

データ基盤から知識基盤へ



国立情報学研究所 所長 / 京都大学 特定教授
黒橋 禎夫

京都大学 研究データ管理シンポジウム (2026/2/27)



研究データ基盤

1990年代後半: デジタル化とデータの爆発

論文だけでなくデータそのものが研究資源

2000年代: オープンアクセス運動と政策化

公的資金による研究データは共有されるべき

2010年代: FAIR原則とRDMの制度化

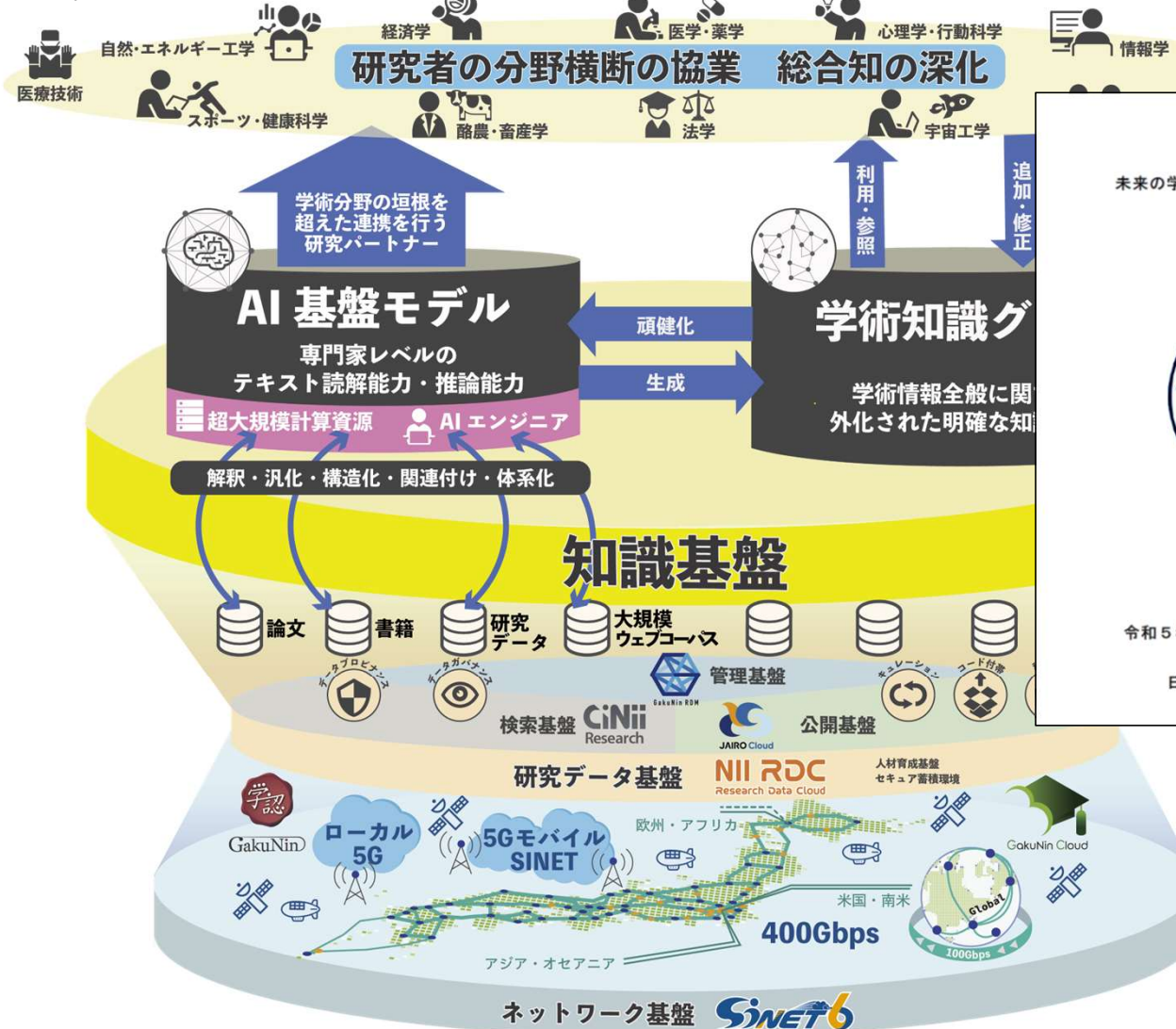
「公開」から「再利用可能」へ

知識基盤
研究データ基盤

LLM
2022.11 ChatGPT

データ基盤から知識基盤へ

日本学術会議「未来の学術振興構想」の策定に向けた「学術の中長期研究戦略」に提案 (2022年12月16日)



提言

未来の学術振興構想 (2023年版)

グランドビジョン^⑩

データ基盤と利活用による学術界の再構築

⑩ データ基盤と利活用による学術界の再構築

概要：あらゆる分野のデータを集約するデータ基盤、さらに深化させた知識基盤の整備を進め、同時に利活用し得る人材を育成することでデジタル変革を推し進め、学術界を再構築していく。

キーワード：学術情報基盤、知識基盤、資料保存、データ駆動型研究、オープンサイエンス

ア 背景

現代では、学術の進化を継続的に支えるためにデータ基盤の構築とその活用が不可欠である。例えば、世界中の研究者が公開データを利用して共通の課題の解決や新発見の存在等、予想外の事象まで洞察できるようになった。しかし、日本の学術界はビッグデータやオープンサイエンスの恩恵を十分に享受できておらず、その主原因として、国内のデータ基盤の未整備が挙げられる。人文や社会科学においてもビッグデータの機械学習等による解析やシミュレーションが重要となっており、利用技術習得が十分に進んでいない。また、コピペに基いた複製形成(以下「図写」)という、1が徐々に可能になってきたものの、日本における図写は遅くであり、今後は政府・地方自治体・民間企業等と連携して、海外からの評価も得ていく必要がある。データ基盤を構築し知識基盤へと進化させ、あらゆる学術分野で「図写」を実現させることにより、日本学術会議の使命である「学術の発展」が実現できる。

イ 目的・目標

学術界のデジタル変革を支援する目的が見られる。日本においても、政府統計調査に基づく全体的データの活用が促され、国際的なデータの活用が可能となってきた。

エ 中長期の学術構想

学術界全体のデジタル変革の具体策は、1)国内のデータ基盤・知識基盤の整備、2)データ基盤・知識基盤を中心とした学術の再構築、3) AI、計算科学、デジタルツイン技術、可視化等の分野融合的人材育成と技術開発、4)知識基盤の世界とのネットワーク化、5)知識基盤を広く開放し、利用を促すアウトリーチ、6)図写の基礎的なデジタルインフラ利用、7)個人情報・個人情報等に関する意識を高めるための教育、8)地域の独自性に配慮したデータ収集と利活用方法の確立である。また、その設備体制として、高性能スーパーコンピュータ群と知識処理をサポートするための計算機群への大容量高速ネットワークアクセスを整備を進める。

ウ 国内の状況

2012年の時代が到来し、様々なデータが蓄積され、その活用が求められるようになった。

図 11 データ基盤とその利活用による学術の発展

(出典) 学術の中長期研究戦略「データ基盤から知識基盤へ」を基に、本提言に独自に作成

AI for X

AI for Science

知識基盤

研究データ基盤

LLM

The screenshot shows a web browser window with two tabs. The active tab is titled 'GPT-5.2 derives a new result in...' and the address bar shows 'openai.com/index/new-result-theoretical-physics/'. The OpenAI logo is in the top left, and a 'Log in' button is in the top right. The main heading is 'GPT-5.2 derives a new result in theoretical physics', dated 'February 13, 2026', with 'Research' and 'Publication' tags. A sidebar on the left lists various OpenAI products. The article text states that GPT-5.2 proposed a formula for a gluon amplitude, which was later proved by an internal OpenAI model. A 'Read the preprint' button is centered below the text. At the bottom, there is a 'Listen to article' button with a 3:54 duration, a 'Share' button, and a ChatGPT prompt: 'We've published a [type of particle interaction many physicists expected would not occur can in fact]'. The browser's address bar also shows a Japanese tab '新しいタブ'.

OpenAI

February 13, 2026 Research Publication

GPT-5.2 derives a new result in theoretical physics

In a new preprint, GPT-5.2 proposed a formula for a gluon amplitude later proved by an internal OpenAI model and verified by the authors.

