4-1-2　AI 研究開発環境「AI データテストベッド」の取組
4-1-2　Efforts on R&D Environment for AI:'AI Data Testbed'

藤井秀明　岩爪道昭
FUJII Hideaki and IWAZUME Michiaki

NICT では、オープンイノベーション型の AI（artificial intelligence）研究開発及びその成果を社会実装するため、多様な AI 関連データセットを格納・管理・検索及び共有・公開する「AI 研究開発環境（AI データテストベッド）」を提供している。本稿では、AI データテストベッドの概要と運用状況を紹介する。

NICT provides R&D environment for AI: AI Data Testbed, which storages / manages / retrieves and shares / publishes a variety of datasets for AI in order to promote Open Innovation styled R&D for AI and implement the results into our society. In this paper, we show a brief overview and operational status of AI Data Testbed.

1　まえがき

情報通信研究機構 NICT では、脳情報通信、サイバーセキュリティ、リモートセンシング分野の人工知能応用技術の研究開発と、翻訳バンクプロジェクト等の各種ビッグデータの取得や利活用を意識したオープンイノベーションを推進している。

現在、様々な分野で活用が進められている人工知能技術はコア技術として大規模な学習データが必要となる深層学習と呼ばれる機械学習技術が活用されているが、NICT では人工知能技術が注目される前からデータの重要性に着目し、多くの研究分野で様々な種類のデータを集積し、研究開発に活用してきた。そのため NICT が保有するデータの中には人工知能技術に活用できるデータも数多く存在している。そこで、人工知能技術の開発に利用可能と考えられる 8 カテゴリのデータを NICT 内外の研究者・技術者がダウンロードして活用できる Web サイトとして、AI データテストベッド公開基盤を構築して 2019 年 5 月末から運用を開始した。現在までに上記 8 カテゴリ 50 件のデータセットが公開され、オープンイノベーションのためのデータの利活用が始まっている。

2　AI データテストベッドの全体構成

2.1　三つの基盤

AI データテストベッドは図 1 に示すように、次の三つの基盤から構成されている。

AI データ管理基盤は、AI や機械学習等に利用可能なデータセットまたは学習済みのモデルを保存・管理する基盤で、ファイルシステム及びデータベース管理システム（Relational DataBase Management System）から成り立っている。

AI データ公開基盤は、AI データ管理基盤に格納されているデータセットや各種情報を発信するための Web サイトである。本基盤については、**3** にて紹介する。

AI データ利活用基盤は、ディープラーニングの学習または研究を実施するための基盤で一般には開放されておらず、NICT 職員や共同研究者による利用を目的としている。こちらについては、**4** にて紹介する。

2.2　計算機基盤

上記三つの基盤は、全 32 台のサーバで構成されるクラスタ上に構築されている。各サーバには GPGPU（General-Purpose computing on Graphics Processing Units）が 2 枚ずつ搭載されており、約 450 TB の共有ストレージにアクセス可能である。

3　公開基盤

本節では、インターネット上で広く一般向けにデータセットを公開している公開基盤の Web サイトについて紹介する。なお、Web サイトへは以下の URL にてアクセス可能である。

