情報処理学会論文誌　デジタルプラクティス　Vol.2　No.2　21–31 (Apr. 2021)

データ駆動型農業に向けた研究データ基盤の構築

川村 隆浩[1]　桂樹 哲雄[1]　小林 暁雄[1]　稲冨 素子[1]　江口 尚[1]

受付日 2020年8月25日, 再受付日 2020年11月6日,
採録日 2020年12月21日

概要：昨今，研究データの有効活用による研究の加速化や新たな知の創造が強く求められており，各国立研究機関（以下，国研）や大学では研究データ基盤の整備が進められている．そこで，農研機構では農研機構統合 DB（以下，統合 DB）および統合 DB と一体的に運用する AI 計算用スーパーコンピュータ（以下，スパコン）「紫峰」を構築し，2020 年 5 月より運用を開始した．統合 DB では，農業に関する様々な研究データ（ゲノムや品種，病害虫や環境に関する情報）に FAIR 原則に基づく共通メタデータを付けてカタログ化する．一方，自然科学特有の複雑に絡み合ったデータ間の関係を RDF（Resource Description Framework）や Property Graph，または RDB 形式で記述し，統計分析や機械学習の適用を容易にする仕組みも提供している．ここでは研究者に出口イメージを持ってもらうため，これらデータカタログから機構内クラウド DB，データ解析用スパコン，さらにはビジネス向けサービスポータルとをパイプラインのように接続した構成としている点に特徴がある．農研機構では，これら研究データ基盤の整備により，ゲノム解析から育種，生産，加工・流通に至るサプライチェーン上の様々なシーンにおいてデータ駆動型農業研究を加速していく．本稿では，研究データリポジトリの動向を概観した後，統合 DB のシステム構成，農研機構共通メタデータや紫峰について紹介し，最後に制度面・ソフト面の施策と今後の課題を示す．

キーワード：研究データ管理，データカタログ，データベース，農業研究

Research Data Platform for Data-driven Agriculture

Takahiro Kawamura[1]　Tetsuo Katsuragi[1]　Akio Kobayashi[1]　Motoko Inatomi[1]　Hisashi Eguchi[1]

Received: August 25, 2020, Revised: November 6, 2020,
Accepted: December 21, 2020

Abstract: Recently, research data are strongly required to be effectively utilized for research acceleration and new knowledge creation, and national research institutes and universities are developing research data management systems. The National Agriculture and Food Research Organization (NARO) constructed NARO Linked Database (narolin DB) and AI supercomputer "Shiho" operated together with the narolin DB, and started its operation in May 2020. In the narolin DB, various research data on agriculture (genome, variety, pest, environmental information, and so forth) are cataloged with common metadata based on the FAIR principle. On the other hand, the relationship between complicated data in natural science is described in Resource Description Framework (RDF), Property Graph, or RDB format, and it provides mechanisms to facilitate the application of statistical analysis and machine learning. In order for researchers to have an incentive to their data registration, our system is unique in that it is connected to the data catalog, the private cloud database, the supercomputer for data analysis, and the service portal for business, like a pipeline. Through the development of an agricultural research data platform, NARO will accelerate data-driven agricultural research at various stages in the supply chain ranging from genome analysis to breeding, cultivation, processing, and distribution. In this paper, after outlining the trend of the research data repository, the system architecture of the narolin DB, the NARO common metadata schema, and "Shiho" are introduced, and finally, we conclude this paper with institutional measures and future issues.

Keywords: research data management, data catalog, data base, agricultural research

[1] 国立研究開発法人農業・食品産業技術総合研究機構
NARO, Chiyoda, Tokyo 100–0013, Japan

1. はじめに

近年，農業分野においても研究環境のICT化が進み，研究データが電子化され，爆発的に増加している．これら研究データを適切に収集・管理し，オープン＆クローズ戦略に沿って必要な範囲で共有し，統計解析や機械学習といったデータ科学やAI技術を適用することで，研究活動の加速化や従来の農業分野・領域を超えた学際的な研究の創出が期待されている．データ駆動型農業とも呼ばれ，これまであまり共有されてこなかった研究データを外部機関，特に国際的な協調のために公開したり（オープンデータ），国家プロジェクト等のコンソーシアム関係者間で共有したり（ディスクローズデータ），あるいは競争的活用のため秘匿したり（クローズデータ）するなど，戦略的に活用することが求められている．一方で，研究データの信頼性確保や研究不正防止の観点からも研究データの保存（10年間）と管理が求められている．不正競争防止法における「限定提供データ」として研究データを保護する観点でも組織として一貫した管理が必要である．このような背景を受けて，内閣府から研究データ基盤整備とオープン＆クローズ戦略に関する提言が出され[1]，各国研・大学は研究データの保存・管理・検索等を支援するシステムの導入を進めている．

一方，国内最大の農業研究機関である農研機構においては，動植物や微生物に関するゲノム情報から育種情報，ウィルスや病害虫対策を含む栽培・生育管理情報，食品の加工・流通情報，ロボットトラクタの設計，気候変動や土壌といった環境情報など農業という研究分野の多様性に加えて，農林水産省管轄の複数の研究所を統合して2016年度に現体制に至った経緯もあり，研究データを分野横断的に連携させつつ，データ駆動型農業研究開発を促進するための統一的な研究データ戦略とそれを支える基盤整備が急務であった[*1]．

そこで今回，農業関連データに関する統合的なデータベース（以下，統合DB）[*2]，およびそれらを機械学習等のAI計算に活用するスーパーコンピュータ（以下，AIスパコン）の整備を実施した[*3]．本稿は，[2]にて概要を紹介した統合DBについて詳述するとともに，AIスパコン，農業データ連携基盤WAGRI，およびセキュリティ対策を加えた農研機構研究データ基盤の構築にあたって，設計上の論点や考え方に関する知見，経験を述べたものである．以下，2章にて研究データに関連するシステム・サービスについて概観した後，3章にて農研機構研究データ基盤の構成や機構共通メタデータなどについて述べる．さらに4章にてAI計算用スーパーコンピュータ「紫峰」について紹介した後，5章にてまとめと今後の課題，およびインフラ整備以外の制度面・ソフト面の施策を示す．