Read the preprint ↗

Listen to article 3:54 Share

We've published a Ask ChatGPT ↑ ype of particle interaction many physicists expected would not occur can in fact

AI for X

AI for Science

知識基盤

研究データ基盤

オープンサイエンス

LLM

透明性・信頼性

研究データ基盤

オープンサイエンスへ

オープンサイエンスの潮流と我が国の施策

2023 G7科学技術大臣会合 (5.12-14 仙台)



G7科学技術大臣コミュニケ

1. 科学研究における自由と包摂性の尊重 およびオープン・サイエンスの推進

G7は、FAIR原則（Findable=見つけれられる、Accessible =アクセスできる、Interoperable=相互運用できる、Reusable =再利用できる）に沿って、科学的知識並びに研究データ及び学術出版物を含む公的資金による研究成果の公平な普及による、**オープン・サイエンスの拡大**のために協力する。これは、世界中の研究者や人々がその恩恵を受けるとともに、新しい知識の創造、イノベーションの促進、社会による知識への…

https://www8.cao.go.jp/cstp/kokusaiteki/g7_2023/2023.html

学術論文等の即時オープンアクセスの実現に向けた基本方針

令和6年2月16日

統合イノベーション戦略推進会議決定

(1) 公的資金による学術論文等の即時オープンアクセスの実施

- 公的資金のうち2025年度から新たに公募を行うい即時オープンアクセスの対象となる競争的研究費を受給する者(法人を含む)に対し、該当する競争的研究費による**学術論文及び根拠データ**の学術雑誌への掲載後、**即時に機関リポジトリ等の情報基盤への掲載を義務**づける。
- ...

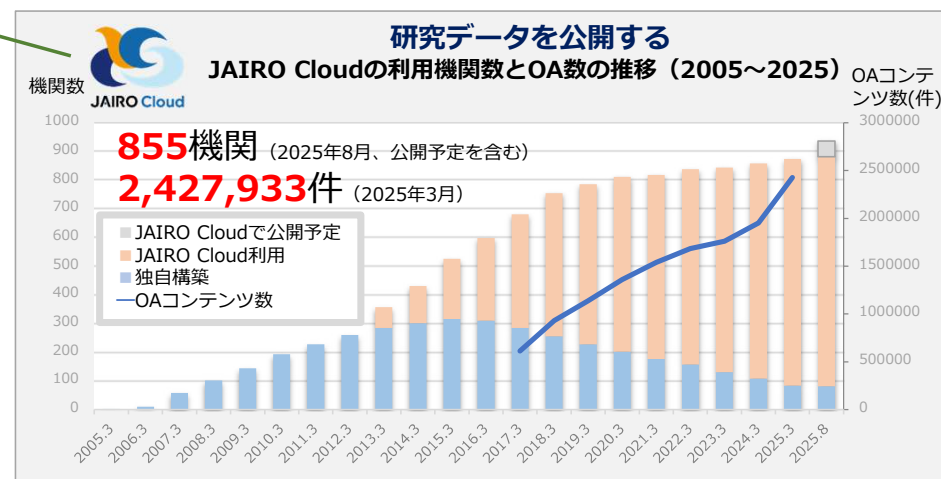
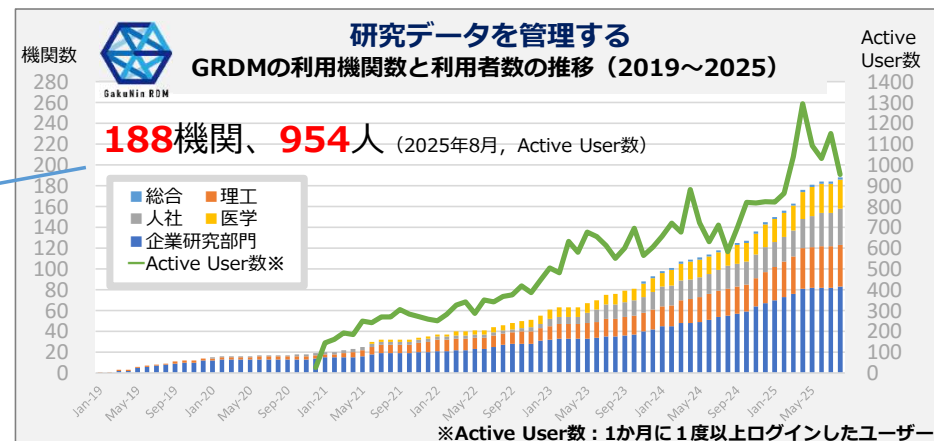
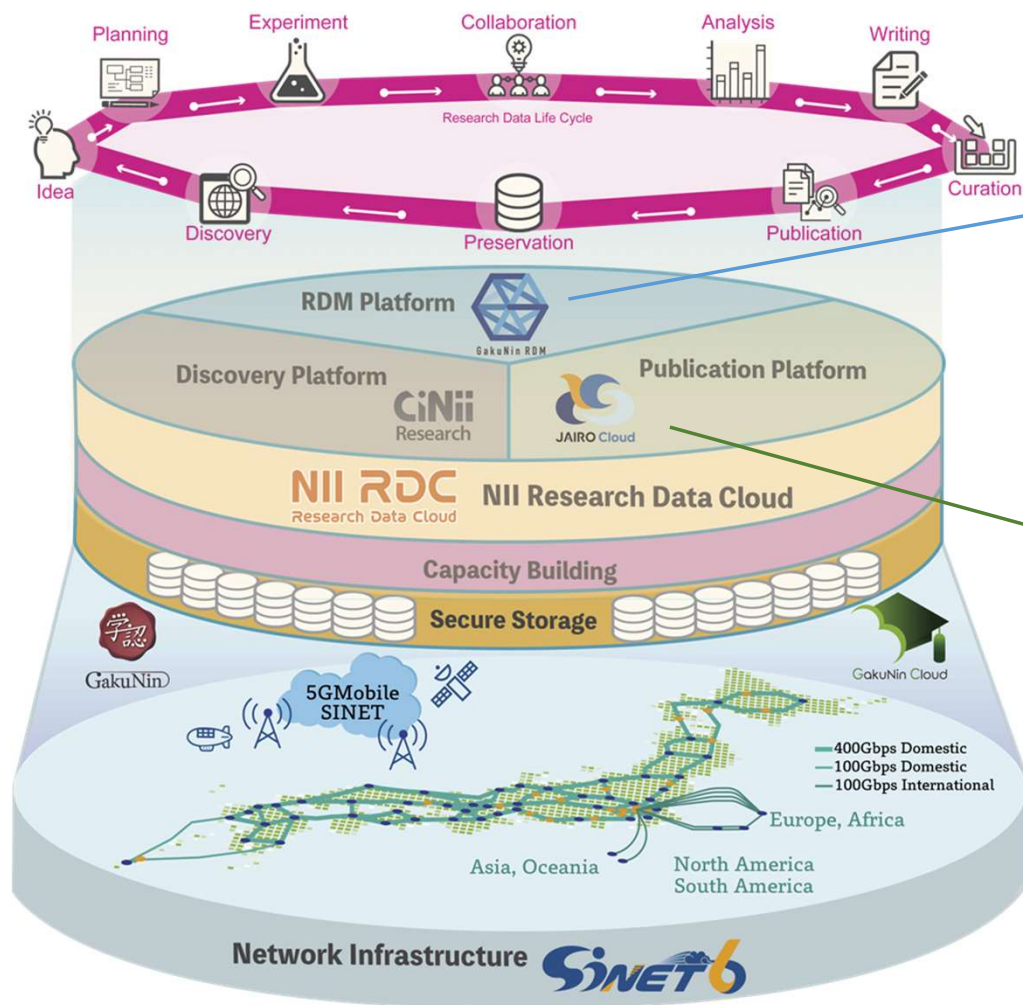
(3) 学術論文及び根拠データの機関リポジトリ等の情報基盤への掲載

- 学術論文及び根拠データの機関リポジトリ等の情報基盤への掲載を通じて、誰もが自由に利活用可能となることを目指す。
- 機関リポジトリ等の情報基盤とは、第6期科学技術・イノベーション基本計画(令和3年3月26日閣議決定)において「研究データの管理・利活用のための**我が国の中核的なプラットフォーム**」として位置付けた**研究データ基盤システム(NII Research Data Cloud)**上で学術論文及び根拠データが**検索可能**となるものとする。
- ...

