Vol.2020-IOT-48 No.27
2020/3/3

オープンサイエンス推進のためのデータ分析プロセス共有

横山 重俊[1,3,a)]　浜元 信州[1]　長久 勝[2]　藤原一毅[3]　政谷 好伸[3]　竹房 あつ子[3]　合田 憲人[3]

概要：研究データの再利用を進めるためには，研究データの公開だけではなく，論文に公表したデータ分析結果にたどり着くまでの分析プロセスを公開し共有する必要がある．その分析プロセスの中には，分析手順だけではなく，その分析手順に従ってデータ分析を行えるデータ分析システム環境として何をどう利用したのかを記述する必要がある．独自にその環境を構築した場合は，それをどう構築するかについて記述した手順も必要である．これらが揃うことで，第三者である別の研究者がいつでも研究データを再利用し，研究の再現検証が可能となるし，そのオリジナル研究に加えて自らの研究を派生させることがスムーズにできるようになる．本項では，オープンサイエンスを推進するためにこれらの再現手順を保存・管理・提供するデータ分析プロセス共有手法を提案する．

Sharing Data Analysis Process in Data Management Platform

Shigetoshi Yokoyama[1,3,a)]　Nobukuni Hamamoto[1]　Masaru Nagaku[2]　Ikki Fujiwara[3]
Yoshinobu Masatani[3]　Atsuko Takefusa[3]　Kento Aida[3]

1. 研究の背景

研究データの公開は，データ活用による研究が主流になる時代の研究推進方法であるとともに，研究に携わる者の責務として広まりつつある．この動きの中で良く言及されるのが「FAIR 原則」である [1]． FAIR は，「Findable（見つけられる），Accessible（アクセスできる），Interoperable（相互運用できる），Reusable（再利用できる）」の略で，データ公開の適切な実施方法を表現しており，データ公開の原則として広まっている．例えば日本では，国立情報学研究所が中心となって，FAIR 原則に沿った研究データ管理を実現するための基盤構築が進められている [2]．Findable は，データ検索基盤（CiNii Research）が，Accessible はデータ管理基盤（GakuNin RDM）が，Interoperable はデータ公開基盤（JAIRO Cloud/WEKO3）がそれぞれ主に担っている．図 1 にこれら三基盤を用いた研究の進め方を示す．

[1] 群馬大学
Gunma University
[2] ライフマティックス株式会社
Lifematics
[3] 国立情報学研究所
National Institute of Informatics
[a)] yoko@gunma-u.ac.jp

オリジナル研究を行っている研究者 α は，これらの基盤を用いて (1) 研究成果を論文として公開する，と同時に相互運用できる形にして (2) 研究に用いたデータを公開することができる．その研究を再現しようとする研究者 β は，その公開されたデータを使うために，公開された論文に含まれる情報を元にオリジナル研究者が使った (3) データ分析環境を復元し，さらに同様にオリジナル研究者が行った (4) データ分析手順を復元する．再構成したデータ分析基盤上でそのデータ分析手順を再実行することでオリジナル研究の中で得られたデータ分析結果を再現することができる（はずである）．

なお，本報告でデータ分析環境という場合，(1) ハードウェア，(2) オペレーティングシステム，(3) 研究分野毎に存在するアプリケーションソフトウェアスタックおよびワークフローエンジン，(4) データ分析ワークフロー群の４レイヤ全体を指すものとする．

