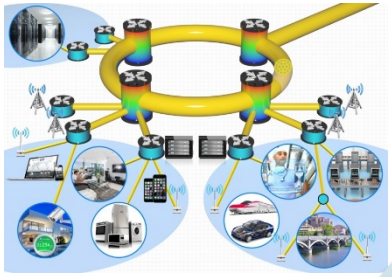
資料１６－４

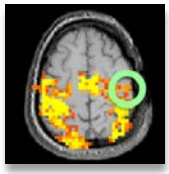
データ利活用を中心とした研究開発の課題と対策

国立研究開発法人
情報通信研究機構

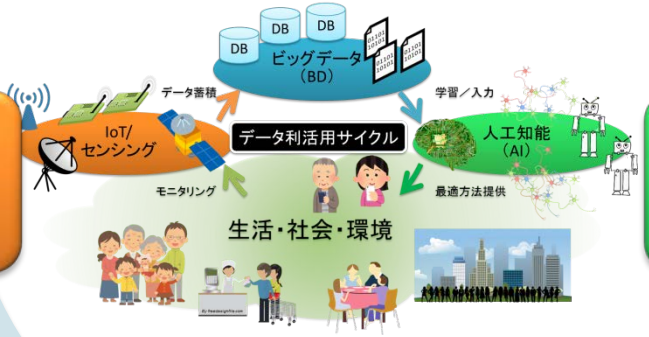
データを利活用する研究開発の概要

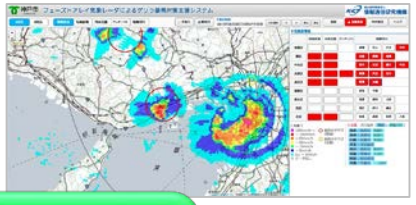
NICTでは、データの「取得・収集」から「流通・管理」、「統合・分析・情報抽出」、「提供・利用」までの各フェーズにおいて研究開発を行い、それらの連携にも取り組んでいる。





フェーズドアレイ気象レーダー





第14回技術戦略委員会NICTプレゼン資料「データ利活用を中心とした研究開発の現状及び課題」より。
各研究の現状については同資料を参照

ICSU WORLD DATA SYSTEM

データを利活用する研究開発についての課題

第14回技術戦略委員会NICTプレゼン資料「データ利活用を中心とした研究開発の現状及び課題」から

産学官の様々な機関が取得するデータは、現在、加速度的に増加している。NICTでも宇宙・地球から脳・微生物、実環境、社会活動に亘る多種多様な対象についてデータを取得・収集している。これらを融合的に活用できれば、学術、産業、公共といった分野で新たな価値の創造につながることが期待される。これを踏まえると、産学官連携で取り組むべき課題として、例えば次のようなものが挙げられる。

- **多種多様なデータとニーズを結び付ける有用なデータベースの構築**が必要
- データは取得のみならず前処理・解析においても各分野の専門的な知見の活用が不可欠であり、**専門性の高いデータを誰でも利用できる方策の検討が必要**
- 出来るだけ多くの良質なデータを共有し、社会全体としてイノベーションを創発するために、**データのオープン化**や**分野・データ毎の特性を踏まえた検討が必要**
- データの利活用にあたっては、**セキュリティ、プライバシー、権利関係等についても技術的、制度的な検討が必要**
- これら**高度なデータの利活用に対応する人材の育成が必要**

課題の全体像イメージ

MRI

データ

（科学データ
ソーシャルデータ
Webデータ）
など

- 多種多様性
- 異なる品質
- オープン性
- 一貫性
- 持続性

など

- データマッチング
- 統合化
- ノイズ処理
- フォーマット統一
- 目的に応じた専門知識、ノウハウの活用

など

データ解析

- ビッグデータ解析
- 画像解析
- 音声処理
- 自然言語処理
- 機械学習

など

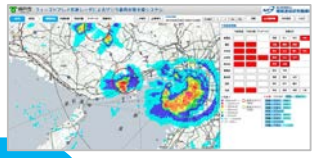
- ビジュアリゼーション
- ユーザマッチング
- サービス品質
- アクセシビリティ