2. 関連動向

研究データリポジトリとは，電子的データの保存・共有等を行う広い意味での情報基盤であり，特に国研・大学等により整備され，公的資金により得られた研究データを再利用できる形で保存・管理し，当該機関内外へデータ利用サービスを供するものを指す[3]．各国研・大学が個別にサービスとして提供しているものを別として，システム・ソフトウェアとして第3者も利用できるものとしては，CKAN，DKAN，Dataverse，Open Science Framework，figshare，data.world，Mendeley Data，Dryad，Open ICPSR，Zenodoなどが挙げられる．2017年7月の記事だが，Dataverseを含む九つの研究データリポジトリの機能比較が[4]にまとめられている．以下，統合DBに特に関連する四つのシステムについて紹介する．

2007年頃よりデータポータルを構築するためのCKANと呼ばれるオープンソースソフトウェアが存在し，世界的なオープンデータサイトDatahub.ioや日本政府のデータカタログサイトdata.go.jpなどに広く利用されている．ただ，CMSとして作られていないことから機能やデザインの変更が難しいという問題点があった．そこで，エンタープライズ向けCMSであるDrupalのディストリビューションとして作られたデータポータルがDKAN（https://getdkan.org）である．細かなアクセス権の設定やメタデータ項目の追加にも対応し，デザインも自由に変えることができるため，米国農務省（USDA）を始め，国内の複数の自治体でもデータ提供プラットフォームとして利用されている．

Dataverse（https://dataverse.org）は，ハーバード大学が開発した研究データを共有，保存，引用，調査分析するためのオープンソースのWebアプリケーションである．長らく研究コミュニティは研究データの作成者に適切な学術的クレジットとビジビリティを与える方法を必要としていたことから，Dataverseではデータの出版や引用，長期的アクセス，再利用を可能とするため，研究データのWeb上での発見可能性の向上や文献管理ツールへのメタデータのエクスポートに対応している．特に，Dublin CoreとSchema.org[*4]の両メタデータに対応しており，これにより後述するGoogle Dataset Searchにも対応してい

[*1] 2019年に機構内で実施した事前調査によれば，研究データの約6割が個人PCやHDDに保存されており，機構内外での共有は遅れていた．

[*2] ここではデータカタログのような研究データリポジトリからRDBMSまでを広く指している．

[*3] http://www.naro.affrc.go.jp/publicity_report/press/laboratory/rcait/135385.html

[*4] Google，Microsoft，Yahoo，Yandexが開始したインターネット上の構造化されたデータのためのスキーマを生成・維持・促進するためのイニシアティブ．

る．フランス国立農学研究所（INRAE）や国際半乾燥熱帯作物研究所（ICRISAT）のデータポータルは Dataverse をベースとしている．

国内では，国立情報学研究所（NII）が米 Center for Open Science の研究データ管理用オープンソースソフトウェア Open Science Framework（OSF，https://cos.io/our-products/osf）を国内向けにローカライズした GakuNin RDM（https://rdm.nii.ac.jp）を 2020 年より大学・公的研究機関向けにリリース予定である．GakuNin RDM は論文を含む研究データを管理するための Web サービスであり，研究室内や共同研究者とデータを共有できる．サイズ制限のある標準ストレージに加えて，所属機関のオンプレミスのストレージを接続する手段やクラウド上のストレージを接続する手段も提供している．また，リポジトリ連携機能を使用して NII の研究データ公開基盤（WEKO3，https://rcos.nii.ac.jp/service/weko3）と接続し，データを公開することもできる．ほかにもバージョン管理機能や，ソースコードリポジトリ，文献管理ツール，図表共有サービス，データ解析，ワークフローエンジンへのアドオンを提供している．

ほかにも，論文のエビデンス情報などオープンアクセスとして公開可能なデータであれば科学技術振興機構が提供する研究データ共有プラットフォーム J-STAGE Data が存在する．これにより，科学技術情報発信・流通総合システム J-STAGE に登録された論文の画面に研究成果のエビデンスとなるデータの表示が可能である．

一方，Google は 2018 年 9 月に科学者やデータジャーナリスト向けに Dataset Search を立ち上げた．学術情報検索サービス Google Scholar と同様に，出版社やデジタル図書館，個人の Web ページなどをクロールし，膨大な Web 情報の中から必要なデータセットの発見を支援している．なお，情報は前述した Schema.org に準拠して整理されている．

今回，これらの既存システム・サービスに対して機構独自の研究データリポジトリを開発した理由は，主に 3.1 節，3.2 節にて示す機能追加・変更等を行うにあたって，オープンソースをベースにローカライズ，カスタマイズ開発を行うよりも，プロプライエタリではあっても開発者に直接アクセスできるソフトウェアをベースとすることを選択したためである．

3.　農研機構統合 DB

本章では，まず農業または農研機構向けに研究データ基盤を構築するにあたっての基本的なアプローチを 2 点述べた後，主要な構成要素，機能等の概要を述べる．

3.1　アプローチ

農研機構統合 DB には，データカタログとしての研究データリポジトリと機構における集約的なデータベースサーバ（組織内クラウド）としての二つの役割を持たせている．2019 年に約 3 ヶ月かけて機構内 18 の研究部門・センターへヒアリングを実施した結果，機構内には多様な分野，形式のデータが約 1–2 PB 蓄積されていることが分かった（**表 1**，**図 1** 参照）．研究データの共有，さらにはオープン化には研究分野ごとの性質に配慮した「戦略的な開放」[5] が必要とされているが，比較的オープン化しやすい農業環境に関するデータから比較的クローズな品種開発，完全にクローズな農創薬まで様々な分野が含まれており，それぞれの研究者の考え方は大きく異なる．また，農業分野に限らないが，外部に対してオープンな分野であっても分野 A と分野 B とで研究者が分かれており，必ずしも交流が多くなかったり，研究者の組織へのロイヤルティが低く，分野個別の学会や団体にデータを直接提出している場合も多い．さらに，フィールドワークが多く，必ずしも研究の現場における ICT 環境は整備されていない．しかし，農業におけるデータ・AI 活用の潜在的な可能性は高く，経済的にも政治的にも生産性向上や高品質化，機能性農産物など高付加価値作物へのニーズは非常に高い．

表 1　農研機構内研究データの分野別分類
Table 1　Disciplines of NARO Research Data.