<https://www8.cao.go.jp/cstp/siryo/haihui071/siryo3-2.pdf>

NIIが提供する学術研究プラットフォーム

(NIIの起源は1976年発足の東京大学情報図書館学研究センター)



AI等の活用を推進する研究データエコシステム構築事業

(文科省 2022-2026年度)

- 全国的な研究データ基盤の構築・高度化・実装
- 研究データ基盤活用のための環境整備 (ルール・ガイドライン整備、データマネジメント人材育成支援等)

中核機関群

研究データ基盤高度化チーム

NII NII RDCを7つの側面から機能拡張

プラットフォーム 連携チーム

RIKEN 理化学研究所
NII RDCと
分野PFの連携

融合・活用 開拓チーム

THE UNIVERSITY OF TOKYO 東京大学
NII RDC利用
のプラクティス

ルール・ガイド ライン整備チーム

NAGOYA UNIVERSITY 名古屋大学
データ管理公開の
組織的役割

人材育成 チーム

OSAKA UNIVERSITY 大阪大学
データ管理公開
の組織的支援

顔が見える地域ネットワーク

延べ97機関が参加 (2026年1月現在)

2025年度開始:

- ・北海道地区 (北海道大学)
- ・東北地区 (東北大学)

2026年度開始予定:

- ・関西地区 (阪大・京大)

2023年度開始:

- ・東海地区 (名古屋大学)
- ・北陸地区 (金沢大学)

2024年度開始:

- ・中国四国地区 (広島大学)
- ・九州地区 (九州大学)

- ・データポリシー策定
- ・セミナー開催
- ・学内アンケート実施・分析
- ・支援チーム、セミナー講師の派遣

NII RDC のサービス概要

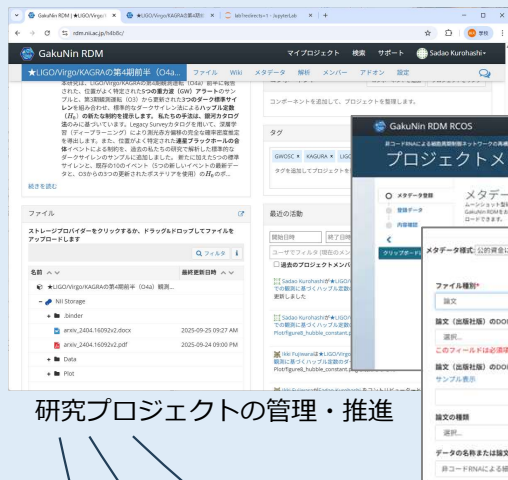


GakuNin RDM
研究の管理・推進基盤

一旦完成した研究の
論文・根拠データ

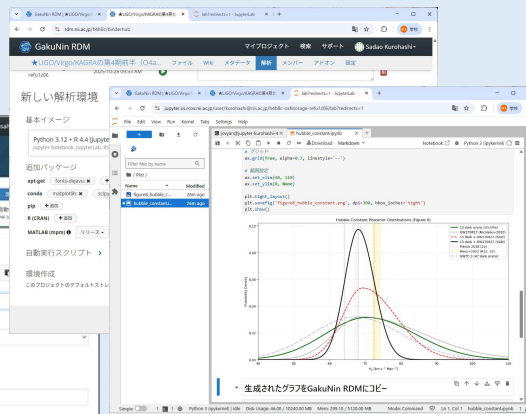


JAIIRO Cloud
研究成果の公開基盤
(850以上の機関リポジトリを運用)

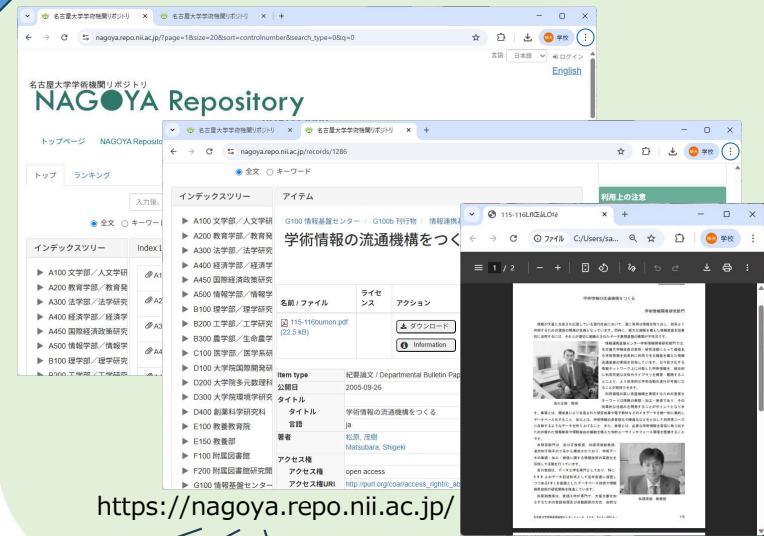


研究プロジェクトの管理・推進

研究プロジェクトのメタデータ (e-Radからインポート)
論文のメタデータ (既発表の論文DOIからインポート)
根拠データのメタデータ (論文と紐付け) 等を管理



紐づけたGPU計算資源等の利用
関連研究論文の実験の再現等



<https://nagoya.repo.nii.ac.jp/>

GRDM用 NIIストレージ
(1人あたり100GB利用可)

機関ストレージ

商用ストレージ

mdx等の計算資源

JCloud用 NIIストレージ
(各機関と有償契約)

The screenshot displays the GakuNin RDM (Research Data Management) web interface. The browser's address bar shows the URL `rdm.nii.ac.jp/h4b8c/`. The page header includes the GakuNin RDM logo and navigation links such as 'マイプロジェクト' (My Project), '検索' (Search), 'サポート' (Support), and '情報犬 ビット' (Information Dog Bit). The main content area features a project titled '★LIGO/Virgo/KAGRAの第4期前半 (O4a) 観測、およびそれまでの観測に基づくハッブル定数のダークサイレン測定' (★LIGO/Virgo/KAGRA O4a observation and Hubble constant measurement based on observations up to O4a). The project is managed by Ikki Fujiwara, Information Dog Bit, and Shimoyama Takeshi, affiliated with the National Institute of Informatics (NII). The page includes a 'Wiki' section with a summary of the research, which discusses the use of gravitational wave (GW) alerts and standard sirens to constrain the Hubble constant (H_0). It also mentions the use of Legacy Survey catalogues and deep learning for redshift estimation. On the right, there are sections for '引用' (Citation), 'コンポーネント' (Component), and 'タグ' (Tag). The 'タグ' section lists various keywords like GWOSC, KAGURA, LIGO, O4a, Virgo, Hubble constant, and gravitational waves. A small video inset in the bottom right corner shows a man in a red shirt speaking at a podium.

Online Do x GakuNin x LIGO/Virgo x GakuNin x anov-grd x NII-DG D x GakuNin x GakuNin x JupyterHu x 新しいタブ x +

← → ↻ 📄 rdm.nii.ac.jp/h4b8c/ ☆ 📄 📄 📄

GakuNin RDM

マイプロジェクト 検索 サポート 情報犬 ビット

★LIGO/Virgo/KAGRAの第4期前半 (O4a) ファイル Wiki メタデータ 解析 メンバー アドオン 設定 証跡管理

NII

★LIGO/Virgo/KAGRAの第4期前半 (O4a) 観測、およびそれまでの観測に基づくハッブル定数のダークサイレン測定

メンバー: Ikki Fujiwara, 情報犬 ビット, Shimoyama Takeshi

所属機関: 国立情報学研究所

rdm.nii.ac.jp/ktb4pからフォーク(2025-10-03 11:53 AM)

作成日時 2025-10-03 11:53 AM | 最終更新日時 2025-10-09 12:27 PM

カテゴリー: 📁 プロジェクト

説明: プロジェクトに簡単な説明を追加してください

ライセンス: ライセンスなし

Wiki

概要

本研究は、LIGO/Virgo/KAGRAの第4期観測運転 (O4a) 前半に報告された、位置がよく特定された5つの重力波 (GW) アラートのサンプルと、第3期観測運転 (O3) から更新された3つのダーク標準サイレンを組み合わせ、標準的なダークサイレン法によるハッブル定数 (H_0) の新たな制約を提示します。私たちの手法は、銀河カタログ法のみに基づいています。Legacy Surveyカタログを用いて、深層学習 (ディープラーニング) により測光赤方偏移の完全な確率密度推定を導出します。また、位置がよく特定された連星ブラックホールの合体イベントによる制約を、過去の私たちの研究で解析した標準的なダークサイレンのサンプルに追加しました。新たに加えた5つの標準サイレンと、既存の10のイベント (5つの新しいイベントの最新デー

