

生命科学分野における高次元計測データを用いた研究活動への仮想データレイク技術の適用について

實本英之¹⁾, 黒川原佳¹⁾, 京田耕司²⁾, 糸賀裕弥²⁾, 木川隆則²⁾, 栃尾尚哉²⁾, 大浪修一¹⁾²⁾, 栗本崇³⁾, 笹山浩二³⁾, 重松光浩⁴⁾, 大黒毅⁴⁾, 伊藤哲郎⁴⁾

1) 理化学研究所 情報統合本部

2) 理化学研究所 生命機能科学研究センター

3) 国立情報学研究所

4) 日本電信電話株式会社

hideyuki.jitsumoto@riken.jp

Applying Virtual Data Lake Technology to Research Activities in the Life Science Field that Utilize High-Dimensional Measurement Data

Hideyuki Jitsumoto¹⁾, Motoyoshi Kurokawa¹⁾, Koji Kyoda²⁾, Hiroya Itoga²⁾, Takanori Kigawa²⁾, Naoya Tochio²⁾, Shuichi Onami¹⁾²⁾, Takashi Kurimoto³⁾, Koji Sasayama³⁾, Mitsuhiro Shigematsu⁴⁾, Takeshi Ohguro⁴⁾, Tetsuro Ito⁴⁾

1) RIKEN Information R&D and Strategy Headquarters

2) RIKEN Center for Biosystems Dynamics Research

3) National Institute of Informatics

4) Nippon Telegraph and Telephone Corporation

概要

本研究では、生命科学分野の研究活動における高次元計測データの利活用を高度化するための課題と、その解決策について検討した。実験データがロケーションや組織にまたがって存在する現状の中で、組織の垣根を越えた研究者の共創を生み出す仕組みが求められている。こうした課題に対応するために、仮想データレイク技術および超低遅延、超高速ネットワークを導入し、その機能性評価を通じて、共創のために必要な、実験データ利活用の高度化を支援する具体的な解決策を示す。

1 はじめに

1.1 生命科学分野における高次元計測データの重要性

生命科学分野において、撮像装置や NMR (Nuclear Magnetic Resonance: 核磁気共鳴分析装置) のような測定機器の技術進展により、大規模な高次元計測データが日々生成されている。これらの高次元計測データは、細胞の動態解析や構造の解明など、さまざまな研究において重要な役割を果たしている。特に、時空間解像度が高いデータは、生体の複雑な現象を理解するために不可欠であり、複数のロケーションや組織でやり取りすることによる共創が求められており、こうした実験データの利活用が可能なインフラの整備が急務となっている。

1.2 データ利活用の現状と課題

現在、高次元計測データ等の実験データの利活用にはいくつかの課題が存在する。まず、組織の垣根を越えた様々な研究者との共創を目的に、顕微鏡撮像データから生成した実験データ等を、ロケーションや組織をまたいで研究者間で共有（公開）することによる、共同研究のさらなる推進が求められている。しかし、現状ではこの実験データを効率的に共有するのは難しく、実験データの存在の発見やアクセスの容易さに課題が残っている。特に、異なるロケーションや組織間での実験データの一貫した管理や共有には、多くの手間や時間がかかり、研究の進行を妨げることがある。

研究活動の成果の最大化・スピードアップ・効率化を目的に、実験活動から実験データの処理までのプロセスを複数の研究者やスタッフで協業、

分業することが求められているが、現在はこのプロセスを支える基盤が十分に整備されていないため、協力がスムーズにいかない場面も見受けられる。効率的な分業や作業の流れを構築することが急務である。

さらに、NMR 共同利用[1]の促進や組織の垣根を越えた様々な研究者との共創を目指す中で、複数の実験装置で生成された実験データをロケーションや組織を超えて利活用するための基盤が現状では不十分である。データの取得や利用が煩雑で、スムーズなデータ利活用が阻害されている。この仕組みを高度化することで、研究のスピードと効率を大幅に向上させる余地がある。

また、実験データの更なる信頼性向上を目指して、実験データと二次データの紐づけやトレーサビリティの確立が必要だが、現時点ではデータの追跡や管理が十分に整備されておらず、実験データの信頼性や透明性に課題がある。この問題を解決するためには、信頼性を担保できる基盤の構築が望まれる。

1.3 本研究の目的と意義

本研究は、これらの課題に対処するために、仮想データレイク技術 [2] を活用した新しい実験データ基盤を提案し、その適用性を評価することを目的としている。仮想データレイク技術により、遍在している実験データを仮想的に統合し、実験データの一元管理と信頼性の向上を図る。また、組織を越えて参照権限のある実験データの一覧を提供することで、研究者間の迅速なデータアクセスを実現する。さらに、実験データのリネージの適切な管理を通じて、実験データの再現性と信頼性を確保し、生命科学分野における実験データ利活用の高度化を目指す。