大分類	データの内容	データセット数（概数）
植物	ゲノム，育種，栽培，画像等	230
動物	ゲノム，家畜診断，管理等	20
昆虫・線虫	害虫，診断方法等	50
微生物	病害，病害写真，病害同定等	60
食品	機能性成分，成分分析等	30
環境データ	気象，施設内環境，土壌等	60
その他	農作業，農業経済，農業用地利用状況，水利，水質，インフラ整備関連，地盤調査，実験ノート等	50

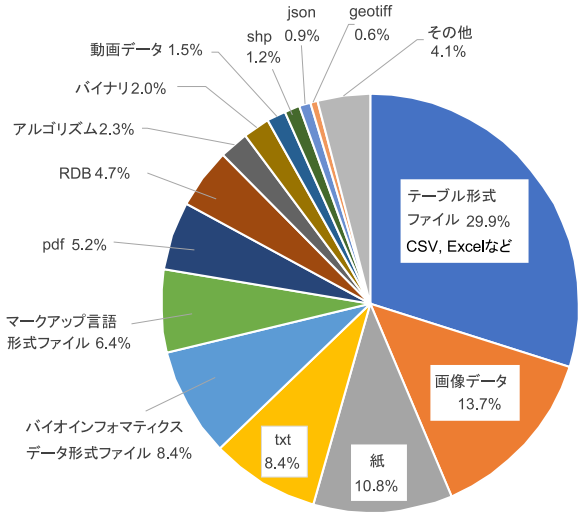


図 1　農研機構内研究データの形式別分類
Fig. 1　Formats of NARO Research Data.

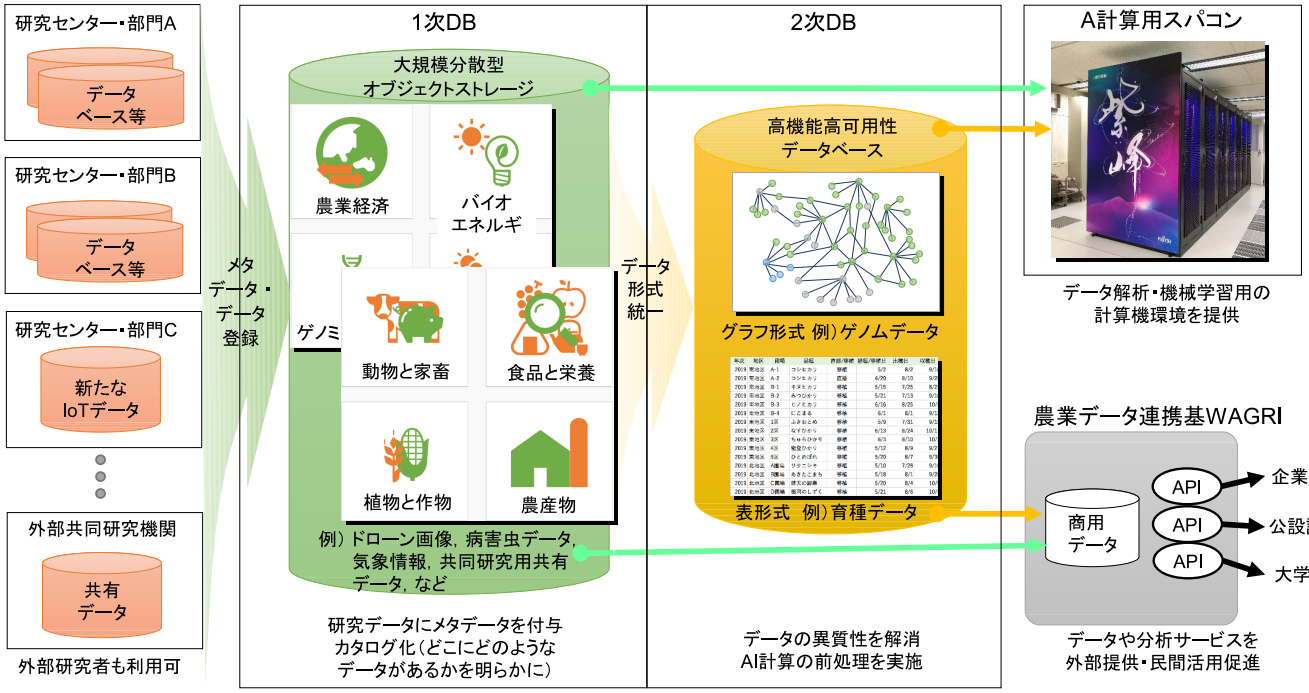


図 2　農研機構研究データ基盤の概念図

Fig. 2　Overview of NARO Research Data Platform.

そこで，まずは全体へ向けた対策として，機構内のどこにどのような研究データが存在するのかを見える化し，類似研究者や他分野への関心と興味を喚起したうえで，次に RDF などの仕組みでデータの相互連携を農業情報研究センターがサポートし，さらに AI 技術の活用によって新たな研究につながる（少なくともこれまでの作業が容易になる）までをストーリとして示すこととした．データを登録するインセンティブとして，研究者の具体的なメリットを意識，イメージしてもらうことが目的である．そのため，農研機構の研究データ基盤は研究データリポジトリとしての 1 次 DB と RDB やグラフ DB 群のプラットフォーム（機構内クラウド）である 2 次 DB，さらにデータ解析用の AI スパコンとビジネス向けの出口である農業データ連携基盤 WAGRI（3.4 節参照）とを接続したパイプラインのような構成として整備した．**図 2** に研究データ基盤の概念図を示す．そのうえで，制度面・ソフト面でビジネス化や分野横断研究を後押ししている．詳しくは 5 章にて述べる．

その他にも各分野への個別対策として，環境系データなどを公開するためのオープンデータカタログサイトの構築や，育種に関係するデータを国内で共有するための内閣府官民研究開発投資拡大プログラム（PRISM）への参画，病害・虫害に関する学習用画像データを WAGRI を介してビジネス利用するプロジェクトの立ち上げなどを進めているが，本稿では割愛する．

また，一般的に研究データリポジトリは研究成果を共有・公開する目的で作られており，研究成果の登録は研究終了時や年度末にのみ行われることが多い．しかし，品種開発や環境問題など研究期間が長期に渡る研究では成果が出て他者と共有するまでに長い時間を必要とし，データ利活用のサイクルが遅くなることが懸念される．また，ある研究の途上で生まれたデータが他者にとって別の目的で有用である場合も多い．そこで 1 次 DB では，研究過程で日常的に利用する共有フォルダと研究成果のリポジトリとを一体的に運用し，分野横断的なデータの利活用をより早いサイクルで回すことを企図している．フォルダやシンボリックリンクの概念やファイルの移動，コピー，リネームといった使い慣れた PC の機能を研究データリポジトリに持たせることで，研究成果としての静的なデータの公開と研究途上のデータの共有とを一つのシステムで実現することとした．

3.2　1 次 DB

1 次 DB は，研究者がアクセス権を設定した範囲でデータを共有することで，有用な研究データの存在をカタログ化，見える化し，新たな研究テーマの発見や既存テーマの新たな展開に役立たせるなど，研究者間の共創促進を目的としている．そのため，データ登録時には農業分野に特化した NARO Commons メタデータ（3.6 節参照）の付与を必須としたうえで，主に以下の機能を提供している（**図 3** 参照）．

- フォルダやファイルの移動，コピー，リネーム，シンボリックリンク機能．
- NARO Commons メタデータに対応したファセット検

図 3　1 次 DB における研究データセットの一覧・検索画面の例

Fig. 3　Screen shots of research data list and search in the first DB.