引用

コンポーネント

コンポーネントを追加 プロジェクトをリンク

コンポーネントを追加して、プロジェクトを整理します。

タグ

GWOSC x KAGURA x LIGO x O4a x Virgo x ハッブル定数 x 重力波 x

気温が上がります
日曜日

検索

15

オープンサイエンス・AI4S時代における研究データ基盤活用の重要性

- 現状、多くの研究者は、DropboxやGoogle Driveなどの商用サービスを利用し、商業出版社等に求められる最低限のデータ提出を行えば十分と考えている。研究データの整備や共有は「余分な作業」と見なされがちである
- 今後、分野横断的な共同研究が拡大しデータ共有の重要性が飛躍的に高まる。公開された研究データは他の研究者やAIによって再利用され、知識創出の原動力となる。また、AIによるデータ活用が進むにつれ、**データ提供や整備への貢献が研究者の新たな評価指標**として重みを増していく
- 個々の研究者やグループが独自に商用ストレージを用いるだけでは、データ活用はスケールせず、持続可能なエコシステムは構築できない。また、商業出版社のもとにデータが集中すれば、**学術情報流通の商業的支配**という現在の課題の再生となる
- 我が国の研究力強化およびオープンサイエンス・AI4S推進のために、メタデータ等を統一的に管理・活用できるNII RDC等の公的な研究データ基盤の整備と利用が重要

LLM研究開発

透明性と信頼性

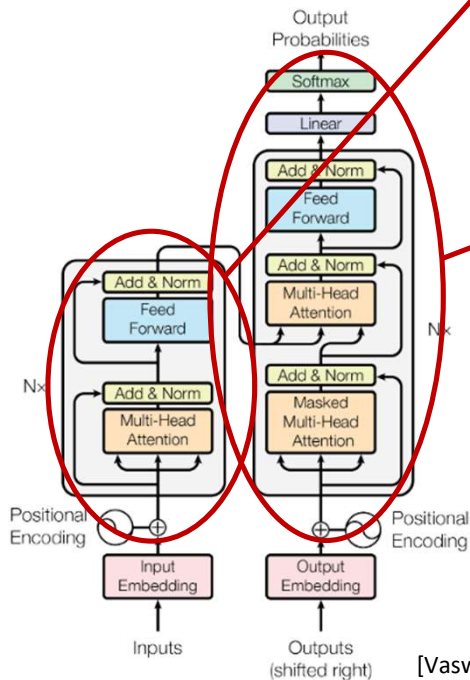
Large Language Model の歴史

2014 Attention

機械翻訳において目的言語の次の語を生成する際に原言語の文のどこに着目するか

2017 Transformer

attentionの精緻化、原言語文内、目的言語内でのself-attention



[Vaswani et al. 2017]

2018 BERT

Transformerのencoder側を単言語の分類問題等に

2018 GPT (1.17億パラメータ)

Transformerのdecoder側を言語モデルに

2019 GPT-2 (15億パラメータ)

2020 GPT-3 (1750億パラメータ)

2022 GPT-3.5 / InstructGPT

2022 ChatGPT

2023 GPT-4 (2兆パラメータ?)

画像も扱える、多言語能力も大幅向上

- 米司法試験で人間受験者の上位10%の成績
- 米大学入試テストSATで1600点中1410点
- 米医師試験USMLEでも合格レベルの点数

LLM・AIエージェントの発展（2023～2025年度）

2023.6 OpenAI APIに**function calling**(関数呼び出し)を追加

2023.12 Mistral AI社が**混合エキスパート**(MoE)方式のオープンLLM Mixtral 8x7Bモデルを公開

2024.3 GPT-4などのLLMで**検索拡張生成**(RAG)が利用可能に

2024.4 Meta社が**オープンLLM**としてLlama3を公開(8B, 70Bモデル。405Bモデルは7月公開)

2024.6 Google社が**コンテキスト長**2MトークンのGemini 1.5 Proを公開

2024.8 SakanaAI社が研究のアイデア生成・実験・論文執筆等を自動化した**AI Scientist**を提案

2024.9 OpenAI社が**推論能力**を強化したモデルo1を発表

2024.11 Anthropic社がAIエージェントのプロトコル**Model Context Protocol (MCP)** を提案

2025.1 DeepSeek社が**低開発コスト**で世界トップレベルの推論能力をもつDeepSeek-R1を公開

2025.2 OpenAI社がGPT-4.5を発表、**Deep Research**モードを搭載

2025.10 論文の執筆・査読にAIが参加する実験的国際会議 **Open Conference of AI Agents for Science**開催

LLMに関する懸念

GPT-4出現時（2023.3）

研究開発が**一部の組織の寡占状態**であり、
モデルの中身は公開されていない
（OpenAIはもはやオープンではない）

**なぜここまで賢いのか（言語を理解し、
妥当な応答ができるのか）** 明らかでない。
一方で**Hallucination**など課題が山積

**GPT-3において日本語テキストは0.11%、
日本語の理解・生成能力は英語に比べて
劣る**

現在

中国系企業なども強いモデルを公開。
しかしオープンモデルと呼ばれるLlama
やDeepSeekでも**学習データは未公開**。
応答にバイアスがある強いモデルが普及
することへの懸念も

LLMの原理・学習過程は依然明らかでない。
Hallucinationの問題もRAGによって
緩和されているが本質的には未解決

広く活用されている大規模ウェブコーパス
FineWeb+FineWeb2（約20T tokens）
の日本語テキストは**1.4%**



LLM-jp (LLM勉強会)

- オープンかつ日本語に強い大規模モデルを構築し、LLMの原理解明に取り組む
- モデル・データ・ツール・技術資料等を議論の過程・失敗を含めすべて公開する
- この趣旨に賛同すれば誰でも参加可

2023.5
自然言語処理の研究者
30名程度による勉強会を開催

2023.10
mdxを用いて
13Bパラメータモデル
LLM-jp-13Bを公開

2023.11
ABCI第2回LLM構築支援
プログラムに参加
175Bパラメータモデルの
学習トライアル

2024.4
**国立情報学研究所に
LLM研究開発センター設置**
文部科学省「生成AIモデルの透明性・信頼性の
確保に向けた研究開発拠点形成」事業
GENIAC環境で
172Bパラメータモデルの
学習開始

2600名超

- **mdx**: データ活用社会創成プラットフォーム. 9大学2研究所が連合して共同運営する、データ活用にフォーカスした高性能仮想化環境
- **ABCI**: AI橋渡しクラウド. 産業技術総合研究所 (AIST) が提供するAI向け計算用で現状国内最大の計算資源
- **GENIAC**: Generative AI Accelerator Challenge. 日本国内の基盤モデル開発力の底上げのために計算資源の提供等を行う経産省のプログラム

LLM研究開発はビッグサイエンス

コーパス構築WG

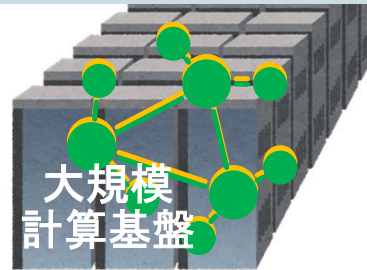


大規模コーパス



河原大輔教授
(早稲田大学)

モデル構築WG



大規模
計算基盤



横田理央教授
(Science Tokyo)



鈴木潤教授
(東北大)



田浦健次郎教授
(東大)

チューニング評価WG



チューニン
グデータ

評価データ



宮尾祐介教授 (東大)

安全性WG



関根聡特任教授
(NII)

マルチモー ダルWG



岡崎直観教授
(Science Tokyo)

実環境インタ ラクションWG



尾形哲也教授
(早稲田大学)

原理解明 WG



大関洋平准教授
(東大)

対話WG



東中竜一郎教授
(名大)