2. 課題

前章に述べたような研究の進め方では，論文に付随するデータが公開されているため，論文のデータ分析の再現が可能である．ただ，そのためには論文情報から，その再現

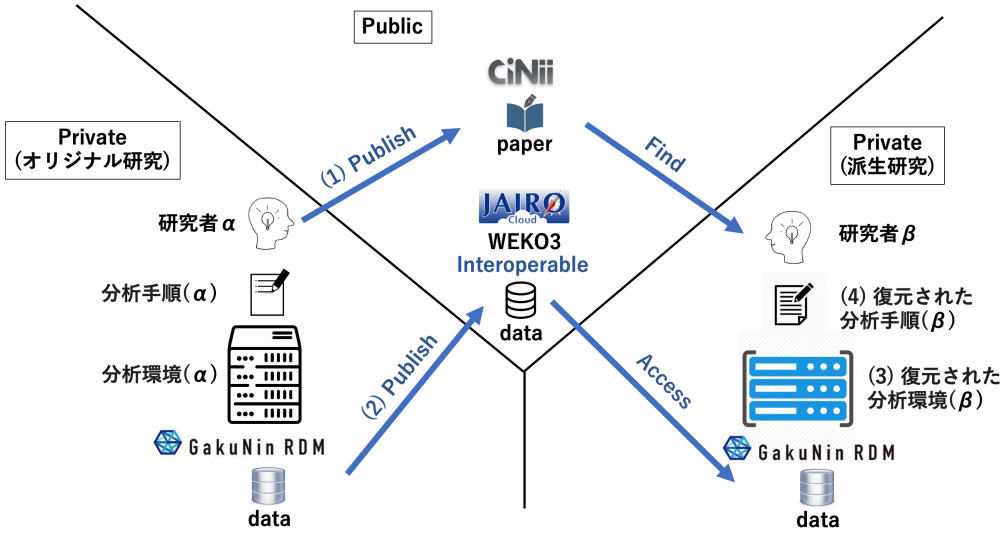

図 1　既存三基盤を用いた研究の進め方

のための手順をデータ分析ノートとして，再現を行う研究者側で正確に復元する必要がある．さらに，その手順を実行するためのコンピュータ環境も論文情報から高精度に復元する必要がある．これらの再構成にかかる復元の手間が大きいことや，論文に記載されている情報の不足などにより再構成自体が不可能である事例も多数あると報告されている [3], [4]．この再構成に関する障害が「FAIR 原則」にもとづく研究データ管理の好循環を創起させる際のバリアになっていると考えられる．

実際，報告者らが論文再現性の問題点を自ら体験することを目的として，バイオインフォマティクス分野のある論文の再現性を試行したことがある．著名な論文誌に掲載された質の高い論文であり，分析対象データは全て公開されており，論文中にも分析環境の再現方法も明確に記述され，利用したソフトウェアのバージョンも完全に記述された，論文の再現性を意識した再現性を試行するには，充分なレベルの論文であった．

この論文の再現性を情報学の専門家とバイオインフォマティクス の専門家がほぼ一ヶ月かけて実施した．これだけの質の高い論文について，これだけのコストをかけて分析結果の再現性について試みたものの，解析結果の完全な再現に成功しなかった．主な原因は二つ．一つはオリジナル論文で使用したソフトウェアと同じバージョンが既に入手困難となっており，再現時点で入手可能な一番近いバージョンを利用せざるを得なかったこと．もう一つは利用ソフトウェア内で乱数を暗に利用した箇所があり，その乱数発生に関する再現が出来なかったことであった．

3. 解決策

本研究は，この課題を解決するために，FAIR 原則の四つ目の構成要素である Reusable（再利用できる）部分に注目する．データの再利用環境を整えることで，派生研究の可能性を広げることができると共に，ひいては研究データの公開へのインセンティブを高める効果も持つ．つまり，「FAIR 原則」もとづく研究データ管理の好循環を創起させるために必要な最後のリンクである Reusable 部分への貢献を研究目的とする．

研究データの再利用を進めるためには，研究対象データの公開だけではなく，論文に公表したデータ分析結果にたどり着くまでの分析プロセスを公開し共有する必要がある．その分析プロセスの中には，分析手順だけではなく，その分析手順に従ってデータ分析を行えるデータ分析システム環境をどう構築するかについて記述した手順も必要である．これらが揃うことで，第三者である別の研究者がいつでも研究データを再利用し，研究の再現検証が可能となり，さらにそのオリジナル研究に加えて自らの研究を派生させることがスムーズにできるようになる．図 2 に示すように，これらの手順を保存・管理・提供する再構成基盤を新たに組み入れることで再構成に必要であった手間を大幅に削減できる可能性がある．