索機能.

- データの全文検索機能（主に研究記録，実験ノート，報告書，論文等の文書データを対象とし，文書構造解析を行った上，文書内のフィールド情報に基づいた検索が可能．ただし，暗号化データ除く）.
- バージョン管理機能.
- コメント・掲示板機能.
- データ利活用状況のダッシュボード機能（ダウンロード数やコメント数の推移など）.
- 業績管理系システムとの連携.

加えて，機構の組織構造や横断的なプロジェクト体制に合わせたアクセス権設定機能を設けている点に特徴がある．2 章で述べたシステムの中にはデータ公開を前提としており，細かなアクセス権設定ができないものもあるが，研究者に所有するデータの登録を促すためには厳密なアクセス権の設定は最も重要な機能である．特に，統合 DB は研究成果のリポジトリとしての役割と研究途上のデータの共有フォルダとしての役割を持たせているため，アクセス権設定は機微な問題でもある．具体的には，縦割りの研究課題[*5]や横断プロジェクト（ディレクトリのトップに相当），それらの配下のデータセット（フォルダに相当），個別のデータファイルに対して，以下の 4 種類のアクセス権を組織上の部門/領域/ユニット/グループ/チーム単位，あるいは任意のユーザ集合単位やユーザ個人単位に付与することができる（権限が重複した場合は，より強い権限が優先される）.

オーナ　プロジェクト，データセット，ファイルを作成したユーザ．下位階層のオブジェクトに対してすべての操作が可能.

編集者　オーナにより編集権限を付与された組織，ユーザ．該当するデータセット・ファイルに対してすべての操作が可能.

利用者　オーナにより利用権限を付与された組織，ユーザ．情報の修正や上書きはできないが，ダウンロードなどの操作が可能.

閲覧者　オーナにより閲覧権限を付与された組織，ユーザ．情報の修正や上書き，ダウンロードはできないが，閲覧が可能.

新しく作成されたデータセットやファイルは，プロジェクトや上位のデータセットの権限を継承し，自動的にアクセス権が設定される．また，編集者権限を持っているユーザはアクセス権の設定を変更することが可能であり，データセットのアクセス権を変更した場合，下位のデータセット，ファイルすべてに自動的に適用される．なお，メタデータは基本的に全ユーザが閲覧可能だが，データの存在自体を隠したいプロジェクトの場合はメタデータ自体を不可視化することもできる．さらに，内部 ID と機構外からの外部 ID（3.5 節参照）の権限管理や，機構内のユーザ管理系システム（LDAP など）と連携し，組織変更に合わせて自動的に ID やアクセス権を変更する機能，などを持っている.

[*5] 農研機構では，研究課題は大課題 18，中課題約 200，小課題約 650 に階層化されている.

1 次 DB のストレージには Amazon S3 に準拠した分散オブジェクトストレージをオンプレミスで導入し，スケーラビリティと事業継続計画（BCP）に備えた分散性を確保した．また，前述した機構内の研究データに関するヒアリングの結果，ストレージサイズは物理サイズ 5 PB（バックアップ等を除く実効サイズ 3 PB）とした．

3.3　2 次 DB および全体構成

データカタログである 1 次 DB に格納されたデータセットのうち，データセット間の連携を進めているプロジェクトや研究計画において連携が計画されているものを中心に，プロジェクトの進捗やデータ量，予算などに応じて優先度を付けて，各データ形式・項目の整備，RDB またはグラフ形式への変換を進め，2 次 DB への移設・新設を行っている．また，内閣府 戦略的イノベーション創造プログラム（SIP）第 2 期「スマートバイオ産業・農業基盤技術」の一環として，ライフサイエンス統合データベースセンターによるゲノム情報に関する RDF DB，TogoGenome（http://togogenome.org）[6] も 2 次 DB の一つとして運用している．

データ分析は前処理に掛かる工数が 9 割を占めるとも言われるが，2 次 DB では事前にオントロジー等を用いて項目を整理し，当該分野内で標準的あるいは相互利用に適したデータ形式に統一するなど，データのセマンティクスの異質性を解消しておくことで統計分析や機械学習による知識発見をサポートしている．例として，以下のようなデータ連携を通して相関関係や因果関係の発見をサポートする試みを進めている．

- 圃場単位で気象情報と土壌情報，病害虫発生状況を連携させることで発生予察につなげる．
- 研究機関や品種を横断してゲノム情報と表現型情報を連携させた RDF データベースを構築する．
- 品種の系譜情報と食品としての機能性成分情報とを結びつけサプライチェーンの上流からの育種開発を行う，など．

なお，2 次 DB にデータベースを設置した場合，当該データベース自体にメタデータを付けてデータベースへのアクセス先（Web サーバなど）を付記した，いわばメタデータだけのデータセットを 1 次 DB に登録している．これにより，2 次 DB 内のデータベースも 1 次 DB にてカタログの一つとして参照することができ，1 次 DB 上でメタデータを介して関連データセット・データベースの一覧性を確保している．そのため，将来的には多くのデータが 1 次 DB から 2 次 DB に移っていくことになると思われるが，あくまで 1 次 DB をユーザへのポータルサイトと位置づけている．

図 4 に，これまで述べてきた 1 次 DB，2 次 DB を含む農研機構研究データ基盤のシステム構成を示す．2 次 DB は外部 ID 認証システムや WAF（Web Application Firewall）をクラウド上に配置し，DB とストレージをクラウドおよびオンプレミスに置くハイブリッドな構成で運用している．他に 2 次 DB のプラットフォーム上にはアプリ

