mdx: データ活用社会創成プラットフォーム

合田 憲人[1)], 遠藤 敏夫[2)], 小野 謙二[3)], 工藤 知宏[4)], 姜 仁河[4)], 小林 博樹[4)],
下川辺 隆史[4)], 菅沼 拓夫[5)], 杉木 章義[6)], 関谷 勇司[4)], 田浦 健次朗[4)], 竹房 あつ子[1)],
田中 良夫[7)], 谷村 勇輔[8)], 出口 大輔[9)], 中島 研吾[4)], 中村 覚[4)], 中村 宏[4)], 中村 遼[4)],
南里 豪志[3)], 塙 敏博[4)], 深沢 圭一郎[10)], 松島 慎[4)], 水木 敬明[5)], 宮嵜 洋[11)],
森 健策[12)], Lee Chonho[13)]

1) 国立情報学研究所 アーキテクチャ科学研究系
2) 東京工業大学 学術国際情報センター
3) 九州大学 情報基盤研究開発センター
4) 東京大学 情報基盤センター
5) 東北大学 サイバーサイエンスセンター
6) 北海道大学 情報基盤センター
7) 産業技術総合研究所 情報技術研究部門
8) 産業技術総合研究所 人工知能研究センター
9) 名古屋大学 情報戦略室
10) 京都大学 学術情報メディアセンター
11) 東京大学 情報システム部情報基盤課
12) 名古屋大学 情報基盤センター
13) 大阪大学 サイバーメディアセンター

kobayashi@ds.itc.u-tokyo.ac.jp

mdx: Platform for Creating a Data Utilization Society

Kento Aida[1)], Toshio Endo[2)], Kenji Ono[3)], Tomohiro Kudoh[4)], Renhe Jiang[4)],
Hiroki Kobayashi[4)], Takashi Shimokawabe[4)], Takuo Suganuma[5)], Akiyoshi Sugiki[6)],
Yuji Sekiya[4)], Kenjiro Taura[4)], Atsuko Takefusa[1)], Yoshio Tanaka[7)], Yusuke Tanimura[8)],
Daisuke Deguchi[9)], Nakajima Kengo[4)], Satoru Nakamua[4)], Hiroshi Nakamura[4)],
Ryo Nakamura[4)], Takeshi Nanri[3)], Toshihiro Hanawa[4)], Keiichiro Fukazawa[10)],
Shin Matsushima[4)], Takaaki Mizuki[5)], Hiroshi Miyazaki[11)], Kensaku Mori[12)], Chonho Lee[13)]

1)Information Systems Architecture Science Research Division, National Institute of Informatics
2) Global Scientific Information and Computing Center, Tokyo Institute of Technology
3) Research Institute for Information Technology, Kyushu University
4) Information Technology Center, The University of Tokyo
5) Cyberscience Center, Tohoku University
6) Hokkaido University Information Initiative Center
7) Information Technology Research Institute,
National Institute of Advanced Industrial Science and Technology
8) Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology
9) Information of Strategy Office, Nagoya University
10)Academic center for Computing and Media Studies, Kyoto University
11) Division for Information and Communication Systems, The University of Tokyo
12) Information Technology Center, Nagoya University
13) Cybermedia Center, Osaka University

概要

mdx: データ活用型社会創成プラットフォーム(以下「データプラットフォーム」という。)
について紹介する。これはデータ中心的な研究分野や、データ解析・データ科学的手法へ
の期待が高い分野に情報基盤を提供し、分野・産学間の連携を促進する取り組みである。

