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，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

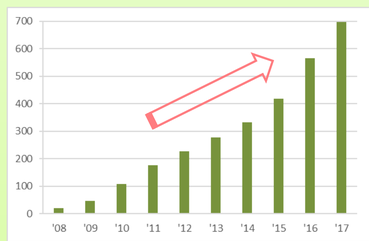
民間企業

NII-SRC

IDR

アカデミアの研究者

研究成果の公開

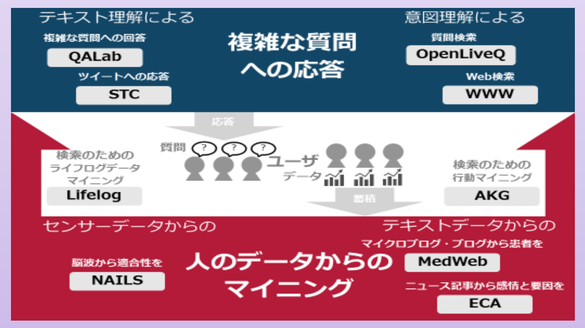


研究成果発表数の推移

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

- 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」の開催

↑データ提供企業登壇のパネル

↑データ利用者のポスター発表

今年度のイベントのご案内

日本DB学会 データチャレンジ

2019年5月～9月 ※参加受付中！

http://dbsj.org/events/project/data_challenge_2019/

NTCIR-14 カンファレンス

2019年6月10日～13日 ※来週開催！

http://research.nii.ac.jp/ntcir/ntcir-14/

IDR ユーザフォーラム2019

2019年11月29日 開催（予定）

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

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!データセット

- 「Yahoo!知恵袋」の質問とその回答

  2019年度提供版：
  - 質問約264万件，回答約611万件
  - 投稿カテゴリ，ベストアンサーフラグ，投稿デバイスなど

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
- 各種アノテーション付きデータ
  - 評価極性タグ付きレビューデータなど

Column	Sample
投稿者	user1
年齢	30
性別	1
商品コード	rakutenstore:123456
商品名	【Rakuten Mug】楽天ロゴ入り マグカップ
店舗名	ラクテンストア
商品URL	rakutenstore/123456/
商品ジャンルID	215903
商品価格	1500
購入フラグ	1
内容	実用品・登録使い
目的	自分用
頻度	はじめて
評価ポイント	4
レビュータイトル	ペアで買いました！
レビュー内容	注文してから2日で届きました。思ったよりも軽くて持ちやすく、使いやすいです。ペアで2つ購入したのですが、2つとも同じ色なので他の色が出ればいいなと思いました。
レビュー登録日時	2010-01-01 0:00:01

ニコニコデータセット

- ニコニコ動画コメント等データ
  - 動画1,670万件に対するメタデータとコメント約38億件
- ニコニコ大百科データ
  - 投稿記事データ，掲示板データ

リクルートデータセット

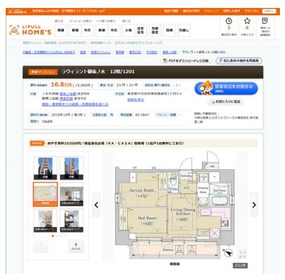
- 「ホットペッパービューティ」の掲載データ
  - 美容室約1万件，ブログ約180万件
  - スタイリスト，メニュー，セットメニュー，クーポン
  - ロコミデータ約36万件

クックパッドデータセット

- レシピデータ
  - 172万件のレシピ，つくれぽ （画像も追加提供予定）
- 献立データ
  - 献立タイトル，各主菜/副菜のレシピ

LIFULL HOME'Sデータセット

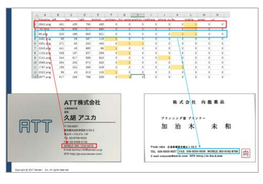
- 賃貸物件スナップショットデータ
  - 2015年9月時点の全国約533万件
- 高精細度間取り図画像データ
- 賃貸・売買物件月次データ（2年分）



不満調査データセット

- 「不満買取センター」への投稿データ
  - 約525万件の不満データ，投稿者の属性情報
- カテゴリ別不満特徴語辞書
  - 投稿カテゴリごとの特徴的な単語リスト（約190万単語）