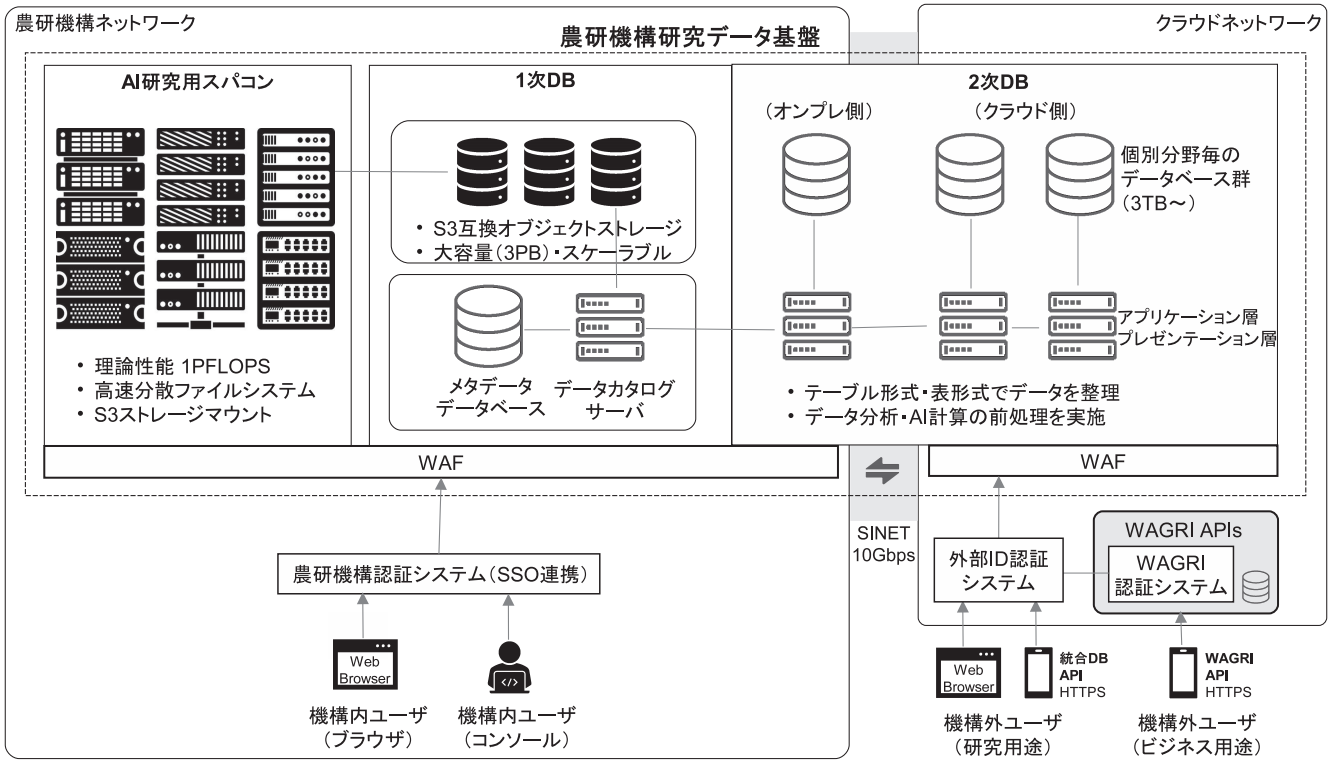


図 4　農研機構研究データ基盤のシステム構成

Fig. 4　Architecture of NARO Research Data Platform.

ケーション層，プレゼンテーション（Web）層として複数のサーバを配置している．現状，各研究部門・センターがOSやデータベースを含めて個別に管理運用している多数のサーバ（Webサーバを含む）を一括して管理することで，コストと保守の負担を軽減するとともに，セキュリティ等に関して統一的なポリシーを適用するためである．なお，DB，Webサーバなど2次DBのシステムは，オープンソースソフトウェアで構成することも可能である．

3.4　農業データ連携基盤WAGRIとの連携

統合DBは，SIP第1期「次世代農林水産業創造技術」の成果である農業データ連携基盤（WAGRI, https://wagri.net）との連携機能も有している．WAGRIは，農林水産省などのオープンデータや農業関連企業のクローズデータをAPIを介して提供（または収集）し，第3者が農業者向けサービスをビジネス展開するためのAPIポータルサイト（会員制）である．クラウド上に構築されており，データベースも一部含まれているが，データ提供者側のデータベースへのアクセスを介して取得されることも多い．いわばWAGRIはデータの蛇口であり，統合DBはデータのダムの一つとしての役割を果たしている．現在，農林水産省のHPで公開されているオープンデータや，県や公設試験研究機関（公設試）のデータを統合DBに格納し，WAGRIからAPIで提供していく計画がある．

WAGRIシステムの詳細は割愛するが，WAGRIを通して統合DBにアクセスする場合，クライアント側は認証用トークンを2重化してRESTでアクセスし，WAGRI認証を通った後，WAGRI APIから転送され，統合DBの外部ID認証（3.5節参照）を通ったうえで統合DB側のAPIを実行することができる．今後，PWA（Progressive Web Apps）によるクライアントアプリのサンプルを配布予定である．これにより，機構内外を問わず圃場や屋内施設，IoTセンサーやドローンなど研究データが発生した時点や必要とされた場所からデータの利活用が容易になることが期待される．

3.5　セキュリティ対策

統合DBとAIスパコンの利用にあたっては，外部との共同プロジェクト，コンソーシアム等でも活用するため，機構外研究者へのIDの払い出しも行っている．機構外ユーザであっても原理的に統合DBにおける全機能を利用可能だが，多くの場合はアクセス可能な範囲（公開範囲）を当該プロジェクト，コンソーシアムのデータに限定されている．2020年11月時点で外部IDは100以下である[*6]．なお，内部IDは機構内の全職員に最低一つ払い出されており，総数は約3000である．内，主な利用者は研究職約1800名である．

外部利用にあたってのセキュリティ対策は以下の3段階から構成される．

第1段　農林水産省研究ネットワーク（MAFFIN）firewall
- MAFFIN全体（農研機構を含む）への外部からの不正アクセスを監視・防御

第2段　農研機構firewall
- 利用者がいる研究拠点からのみアクセスを許可
- 外部だけでなく機構内部からの不正アクセスを監視・防御

第3段　統合DB・スパコンにおけるセキュリティ対策
- 利用者がいる研究拠点からのみアクセスを許可（第2段と二重設定）
- 特定のPCだけをアクセス許可（公開鍵認証方式）
- 不正侵入検知ツールを導入
- 利用者の操作を記録する監査ログを取得
- サーバのファームウェアの改ざん検知し正常なファームウェアに切り替える機能を導入，など