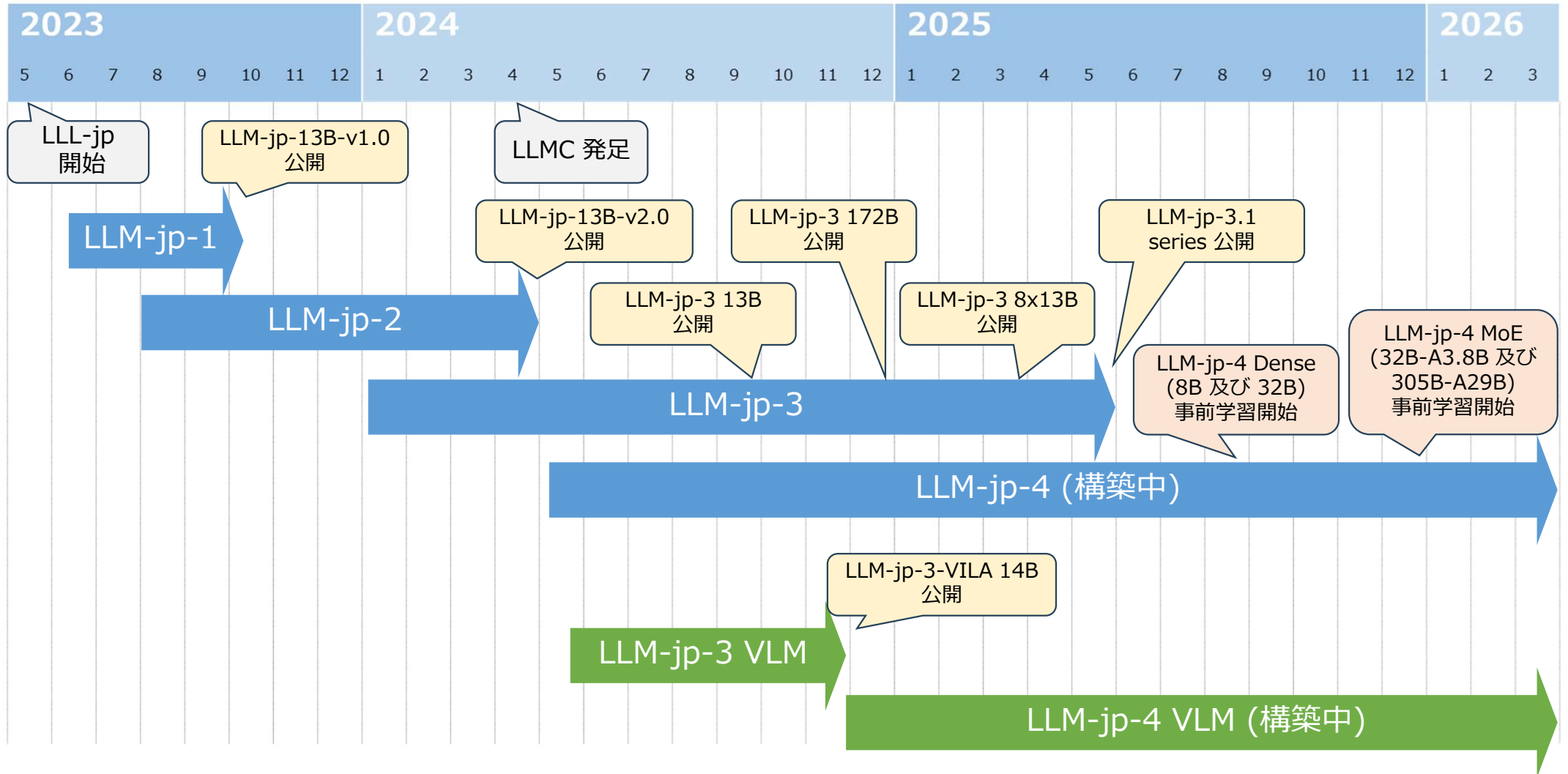
学術ドメイン WG



相澤彰子教授
(NII)

NII/LLMC研究員 約20名
Research Assistant 約50名

モデル構築・公開のタイムライン

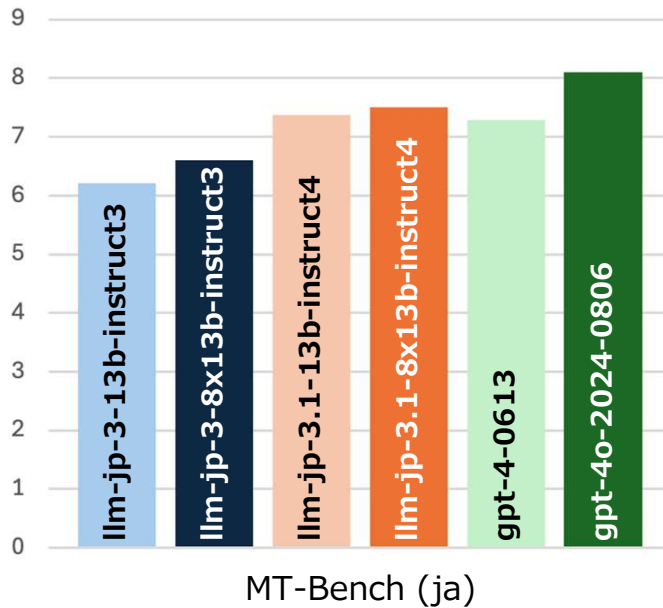


LLM-jp-3 Modelsの性能

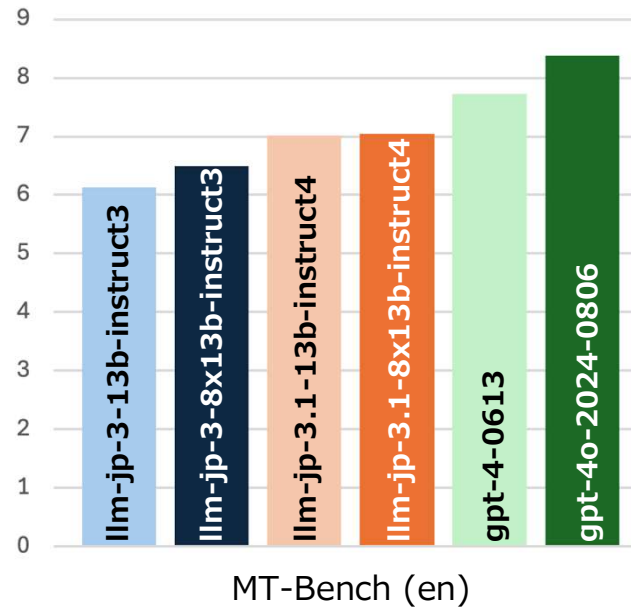
- **llm-jp-3.1-8x13b-instruct4** は日本語有用性で **gpt-4-0613** を上回り、日本語安全性では **gpt-4o-2024-0806** をも上回る

AI Safety Institute
(AISI)と密に連携

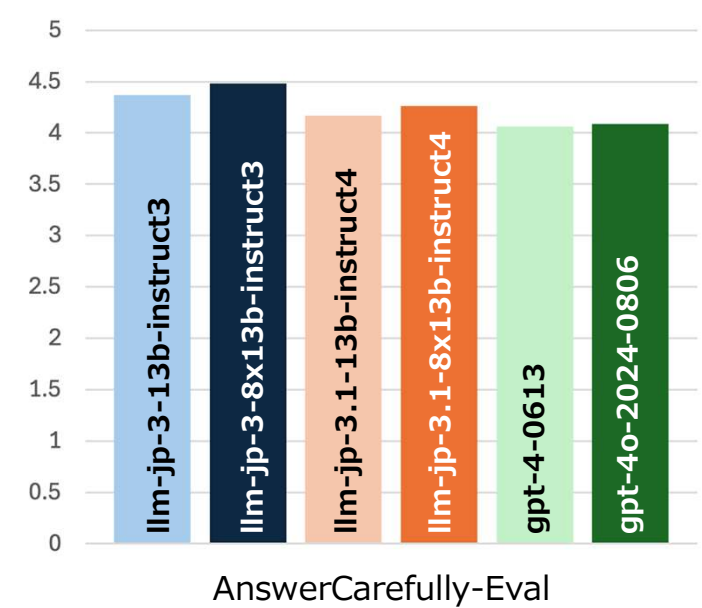
日本語有用性



英語有用性

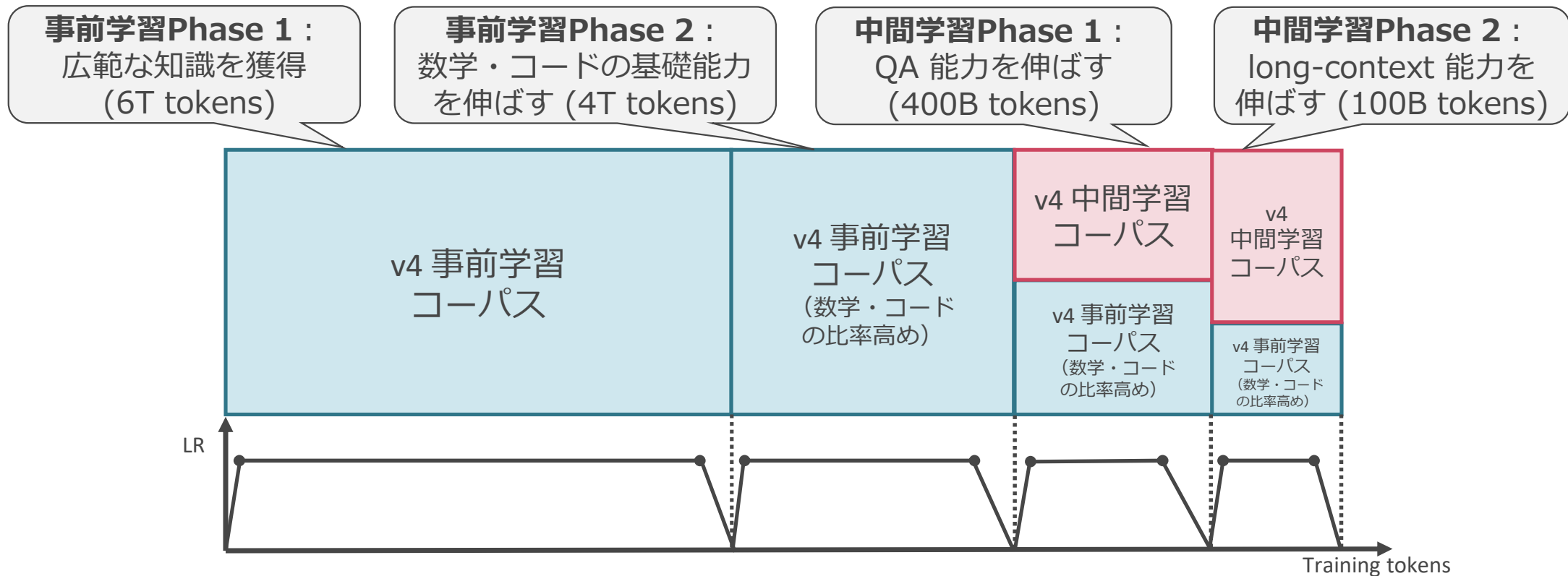


日本語安全性



Training Recipe of LLM-jp-4 Models

(Inspired by SmolLM3 [Bakouch+ 2025])