2 仮想データレイク技術

2.1 仮想データレイク技術の概要

仮想データレイク技術は、データそのものではなくメタデータを収集することで、遍在するデータを仮想的に集約・一元化し、データ利用者がオンデマンドに必要なデータのみを効率良く取得して分析や解析処理に活用するための技術である。この技術は、従来のデータレイクが物理的に全てのデータを単一のロケーションに集約するのに対し、データ原本を遍在させたまま、仮想的に統合する点に特徴がある。これにより、データの移動

や複製を最小限に抑え、複数の組織やロケーションの垣根を超えた利活用を可能にする。

2.2 仮想データレイク技術の機能と利点

仮想データレイク技術には、いくつかの主要な機能と利点がある。まず、①仮想統合により、複数のデータソースを統一的に管理し、データ原本を遍在させたまま、データへの一貫したアクセスを提供することが可能である。この時、遍在させたまま都度転送するため、超低遅延、超高速ネットワークが必要となる。

また、②データセット管理により、論理的に構造化されたデータセットとしてデータを扱うことで、研究者は必要な情報に素早くアクセスすることができ、データの相互依存性や前後関係を把握しやすくなり、データの利用価値を最大化することができる。

さらに、仮想データレイク技術は③認可により、データの参照権限を一元的に管理し、組織を越えたデータ利活用を容易にする。これにより、適切なユーザーにのみデータを提供することができる。

加えて、仮想データレイク技術では、データリネージの管理が重要な役割を果たす。④リネージ登録により、データリネージを適切に管理することで、データがどのように生成され、加工され、利用されたかを追跡することが可能となり、研究の透明性と再現性を確保することができる。これは、データの誤用や改ざんを防ぎ、データの信頼性を高めるために不可欠な要素である。

仮想データレイク技術の一元的な管理とアクセス制御の機能により、データの整合性を維持し、誤った使用を防ぐことが可能である。これにより、研究者はデータの品質と信頼性を確保しながら、効率的なデータ利用を行うことができる。

3 適用性評価とその結果

3.1 適用性評価の方法とプロセス

本研究では、仮想データレイク技術の適用性を評価するために、複数の研究室間での実験データの利活用における効果を検証する方法を採用した。また、ロケーションが異なる複数の研究室に実験データを遍在させたまま都度転送するために、超低遅延、超高速ネットワークとして、ダイレクト光パスで接続するIOWN(Innovative Optical and Wireless Network)

APN(All-Photonics Network)を利用した[4]。その上で第1章で述べた課題に対応するためのシナリオを2つ設計した。

シナリオAでは、組織の垣根を越えた多様な研究者間での共創を促進するために、実験データ（顕微鏡撮像データから生成した、クラウドに最適化された次世代のバイオイメーシングデータ・フォーマットであるZARR形式実験データ[1]など）を複数のロケーションや組織をまたいで共有し、共同研究活動の効率化を図ることを目的とした（図1. シナリオAの構成）。具体的には、実験データの参照のみを行う場合を想定し、複数のデータソースを統一的に扱う利便性の効果を評価した。

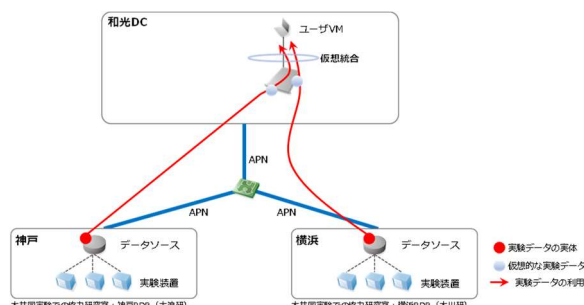


図1. シナリオAの構成

シナリオBでは、NMRデータの共同利用促進や組織を超えた研究者間での実験データ利活用の高度化を目的とし、複数の実験装置で生成された高次元計測データなど実験データを、他ロケーションからも利活用できる仕組みを評価した（図2. シナリオBの構成）このシナリオでは、実験データの参照と更新の両方を行う場合を想定し、シナリオAでの観点に加え、データセットを介した論理的な単位での実験データの管理、リネージ登録やアクセス制御の各機能の実現による実験データの信頼性向上を評価観点とした。

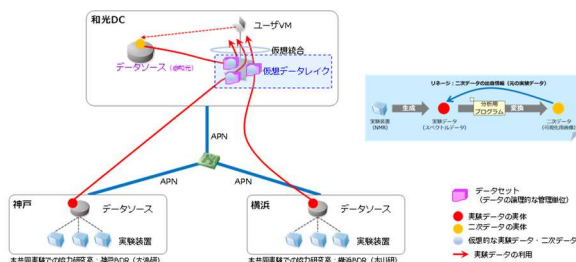


図2. シナリオBの構成

仮想データレイク技術および超低遅延、超高速ネットワークの評価環境を構築した[4]後、対

象研究室で各シナリオの機能の評価した。データの取得、共有、更新の効率や利便性について、第2章第2節で挙げた期待効果に沿って事前に設定した評価項目に基づき定性的に評価を行った。