オリジナル研究を行っている研究者 α は，これらの基盤を用いて (1) 研究成果を論文として公開する，と同時に相互運用できる形にして (2) 研究に用いたデータを公開することができる．それに加えて (3) 分析ノートおよび分析環境構築ノートも公開する．その研究を再現しようとする研究者 β は，その公開されたデータを使うために，(4) 公開された分析環境構築ノートを使って分析環境を再構築する．さらに同様に (5) オリジナル研究者が公開した分析ノートを実行し分析の再現を行う．再構成したデータ分析基盤上でそのデータ分析手順を上記の通り再実行することで，オリジナル研究の中で得られたデータ分析結果を再現することができる．

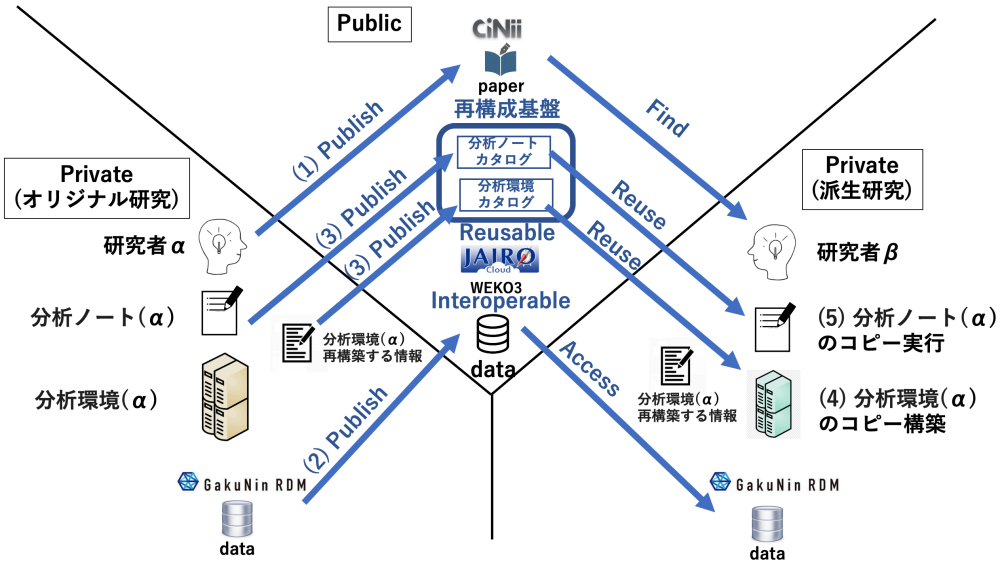

図 2　再構成基盤も含めた四基盤を用いた研究の進め方

これらデータ分析手順とデータ分析環境構築手順という二種類の手順を記述したノート（分析ノートと分析環境構築ノートと呼ぶ）をメタデータと共に，それら自身も，「研究データ」として研究データ管理基盤に登録する．この「研究データ」に DOI なども付与することで，他の研究データと同様の扱いを研究データ管理基盤内で受けることができるようにすることが，既存の三基盤との連携を考慮した，研究データ共有基盤を構成することが本研究のコアアイデアであり，その実現性と有効性の確認が本研究の主な目標となる．

図 3 に本研究で想定するプライベート空間での研究とパブリックな空間の関係を示す．プライベートな空間では研究のフェーズ毎に様々な中間的な分析対象データや分析結果データなどの成果物が存在するけれど，論文を公開する際には，それらを整理した後にパブリックな空間に公開するという研究プロセスを想定する．

4. 関連研究

公開情報を元にデータ分析環境を再現する取り組みとして，Materials Cloud[5] と BinderHub[6] について紹介する．Materials Cloud は研究データとともにデータ分析を再現するために必要な情報を合わせてアーカイブするクラウドサービスである．実際の再現環境自身はそれらの情報を元に再現者が自ら構築することを前提としている．一方 BinderHub は分析ノートが Jupyter Notebook[7] で書かれていることを前提として，その分析ノートを実行できる環境を提供するクラウドサービスである．

4.1 Materials Cloud

Materials Cloud は，研究論文に関連するデータとデータ分析に関する情報をアーカイブするツールを提供するオープンサイエンスプラットフォームである．アーカイブではシミュレーションソフトウェア，サービスおよび生データを提供している．これは非営利のサービスである．