https://ai-data.nict.go.jp/

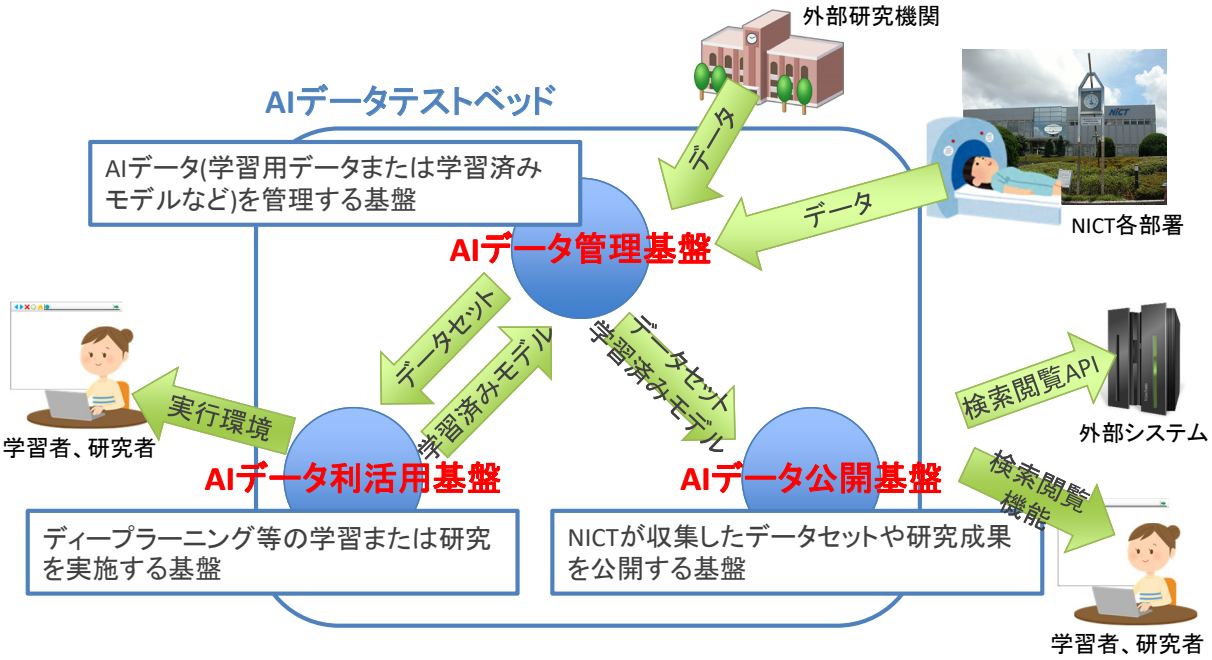

図 1　三つの基盤の概念図

3.1　公開しているデータセットについて

AI データテストベッドでは、NICT における研究過程で取得または生成された 8 ジャンル 50 件のデータセットを公開している。ジャンル名及びそれぞれのデータセット数を表 1 に示す。

これらデータセットは、データ本体が AI データテストベッド（管理基盤）に格納されているものと、外部サイトで公開されているものがある。後者は、AI データテストベッド公開以前に既に公開されていたデータセット等であり、これらについては AI データテストベッドよりリンクを張ることで、ユーザーに当該データセットを公開している外部サイトへ遷移しダウンロードしていただくよう誘導している。

表 1　ジャンルと含まれるデータセット数

ジャンル名	件数
言語資源	21
音声資源	8
バイオ関連	2
脳情報関連	15
大気環境関連	1
宇宙天気関連	1
サイバーセキュリティ関連	1
機械学習・量子機械学習	1

3.2　データセットの利用について

AI データテストベッドでは利用規約を定めており、データセットの利用は原則として無償であり、研究開発目的に限って利用することを許諾している。

ただし前述のように、外部サイトで公開されているものについては、当該サイトで規定されている規約が優先される形式となっている。

また、AI データテストベッドに格納され、ここよりダウンロードされるデータセットについても、個別規約を設けることで、同様に優先される仕組みとしている。これはデータセットごとの事情に応じて柔軟な公開を可能とすることを目的としたものであり、これによってデータ公開にかかわる障壁を緩和することを意図している。

3.3　Web サイトの機能について

機能面では基本的な検索機能とデータセットのダウンロード機能に加えて、ユーザー登録しログインすることで、興味を持ったデータセットをブックマークできるお気に入り機能がある。

検索機能については、あいまい検索の機能を備えており、ユーザーが希望するデータセットをより検索しやすくしている。例えば図 2 に示すように、脳情報関連のデータセットを探しているケースで、検索ワードとして「前頭葉」と入力して検索した場合、「前頭葉」以外にも「大脳」や「前頂葉」等のワードを含むデータセットも検索結果としてヒット可能としている。