など

利用者

- ニーズに対して最適なサービスの提供
- ユーザ行動支援
- 有効な選択肢の紹介
- 新たな発想の提供

など



ICSU
WORLD DATA SYSTEM

データのセキュリティ、プライバシー、権利関係、秘匿への対応

高度なデータの利活用に対応する人材の育成

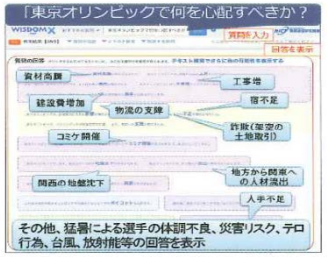
課題への対応方策について(1)

課題:「多種多様なデータからニーズとシーズを結び付ける有用なデータベースの構築が必要」

対策例

■　社会知解析技術

社会に溢れる膨大な情報や知識のビッグデータ(社会知)を基に、誰でも専門的な知識に容易にアクセスして様々な意思決定において有用な知識を得るための技術を研究開発します。インターネット上の災害に関する社会知を、各種の観測情報とともに分かりやすく整理し、リアルタイムに提供する基盤技術の研究開発も行います。

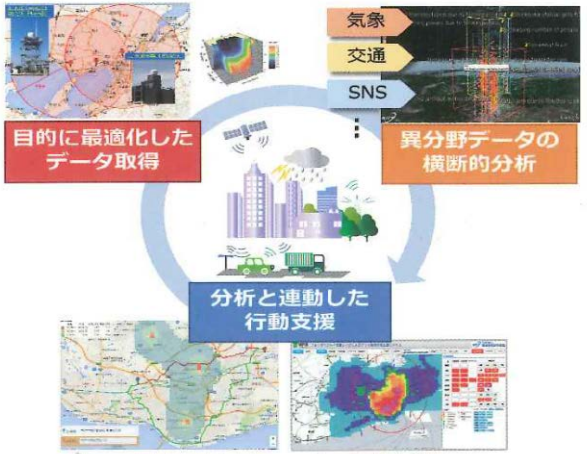


現在公開中の大規模 Web 情報分析システム WISDOM X
http://wisdom-nict.jp/

現在公開中の対災害 SNS 情報分析システム DISAANA の実行結果例
http://disaana.jp/

■　実空間情報分析技術

様々なセンサーやデバイスから取得された環境データや社会生活に関連するソーシャルデータなどの実空間情報を横断的に収集・分析し、ゲリラ豪雨や環境変化による交通等の具体的社会システムへの影響をモデルケースとして分析できるようにするためのデータ取得・解析技術やデータマイニング技術の研究開発を行います。また、分析結果をフィードバックし高度な状況認識や行動支援を行うことで社会システムの最適化・効率化を実現するための基盤技術を開発・実証します。