Sansanデータセット



- サンプル名刺データ
  - データ分析コンテストで使用されたダミーの名刺画像約3,500枚
  - 画像中の領域の座標とその項目ラベル

インテージデータセット

2018年6月
提供開始！

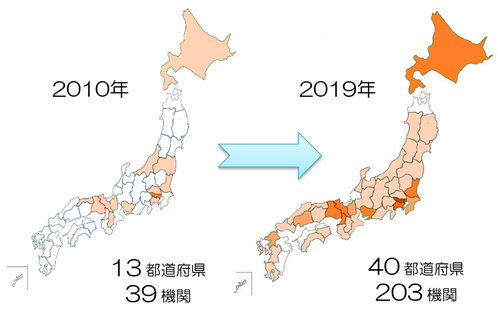
- i-SSPデータ
  - 同一モニターから収集した，テレビ・PC・スマートフォンでのメディア（TV番組・TVCM・webサイト・アプリ）接触データ
  - 関東 約6,000人1か月分，関東＋関西＋中京 約9,500人3か月分

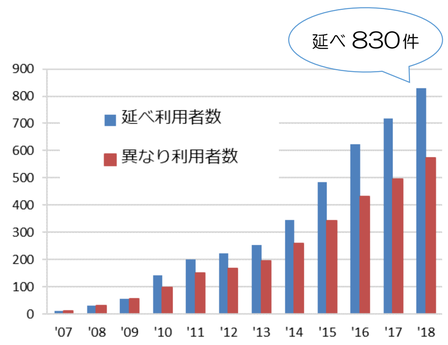
○○○○データセット

- 満足度アンケート調査データ

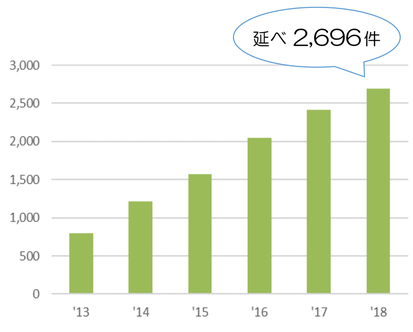
近日提供
開始予定！

＜提供実績＞（～2019.3）





研究室単位提供データセット



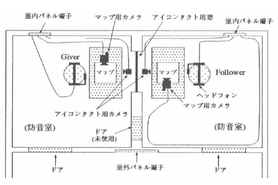
個人単位提供データセット
（ニコニコ，不満辞書，Sansan）

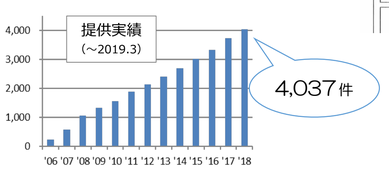
研究者構築のデータセット

～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～

音声コーパス

43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，プロ
- 方言，多言語，感情音声
- 雑音下，残響下，などなど…





※音声資源コンソーシアム(SRC)

SRCサイト：http://research.nii.ac.jp/src/

映像コーパス（準備中）

- **手話コーパス**

  日本各地の手話表現を集めた映像データ

  ※現在のところ
  茨城・群馬・富山・石川
  奈良・福岡・長崎にて収録

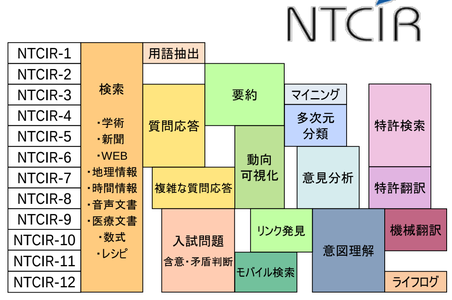
日本手話話し言葉コーパスプロジェクト
http://research.nii.ac.jp/jsl-corpus/public/

- **会話映像コーパス**
  - サイエンスコミュニケータと来館者との対話映像データ
  - 対話システムとの対話実験映像データ　など
- **グループディスカッションコーパス**（近日公開予定）

NTCIRテストコレクション

情報アクセス技術の評価を行うワークショップ

→過去のタスクで構築されたテストコレクションを一般研究目的用に配布



- **タスクデータ32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**

NTCIRプロジェクト： http://research.nii.ac.jp/ntcir/

※NTCIR-14 の最終成果報告会（国際会議）を来週開催！

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

NII

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