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

Center for Dataset Sharing and Collaborative Research

大山敬三，神門典子，佐藤真一，宮尾祐介，山岸順一，大須賀智子

どんな活動？

情報学の研究では大量の実データが欠かせませんが、本当に欲しいデータはなかなか手に入りません。そこで、データセット共用プラットフォームを構築して、企業などが様々なデータを提供しやすく、研究者にも使いやすい、循環型の環境づくりを目指します。

何ができる？

データセットの共同利用を通じて、研究の効率化を図るとともに、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。またデータや課題を共有する評価ワークショップを実施し、技術の深化とコミュニティの創生や活性化を促進します。

センターの活動内容

評価型ワークショップ(国際会議)の企画運営

NTCIR

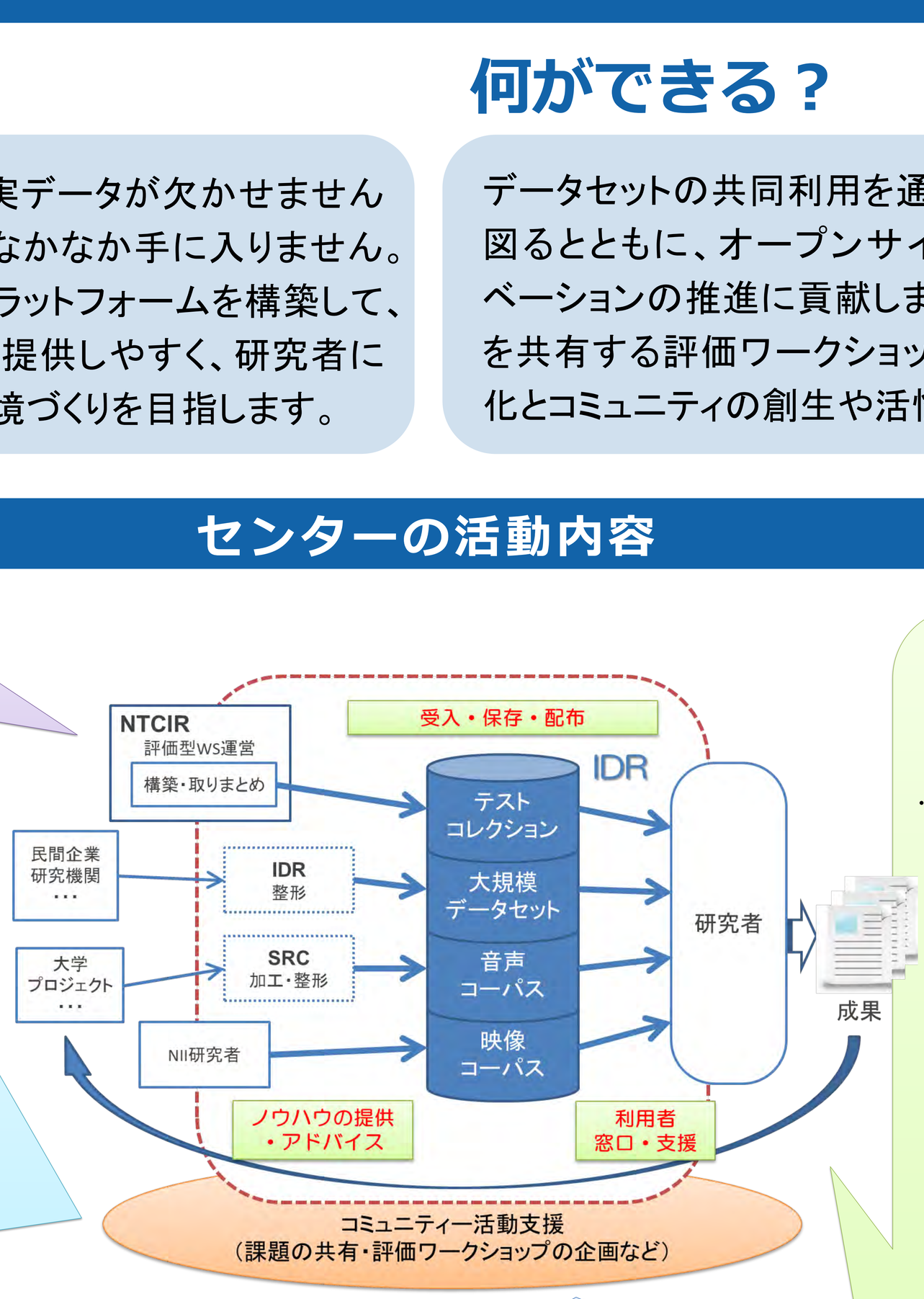
詳細は隣のポスターC12へ

ユーザフォーラム開催

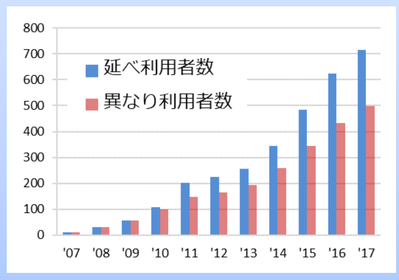
データ提供者と利用者が一同に会し意見交換できる場の提供

↑データ提供企業登壇のパネル

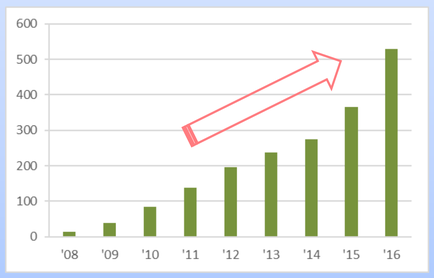
↑データ利用者のポスター発表



産学界のデータの研究者への提供



利用者数（研究室単位）の推移



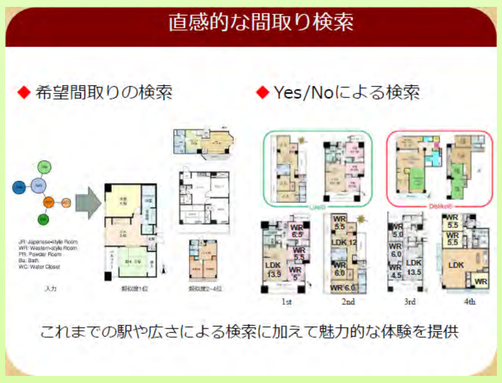
研究成果発表数の推移

研究成果の公開

利用者による発表論文一覧をJAIRO Cloud 上で公開中

・楽天レシピデータを用いた研究の例
（関西学院大学・角谷研究室
& 九州大学・牛尼研究室）

・HOME'Sデータを用いた研究の例
（東京大学・山崎研究室）



今年度の開催イベントのご案内

JOSS2018 -セッションB4

2018年6月19日 開催（済）

http://joss.rcos.nii.ac.jp/

IDR ユーザフォーラム2018

2018年11月開催（予定）

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

NTCIR-14 ※タスク参加者募集中！

