4. その他の事項

4-3. 利用可能な大規模データセット

令和３年２月１５日 作成

Q13.　研究開発に必要なデータの取得・収集・管理について、問題となっている事項はありますか。

課題等の概要	課題対応等の詳細
● 機械学習用に研究者が自由に利用できるセキュアなデータベースが必要。 ● 広範に利用できる大規模データセットを、各研究室レベルで準備できない。	● AI Japan 中核 3 センター（産総研 AIRC・理研 AIP、NICT）が保有する公開中または公開可能な AI 研究用の汎用データは、欄外「AI 関連公開データリスト」のとおりです。 ● 上記のデータセットについては、各中核センターのポータルサイトで公開されています。また、AI Japan ウェブサイトにてカテゴリ分けをしたリンク集として公開されています（※１）。 ● 今後、人工知能研究開発ネットワーク（AI Japan R&D Network）会員機関等において提供されている、外部の方が利用可能なデータセットの情報につきましては、AI Japan ウェブサイトにおいて発信していくことを検討中。 ※１　AI 学習用の公開データセット https://www.ai-japan.go.jp/ai-open-dataset

■AI 関連公開データリスト

産業技術総合研究所 人工知能研究センター（産総研 AIRC）

番号	分類	名称	概要	リンク先
1	知識データ	介護の構造化マニュアルの例	8 種類の介護行為に関する知識を、目的指向の観点から記述。	https://www.airc.aist.go.jp/achievements/ja/p-035.html
2	画像データ	日用品データセット＋ 3D データベース	日用品やコンビニ商品を効率的に認識するための学習用 3D データ	https://www.airc.aist.go.jp/achievements/ja/p-007.html
3	画像データ	ABCD: AIST Building Change Detection	津波による建物の変化検出学習用データ	https://github.com/gistairc/ABCDdataset
4	画像データ	MUSIC: Multiband Satellite Imagery for object	太陽光発電所（メガソーラー）検知用の衛星画像教師データ	https://github.com/gistairc/MUSIC4P3
5	画像データ	MIRO: Multi-view Images of Rotated Objects	多視点から撮影した日用品画像データ	https://github.com/kanezaki/MIRO
6	動画データ	STAIR Actions: A Large-Scale Video Dataset of Everyday Human Actions	100 種類の日常動作の認識学習用の大規模ラベルつき動画データ	https://stair-lab-cit.github.io/STAIR-actions-web/
7	動画データ 言語データ	STAIR Actions キャプション：STAIR Actions 動画のキャプション（説明文）データ	STAIR Actions 動画の一部に日本語の説明文（誰がどこで何をしている）を付与したデータセット ※動画の説明文データセットとしては世界最大規模（2019/3 公開時）	https://actions.stair.center/
8	画像データ	Patterns_of_Beauty	絵画の美しさの学習用データセット	https://github.com/chupibk/PoBDB_Patterns_of_Beauty
9	画像データ	Part-Function Dataset:	日用品や茶道具の 3 次元データの部分に、機能属性（にぎる、すくう、など）を付与したデータセット	http://isl.sist.chukyo-u.ac.jp/archives/nedopro/
10	行動データ	高齢者行動ライブラリ	人の認知・身体機能に紐づけされた高齢者のビデオ・データ	http:/ www.behavior-library-meti.com/behaviorLib/homes/about

情報通信研究機構（NICT）

AI データテストベッド（https://ai-data.nict.go.jp/）にて下記データセットを公開中。

(*)付のデータセットは NICT ユニバーサルコミュニケーション研究所が事務局を務める高度言語融合フォーラム(ALAGIN)のサイトから ALAGIN 会員向けに公開中です。

番号	分類	名称	概要
1	音声資源	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です。
2	音声資源	日本語音声データベース(*)	ATR にて開発された、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です。
3	音声資源	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です。
4	音声資源	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です。
5	音声資源	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です。
6	音声資源	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です。
7	音声資源	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です。
8	音声資源	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です。
9	言語資源	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です。
10	言語資源	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です。
11	言語資源	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 件を収録したものです。

12	言語資源	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です。
13	言語資源	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、前者は京都観光を中心とした執筆者 47 名・合計 1041 記事(平均約 480 字)から構成される日本語ブログ記事のデータベースです。後者は前者に対して評価情報(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です。
14	言語資源	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収集した日・英・中・韓国語の固有名詞辞書を NICT で整備した辞書です。
15	言語資源	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されたモデルです。
16	言語資源	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です。
17	言語資源	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)から構成されています。
18	言語資源	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です。
19	言語資源	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です。
20	言語資源	アジア言語ツリーバンク（クメール語品詞データ）	英語のウィキニュースから無作為に抽出した 20,000 文をクメール語に翻訳したものについて、単語分割・品詞付与を適用したツリーバンクです。 本ツリーバンクは、Asian Language Treebank（ALT）プロジェクトの一環として作成されました。
21	言語資源	NICT BERT 日本語 Pre-trained モデル（*）	日本語 Wikipedia を対象に情報通信研究機構 データ駆動知能システム研究センターで事前学習を行った BERT モデルとなります。
22	言語資源	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です。
23	言語資源	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、合計約 69,000 名詞句から成る階層的シソーラスです。