1　背景: データ科学・利活用の周辺状況

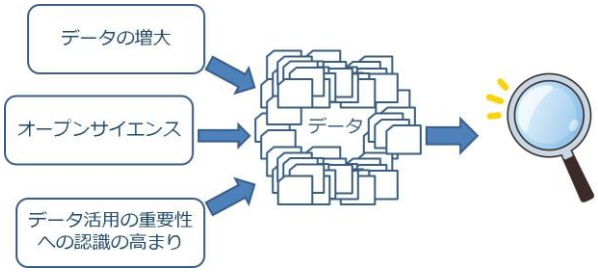

図 1　データ科学・利活用のイメージ

Society 5.0 は、地域、年齢、性別、言語による格差等の課題を解消し、地域の特色を活(い)かした多様な産業の活性化に貢献する社会を目指している[1]。その実現にむけてこれからの社会ではあらゆる分野でデータ活用が必須であり、学術コミュニティにおいても、社会実装を指向した研究分野が重要である（図 1）。実際に様々な研究分野でデータが研究における重要な資産となってきた[2]。背景として次の 5 つがあげられる。

1. 大量データの取得や利用が可能なった
2. 機械学習技術の進展により高次な情報抽出が汎用的かつ高精度に実現
3. 研究成果が少数の法則の発見ではなく大量のデータそのものという分野の出現
4. 従来の物理法則の忠実なシミュレーションと、大量のデータから計算結果を推論する手法の相乗効果
5. オープンサイエンスの流れで論文だけでなくデータも成果として共有する考え方

しかし、データ活用においては課題がある。必要となる異分野のデータの把握、それらのデータを活用する解析ノウハウの獲得である。更に解析に用いる IT インフラの構築を行う必要があり、データ活用に至る障害の大きさから容易ではない。Society 5.0 の実現にはこうした課題を解決して、「データプラットフォーム」の構築と推進を行う必要がある。そのためには Proof-of-Concept（PoC）を積み重ね、その効果を実証することが重要である。そこでデータプラットフォーム推進における課題に次の 4 つがあげられる[2]。

1. データが発生または活用する現場の（地方を含む）人材不足、あるいは、データのコミュニティがサイロ化している
2. データ活用のインフラ（基盤）の未整備と方法論の確立に関する課題。大量データの処理、リアルタイム処理に利用可能なインフラが存在しない。
3. データ流通における課題。個人情報保護とデータ活用を両立するためのポリシー等の方策が未整備。適切な活用を促進が困難な現状。人材不足により、どのようなデータが利用可能であるか、そのデータをどうすれば入手できるのかといった情報が広く共有されていない。データを他に活用させることへのインセンティブ
4. エコシステムの機能に関する課題潜在的なデータ活用の需要を拾い上げるエコシステムが存在していない。

mdx データ活用型社会創成プラットフォーム計画は、これらの潮流に基づき、広くは大学・研究機関、その中で情報系研究者、さらにその中で全国の情報基盤センターのような機関が何をすべきかを考えて生まれた。本計画は以下を具体的なアクションとしている[1]。

1. 情報基盤の設計と提供
2. 全国的研究コミュニティの創出
3. 将来像としての、大学間が有機的に連携した産・地域との連携

本論文では 1. と 2. について述べていく。

2　mdx:データプラットフォームと特徴

mdx:データプラットフォームについて紹介する。これはデータ中心的な研究分野や、データ解析・データ科学的手法への期待が高い分野に情報基盤を提供し、分野・産学間の連携を促進する取り組みである。国立情報学研究所(NII)との綿密な連携の元、2018 年初頭から構想を始め、全国の大学との連携・協力関係を築きつつある。全国利用を前提としたデータプラットフォームの先導的シ