- コーパスv4 (19.5T、うち日本語0.7T)を整備
- 高品質な数学データセットMegaMath-Web-Pro-Max (Wang et al., 2025) を、Llama 3.1によるライセンス制約を避けるためにCommon Crawlデータから再構築
- LRスケジューラは学習管理の柔軟性からWSD (Warmup-Stable-Decay)を採用

SIP 統合型ヘルスケアシステムの構築 (2023-)

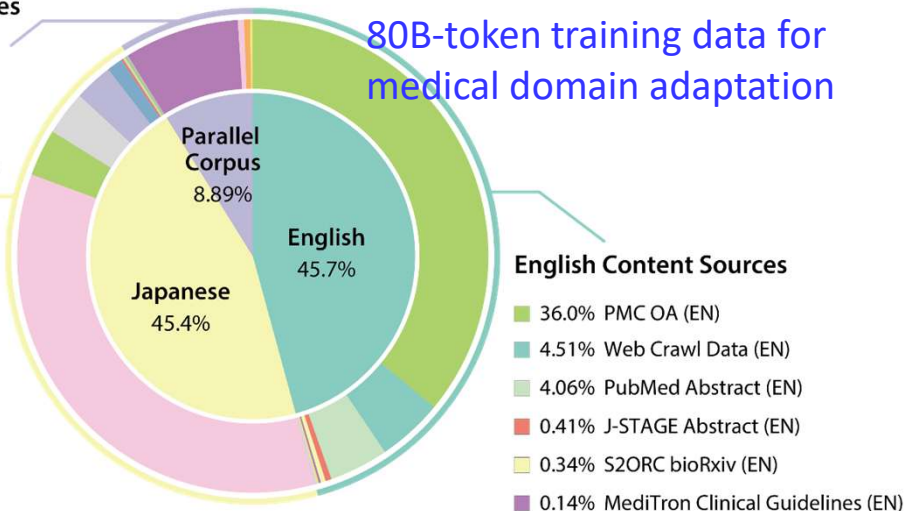


Parallel Corpus Content Sources

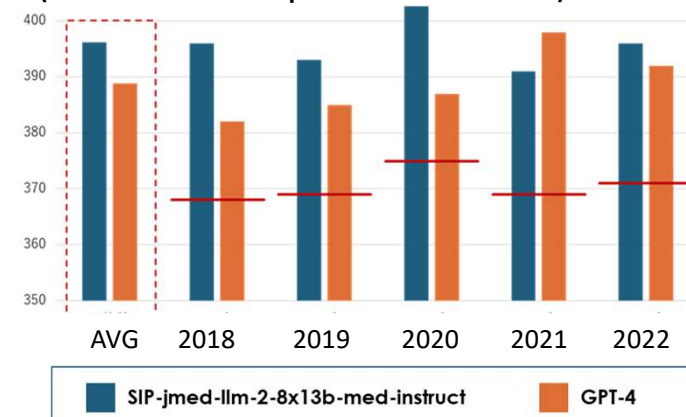
- 7.88% PubMed Abstract (JA/EN)
- 0.42% KAKEN (JA/EN)
- 0.42% J-STAGE Abstract (JA/EN)
- 0.16% JP Medical Abstract (JA/EN)

Japanese Content Sources

- 34.8% PMC OA (JA)
- 3.23% J-STAGE Full Text (JA)
- 3.03% J-GLOBAL Abstract (JA)
- 2.63% Web Crawl Data (JA)
- 1.08% Ichushi Web Abstract (JA)
- 0.15% J-STAGE Abstract (JA)



National medical licensing exam scores (SIP model outperforms GPT-4)



SIP-JMED-LLM-3 デモ (NII 小林和馬特任准教授作成)

≡ SIP-jmed-llm-8x13B-R1-Experimental ▾ +

デフォルトに設定



🕒 SIP-jmed-llm-8x13B-R1-Experimental

今日はどのようにお手伝いしましょうか？

+  Code Interpreter



🔗 提案

医療関連文書の要約

医療関連文書を自由文で要約する

医療関連文書の日本語訳

英語の医療関連文書を日本語に翻訳する

医療関連文書からの薬剤情報抽出

医療関連文書から薬剤情報を抽出する

AI for X

AI for Science

知識基盤

研究データ基盤

オープンサイエンス

LLM

透明性・信頼性

学術AIエージェントの開発

学術文書解析

2025.11 論文中心のPDF/XML文書からなる約120B
トークンの学術コーパス ver.1 構築



学術ドメインLLM構築

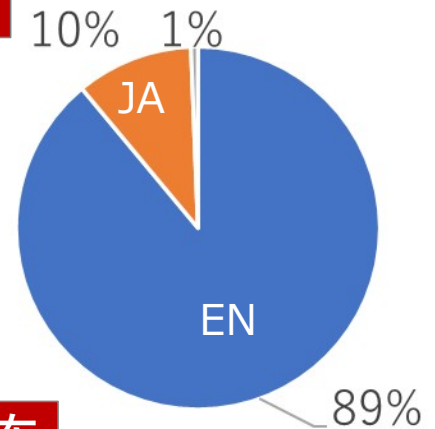
2026.3予定 llm-jp-4-8Bの継続事前学習および
チューニングによる学術ドメインへの適応



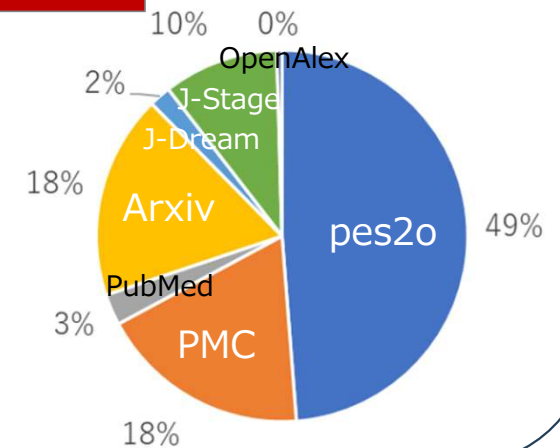
学術AIエージェント

2026.3予定 指示チューニングデータセット ver.1 の
設計・構築し、学術AIエージェント機能を実現

言語比率



ソース分布



AI for Scienceを支える情報基盤の高度化

令和7年度補正予算額（案）

5億円



現状・課題

- 世界的なAI for Scienceの流れ等により、研究データの重要性がこれまで以上に増しており、我が国の資産として有効に活用していくためには、AI readyな形で研究データを整えて保存・管理することが求められている。しかし、現状はメタデータ付与に多くの負担がかかっている状況であり、AI for Scienceの促進のためには、AI時代に即した研究データ基盤の構築が不可欠。
- また、「学術論文等の即時オープンアクセスの実現に向けた基本方針」（令和6年2月16日統合イノベーション戦略推進会議決定）では、2025年度から新たに公募を行う科研費等の特定の競争的研究費に対して論文及びその根拠データの即時オープンアクセス（OA）化が義務化されており、今後、研究データ基盤に膨大な研究データが集約されてくる見込み。
- 研究データは我が国の貴重な財産であり、それらを死蔵させないためにも、研究者の負担を軽減できる研究データ基盤の高度化が急務。