アーカイブに蓄積されている各研究論文に関連する情報に付与したデータ記述子の公開により，データの作成者とキュレーターは自分の作業に適切なクレジットを獲得できるなど，再現可能な研究を促進できる仕組みがある．

4.2 BinderHub

BinderHub を使用すると，Git リポジトリに Docker イメージを構築するために必要な情報を登録し，JupyterHub と接続して，研究者が分析ノートを実行できる Jupyter Notebook 環境を docker コンテナとして提供することができる [8]．この際 Repo2Docker[9] と呼ばれるツールを利用することで Git リポジトリにある情報を利用して Docker イメージを動的に構築する仕組みが利用されている．

Materials Cloud の場合，再構築手順は研究データとともにストレージサービスである Zenodo[10] に格納されている．BinderHub の場合，再構築手順は Github に格納されており研究データとは別に管理され，それらの間はリンク情報で連携されている．Materials Cloud は再現環境自身はそれらの情報を元に再現者が自ら構築する必要があり，この再構築手順は標準化されていないため，論文毎に様々で再現者の研究再現の障害となる可能性がある．また，BinderHub は一つのコンテナイメージでデータ分析ノートが実行される場合には有効であるけれど，マルチノードからなる計算機クラスタのような実行環境が必要な場合には適用できないという制約がある．

本報告の取り組みは，複雑な計算機環境を必要とする分析ノートも実行できる環境構築を再現者がアーカイブサー

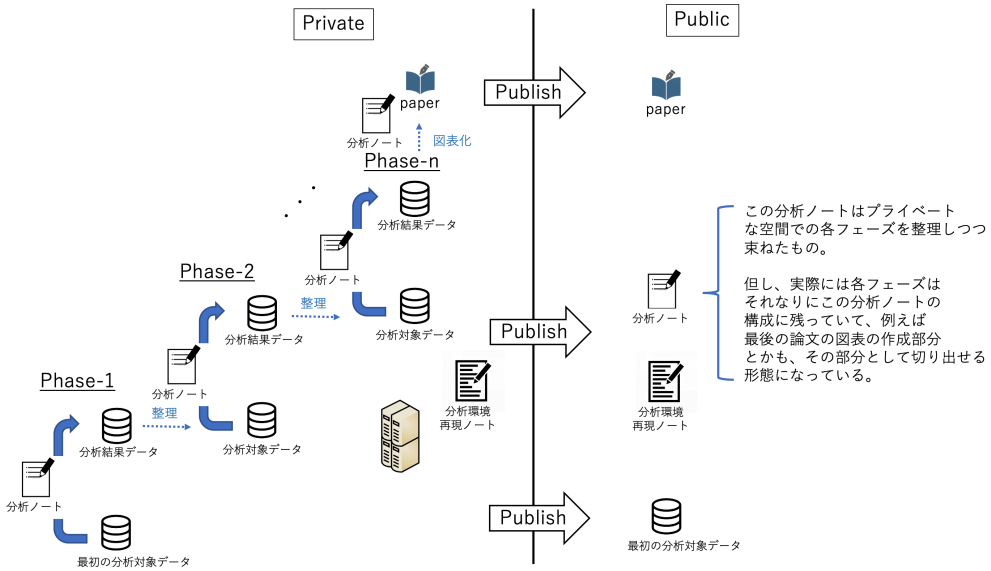

図 3 プライベート空間での研究とパブリックな空間の関係

ビスからシームレスに実施でき，スムーズにその環境の中でデータ分析を行う環境へ移行できる仕組みを実現することを，オープンサイエンス推進のためのデータ分析プロセス共有の目標とする．

5. 研究内容

第 3 章で述べた解決策の実現を目指したデータ分析再構成基盤として，分析ノートおよび分析環境構築ノートを保存・管理・公開する分析ノートカタログと分析環境カタログに加えて，それらを使って実際に分析を再現するための機能を持つ必要がある．すなわち，分析手順に従ってデータ分析が行えるデータ分析システム環境がその分析を実行したい研究者の必要な時にいつでも，彼らが使えるどのコンピュータリソース上にも構築できるために「可搬性の高いデータ分析基盤の構築方式」を研究する必要がある．