セキュリティ上の懸念から詳細を述べることは割愛するが，上記3段階はフィルタのような位置づけであり，段階が上がるにつれてセキュリティはより厳しくなっている．ネットワーク内のサーバ配置によっていずれかの段階を選択することは可能だが，統合DBとAIスパコンはいずれも3段階目に位置づけられており，すべてのセキュリティ対策が適用されている．なお，統合DB内のデータは，アクセス権設定機能によって公開制限レベル1〜3（非公開，制限共有，制限公開）のいずれかに設定することができるが，レベル4（非制限公開）に向けてはデータ公開用の別サーバ（3.1節のオープンデータカタログサイト）を準備中であり，こちらはセキュリティ対策の2段階目に置かれる予定である．

3.6　メタデータスキーマ

メタデータは，データの中身を説明するために必須なデータであり，内閣府の研究データリポジトリ整備・運用ガイドライン[3]においても，データ管理システムは研究データを共有するための国際的な基準であるFAIR原則（FAIR Data Principles，Findable，Accessible，Interoperable，Re-usable）に従い，“データ再利用を促進する付帯情報としてのメタデータを整備することなどにより，研究データの相互運用性を確保し，研究データの共有（公開を含む）を図ること”が求められている．統合DBでは，Data Catalog Vocabulary（DCAT, https://www.w3.org/TR/vocab-dcat-2/）やDataCite Metadata Schema（https://datacite.org/）といった一般の研究データ用メタデータをベースに，米USDA Ag Data Commonsや仏

[*6] 2020年11月時点では，申請ベースで必要性に応じて登録している．

表 2　NARO Commons メタデータスキーマ
Table 2　NARO Commons metadata schema.

項目名	説明	入力方法	必須
Title	データのタイトル	自由記述	○
Alternative name	データの別名	自由記述	
Author Name	作成者名	自動取得	○
Author ID	機構ID，Google Scholar, ORCID, Researchmap	自動取得	
Affiliation	作成者の所属機関	自動取得	
Project	データセット作成した研究課題名	プルダウンから選択	○
Author Contact address	作成者のメールアドレス	自動取得	
Contact person	連絡窓口	自由記述	○
Contact person ID	機構ID，Google Scholar, ORCID, Researchmap	自動取得	
Affiliation	責任者の所属機関	自動取得	
Project	データセット作成した研究課題名	プルダウンから選択	○
Contact address	作成者のメールアドレス	自動取得	
License holder	ライセンス保持者名	自由記述	○
License holder ID	機構ID，Google Scholar, ORCID, Researchmap	機構所属であれば自動取得，自由記述	
Affiliation	ライセンス保持者の所属機関	機構所属であれば自動取得，自由記述	
Project	データセット作成した研究課題名	プルダウンから選択	○
Contact address	作成者のメールアドレス	機構所属であれば自動取得，自由記述	
Coworker ID in NARO	農研機構に所属する協力者．機構 IDで入力	複数人の場合は",（半角カンマ）"で区切る	○
Coworker ID in other insititute	農研機構外の協力者（例えば公設試）．Researchmap IDで入力	複数人の場合は",（半角カンマ）"で区切る	
Coworker name in other insititute	農研機構外の協力者名（例えば公設試）	複数人の場合は",（半角カンマ）"で区切る	
Subject	データ分類	プルダウンから選択	○
Keyword	データのキーワード	入力を制御しながらKeywordを入力	○
Taxonomy Name	データベースが対象としている生物種をTaxonomy Nameで表記	生物種，Taxonomy ID，"NA"から選ぶ	○
Description	データベースの内容の説明	自由記述	○
Deposit Date	データをデータベースにアップロードした日	自動取得	○
Modified Date	データベースを更新した日	自動取得	○
Type	データセットのタイプ	データセット，データセット以外の2択	○
Format	データの形式	自動取得	○
Dataset Persistent ID	データに固有のID	自動取得	○
Language	言語	日本語，英語の2択	
Reference	データを取得した論文，データを活用した文献があれば記載	自由記述	
Database URL	データベースとして公開されている場合のURL	自由記述	
Temporal Coverage		自由記述	
Spatial / Geographical Coverage Area	データを取得した地域 POINT/POLYGON	自由記述	
License	データのライセンス	プルダウンから選択	○
Funding	予算的背景	プルダウンから選択もしくは自由記述	○
Classsified dataset	機1・機2・機3	プルダウンから選択	○

INRAE，ICRISAT における農業系のメタデータ項目を取り込んで，独自の NARO Commons メタデータスキーマを作成した．**表 2** にメタデータの項目を示す．国内では，オープンアクセスリポジトリ推進協会（JPCOAR, https://jpcoar.repo.nii.ac.jp/）が，メタデータの相互運用性を向上させ，学術的成果の流通を図ることを目的とし，JPCOAR スキーマというメタデータ項目を定めているが，NARO Commons 37 項目は JPCOAR スキーマで必須とされている 30 項目を含んでいる．そのうえで，農業研究および農研機構に必要な生物種や圃場の情報，課題番号などを加えた形となっている．**図 5** に NARO Commons メタデータの他スキーマとの主な関係性を示す．

また，ファセットによる検索性を向上するため，先に述べた機構内ヒアリングの結果を踏まえて以下の統制語を定義している．データ分類の大分類は，ゲノミクスと遺伝学，植物と作物，農産物，食品と栄養，農業生態系と環境，バイオエネルギー，動物と家畜，農業経済学，地図とマルチメディア，オンラインデータベースの 10 種である．

- 研究分野を表すデータ分類（大分類 10，中分類 60，小分類 106）．
- 日英対訳で 21,016 語のキーワード．
- 統制語 14,604 種の生物種．

なお，メタデータの入力を省力化するため，上位のフォルダのメタ情報を下位のフォルダ，データに自動的に継承している．また，メタデータを別途，Excel ファイルで記載し，データと一緒にアップロードすることで登録させる機能も提供している．これによってメタデータの再利用が可能となっている．