事業内容

● 先行事例やニーズの調査研究

0.5億円

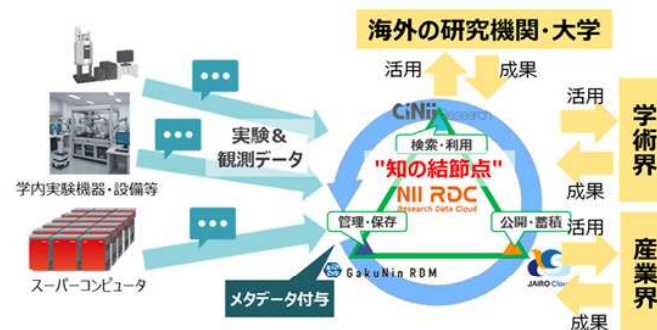
AI時代に即した研究データ基盤の構築のために必要な海外における研究データ基盤等の情報基盤の状況やAIへの対応状況等の先行事例及び情報基盤の構築・整備・管理に係る費用・工数や研究現場のAI活用に関する情報基盤へのニーズ等に係る調査研究を実施し、高度化する情報基盤の概念設計を行う。

● 次世代研究データ基盤の構築

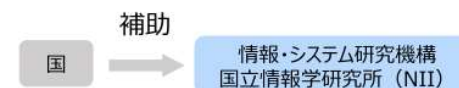
4.9億円

全国の研究者がAIやデータを最大限に活用できるよう、研究者の負担となっているメタデータ付与について、AIによるメタデータ付与支援機能を先行して開発を進めることで、いち早く研究データ基盤の高度化を実施し、AI for Scienceの促進及び我が国の研究力強化・産業競争力強化を目指す。

【目指すべき姿】



【事業スキーム】



（担当：研究振興局参事官（情報担当）付）

65

「AIによるメタデータ付与」の具体イメージ



研究者

研究メモ、
研究データ



手動整理

「メタデータ登録ツール」 (Excelシート)

AMSメタデータ項目リスト							
プロジェクトメタデータ							
番号	項目枠名	項目名	説明	入力方法	必須	入力1	入
1		データセットの名称 (日本語) *		入力BOX	○		
2		Title of Dataset (English)		入力BOX			
3		MSプロジェクト名 *		選択	○		
4		データID		入力BOX			
5		メタデータ登録日 *		入力BOX	○		
6		メタデータ更新日 *		入力BOX	○		
7		実験 (もしくは解析結果・解析ツールなど) の目的 (日本語) *		入力BOX	○		
8		Purpose of experiment (Purpose of Analytical results or Analytical tools) (English)		入力BOX			
9		実験状況など (もしくは解析結果・解析ツールの背景など) の説明 (日本語) *		入力BOX	○		
10		Description of experimental condition (Description of Analytical results or Analytical tools) (English)		入力BOX			
キーワード							
11		キーワード (日本語)		入力BOX	○		
12		Keywords (English)		入力BOX			

実験データセットのメモ 2025年2月20日
高脂肪食メタボマウス(C57BL/6マウス)の遺伝子発現量データの説明
最終更新日: 2026/2/10
概要: このデータはNIIの込山グループが取得した、次世代シーケンサー (NGS)による高脂肪食メタボマウス(C57BL/6マウス)の遺伝子発現量データです。マウスがメタボ発症前の未病状態および発症後の状態において、13臓器、16時点の遺伝子発現量をまとめたデータセットです。
データのURL: ...
対象生物: 高脂肪食メタボマウス(C57BL/6マウス)。高脂肪食を与える開始時点から、8週間の間にメタボ発症状態となります。
計測時点: 8週間の16時点
個体数: 10匹 (雄)
対象臓器: 以下の12臓器
脳、胸腺、心臓、肺、肝臓、脾臓、腎臓 (R,L)、精巣 (R,L)、精巣上体 (R,L)
計測法: 次世代シーケンサー(NGS)

実験メモ (「未病医療」領域での作例データ)



手動コピー＆ペースト



研究データ管理基盤GRDM

「AIによるメタデータ付与」の具体イメージ



研究者

研究メモ、
研究データ

AIエージェント
SIP-jmed-llm-3-8x13b-
AC-32k-instruct*

実験データセットのメモ 2025年2月20日
高脂肪食メタボマウス(C57BL/6マウス)の遺伝子発現量データの説明
最終更新日: 2026/2/10
概要: このデータはNIIの込山グループが取得した、次世代シーケンサー(NGS)による高脂肪食メタボマウス(C57BL/6マウス)の遺伝子発現量データです。マウスがメタボ発症前の未病状態および発症後の状態において、13臓器、16時点の遺伝子発現量をまとめたデータセットです。
データのURL: ...
対象生物: 高脂肪食メタボマウス(C57BL/6マウス)。高脂肪食を与える開始時点から、8週間の間にメタボ発症状態となります。
計測時点: 8週間の16時点
個体数: 10匹(雄)
対象臓器: 以下の12臓器
脳、胸腺、心臓、肺、肝臓、脾臓、腎臓(R,L)、精巣(R,L)、精巣上体(R,L)
計測法: 次世代シーケンサー(NGS)

実験メモ(「未病医療」領域での作例データ)

メタデータ項目	AIによる出力
データセットの名称	名称: 高脂肪食メタボマウス(C57BL/6マウス)の遺伝子発現量データ 説明: データセット全体を表す名称
プロジェクト名	NIIの込山グループ
データの種類	遺伝子発現量データ 説明: 遺伝子発現量データであること(次世代シーケンサー(NGS)で計測したもの)
測定対象	C57BL/6マウス、13臓器(脳、胸腺、心臓、肺、肝臓、脾臓、腎臓、精巣、精巣上体)
同種のファイル個数	記載がありません 計算したファイル個数: 記載がありません
測定装置の分類	次世代シーケンサー(NGS)
データセットの分野	ライフサイエンス
関連する器官名など	脳、胸腺、心臓、肺、肝臓、脾臓、腎臓、精巣、精巣上体

AIエージェント
による自動登録



研究データ管理基盤GRDM

*SIP-med-LLM/SIP-jmed-llm-3-8x13b-AC-32k-instruct・Hugging Face

「AIによるメタデータ付与」の具体イメージ

NII



NII平木研究員が京大理学部
博士課程時代を回想した作例

実験&解析時のメタデータ統合&README 作成 (ソフトマター物理分野)

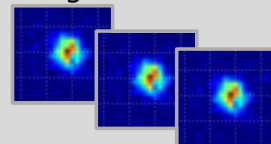
ラボノート (非デジタル)

509368 full power + 150 fs
Delay: -6000 pps ND: -18,000 pps
Stag. pres. 2,954 MPa T_0 : -0.370 mV
Source: 9.32
main: 6.43 509369 continue
Stag. Press (diff): 2,894 MPa, T_0 : -0.369 mV
Source: 8.96×10^{-1} Pa, diff: 2.75×10^{-4} Pa
main: 5.97×10^{-4} Pa

光源・Runデータ in SACLA DB

Run: 509368
Tag: 506609122
BM2_Coulomb:
7.7E-9
hv_eV: 11104.12

Image:

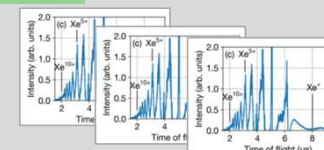


SACLA
MPCCD



持込機器計測データ in SACLA HPC storage

Tag: 506609122
Ion signal:



TOF detector



実験中に変換、統合
(free-text → tidy に)

実験後に変換、分析、統合

作成補助

出力 (Excel)

Run: 509368
 P_0 : 30 bar
 T_0 : 290 K
XFEL: 11 keV
NIR: 500 mJ
XFEL Delay: +150 fs
Run 概要: NIR full power

再整理

解析 DB (各種条件+解析結果)

Tag: 506609122
XFEL power: 400 PW/cm²
XFEL hv: 11.1 keV
NIR power: 40 PW/cm²
XFEL Delay: +150.24 fs
Bragg spot 面積: 10.9 pixel
Bragg spot 強度: 9.2E4 eV
Bragg spot 光子数: 230
Bragg spot 幅x: 2.6 pixel
Bragg spot 幅y: 1.9 pixel

作成補助

README 等の自動生成

SACLA実験データ解析データベース

1. 基本情報

- 作成日: 2025/12/9
- 作成者: 平木俊幸
- 最終更新日: 2025/12/9
- 最終更新者: 平木俊幸

2. DBの概要

本データベースは、SACLAにてXeクラスターを試料とした、NIR-pump-XFEL-probe 時分割計測実験における実験条件および解析結果を統合したものです。



NII平木研究員が京大理学部
博士課程時代を回想した作例

SACLA実験データ解析データベース

1. 基本情報

- 作成日: 2025/12/9
- 作成者: 平木俊幸
- 最終更新日: 2025/12/9
- 最終更新者: 平木俊幸

2. DBの概要

本データベースは、SACLA にて Xe クラスターを試料とした、NIR-pump-XFEL-probe 時分割計測実験における実験条件および解析結果を統合したものです。

