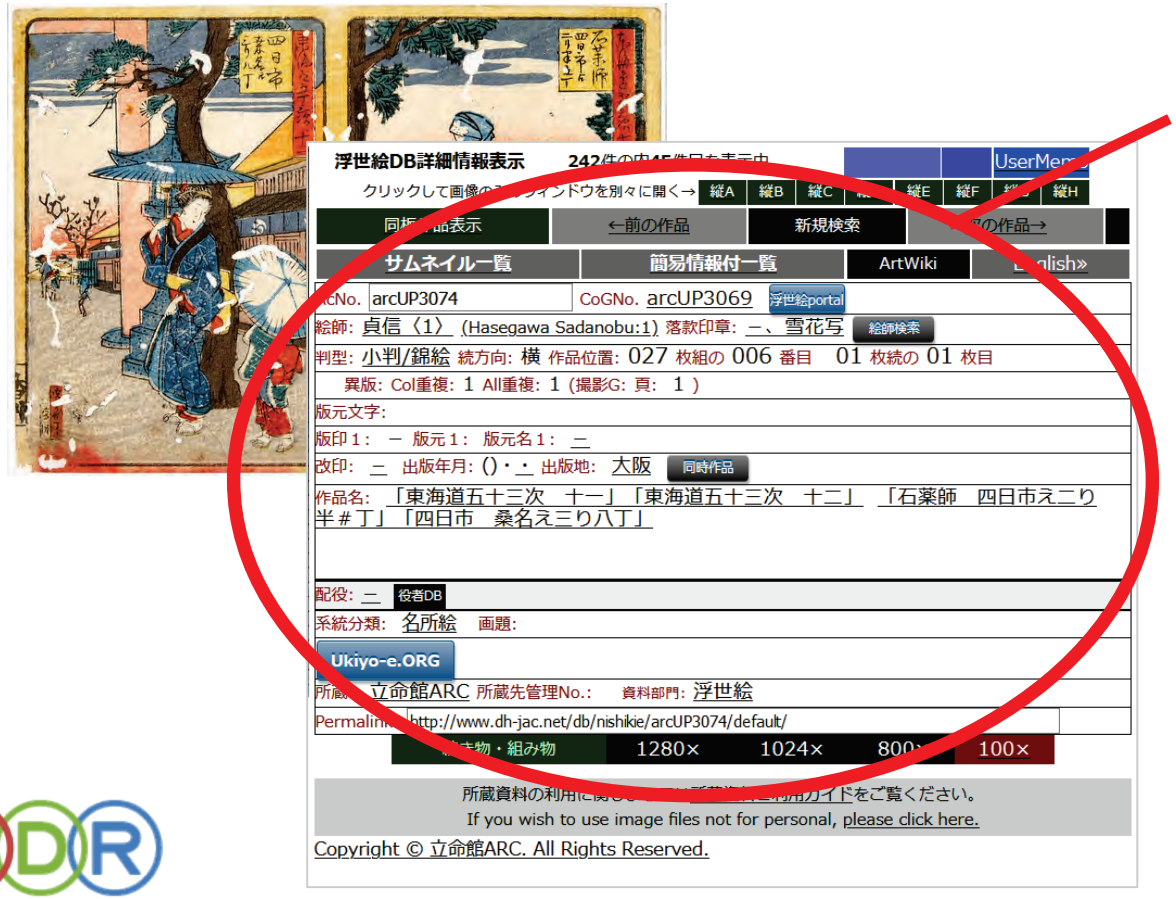
研究者等提供データセットの受入を開始

- 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

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



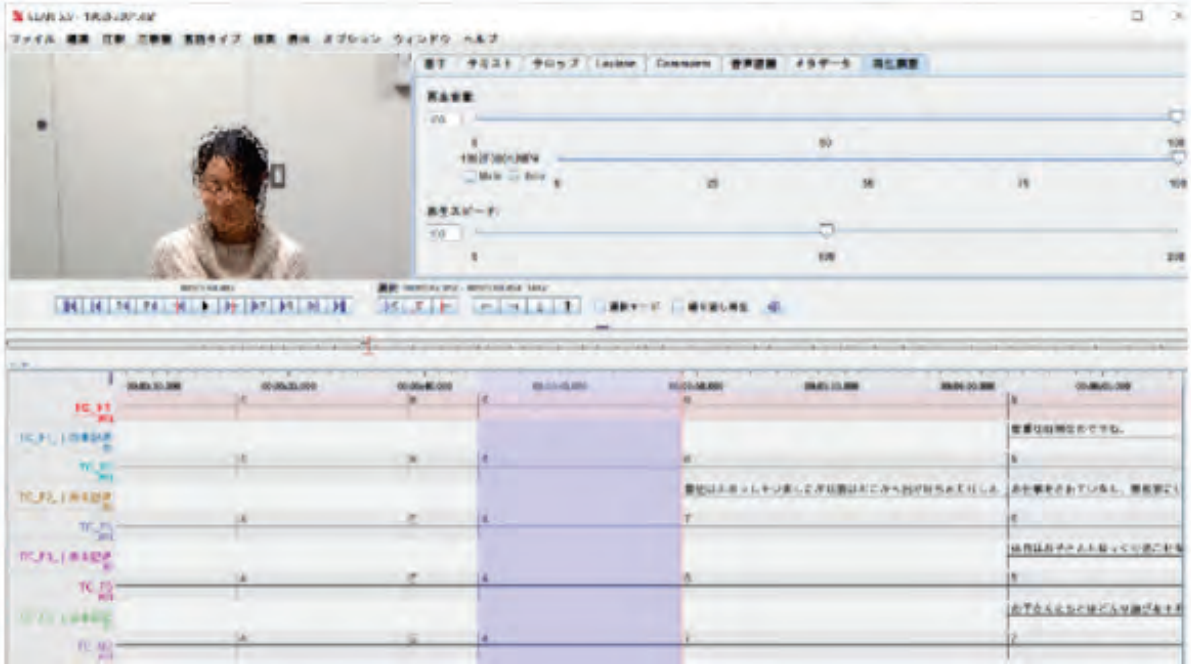
浮世絵の書誌情報
約11,000件
（画像データURL付き）

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