また，実際のデータ分析の実施の際には別の研究データ管理基盤である「研究データ保存サービスとの連携」が重要となり，その連携方式に関する研究も行うことが重要である．

再構成基盤は，現状欠けている分析手順の公開，それとその手順に従って分析を実施するデータ分析環境の構築手順の公開を推進するために，前者を保存・管理する分析ノートカタログと，後者を保存・管理する分析環境カタログ二つの部分から構成される研究データ管理基盤として構成し，既存基盤との連携も含めた研究データ管理基盤のプロトタイプを開発する．そのプロトタイプ上でいくつかの具体的な研究における分析プロセスの再現を実施することを通じて研究対象である「可搬性の高いデータ分析基盤の構築方式」と「研究データ保存サービスとの連携」を評価する．プロトタイプに適用する各方式について本章で説明する．

5.1 可搬性の高いデータ分析基盤の構築方式

分析環境カタログに保存されている分析環境再構成手順に従って，研究再現を目指す研究者がデータ分析環境を自らが利用可能なコンピュータリソース上にデータ分析環境を再構築し，その分析環境の中にデータ分析ノートを持ち込み，データ分析作業を再現するための方式を，可搬性の高いデータ分析基盤の構築方式と呼び今回の研究対象の一つとする．

可搬性の高いデータ分析基盤の構築方式として，学認クラウドオンデマンド構築サービス [11], [12] と Literate Computing for Reproducible Infrastructure(以下 LC4RI) [13] を組み合わせる方式を提案し，その有効性を再構成基盤のプロトタイピングと試用により検証する．

学認クラウドオンデマンド構築サービスは，オンデマンドにクラウド環境を構築するソフトウェアサービスであり，クラウド環境構築用テンプレートを指定して起動するだけで，クラウド環境の構築・再構築が可能となる．

特徴はクラウドプロバイダごとの操作方法の違いを吸収し，複数のクラウドプロバイダの同時利用や切り替えの煩雑さを軽減することである．一方，LC4RI は，計算機システムの構築・運用の方法論であり，システムの構築・運用作業に必要となる自然言語による技術情報，実行コード，さまざまな状況で起こり得る複数の実行結果を Jupyter Notebook として 1 つの文書にパッケージし，実行可能な手順書を実現している．

学認クラウドオンデマンド構築サービスの持つ基盤構築用の SDK を利用した LC4RI による分析環境構築手順書を作成することで，可搬性の高いデータ分析基盤の構築が可能となると期待できる．この分析環境構築手順書を学認クラウドオンデマンド構築サービスでいうアプリケーションテンプレートとしてその Jupyter Notebook を分析環境カ