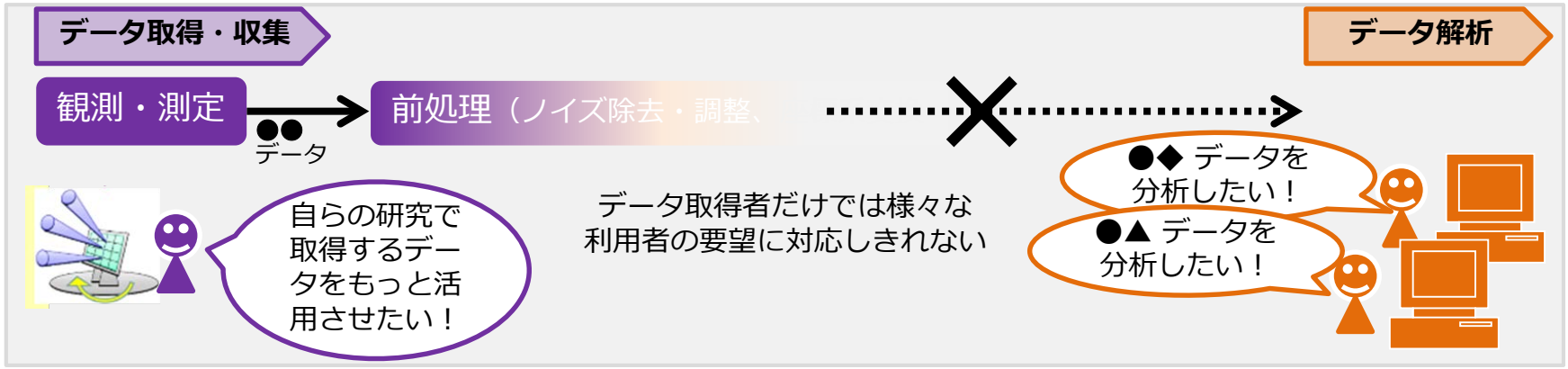


環境リスクを軽減する経路案内　　ゲリラ豪雨対策支援システム
実空間情報分析技術

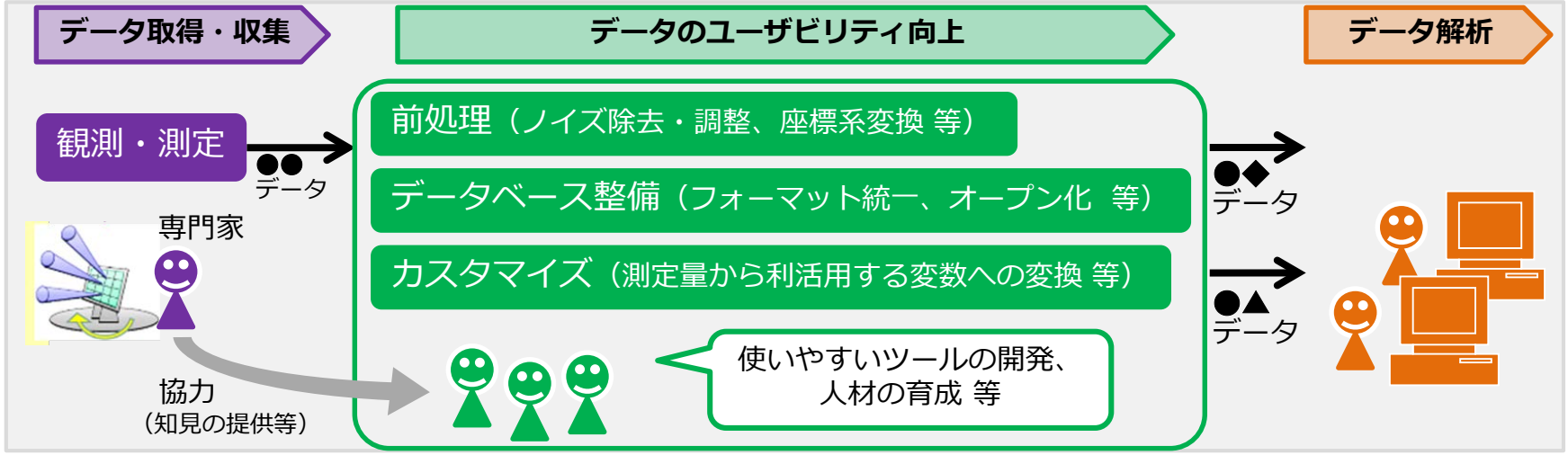
課題への対応方策について(2)

課題:「データは取得のみならず前処理・解析においても各分野の専門的な知見の活用が不可欠であり、
専門性の高いデータを誰でも利用できる方策の検討が必要」

現状



対策



データビリティ

ビッグデータの利用に関する
- sustainability
- responsibility

\+
- usability

↓

ビッグデータ利用の促進

課題への対応方策について(3)

課題:「出来るだけ多くの良質なデータを共有し、社会全体としてイノベーションを創発するために、データのオープン化や分野・データ毎の特性を踏まえた検討が必要」

言語情報データや脳情報モデル等の共有利用に関して、NICTでは「最先端AIデータテストベッド」の構築(→「参考」参照)や、言語情報データに係る産学官のフォーラム(ALAGIN)活動を推進

対策

プライバシー性

高

低

無

データ取得・収集

- 脳機能計測
- 音声コーパス収集

- Web・SNSデータ収集
- サイバー攻撃観測
- IoTセンサーによるデータ取得

- 航空機搭載レーダーによる地表画像取得
- 衛星や地上レーダーによる気象・環境観測
- 宇宙環境計測(太陽、電離圏)