ただし，NARO Commons は多岐に渡る機構内データの共通セットとして作成したものであるため，今後分野ごとに拡張セットを追加していくことも検討している．

4.　スーパーコンピュータ紫峰

筑波山の雅名である「紫峰」と名付けた本スパコンの導入にあたっては，産業技術総合研究所のスーパーコンピュータ（AI 橋渡しクラウド）など国内の先進的な事例を参考にするとともに，機構内での計算機資源の利用状況を踏まえて必要な計算能力を想定した．紫峰は AI 計算において多用される行列演算性能に優れた NVIDIA 社 Tesla V100 を GPU に採用し，1 計算ノード内に NVLink で接続した 8 基の GPU を搭載している．また，全 16 計算ノードを InfiniBand 100 Gbps で接続し，計 128 基の GPU で理論ピーク性能 1 PFLOPS を実現している．国内農業系研究機関において PFLOPS クラスの計算機の導入は初である．ただし，紫峰は計算機としての性能を誇るためのものではなく，あくまで農業・食料という分野の研究者が利用しやすいデータ処理環境の提供を目指したものであり，コマンド，バッチ操作を主とする従来の利用に加えて，Web ブラウザから操作できるインタフェース（juypyter notebook や R studio など）や利用者のパソコンから遠隔操作できるインタフェースを通して，PC 操作のような

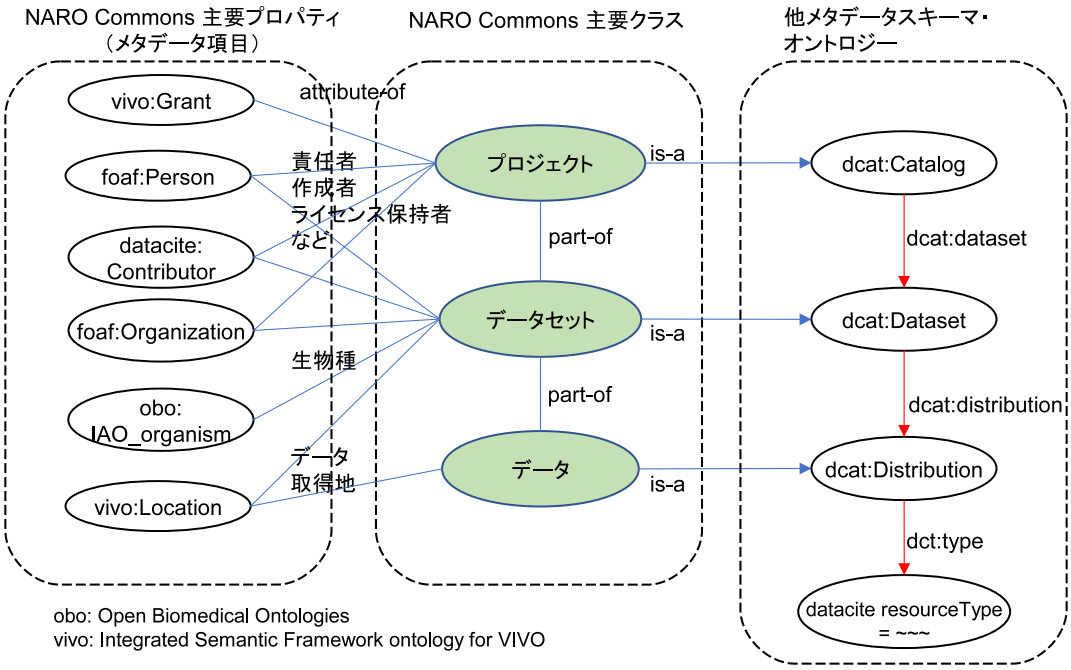


図 5　NARO Commons メタデータの他スキーマとの関係性

Fig. 5　Relationship of NARO Commons and other metadata schema.

ユーザビリティを提供している．最新の機械学習用のパッケージをあらかじめインストールした仮想化環境も提供し，利用者がそれらをすぐに利用できるようにもしている．

なお，1 次 DB 上のデータは，メニューから AI スパコンへのエクスポートを選択することで，AI スパコンのファイルシステムにコピーすることができる．今後は全国 20 拠点以上におよぶ機構内の各研究部門・センターを順次 SINET 経由で 10 Gbps 接続し，統合 DB と AI スパコンの利便性を高めていく計画である．

5.　まとめと今後の課題

本稿では，オープン＆クローズ戦略に沿って研究データを適切に共有し，データ駆動型農業を推進することを目的として，農研機構が 2020 年度より運用を開始した研究データ基盤について述べた．研究者はデータファイルのバックアップ先を統合 DB とすることで，データ管理の手間と義務から解放され，データの永続的で安全な保存（研究データの 10 年保存）に役立てることもできる．

今後，統合 DB に実装すべき機能としては，統一的な Identity Provider による国内研究者のシングルサインオンや，JaLC 等を介した研究データへの DOI 付与機能などがあげられる．さらに，外部研究データリポジトリとのメタデータ連携機能もあげられる．すでに JaLC から DOI を取得する際に JaLC が提供する REST API[*7] でメタデータを自動的に登録する機能を開発しているが，現在，いくつかのムーンショット型研究開発プロジェクトにおいて農研機構統合 DB の利用が検討されているため，前述した GakuNin RDM とのメタデータ連携も必要と考えている．学術機関リポジトリデータベース（国内各研究データリポジトリに登録されたメタデータを収集，一括検索を提供するサービス）への対応と併せて，今後の改修を検討していく．また，現在はユーザ登録なしでの外部アクセスを許可していないため，外部の検索サービスによるクローリングは許可していないが，先述したレベル 4（非制限公開）データを格納するためのデータ公開用サーバは外部サービスによる検索も許可する予定である．

一方で，研究データの登録と共有，および積極的な利活用をどのように研究者に根付かせていくかは依然大きな課題である．研究データ基盤というインフラを用意するだけではなかなか利活用が進まないことは周知の事実である．そこで，農研機構では制度面，ソフト面の施策として，前述した内閣府研究データリポジトリ整備・運用ガイドラインに沿って 2019 年度に農研機構版研究データベース運用ガイドラインを策定した．この中では，研究成果とする研究データは機構内で原則共有することを明記している[*8]．一方で，標準的な研究データ利用規約も定め，データ作成者のオーナーシップを明確にし，論文発表や特許提案時の権利保護を明記した．さらに，2021 年度からは研究管理等を定める機構内の規定に研究成果等管理ガイドラインを追加し，研究管理の一環として研究データを含む研究記録