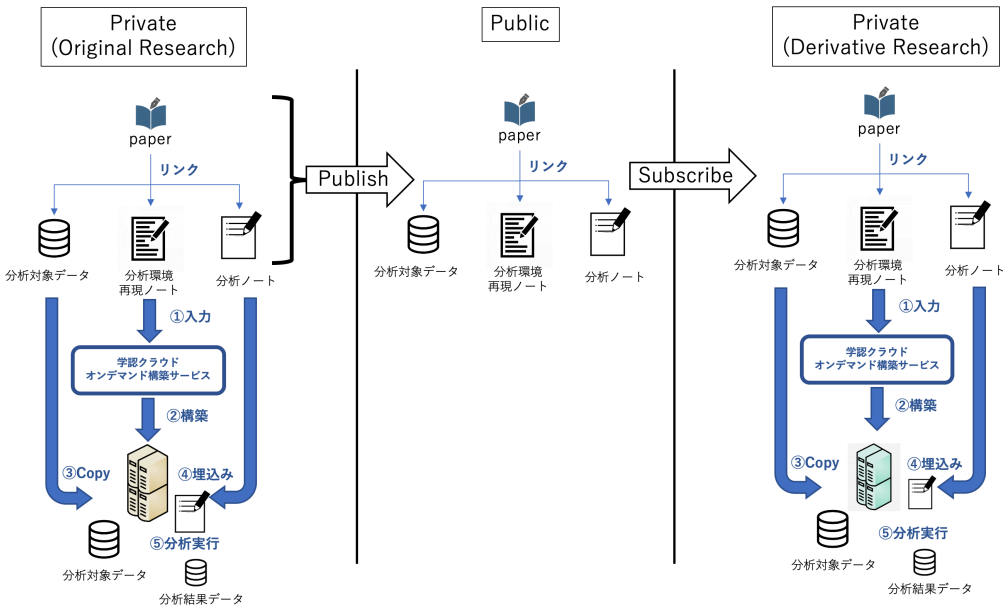

図 4　可搬性の高いデータ分析基盤の構築方式

タログに Publish することが研究成果の公開作業の一部として定着することが重要である．この作業はオリジナル研究者自身が自らの研究における分析環境を構築する際に，その簡便さのために，元々プライベートな情報としても同様の Jupyter Notebook を利用していると今回は仮定して取り組みを進める．

データ分析ノートは Jupyter Notebook で記述されているものを対象とし，そのデータ分析ノートには分析対象となる研究データの保存先および処理手順が記述され，さらには分析後のデータ保存先も記述されているとする．

このデータ分析ノートを使って実際にデータ分析することのできる環境を作成するための手順書も LC4RI に則り Jupyter Notebook で記述されているとする．この Jupyter Notebook は学認クラウドオンデマンド構築サービスを活用することで，様々なクラウド基盤上に動的にその環境を構築できるアプリケーションテンプレートとして実現しておく．さらには，後者の分析環境を再構築する Jupyter Notebook を実行することで構築されたデータ分析基盤に前者のデータ分析ノートなどの必要データを取り込むための手続きも環境構築の一部としての後者の Jupyter Notebook の一部として用意する．

これら二種類の Jupyter Notebook をメタデータと共に，「研究データ」として登録する．この「研究データ」に DOI なども付与することで，他の研究データと同様の扱いを研究データ管理基盤内で受けることができることが，既存の三基盤との連携を考慮した本取り組みのコアアイデアである．可搬性の高いデータ分析基盤構築に関する提案方式を図 4 に示す．

5.2　研究データ保存サービスとの連携方式

研究データ保存サービスとの連携方式については，研究データ保存サービスとして JAIRO Cloud/WEKO3 を例として，構築した可搬性の高いデータ分析基盤からの連携を実現できるように，データ分析基盤構築時にはそれらの連携ができる認証連携などを組み込む．さらにはデータ分析ノート内には JAIRO Cloud/WEKO3 内のデータへのアクセス情報が格納されていることで，そのデータをデータ分析基盤へ自動コピーできる仕組みも持つ．認証連携には学認基盤を用い，さらにデータのコピーについてはすでに実績のある JupyterHub WEKO 拡張機能を活用する方式を基本に，さらに使い易く，コピーすることが現実的でない程度の大容量のデータにも適用できる連携方式を探る．

具体的には図 5 に示すように既存の WEKO と JupyterHub の連携のために開発された JupyterHub WEKO 拡張機能を流用する．この流用先として，分析環境再現ノートを WEKO3 から取り込み学認クラウドオンデマンド構築サービスへの入力とする部分と，オンデマンドで構築した分析環境に分析ノートを埋め込む部分の二箇所とする．このことで，新たなソフトウェア開発・保守することなく研究データ保存サービスの連携範囲を拡充できる．論文に紐付けられた，分析対象となる研究データおよび，可搬性の高いデータ分析基盤の構築方式で扱う分析環境再構築のための「研究データ（データ分析環境再構築のための Jupyter Notebook）」と分析手順再現のための「研究データ（Jupyter Notebook で記述された分析ノート）」を組み合わせて一気に研究再現を可能とする．

