研究用データセットのシェアリング文化を創る！

情報学データ資源の共同利用

Shared Use of Informatics Data Resources

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

どんな活動？

情報学研究に有用なデータを民間企業や大学等から受け入れて研究者に提供したり，データや課題を共有する評価ワークショップを実施しています。

何ができる？

データセットの共同利用を通じて，オープンサイエンス，オープンイノベーションの推進に貢献します。また研究コミュニティの創生や活性化を促進しています。

センターの活動内容

産学界のデータを学術研究目的で提供

オープンデータにできないデータを適正な管理の下に提供

詳細は隣のポスターへ

大学等

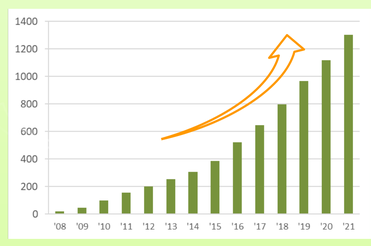
民間企業

NII-SRC

IDR

アカデミアの研究者

研究成果の公開

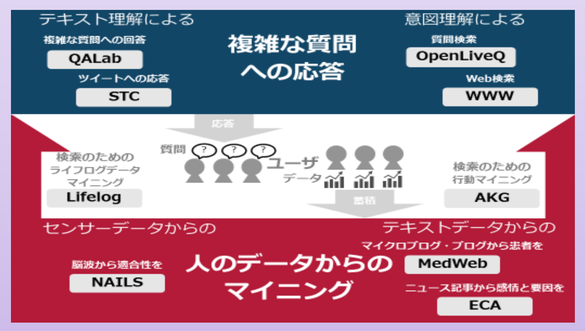


研究成果発表数の推移

評価型WSの企画運営

NTCIR

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



↑NTCIR-13の例

- 1年半サイクルのタスク
  - 回次ごとに複数のタスクを実施
  - 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

交流の場の提供

「IDRユーザフォーラム」の開催

↑データ提供企業のセッション

↑データ利用者によるポスター発表

今年度のイベントのご案内

IDRユーザフォーラム2023

2023年12月上旬開催予定

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

NTCIR-17 カンファレンス

2023年12月12～15日

https://research.nii.ac.jp/ntcir/ntcir-17/

詳細は **D05** のポスターへ

NII

連絡先：国立情報学研究所　データセット共同利用研究開発センター

URL：https://www.nii.ac.jp/dsc/　　Email：dsc@nii.ac.jp

IDR：情報学研究データリポジトリ

Informatics Research Data Repository

どんな活動？

企業活動により得たデータや、研究用に構築したテキスト・音声・映像データなど、
大量のデータを有する産学界と、データを使いたい研究者の橋渡しをしています。

民間企業提供のデータセット

～リアルビジネスで構築されたデータの研究利用～

Yahoo!知恵袋データ

2022年度提供版：
- 質問約247万件
  回答約649万件
- 投稿カテゴリ，
  ベストアンサーフラグ，
  投稿デバイス，など

https://chiebukuro.yahoo.co.jp/

楽天データセット

- 楽天市場
- 楽天トラベル
- 楽天GORA（ゴルフ）
- 楽天レシピ
- アノテーション付きデータ
  （評価極性タグ付きレビューなど）

LIFULL HOME'Sデータ
アットホームデータ

不動産物件データ
- 賃料，間取り，築年，
  立地，諸設備 など
- 月次掲載情報
- 間取り図画像や室内写真

https://www.homes.co.jp/

https://www.athome.co.jp/

メルカリデータセット

フリマ商品データ
- 出品状態（出品中/取引中/売切）
- 品名，ブランド，商品の状態
- 販売価格，発送方法 など
- いいね！やコメントの数
- コメント
- 画像（写真）

（※1年分の取引データ）

https://www.nii.ac.jp/news/release/2022/0831.html

クックパッドデータ

https://cookpad.com/

- レシピデータ
  - タイトル
  - 材料
  - 手順
  - コツ，ポイント
  - 生い立ち
  - つくれぽ
- 献立データ
  - 主菜／副菜

JASTメディカルデータ

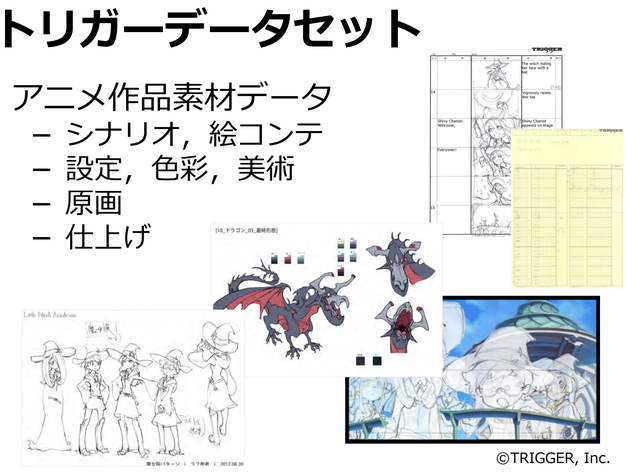
レセプト集計データ
- 疾病ごとの患者数や医療費
- 性別・年代・都道府県別

項目名称（和名）	変数名称	項目説明
診療年月	medtreat_month	201804、201805、201806…
ICD-10 章分類	icd10_dai	
ICD-10 中間分類	icd10_chu	ICD-10に準拠した設定値
ICD-10 3桁分類	icd10_sho	
性別	sex_type	01_男性、02_女性、99_その他
年代	age_kbn	01_9歳以下、02_10代…
医療機関都道府県	pref_type	01_北海道、02_青森県、03_岩手県…
レセプト件数	rezept_count	当該傷病が記載されたレセプト件数
患者数	patient_count	当該傷病と診断された患者数
主傷病寄せ医療費	syu_medical_cost	当該傷病を主傷病とした際の医療費
主傷病フラグ	syu_flag	主傷病として判定されたレセプトが1件以上ある場合のみ"1"