↑NICTの例

データのユーザビリティ向上

- データ取得者(専門家)の協力を得つつ、データの扱いやすさ・有用性を向上
- データベースの構築とオープン化

人材連携

データ解析

AI手法等による最適化やデータマイニング、現象の予測等

共同研究先など限定された範囲内にてデータを利用
(提供者の同意があるものや匿名加工情報は提供可)

分析結果の提供

社会に対して、最適解の提供や災害予測情報等の提供

企業・大学・研究機関

データ連携

アプリ開発

草の根によるデータ分析

参考：NICTにおけるAIデータテストベッドの推進

「人工知能技術戦略」から
（人工知能技術戦略会議　平成29年3月31日）

- AI技術の利活用にはデータが不可欠であり、**データそのものが競争力**となりつつある
- AI技術の技術開発にはデータが不可欠である。健康、医療、介護、交通、農林水産分野など**社会ニーズにつながっているデータの活用、環境整備を行っていく必要**がある
- また、データ自体だけでなく、データからAIで生成される学習済みモデルはより重要な価値を持つ。**学習済みモデルを流通させる仕組みを構築することは重要な課題**である
- 大学や研究機関でデータ整備・提供を行うことは大きな負担が伴う。必要とされるデータを見極め、効果的に**データ整備・管理を行う支援体制の整備・強化が必要**である

※ 一部抜粋の上、太字・下線はNICT加筆

このような認識のもとに推進

知能科学融合研究開発推進センター（AIS）の設立

- オープンイノベーション型のAI関連研究の戦略立案
- データのカタログ化等によるワンストップ窓口

今後の活動方針

- ◆ **「最先端AIデータテストベッド」構築を推進**
  言語情報データ、脳情報モデル等について、NICTの実証ネットワーク（JGN）を通じて全国規模で利用可能とし、研究開発と実証を加速
- ◆ データを切り口としたオープンイノベーション型プロジェクトの推進
- ◆ データ利活用の課題への対応

※ AIS：　AI Science R&D Promotion Center

課題への対応方策について(4)

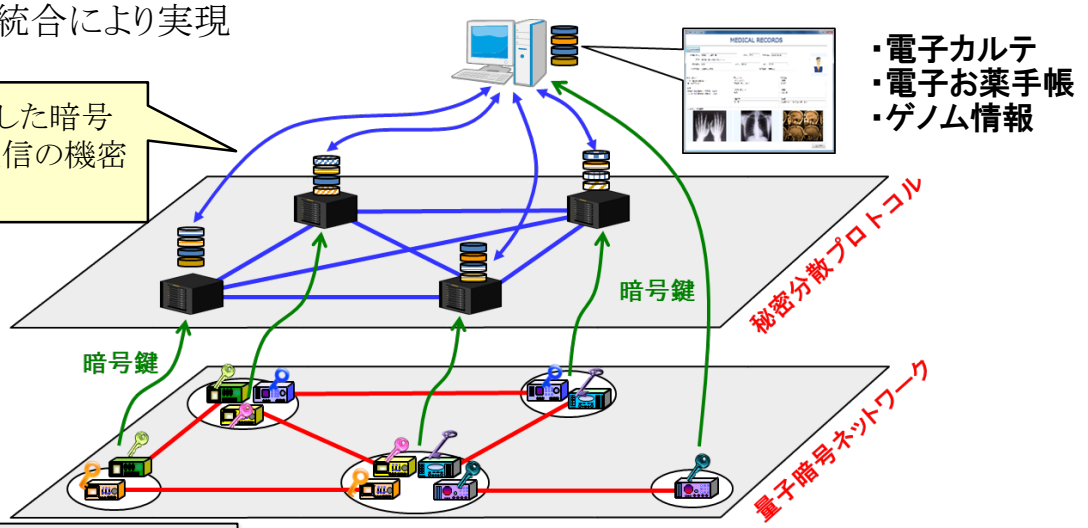
課題:「データの利活用にあたっては、セキュリティ、プライバシー、権利関係等についても技術的、制度的な検討が必要」