（カンファレンス：2019年6月開催）

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

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：http://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？

情報学の最新の研究分野では、音声や映像、Web上にあるテキストなど、大量のデータを必要としています。
IDRでは、これらのデータを持っている産学界と、データを使いたい大学等の研究者の橋渡しをしています。

現在提供中のデータ

Yahoo!データセット

Q&Aサイト「Yahoo!知恵袋」に投稿された質問とその回答
- 質問約1,600万件
- 回答約5,000万件
（ベストアンサー，それ以外）
- 投票数，評価数など

http://chiebukuro.yahoo.co.jp/

- **Yahoo!データセット**
  - Q&Aサイト「Yahoo!知恵袋」の投稿データ

ニコニコデータセット

・「ニコニコ動画」約1,400万件のメタデータとコメント約35億件

http://www.nicovideo.jp/

・「ニコニコ大百科」に投稿された記事データとその掲示板データ

http://dic.nicovideo.jp/

- **ニコニコデータセット**
  - ニコニコ動画のメタデータ＆コメントデータ
  - ニコニコ大百科データ

クックパッドデータセット

・「クックパッド」に2014年9月までに掲載された約172万品のレシピ
- タイトル
- 材料
- 手順
- コツ・ポイント
- このレシピの生い立ち
- つくれぽ
- カテゴリ　など

・上記レシピらからなる献立のタイトル，主菜/副菜などのラベル

http://cookpad.com/

- **クックパッドデータセット**
  - レシピデータ，献立データ

不満調査データセット

・「不満買取センター」に投稿された約25万件の不満に関するデータ
・それから作成したカテゴリごとの辞書データ

http://fumankaitori.com/

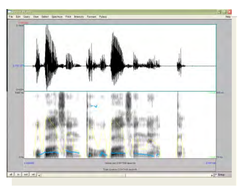
- **不満調査データセット**
  - 不満買取センターへの投稿データ
  - カテゴリ別不満特徴語辞書データ **UPDATE!!**