図 2　あいまい検索機能

図 3　あてっこゲーム

3.4　デモアプリについて

データセットを公開しても、使い方が分からないと実際にダウンロードをして利用しづらいものである。このため、利用促進を目的として、Web サイトで公開しているデータセットやライブラリを利用したデモアプリも公開している。

まだデモアプリの数は少ないが、Web サイトに掲載されている「量子インスパイア正準相関分析（qiCCA）」[1][2] を利用して、図 3 に示す手書き数字画像の左半分から右半分の画像を推定する「あてっこゲーム」を公開

図 4　e ラーニング環境　学習条件設定画面

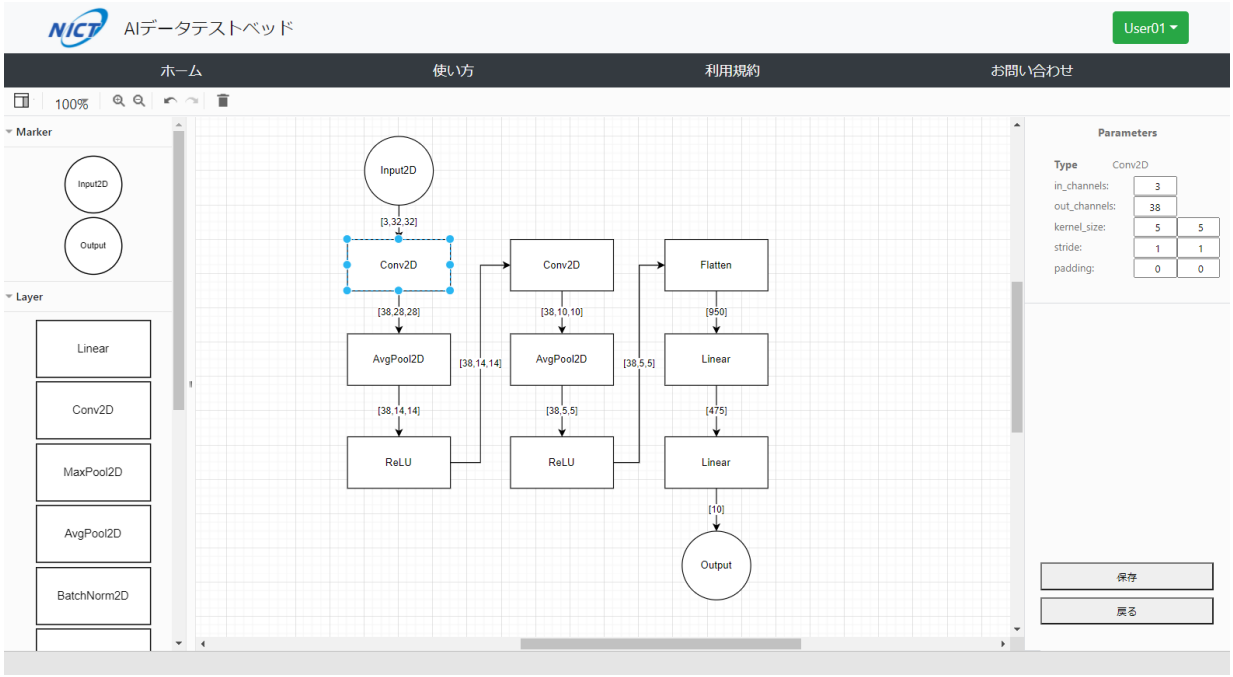

図 5　e ラーニング環境　モデル作成ツール

している。また、あいまい検索機能も掲載データセットである「異表記対データベース」[3] 及び「基本的意味関係の事例ベース」[4] 使用して実装された機能となっている。