(1)超長期(世紀単位)にわたり重要データを安全に保管・活用するための技術の研究開発

- 脅威:データを盗聴・保存しておき、将来、高度な計算機で解読 “Store now, read later”
- 課題:脅威に対応したデータ通信とデータ保存技術の開発
- 対策例:秘密分散技術と量子暗号技術の統合により実現

- 計算量には拠らない、超長期(世紀単位)の機密性
- 分散シェアの一部が失われてもデータを復元可能
- 秘匿化したまま情報処理が可能

量子暗号で配送した暗号鍵によりデータ通信の機密性を確保



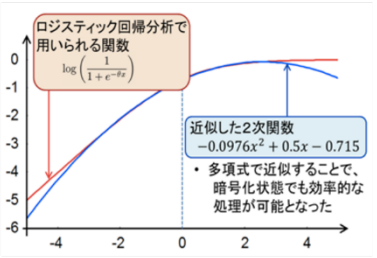
(2)IoT機器への搭載に適した暗号技術の研究開発

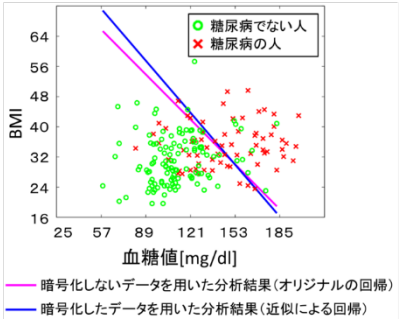
- 脅威:暗号化機能の無いIoT機器への不正アクセス、改竄、情報漏えい
- 課題:省電力、省スペースでも実装可能な鍵生成、暗号化技術の開発
- 対策例:IoT機器内の微細素子から真性乱数を生成する超小型の物理乱数源、公開鍵暗号・共通鍵暗号の軽量実装技術、これらの統合化技術の研究開発

(3)暗号化した状態での大量データ処理技術の研究開発

暗号化したままビッグデータ分類

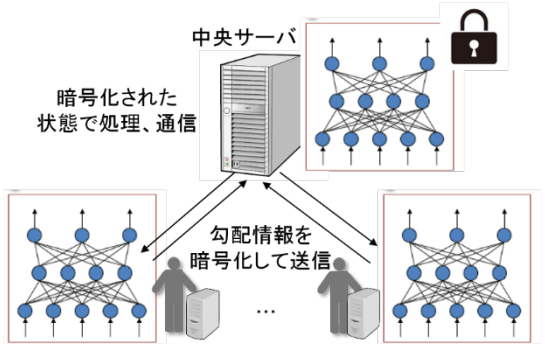
» ビッグデータ解析で多用されているロジスティック回帰分析をデータを暗号化したまま計算可能に





暗号化したままディープラーニング

» 多数の参加者が持つデータセットを　互いに秘匿したままディープラーニングを行うプライバシー保護深層学習システム



N人の参加者と中央サーバ1台による深層学習
(分散協調学習)

(4)改正個人情報保護法への対応

- 匿名加工情報・非識別加工情報の導入
- 匿名加工技術の評価技術(有用性指標と安全性指標)

など

（参考）AIデータ利用にあたってのルール（現行版概要）

● 禁止事項

①法令、条例又は公序良俗に反する利用 、②国家・国民の安全に脅威を与える利用、
③Webサーバに負荷を与える利用

● 第三者の権利侵害に関する注意

NICT以外の第三者が著作権その他の権利を有している場合があるため、特に権利処理済であることが明示されているものを除き、利用者の責任で当該第三者から利用の許諾を得ること。

また、外部データベース等とのAPI連携等により取得しているコンテンツについては、その提供元の利用条件に従うこと。

● 免責事項等

利用者がデータを用いて行う一切の行為（データを編集・加工等した情報を利用することを含む。）についての免責、公開データの完全性・正確性・網羅性・特定の目的への適合性等についての無保証、事前の予告なしのデータの変更・移転・削除等。