ステム整備が 2020 年度末に東京大学柏 II キャンパスになされる予定である。mdx は次の 3 つの特徴を有している。

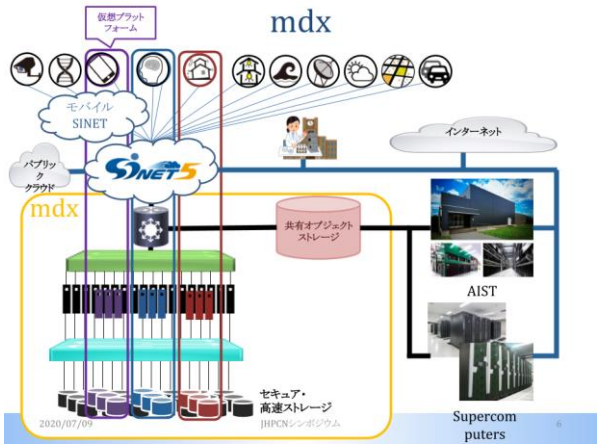

図 2　mdx:データプラットフォームのインフラ構成

① データ収集機能[3]：IoT データや大規模リアルタイムデータを円滑に扱えるセキュアな大容量通信回線の提供。現在、国立情報学研究所で各都道府県を 100Gbps 以上の帯域で接続した学術ネットワーク SINET 5（以下「SINET5」という。）を運用している。また SINET 5 に接続したモバイル網を「SINET 広域データ収集基盤」として提供している(https://www.sinet.ad.jp/wadci)。これは、各プロジェクトごとに、モバイル仮想閉域網(Mobile Virtual Private Network; Mobile VPN)を提供し、その VPN をクラウドなど計算基盤まで延伸することを可能にしたものである。データプラットフォームもこの広域データ収集ネットワークと連携し（図２）、日本全国からのデータの収集から蓄積・処理までをプロジェクトごとに閉じたセキュアな環境で行うことを可能にする。

② データ解析機能[3]：高度・高速な解析を実現する高性能計算環境やストレージの整備。mdx では、IaaS 環境と同様に、仮想化技術を用いて複数のプロジェクトに用途ごとに分離された管理者権限で自由にソフトウェアの設定が可能なプライベート環境であるテナントを提供する（図２）。個々のプロジェクトには一つまたは複数のテナントが割り当てられる。mdx は、広域ネットワークと連携し、共通のインフラを用いて使いたいときに短期間で広域ネットワーク、計算機、ストレージなどから構成される広域にまたがるテナントをプロジェクトに割り当てる。これは、利用する個々のプロジェクトから見ると、専用のインフラが整備されたかのように使える。mdx 全体の管理を行う管理者を「システム管理者」、mdx に資源を要求しテナントの割り当てを受けて利用するユーザを「テナント管理者」と呼ぶ。利用者の管理はテナント管理者に任されており、例えば利用者になんらかの形でテナント上に構築した情報システムへのログインを許す運用もあり得る。テナント管理者がポータルを介してテナントを要求すると、資源管理ソフトウェアとシステム管理者によりテナントが割り当てられる。仮想インフラ上の VM へ他のネットワーク（インターネットや、他のテナント、SINET5 上の VPN など）からのアクセスを許すかどうかは、テナント管理者がポータルを介して設定する。

③ 応用開発基盤[3]：多様な応用を実現する基盤ソフトウェアと共用データの提供。データプラットフォームはこれまでのスーパーコンピュータとは大きく異なる分野、異なる形態での利用が想定される。スーパーコンピュータは主に大規模な並列計算を高性能に行うためのものである。もちろんそれ自身はデータ処理性能も優れており、大規模データ処理や深層学習にも適している。そのレベルで計算機の構成を抜本的に変える必然性はないのだが、これまでの運用方法ではサポートできない利用形態が存在する。

一つの形態は「プラットフォームのプラットフォーム（メタプラットフォーム）」である。それは、データプラットフォームの「ユーザ」とは、単にデータ処理をするための計算資源としてプラットフォームを利用することにとどまらない。分野のデータレポジトリを整備しそれを分野研究者に公開する、つまり「プラットフォーム構築」をするユーザであり得るということである。このようなユーザをサポートするには、これまでのスーパーコンピュータの設計・運用とは異なる要素を取り入れる必要がある。外部ネットワークへの接続、恒久的な資源の割当て、分野プラットフォーム構築のために柔軟にシステムソフトウェアを構成できることが必要で、それと合わせて大規模機械学習処理やデータ同化シミュレーションなどのために、高性能な計算資源と連携できる必要がある。

また、分野データプラットフォームとしての利用のためにはこれまでの単年度ごとの申請よりも永続性を持った利用形態にする必要がある。またそもそも何に対して負担金を課すのか？という点も検討の余地がある(貴重なデータを共有するイ