3.5　運用状況について

2020 年 4 月に公開した Wikipedia で学習した自然言語処理モデルである BERT（Bidirectional Encoder Representations from Transformers）のデータ（パラ

メータ 1.6 億個）[5] は、他機関から公開済みの日本語 BERT よりも高性能であると評価され、専門知識が要求されるにもかかわらず、公開後約 10 か月で 2,743 件のダウンロードを達成しており、社会における AI データテストベッドへの注目と期待の大きさが分かる成果を得ている。

また、政府の AI 戦略 2019 を踏まえ、人工知能（AI）の研究開発に関する統合的・統一的な情報発信や、AI 研究者間の意見交換推進等を行い、AI 研究開発の活性化を図ることを目的として 2019 年 12 月に設置された人工知能研究開発ネットワークの公式 Web サイト「AI Japan」[6] においても本サービスへのリンクを掲載していただいている。

4　利活用基盤

利活用基盤としては、AI に習熟しており自ら AI を利用できる研究者をターゲットとする「共用計算サーバ」と、逆に AI についての知識を持たない初学者向けのサービスを目指す取組について紹介する。

4.1　共用計算サーバについて

共用計算サーバは **2.2** に記載のクラスタのうち 27 台のサーバ（GPGPU54 枚）で構成される。ユーザーは、同クラスタ上で稼働するスケジューラ経由で、ディープラーニングの学習や推論を実行し研究することができる。

4.2　e ラーニング環境について

前項の共有計算サーバが AI 中・上級者向けのサービスである一方、こちらは AI 初学者でもディープラーニング等の AI 技術を享受できることを目的としている。

具体的には、Web ベースのアプリケーションとしてノンプログラミングでディープラーニングの学習や推論が実行できることを目指すものである。

広く一般に知られている画像認識モデル（Convolutional Neural Network: CNN）や画像生成モデル（Generative Adversarial Network: GAN）等のモデルに対して条件を指定して学習（図 4）し、学習したモデルを利用した推論ができる。また、マウスによるドラッグ＆ドロップで自らのモデルを作成することもでき（図 5）、これを用いて学習及び推論できる機能を研究開発している。

まだプロトタイプ段階であり、操作性や利用容易性の点で改善の余地があるが、今後はモニターによる PoC（Proof of Concept）等の概念実証評価を行い、更なる改善を目指している。

5　おわりに

AI データテストベッドは今後もデータの拡充と品質向上、検索機能の充実などの利便性向上を図り、NICT のデータ指向型オープンイノベーションの中核基盤に発展させていきたい。その一環として、DCCS（Data Centric Cloud Service）と連携することで、DCCS が提供する API（Application Programming Interface）を経由したデータ提供を計画している。

【参考文献】

1　“量子インスパイア正準相関分析 (qiCCA),” https://ai-data.nict.go.jp/dataset/detail/?id=52
2　N. Koide-Majima, and K. Majima, “Quantum-inspired canonical correlation analysis for exponentially large dimensional data,” Neural Networks, vol.135, pp.55-67, March 2021.
3　“異表記対データベース,” https://ai-data.nict.go.jp/dataset/detail/?id=7
4　“基本的意味関係の事例ベース,” https://ai-data.nict.go.jp/dataset/detail/?id=9
5　“NICT BERT 日本語 Pre-trained モデル,” https://ai-data.nict.go.jp/dataset/detail/?id=46
6　AI Japan 公式サイト, https://www.ai-japan.go.jp/

藤井秀明　（ふじい　ひであき）
ソーシャルイノベーションユニット
総合テストベッド研究開発推進センター
テストベッド研究開発運用室
有期研究技術員

岩爪道昭　（いわづめ　みちあき）
業務企画部
DX 企画推進室
室長／
ソーシャルイノベーションユニット
総合テストベッド研究開発推進センター
テストベッド連携企画室
マネージャー（兼務）
博士（工学）
人工知能、知識工学