手書きのラボノート、SACLA データベース（光源情報・画像）、およびSACLA HPC ストレージ（持込機器計測データ）の3つの異なるソースから情報を集約し、ショット毎の解析結果（Bragg spot の特性など）と実験パラメータを紐付けています。

3. DBの仕様

- DBの種類: SQLite 3
- ファイル名: analysis.db

スキーマ構成

主要テーブル: analysis_results

カラム名	データ型	説明	例 (図中より)
tag	INTEGER	ショット識別用の一意なタグID (Primary Key)	506609122
xfel_power_pw_cm2	REAL	XFELのピークパワー密度 (PW/cm^2)	400

4. DBの生成方法とデータ出典

本DBは、以下のソースから取得した値を Python スクリプト `create_analysis_db.py` を用いて結合・変換・解析して生成されました。

データソースの内訳

各カラムの値は、以下のソースに基づいています。

1. ラボノート (非デジタル)

- 使用データ: 実験の基本設定値、各 Run (測定の単位) の概要。
- DBへの反映:
 - `nir_power_pw_cm2`: ラボノートの「NIR: 500 mJ」および「Run概要: NIR full power」を元に、ビームサイズ等を考慮して算出。
 - `xfel_delay_fs`: ラボノートの「XFEL Delay: +200 fs」を初期値とし、解析により補正された実測値 (+170.24 fs) を格納。

2. SACLA DB (光源・Runデータ)

- 使用データ: ショット毎のタグ、XFEL モニタ値、検出器画像データ。
- DBへの反映:
 - `tag`: 「Tag: 506609122」等のIDをキーとして使用。
 - `xfel_hv_kev`: 「hv_eV: 11104.12」等のモニタ値を keV 単位に変換・丸め処理 (11.1 keV)。
 - `xfel_power_pw_cm2`: 「BM2_Coulomb」やパルスエネルギー情報を元に算出。
 - 解析結果 (右列): 検出器画像 (Image) を解析アルゴリズムにかけ、Bragg spot の形状 (面積、幅) や強度を算出。

3. SACLA HPC storage (持込機器計測データ)

- 使用データ: TOF (Time of Flight) などの波形データ。
- DBへの反映:
 - `tag`: データの同期 (Sync) 確認に使用。
 - ※本テーブルには直接的な数値は格納していませんが、解析時のイリング (Ion signal の有無など) に使用されています。

理分野)

データ
storage

TOF detector



補助

IME 等の自動生成

SACLA実験データ解析データベース

1. 基本情報

- 作成日: 2025/12/9
- 作成者: 平木俊幸
- 最終更新日: 2025/12/9
- 最終更新者: 平木俊幸

2. DBの概要

本データベースは、SACLA にて Xe クラスターを試料とした、NIR-pump-XFEL-probe 時分割計測実験における実験条件および解析結果を統合したものです。

XFEL: 11 keV
NIR: 500 mJ
XFEL Delay: +150 fs
Run 概要: NIR full power

NIR power: 40 PW/cm²
XFEL Delay: +150.24 fs
Bragg spot 幅x: 2.30
Bragg spot 幅y: 2.6 pixel
Bragg spot 幅y: 1.9 pixel

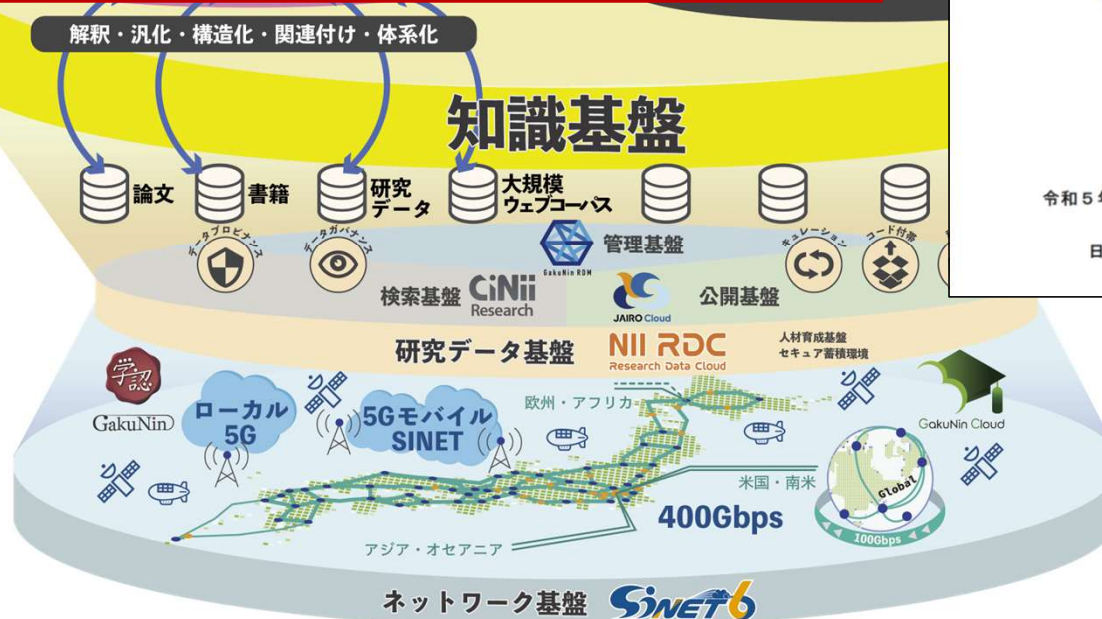
データ基盤から知識基盤へ

日本学術会議「未来の学術振興構想」の策定に向けた「学術の中長期研究戦略」に提案 (2022年12月16日)



研究コミュニティを支援する学術AIエージェント

- 研究データのメタデータ付与、構造化、自動処理、統合支援
- 科学的根拠や新規性を検証するDeep Research機能
- 学際研究テーマの提案や共同研究者探索



提言

未来の学術振興構想 (2023年版)

グランドビジョン^⑩

データ基盤と利活用による学術界の再構築

日本学術会議

令和5年(2023年)

日本学術会議

データ基盤と利活用による学術界の再構築

【概要】あらゆる分野のデータを集約するデータ基盤、さらに深化させた知識基盤の整備を進め、同時に利活用し得る人材を育成することでデジタル変革を推進し、学術界を再構築していく。

【キーワード】学術情報基盤、知識基盤、資料保存、データ駆動型研究、オープンサイエンス

ア 背景

現代では、学術の進化を継続的に支えるためにデータ基盤の構築とその活用が不可欠である。例えば、世界中の研究者が公開データを利用して共通の課題の存在や意外な発見の存在、予想外の事象まで認識できるようになった。しかし、日本の学術界はビッグデータやオープンサイエンスの恩恵を十分に享受できておらず、その主原因として、国内の統一データ基盤の利活用が滞っている。人文や社会科学においてもビッグデータの機械学習等による解析やシミュレーションが実用化しつつあるが、利用技術習得が十分に進んでいない。また、著作権に基づく複製形成(以下「権利」という。)が徐々に可能になってきたものの、日本における権利は道徳に由来し、今後には政府・地方自治体・民間に属する権利を明確に整理し、海外からの評価も得ていく必要がある。データ基盤を構築し知識基盤へと進化させ、あらゆる学術分野で「権利」を確保し、日本学術会議が推進する学術界の再構築が急務である。

イ 目的・目標

学術界の再構築から学術界の発展とイノベーションの創出を促進し、日本における学術界の再構築が可能となること。

エ 中長期の学術戦略

学術界全体のデジタル変革の具体策は、1)国内のデータ基盤・知識基盤の整備、2)データ基盤・知識基盤を中心とした学術の再構築、3) AI、計算科学、デジタルツイン技術、可視化等の分野融合的人材育成と技術開発、4)知識基盤の世界とのネットワーク化、5)知識基盤を広く開放し、利用を促すアトリビュート、6)IPRの基礎的・法的・デジタルインフラ利用、7)個人情報・法人情報等に関する意識を高めるための教育、8)地域の独自性に配慮したデータ収集と利活用方法の確立である。また、その設備体制として、高性能スーパーコンピュータ群と知識処理をサポートするための計算機群への大容量高速ネットワークアクセスを整備を進める。

ウ 国内の状況

2012年の時代が到来し、様々なデータが蓄積され、学術界の再構築が可能となる。

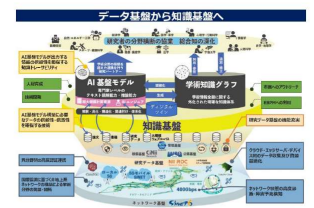


図11 データ基盤とその利活用による学術の発展
(出典: 学術の中長期研究戦略「データ基盤から知識基盤へ」を基に、本提言に独自に作成)