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

NII

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

大学等研究者作成のデータセットの受入
- テキストコーパス
- 音声コーパス
- 映像コーパス
- アノテーションデータ

など，受入要項を公開

（※民間企業データは個別相談）

大学等

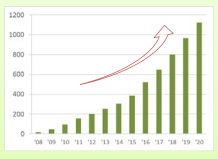
民間企業

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

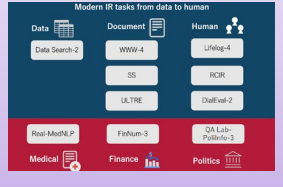
https://dsc.repo.nii.ac.jp/

NTCIRの詳細は左のポスターへ

NTCIR-16 カンファレンス
2022年6月14～17日
※今月開催！
https://research.nii.ac.jp/ntcir/ntcir-16/

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

交流の場の提供 「IDRユーザフォーラム」

← 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 →

提供中のデータセットの具体例は右のポスターへ

IDRユーザフォーラム2022
2022年11月頃開催予定
https://www.nii.ac.jp/dsc/idr/userforum/

IDR：情報学研究データリポジトリ

Informatics Research Data Repository

NII

民間企業提供のデータセットの例

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/

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	mecitreat_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
主傷病告知医療費	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"1"

トリガーデータセット

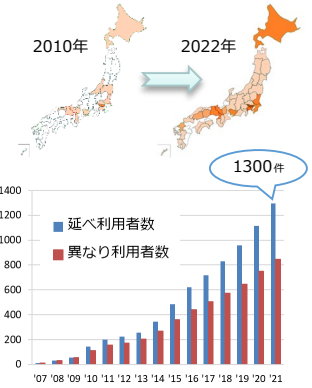
アニメ作品素材データ
- シナリオ，絵コンテ
- 設定，色彩，美術
- 原画
- 仕上げ

©TRIGGER, Inc.

IDR

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

＜提供実績＞（～2022.3）



（研究室単位提供データ）

研究者構築データセットの例

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

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（46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

NII-SRC
Speech Resources Consortium