● 出典の記載

データ利用の際の出典の明記、データを編集・加工等して利用する場合の編集・加工等の追加明記等。

● 個別の利用条件

一部のデータに関し、利用の際に追加的な個別の制約条件（有償、物理上・組織上のアクセス限定、利用者の法人格、利用方法等を想定）がかかることがあるため、当該データの利用規約等の遵守を明記。

● その他

※ AISウェブサイト：
http://www2.nict.go.jp/ais/ais_policy.html

課題への対応方策について(5)

課題:「高度なデータの利活用に対応する人材の育成が必要」

(専門家育成の例)
データサイエンスの知見・技術を活用し、データの素性およびニーズを考慮しつつ、ツールの開発やデータベースの構築を行うなど、データのユーザビリティを向上させる人材の育成

上記の専門家以外に、一般の方に対するデータ利活用方法の普及も重要

- アイデアソン、ハッカソンを通じた知識や技術の交流
- OJTやテストベッド等による機会の提供
- 系統立てた習得方法の提供
- デザイン思考といった手法や、複雑性、曖昧性、不確実性などの観点を取り入れた対応が必要

など

参考情報

「知能科学融合研究開発推進センター（AIS）」の設立

※ AIS: AI Science R&D Promotion Center

- 知能科学融合研究開発推進センター（略称：AIS※）は、従来からNICTが蓄積してきたデータを含め、産学官が利用しやすい研究開発環境を整備するとともに、知能科学領域における次世代研究開発を推進するオープンイノベーション型の戦略的な研究開発推進拠点として本年4月に設立。
- 今後AISを中心として、NICTをはじめ様々な関係者が保有するデータや知見を集め、大規模脳情報データ利活用環境の整備をはじめとするAIデータテストベッド等を活用した実証に取り組むことでイノベーションの創出を推進。

知能科学融合研究開発推進センター

連携

- 統合ビッグデータ研究センター
- 総合テストベッド研究開発推進センター

ユニット内他センター

・AIデータテストベッド管理・運用（サーバー・ネットワーク管理等）
・AI関連以外の機構内ビッグデータ管理・運用

JGN　StarBED

○ 第4期中長期計画のビジョン「オープンイノベーション拠点機能の強化」を、AI分野において具体化
○ AI研究開発の戦略化による社会実装の推進と、経済成長への貢献

○ 企画立案機能
・広範な産学官連携・外部資金獲得等に向けたAI関連研究の戦略立案
・機構内AI関連研究の連携強化
○ 外部からのワンストップ窓口
・ 公開可能データ・設備の「カタログ化」
・ 社会実装に至るまでの「コンサルティング」