24	言語資源	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です。
25	言語資源	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です。
26	言語資源	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です。
27	言語資源	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です。
28	言語資源	アジア言語ツリーバンク（ミャンマー語）（*）	英語のウィキニュースから無作為に抽出した 20,000 文をミャンマー語に翻訳したものについて、単語分割･構文解析を適用したツリーバンクです。本ツリーバンクは、Asian Language Treebank（ALT）プロジェクトの一環として作成されました。
29	言語資源	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です。
30	脳情報関連	Brain Viewer 2012	Brain Viewer 2012 は、人が知覚する様々な物体や動作カテゴリが大脳皮質のどこでどのように表現されているかを可視化する Web インターフェースです。様々な動画を視聴している際の全脳活動記録（fMRI 記録）データを解析することで、約 1,700 種類の物体・動作カテゴリの大脳皮質上の表現分布などを見ることが出来ます。
31	脳情報関連	自然動画視聴下ヒト脳活動データ（リンク先英文）	このデータセットは約 3 時間分の動画を視聴している際のヒト視覚関連領域（後頭葉）における脳活動記録（fMRI 記録）を提供するものです。 データセットはヒト 3 名分の脳活動データ、刺激動画データ、機能領野位置情報データ（例：大脳左半球一次視覚野）等を含みます。
32	脳情報関連	SIPS Probabilistic Atlas（繊維束のアトラスデータ）	このデータは、94 名の被験者から同定したヒト頭頂葉の線維束(stratum proprium of interparietal sulcus; SIPS)の MNI 標準脳座標系での位置を示すアトラスデータであり、NIFI-1 フォーマットで記述されています。
33	脳情報関連	サルとヒトにおける視覚的な奥行き手がかり統合のデコーディング精度マップ(リンク先英文)	このデータセットは、2 つの立体視の手掛かり（両眼視差と運動視差）がサルとヒトの脳内のどこでどのように統合されるのか、その視覚情報処理の相違を明らかにした学術論文 Armendariz, Ban, Welchman, Vanduffel 2019 PLOS Biology の結果を再現可能なデータとソースコードを提供するものです。 2 つの異なる手がかり（両眼視差と相対運動）を組み合わせた刺激の奥行きを fMRI 脳活動計測データからデコードした大脳皮質上の予測精度マップを含みます。

34	脳情報関連	ヒトの立体視力の個人差に対応した神経線維束の走行および神経組織密度データ（リンク先英文）	このデータセットは、ヒトの立体視力の個人差と関連する神経線維束の存在を明らかにした学術論文 Oishi, Takemura, Aoki, Fujita, Amano 2018 PNAS の結果を再現可能なデータとソースコードを提供するものです。拡散強調 MRI 法で調べた Vertical Occipital Fasciculus (VOF) と呼ばれる線維束の走行データと定量的 MRI 手法で調べた VOF の組織高分子量データの一部を含みます。さらに、両眼による立体視力を調べた心理実験の結果と実験刺激プログラムを提供します。
35	脳情報関連	網膜神経節細胞障害後のヒト視覚白質経路における組織特性データ（リンク先英文）	このデータセットは、視覚障がい（レーベル遺伝性視神経症）によってヒト視覚白質経路に生じる組織特性の変化を構造 MRI 計測を用いて明らかにした学術論文 Takemura et al. 2019 NeuroImage: Clinical に関して、その結果を再現可能なデータとソースコードを提供するものです。拡散強調 MRI 手法と定量的 MRI 手法で調べた視覚障がい群と対照群それぞれにおける視覚白質経路の拡散異方性(FA 値)と縦緩和時間(T1 値）のデータを含みます。
36	脳情報関連	ヒト大脳皮質における階層的運動系列の脳内表現データ（リンク先英文）	このデータセットは、ピアノ演奏のような複雑な指運動が脳内で階層的に表現される様子を明らかにした学術論文 Yokoi, Diedrichsen 2019 Neuron の結果を再現可能なデータとソースコードを提供するものです。被験者が予め暗記した複数のキー押しの系列を実行している際の脳活動データに対して、RSA 多変量 fMRI 解析法を適用して得た「脳内運動情報地図」(各階層モデルの対数周辺尤度比の大脳皮質マップ）を再現できます。
37	脳情報関連	ヒト脳の視覚野のレチノトピー(網膜部位再現性)構造計測データ	標準的なレチノトピー刺激(回転あるいは拡大・縮小するチェッカーボードパターン、および視野を縦横に横切るバー)に対するヒト fMRI 脳活動データセットです。
38	脳情報関連	対戦型テレビゲーム中の脳波データ	人がテレビゲームにおける対戦型野球ゲームを行っているときの 4 人分の脳波データです．具体的な野球ゲームの内容に関しては，Yokota et al., ,PLOS ONE 14(2), e0212483 (2019)をご参照ください．脳波データに関しては，空振りストライクの時のデータ，見逃しストライクのデータ，ボールの時のデータが含まれています．
39	脳情報関連	動画などを入力として動的な視覚特徴（運動エネルギー）を出力するモデル	動画等の視覚素材を入力として動的な視覚特徴（運動エネルギー）を出力するモデルです。特徴抽出は局所的な時空間ガボールフィルタアレイによって行われ、フィルタアレイを定義する各種パラメータ（色空間、時空間周波数特性、非線形性等）を指定できます。利用した時空間フィルタの可視化にも対応しています。このコードは学術論文 Nishimoto and Gallant 2011 Journal of Neuroscience 等で利用したモデルを一部改変したものです。
40	脳情報関連	Def Muscle 形式の筋骨格形状データ	脳情報通信融合研究センターで開発しているデフォーマブル筋骨格モデル (Def Muscle) の筋形状は、BodyParts3D（Copyright© 2008 ライフサイエンス統合データベースセンターlicensed by CC 表示ー