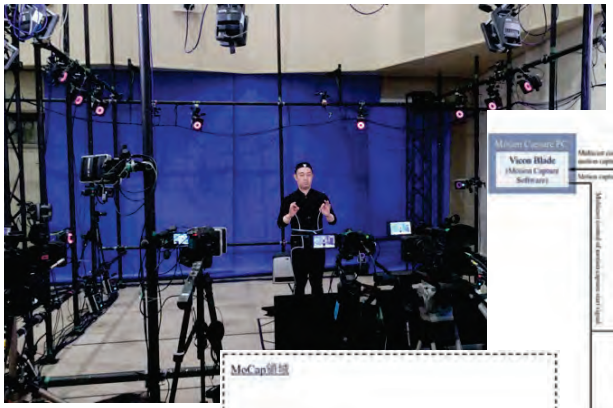


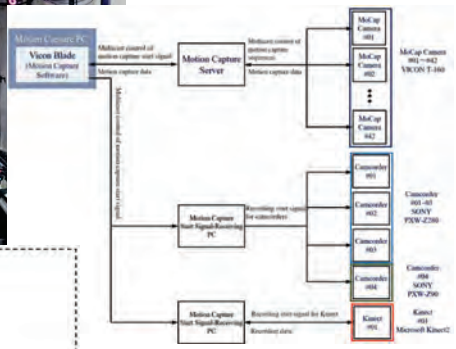
大阪大学のリリース（2020/10/20）より引用

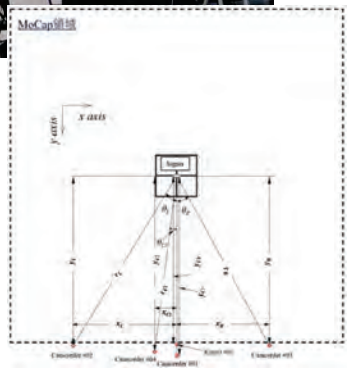
研究者等提供データセットの例（3）

- 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

NTCIR

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

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