AI関連研究拠点

UCRI　未来ICT研究所　DIRECT　サイバーセキュリティ研究所　ASTREC　電磁波研究所　CiNet

・AI基盤技術研究開発
・データ利活用　基盤技術研究開発
・計測基盤技術研究開発
・テストベッド管理運用技術研究開発
・AI関連各種データ取得（生データ管理）

連携
- イノベーション推進部門
- グローバル推進部門
- デプロイメント推進部門

産学官連携強化
企業等　独法等　大学等

AIの社会実装（出口化）
自動運転　物流　食品・薬品　農業・建設土木　スマート工場

外部公開済みのデータ例①

言語資源	概要
文脈類似語データベース	約100万の見出し語それぞれに対して、Web文書上での出現文脈が最も類似している名詞最大500個を類似度とともに列挙したもの
動詞含意関係データベース	含意関係が成立している動詞のペア(52,689ペア)と含意関係が成立していない動詞のペア(68,819ペア)の計121,508ペアを列挙したもの
負担・トラブル表現リスト	「災害」「心理的ストレス」「アスベスト汚染」など社会活動に負荷を与えたり、マイナス効果をもたらす問題や障害に関係する表現、20,115件を収録したもの
上位語階層データ	上位下位関係抽出ツールによって日本語Wikipedia(2007/03/28版)から自動獲得した上位下位関係の上位語を人手で階層化したものであり、合計約69,000名詞句から成る階層的シソーラス
単語共起頻度データベース	各単語に対して、それとの意味的関連を表す共起スコアの高い単語を、スコアの高い順に、スコアとともに列挙したもの
日本語パターン言い換えデータベース	文の係り受け解析の結果を利用して、「AはBが豊富です」のような、一文中で任意の名詞AとBを結ぶパターンに対して、言い換えが可能な別のパターンを収集したもの
異表記対データベース	文字レベルの編集距離の近い、日本語の語句の異表記対(あるいは「表記揺れの対」)の正例と負例を集めたもの
日本語係り受けデータベース	大量の日本語文書を係り受け解析した結果から係り受け関係を抽出し、その頻度を収録したもの
基本的意味関係の事例ベース	約1億ページのWeb文書上において文脈の類似度が高い2語間の意味的関係を人手で分類し、ラベル付けした102,436語対を収録したもの
京都観光ブログの評価情報付与データ	「京都観光ブログ」と京都観光ブログの「評価情報付与データ」から構成される。「京都観光ブログ」は、日本語ブログ記事のデータベースである。京都観光を中心とした内容で、執筆者は47名、合計1041記事(平均約480字)から構成される。「評価情報付与データ」は「京都観光ブログ」に対して評価情報(評判・意見)が人手で抽出され、評価保持者、評価表現、評価対象などが付与されたデータ
実証実験コーパスを用いた言語モデルおよび辞書	大規模音声翻訳実証実験において収集された日英中韓4か国語の実利用音声データを書き起こした約17万発話を形態素解析処理したものから作成したNグラム頻度(4グラム)データおよび、音声認識に用いるための発音辞書

※ AISウェブサイト：
http://www2.nict.go.jp/ais/ais_data.html

外部公開済みのデータ例②

音声資源	概要
日英翻訳エンジン学習・評価用対訳コーパス	IWSLT（International Workshop on Spoken Language Translation）の2005年評価キャンペーンの日英翻訳で使用された基本旅行会話データセットに基づいて作られたコーパスで、翻訳機器学習用データ20,000文、評価用データ1,500文（日英対訳文）から構成
音声翻訳実証実験固有名詞対訳辞書	平成21年度「地域の観光に貢献する自動音声翻訳技術の実証実験」で採択された5つのプロジェクトにおいて、日・英・中・韓国語の固有名詞辞書を収集したものをNICTで整備した辞書
日中特許用語辞書	日中特許用語辞書を、日中特許対訳コーパスを元に、各種自然言語処理ツールを用いて自動構築し、最後に人手による修正作業を行って整備したもの
CNP用中国語解析モデル	オープンソースソフトウェアとして配布している係り受け解析器（A Chinese Dependency Parser、略称CNP）のための中国語解析用モデルパラメータ
JPO・NICT英日対訳コーパス	英語と日本語の対応する公開特許公報の対（パテントファミリー）をもとに、日本国特許庁（JPO）及びNICTが共同で作成したデータ
JPO・NICT韓日対訳コーパス	韓国語と日本語の対応する公開特許公報の対（パテントファミリー）をもとに、日本国特許庁（JPO）及びNICTが共同で作成したデータ
意見（評価表現）抽出ツール用モデル	オープンソースソフトウェアとして配布されている「意見（評価表現）抽出ツール」のための意見解析用モデルファイルと評価表現辞書から構成
日本語高齢者音声データベース	日本語を母国語とする60歳以上の話者の読み上げ音声
中国語音声データベース	中国各地域出身の母国語話者による中国語（普通話）読み上げ音声および自由発話音声
日本語音声データベース	ATRにおいて開発された、音素バランス文などの文や定形単語を発話内容とする、プロナレータによる多数話者日本語音声データベース
ノンネイティブ英語音声データベース	非母語話者の英語読み上げ音声
京都観光案内対話データベース	プロの観光ガイドと、旅行者を模した被験者の2名による対面対話を収録し、書き起こしたデータ
日本語小学生音声データベース	音響モデル学習用の、小学校1年生から4年生までの話者が読み上げた旅行会話及び音素バランス文章
日英・日中バイリンガル独話音声データベース	日英または日中のバイリンガルである声優または一般人が発声した音声コーパス
NICT声優対話コーパス	声優による、台本に基づいた掛け合いを収録した音声コーパス