			継承 2.1 日本）を基に作成しています。BodyParts3D では筋の表面形状のみが表現されていますが、Def Muscle では筋の内部にも質点を配置しています。 具体的には 3 次元格子状に質点を配置し、質点のインデックスを n = i + (Ni*j) + (Ni*Nj*k)で管理しています。Ni, Nj, Nk は 3 方向の質点数で、筋によって異なります。筋ごとの質点数は、MuscleList.csv で確認することができます。質点数やインデックスの詳細については、「運動解析装置および運動解析方法（特開 2017-037553)」をご覧ください。 今回公開するデータには、上肢の筋群（50 要素）が含まれています。
41	脳情報関連	Brain Viewer 2013	Brain Viewer 2013 は、人が知覚する様々な物体や動作カテゴリが大脳皮質のどこでどのように表現されているか、またそれが認知状態（人や乗り物に注意を向けている状態）によってどのように変化するかを可視化する Web インターフェースです。人あるいは乗り物に注意を向けている状態で様々な動画を視聴している際の全脳活動記録（fMRI 記録）データを解析することで、約 1,700 種類の動作・物体カテゴリの大脳皮質上の分布について認知条件毎に見ることが出来ます。
42	脳情報関連	MP2RAGE 脳構造計測データ	7 テスラ MRI で撮像した高解像度脳構造画像データです。 MP2RAGE シーケンスにより撮像され、灰白質、白質、脳脊髄液の構造分離する際に使用される T1 強調画像、T1 画像が含まれています。
43	脳情報関連	自然動画刺激下 2 光子イメージングデータ	このデータセットは約 1 時間分の動画を視聴している際のマカク一次視覚野神経細胞カルシウム応答データを提供するものです。データセットは大脳皮質一次視覚野 2/3 層の約 130 個の神経細胞およびグリア細胞について、2 光子カルシウムイメージング法で記録した動画刺激下におけるカルシウム応答データ、および各細胞の位置情報データ等を含みます。
44	脳情報関連	攻撃行動に加担する程度と機能結合のデータ	安静時 fMRI で計測した被験者の機能結合行列、攻撃行動課題中の各被験者の行動パラメーター(加担を含む)とパーソナリティスコアが含まれています。 安静時 fMRI を標準化した EPI データから 260 個の関心領域内の平均時系列を抽出し、その相関行列を作成しました。提供するデータはこの相関行列、および被験者が攻撃行動課題で攻撃に加担した程度をあらわすパラメーターならびに社会性に関するアンケート結果をセットにしたものです。
45	バイオ関連	細胞内タンパク質局在データ	分裂酵母タンパク質の蛍光タンパク質タギングによる細胞内局在データです。
46	バイオ関連	遺伝子発現プロファイルデータ	分裂酵母遺伝子の窒素源枯渇、接合フェロモン投与時における遺伝子発現プロファイルです。
47	サイバーセキュリティ関連	ダークネット・データセット 2019	"本データセットには，下記参照論文の解析に用いたダークネット統計データ及び解析結果データが含まれています。 ダークネット統計データ: NICTER ダークネット・トラフィックデータから作成した統計データ

			解析結果データ:上記のダークネット・パケット統計データの解析で検出したインシデント情報
48	大気環境関連	環境×健康スマート IoT 地図データ	大気環境を可視化したデジタル地図のデータベースです。 同時刻かつ同地点における環境データ及び体調データのレーダーチャート並びにルート探索中に撮影した写真及びコメントを合わせて作成しています。
49	宇宙天気関連	Deep Flare Net 特徴量データベース	太陽フレアの発生予測モデル Deep Flare Net (DeFN)の特徴量データベースです。学習用・テスト用データセットとして使用することができます。