[*7] https://japanlinkcenter.org/top/doc/190207_t_01_jst.pdf

[*8] システム上でアクセス制限されていても，所定の手続きをすれば入手できる．

の登録および責任者による定期的な確認も定める．これらと並行して，2020 年度より機構内で AI スパコンと統合 DB を用いたハンズオンを主体とした AI 教育プログラムを開始した．大学における半期 15 回分に相当し，データ科学や AI，特に機械学習に関する授業を農業研究を題材として予備コース，初級コース，中級コースを通して実施する．本プログラムは，2021 年度から県や公設試など機構外へも提供し，オープンサイエンスの考え方やデータ駆動型農業を広めていきたい．他にも，農水省 農業分野におけるデータ契約ガイドライン [7] に沿った産学官におけるデータ利活用に関する仕組みやビジネスモデルの検討，研究データ基盤や研究データ利活用に関する国内外の枠組・組織（RDA 等）への参画を通した他機関との協働なども検討している．

研究データ基盤の整備においてはインフラ構築だけでなく，制度・ソフト面の施策を同時に実施していくことが重要であり，それによって研究データを介した研究者間の共創が広がり，オープンサイエンスへ繋がっていくと考えている．

参考文献

[1] 研究データ基盤整備と国際展開ワーキング・グループ報告書，内閣府 研究データ基盤整備と国際展開ワーキング・グループ：(2019)，〈https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/houkokusho.pdf〉(accessed: 2020–10–28).

[2] 川村隆浩，桂樹哲雄，稲冨素子，鐘ヶ江弘美，江口尚：農業研究データ基盤整備に向けた統合データベースの構築，第 34 回人工知能学会全国大会論文集，2O4-GS-13-04（2020）.

[3] 研究データリポジトリ整備・運用ガイドライン，内閣府 国際的動向を踏まえたオープンサイエンスの推進に関する検討会：(2019)，〈https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/guideline.pdf〉(accessed: 2020–10–28).

[4] A Comparative Review of Various Data Repositories, Dataverse Project: (2017), 〈https://dataverse.org/blog/comparative-review-various-data-repositories〉(accessed: 2020–10–28).

[5] 林　和弘：統合イノベーション戦略におけるオープンサイエンス，STI Horizon, Vol.4, No.3, pp.42–47（2018），〈http://www.nistep.go.jp/wp/wp-content/uploads/NISTEP-STIH4-3-00145.pdf〉(accessed: 2020–10–28).

[6] T. Katayama, S. Kawashima, S. Okamoto, et al.: TogoGenome/TogoStanza: modularized Semantic Web genome database, Database (Oxford), pp.1–11 (2019).

[7] 農業分野におけるデータ契約ガイドライン，農林水産省：(2018)，〈https://www.maff.go.jp/j/kanbo/tizai/brand/b_data/attach/pdf/deta-50.pdf〉(accessed: 2020–10–28).

川村 隆浩

1994 年早稲田大学大学院理工学研究科電気工学専攻修士課程修了．同年，株式会社東芝研究開発センター入社．2001–2002 年米国カーネギー・メロン大学ロボット工学研究所客員研究員兼．2003–2018 年電気通信大学大学院情報理工学研究科客員准教授兼．2007 年より大阪大学大学院工学研究科非常勤講師兼．2015–2018 年科学技術振興機構情報分析室主任調査員，2019 年特任フェロー兼．2018 年法政大学理工学部非常勤講師兼．2019 年より農業・食品産業技術総合研究機構企画戦略本部データマネジメント統括監．2020 年より産業技術総合研究所人工知能研究センター招聘研究員兼．博士（工学）．2012 年国際会議 ISWC 10-Year Award 受賞．2013 年，2019 年人工知能学会研究会優秀賞受賞．人工知能学会理事，代議員，研究会主査などを歴任．主に非構造化データからの知識抽出とナレッジグラフ構築および分析・活用に従事．

桂樹 哲雄

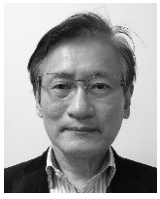
2005 年大阪府立大学大学院工学研究科機械系専攻海洋システム工学分野博士前期課程修了．2005–2011 年大阪府立大学大学院工学研究科航空宇宙海洋系専攻海洋システム工学分野博士後期課程．2011–2014 年奈良先端科学技術大学院大学情報科学研究科情報科学専攻博士後期課程．2014–2015 年日本学術振興会特別研究員（DC2，PD）．2015–2019 年豊橋技術科学大学情報・知能工学系助教．2019 年より農業・食品産業技術総合研究機構農業情報研究センター主任研究員．博士（工学）．2014 年国際会議 Metabolomics 2014 The Best Plant Metabolomics Poster Awards 受賞．2015 年 Molecular Informatics 誌 Best Paper Award 2014 受賞．海洋流体の数値計算手法に関する研究の後，バイオインフォマティクス，ケモインフォマティクスの分野での研究を経て，現在は農業データ利用促進のためのデータベースに関する研究・開発に従事．

小林 暁雄

2012 年豊橋技術科学大学大学院工学研究科電子・情報工学専攻修了．同年より同学 RA．2013 年豊橋技術科学大学情報・知能工学専攻助教．2017 年理科学研究所革新知能統合研究センター研究員．2020 年より農業・食品産業技術総合研究機構農業情報研究センター主任研究員．博士（工学）．自然言語処理，主に言語資源の自動構築に関する研究および，農業情報管理に関する研究に従事．

稲冨 素子

2004 年岐阜大学連合農学研究科博士課程修了．海洋研究開発機構フロンティア研究センター，茨城大学農学部，神奈川県自然環境保全センター，森林研究・整備機構森林総合研究所研究員を経て，2019 年より農業・食品産業技術総合研究機構農業情報研究センター上級研究員．博士（農学）．農研機構統合 DB の開発および普及に従事．



江口 尚

1980 年農林水産省入省．1989–2008 年，2014–2018 年農林水産研究情報総合センター研究ネットワーク（MAFFIN）および各種研究システムの企画・設計・構築に従事．2010 年放送大学大学院文化科学研究科政策経営プログラム専攻修了．2018 年より農業・食品産業技術総合研究機構農業情報研究センター専門職．修士（学術）．2007 年テクノロジーショーケース・イン・ツクバ 2007 年ベストインデクシング（ベスト・アイディア）賞受賞．JPNIC 評議員．スパコン，ネットワーク，データベース等のシステム設計・構築に従事．