https://www.nii.ac.jp/news/release/2022/0114.html

トリガーデータセット

アニメ作品素材データ
- シナリオ，絵コンテ
- 設定，色彩，美術
- 原画
- 仕上げ

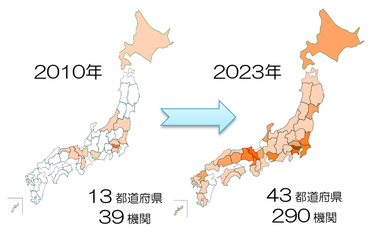


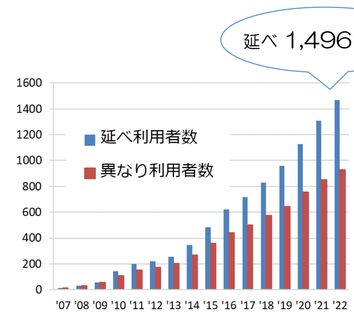
©TRIGGER, Inc.

他にも…

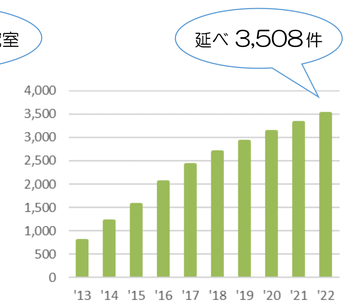
- ニコニコデータセット
- リクルートデータセット
- 不満調査データセット
- Sansanデータセット
- インテージデータセット
- オリコンデータセット
- ダイエットロコミデータセット
- 弁護士ドットコムデータセット
- みんなの評判口コミデータセット
- 地球の歩き方旅行記データセット

＜提供実績＞（～2023.3）





研究室単位提供データセット

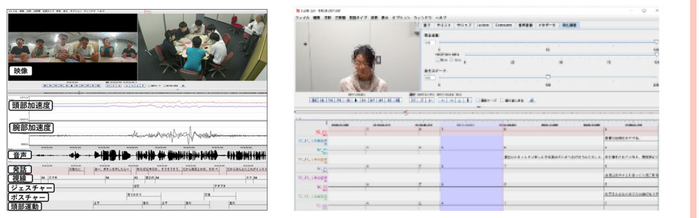


個人単位提供データセット
（ニコニコ，不満辞書，Sansan）

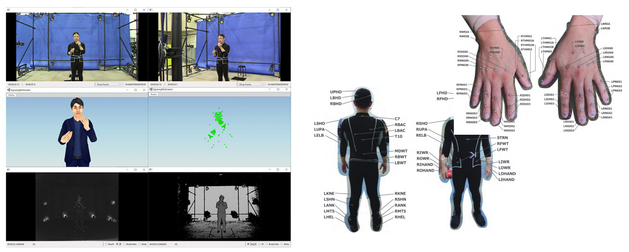
研究者構築のデータセット

～研究用に構築された音声/映像コーパス，テストコレクションの研究利用～

- グループコミュニケーションコーパス
- 大阪大学マルチモーダル対話コーパス

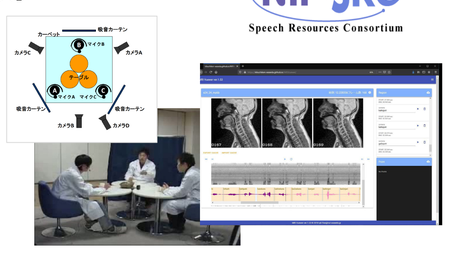


- 工学院大学 多用途型日本手話コーパス



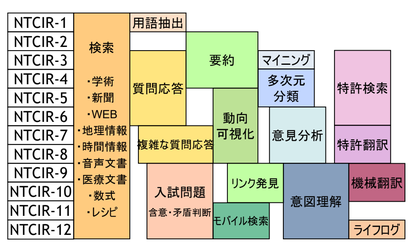
- 立命館ARC所蔵浮世絵データベース
- 理研記述問題採点データセット

音声コーパス（50種類）



- 読み上げ／講演／演技／対話
- 単語／短文／長文
- 成人／幼児／高齢者
- ナレータ／声優／非母語話者
- 雑音下／残響下／車内
- 方言，多言語，感情音声 など

NTCIRテストコレクション



過去のタスクで構築されたデータセットを，研究目的用に参加者以外にも配布中

- **タスクデータ（35種類）**
  - 検索課題と正解判定
  - 質問と解答
- **WEB文書データ（2種類）**
- **特許，医療文書，議会会議録** など

データの入手をご希望の方は…

それぞれ利用申請 → （審査） → 契約手続きが必要です。まずはIDRサイトの案内に従い利用申請をしてください。

NII　IDR　検索

データのご提供もお待ちしています

実情に応じた提供方法をご提案致します。まずはご相談下さい。

メール（ IDR事務局）：idr@nii.ac.jp

連絡先：国立情報学研究所　IDR事務局

URL：https://www.nii.ac.jp/dsc/idr/　　**Email：idr@nii.ac.jp**