ンセンティブを作るためには、使用ストレージ容量での課金に再考の余地がある)。そこでデータプラットフォームでは VPN 構築や仮想化・コンテナなどの技術を組み合わせる。ユーザにカスタマイズ可能なプラットフォームを提供しつつ、運用面の検討を行ってこのような利用モデルをサポートできるようにする。

3　全国的な研究コミュニティの創出

データプラットフォームを構築し提供する目的は、それを通じて分野を介した共同研究や産学連携を促進である。東京大学情報基盤センターは JHPCN という全国共同利用・共同研究拠点を全国 7 大学の基盤センター等と共に運営している。共同研究の対象分野は多岐に渡っているが大規模なスーパーコンピュータを用いた計算科学やその萌芽的な研究が中心である。その意義は、単に計算機を提供しているということではなく、計算科学の様々な分野の研究者が、情報学・高性能計算の研究者と、または異なる分野間で交流できることにある。

流体のようにシミュレーションの対象が違っても基づく物理・支配方程式が共通であったり、支配方程式が違っても共通の計算手法(例: 格子法や粒子法)を使うことはしばしば見受けられる。そして計算手法を高性能に実装する際の知見も分野間で共通部分が大きい。このようにスーパーコンピュータと高性能な計算手法、高性能システムの専門家を配した拠点（コミュニティ）には極めて大きな意義がある。

そこでデータプラットフォームでは全国的な研究コミュニティの創出を目指す。データの整備・流通を図るとともに、データ活用のコンサルティング機能、データ提供者/データ解析技術者/データ利用者のマッチング機能（図３）を全国に構築する。全国展開に当たっては、先導的機関による取り組みからユースケースとしての事例を蓄積するとともに、全国の大学がハブとなり、地方自治体や企業も含め展開するコミュニティの創出を目指す。

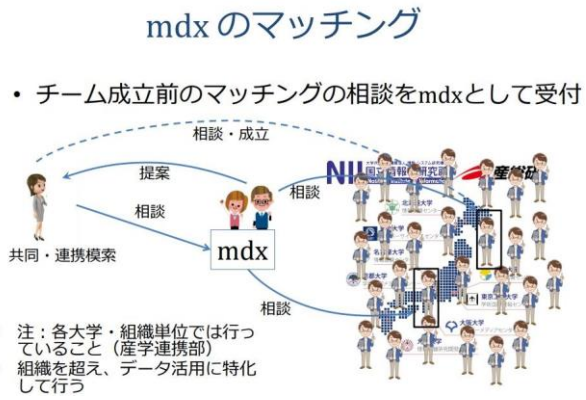

図３　大学間が有機的に連携した産・地域との連携（mdx のマッチング）のイメージ

4　将来

全国利用を前提としたデータ活用社会創成プラットフォームの先導的システム整備が 2020 年度末に東京大学柏第２キャンパスになされる予定である。今後はデータ活用を推進する体制の構築について早急に検討を進める必要がある。

5　まとめ

mdx: データ活用社会創成プラットフォームは Society 5.0 の実現に向けてデータ活用アプリケーションやデータ科学のために設計された基盤である。これまでのスーパーコンピュータと同様に、高速計算能力に優れるほか、「プラットフォームのプラットフォーム」とでも言うべき、データを活用するための基盤となるために、それ以外の特徴も備えている。

参考文献

1. 田浦健次朗、データ活用社会創成プラットフォーム計画について, 東京大学情報基盤センター年報第 20 号（2018 年度）, 54-59, 2019.
2. 文部科学省, データ活用社会創成プラットフォームの推進に向けた当面の整備方策（概要）: https://www.mext.go.jp/b_menu/shingi/chousa/shinkou/059/siryo/001.html
3. 中島研吾ら, Society 5.0 実現に向けた（計算＋データ＋学習）融合, 255-257, 大学 ICT 推進協議会 2019 年度, 2019.