3.2 評価結果と考察

評価の結果、仮想データレイク技術の導入により、いくつかの重要な効果が確認された。

まず、シナリオAにおいて、仮想データレイク技術を用いた実験データの仮想統合により、組織の垣根を超えた、閲覧可能な実験データの一覧提供が実現した。これにより、異なるロケーションや組織との間でのデータアクセスが容易になり、従来必要だった実験データの手動コピーや個別のアクセス許可取得の手間が大幅に削減された。この結果、実験活動から実験データの処理までのプロセスを、複数の研究者やスタッフとで協業、分業できることが期待でき、研究活動の成果の最大化、スピードアップ、効率化に寄与し得ることが確認された。2章2節に記載した機能のうち、①仮想統合の有効性が確認できたことになる。

シナリオBでは、NMRデータの共同利用の促進に向け、ロケーションや組織をまたがった実験データの統一利用ができることを確認できた。特に、仮想データレイク技術のデータリネージ管理機能により、実験データと二次データの紐づけが容易に行われることが確認できた。これにより、実験データの生成過程や利用履歴が正確に追跡可能となり、実験データの透明性と信頼性向上し得ることが確認できた。同じく第2章第2節に記載した機能のうち、①仮想統合、②データセット管理、③リネージ登録、④認可の有効性が確認できたことになる。

仮想データレイク技術の機能、導入による期待効果とその評価結果を、「表1. 期待効果と評価結果」にまとめる。

表 1. 期待効果と評価結果

機能	期待効果	評価結果
①仮想統合	Before: 複数のデータソースをバラバラに扱う After: 仮想データレイクで統一的に扱う (利便性向上)	実験データを統一的に活用でき、ストレスなく操作できることを確認
②データセット管理	データの管理・参照: 何らかの手段により Before: dir/file 単位で直接 After: データセットを介した論理的な単位 (カタログ化・属性登録の利便性向上)	実験データを論理的な単位で管理できることを確認
③リネージ登録	リネージの登録: Before: dir/file 単位に別手段で管理 After: データセットを介して管理 (データ管理性向上)	実験データを論理的な単位で認可制御できることを確認
④認可	認可ポリシーに基づくアクセス制御: Before: dir/file 単位で個別に After: データセット単位で仮想データレイクで (ガバナンス改善)	実験データと二次データの紐づけの登録・管理ができることを確認

これらの評価結果は、仮想データレイク技術が研究活動におけるデータ利活用の高度化に有効であることを示している。

4 まとめ

4.1 本研究の成果と意義

本研究では、生命科学分野における高次元計測データ等の実験データ利活用の高度化を目指し、仮想データレイク技術の適用性を評価した。

その結果、遍在する実験データの統一的な利活用が仮想データレイク技術によって可能となり、組織の垣根を越えた研究者の共創に寄与し得ることを確認できた。

また、仮想データレイク技術の適用性に留まらず、実験データ利活用における重要なポイントを抽出することができた (以下に示す)。

- ・ 実験データを論理的な単位でまとめて管理できること。また、そのリネージを管理できること
- ・ 実験データの真正性を保証した上で永続化できること
- ・ 適切な認可制御をした上で、組織の垣根を越えて参照可能な実験データの一覧が提供できること
- ・ 量的に増大する実験データや、サイズの大きな実験データをストレスなく取り扱えること
- ・ 実験データの管理性向上には、遍在したままの管理方式と集中管理方式との併用が必要であること

4.2 今後の展望と研究の方向性

本研究の結果を踏まえ、実験データ利活用の高度化に向けた検討課題も抽出できた。

- ・ 実験データの管理方式 (遍在したままの管理方式、集中管理方式の適性を踏まえた併用方法)
- ・ 研究プロセスのシステム化範囲 (効率性・利便性の向上だけでなく、実験データの信頼性向上も狙う)

今後の展望として、上記の課題の検討をさらに進めることで、仮想データレイク技術が生命科学分野における高次元計測データ等の実験データ管理と共有のひとつのモデルケースとなり、組織の垣根を超えた研究者の共創を支える基盤としての役割を担うことが期待される。

参考文献

- [1] 理化学研究所 NMR 研究基盤, 理研 NMR 研究基盤 YNMR, <https://www.ynmr.riken.jp>
- [2] 大村圭, ジェイホンジェ, 片山翔子, 河井彩公子, 柏木啓一郎, 馬越健治, 除補由紀子, 木村達郎, 組織を越えたデータ利活用を安全・便利にする次世代データハブ, NTT 技術ジャーナル vol.34, No.2, pp.9-13, 2022
- [3] Josh Moore, et al., OME-Zarr: a cloud-optimized bioimaging file format with international community support, *Histochemistry and Cell Biology*, Volume 160, Number 3, pp.223-251, 2023
- [4] 竹内規晃, 伊藤哲郎, 坂本誠治, 樋田基紘, 竹内太郎, 漆谷重雄, 栗本崇, 藤本幸洋, 窪

田佳裕，齊藤麻友子，黒川原佳，小林克志，
實本英之，2024，IOWN による大規模データ
利活用実験のご紹介，ADVNET2024 報告原稿