音声コーパス

※音声資源コンソーシアム(SRC)より提供中

43種類の音声データベース
- 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…

SRCサイト：http://research.nii.ac.jp/src/



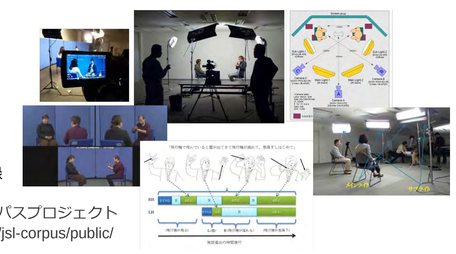
- **音声コーパス** 随時更新中！

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/



- **映像コーパス**
  - 日本手話話し言葉コーパス（準備中）
  - 会話映像コーパス（準備中）

- **NTCIRテストコレクション** 随時更新中！
  - 検索課題･正解／質問･解答データ
  - 文書データ（Webアーカイブ）

楽天データセット

「楽天市場」の約1億5,600万の商品データと約6400万のレビューデータ，
その他，楽天トラベルや楽天レシピ，アノテーションデータなど複数

Column	Sample
商品名	【Rakuten T-Shirt】 楽天ロゴ入り 着心地満点の Tシャツ　様を綺麗に合わせない 幹ぜる生地とデザイン
商品コード	rakutentore:12345678
商品価格	4500
商品説明文	上質の素材を使用し、シルエットにも木を使ったデザインです。ストリートだけでなく、有名百貨店でも取り扱われるようになりました。人気急上昇のTシャツです！ベーシックでなおスマートな為、年齢、性別を問わず楽しんでいただけるデザインになります。すこしセクシーに、すこしカジュアルにもコーディネイトに最適です。飽きやすいTシャツだけでもなく、着心地にもこだわりました。
販売方法別説明文	セール価格含む
商品 URL	r/12345678
商品画像 URL	http://im.api.rakuten.co.jp/00_0mall/rakutentop/r/12345678.jpg
レビュー件数	10
レビュー平均	4.2
店舗コード	rakutentore
ジャンル ID	4038462

http://rit.rakuten.co.jp/opendataj.html

Column	Sample
投稿者	user1
年齢	30
性別	1
商品コード	rakutentore:1234456
商品名	【Rakuten Mug】 楽天ロゴ入り マグカップ
店舗名	ウケテンショップ
商品 URL	rakutentore/1234569/
商品ジャンル ID	210302
商品価格	1600
個人フラグ	1
内容	定価品・管理使い
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	ペアで買いました！ 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。ペアでよっ購入したのですが、2つ共も同じ方なので色の色が別れいいなと思いました。
レビュー登録日時	2010-01-01 00:00:01

（↓レシピ画像の例）

- **楽天データセット**
  - 楽天市場の全商品＆レビューデータ
  - 楽天トラベルの施設＆レビューデータ
  - 楽天GORAのゴルフ場＆レビューデータ
  - 楽天レシピのレシピ情報＆画像データ
  - PriceMinisterのレビュー有効性情報
  - アノテーション付きデータ **UPDATE!!**

リクルートデータセット

「ホットペッパービューティ」に掲載された全国の美容室，サロンのデータ
- 店舗データ（約1万件）
- クーポンデータ（約15万件）
- 店舗ブログデータ（約180万件）
- メニューデータ（約52万件）
- スタイリストデータ（約9万件）
- 口コミデータ（約36万件）

https://beauty.hotpepper.jp/

- **リクルートデータセット**
  - ホットペッパービューティデータ

LIFULL HOME'Sデータセット

不動産・住宅情報サイト「LIFULL HOME'S」に2015年9月時点で掲載されていた全国約533万件の賃貸物件データとその画像データ約8,300万枚
- 賃料，管理費
- 敷金，礼金，更新料
- 面積，間取り
- 立地（市区町村）
- 最寄り駅，徒歩分
- 築年数
- 建物構造
- 諸設備　など

※間取り図に関しては高精細度の画像データも追加提供開始

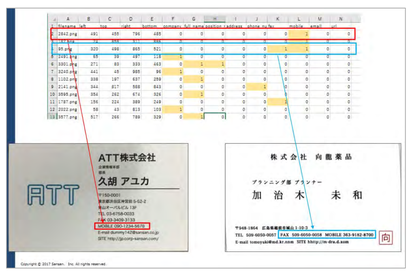
http://www.homes.co.jp/

- **LIFULL HOME'Sデータセット**
  - 賃貸物件データとその画像データ

Sansanデータセット

データ分析コンテスト「人工知能は名刺をどこまで解読できるのか」で使用されたダミーの名刺データ3,481枚
（※氏名等は実在しないサンプル）

各領域に以下のラベルが付与
- 会社名
- 氏名
- 役職
- 郵便番号，住所
- 電話・FAX・携帯番号
- E-mailアドレス
- HPのURL



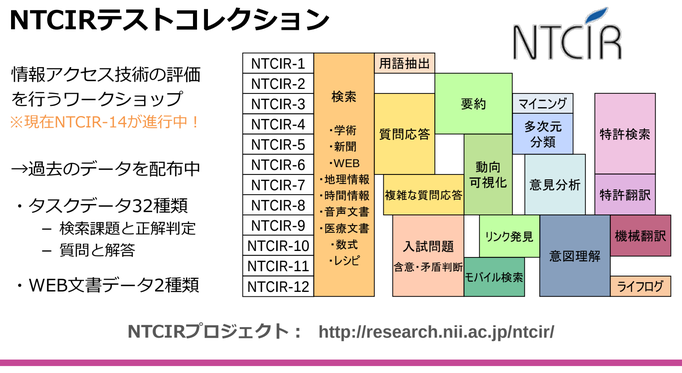
- **Sansanデータセット**
  - ダミーの名刺画像データ

※新規データセット近日リリース予定

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
※現在NTCIR-14が進行中！

→過去のデータを配布中
- 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/

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：http://www.nii.ac.jp/dsc/idr/　　Email：idr@nii.ac.jp