6.　今後の進め方

これらの方式に従って再構成基盤のプロトタイプを実装

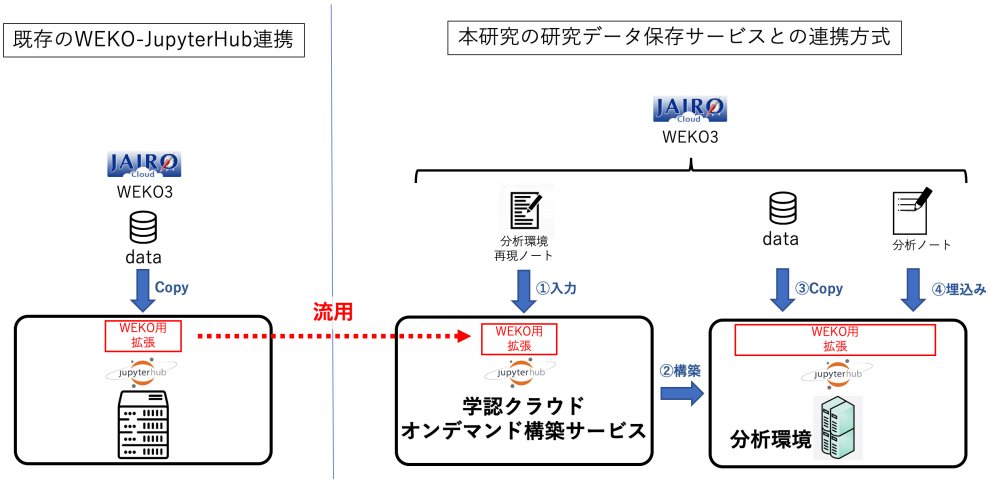

図 5　研究データ保存サービスとの連携方式

し，具体的な研究データおよびそのデータの分析事例を決め，研究データ分析作業の再現性について実証実験を行う．実証実験で利用する研究データやデータ分析事例候補としては，現在実績がある生物学分野や人文科学分野に加え，新たな領域として数学・数理科学分野を想定している．

この運用を通じて本研究で提案する可搬性の高いデータ分析基盤の構築方式および研究データ保存サービスとの連携方式について評価する．主な評価項目としては，研究再現の精度および研究再現や派生研究を行う研究者に対する利便性を考えている．

この取り組みでは分析環境として Jupyter Notebook を分析ノートとする JupyterHub 系列のものを対象としている．しかしながら，分析環境再現ノートを使って再現できる分析環境としては特にこの限りではない．例えば，ゲノム分析分野でよく利用されているワークフローエンジン Galaxy を主体とする系列の分析環境など幅広い研究分野の分析環境も今後の研究対象として横展開が行える．

この際にも，その研究分野特有のワークフロー の保存形式に沿った情報を共有することで本報告と同様の議論が可能である．そういう意味では二つのレイヤの Jupyter Notebook のうちで本報告で本質的なのは，分析手順再現のための「研究データ（Jupyter Notebook で記述された分析ノート）」である．

プロトタイプを開発するための具体的な検証用プラットフォームとしては，現在実証実験環境として国立情報学研究所から提供されてる GakuNin RDM システム [14] を活用して，サンプルのデータ解析ノートおよびデータ解析環境構築ノートの両方を GakuNin RDM システムからアクセスできるストレージサービス内に格納し，それらを使ってクラウド内に研究の再現環境を作れることを検証・評価する予定である．

参考文献

[1] FORCE11: FAIR 原則，https://www.force11.org/fairprinciples (accessed on 01-07-2020).
[2] 山地一禎: 日本の研究データ基盤の構築，https://rcos.nii.ac.jp/document/ (accessed on 01-07-2020).
[3] Feitelson, D. G.: From repeatability to reproducibility and corroboration, *ACM SIGOPS Operating Systems Review vol. 49, no.1*, pp. 3–10 (2015).
[4] Gruning, B. and et al.: Practical Computational Reproducibility in the Life Sciences, *Cell Systems volume 6, issue 6*, pp. 631–635 (2018).
[5] AiiDA: Material Cloud, https://materialscloud.org/ (accessed on 01-07-2020).
[6] Jupyter: BinderHub, https://mybinder.org/,https://github.com/jupyterhub/binderhub (accessed on 01-07-2020).
[7] Jupyter: Jupyter Notebook, http://jupyter.org/ (accessed on 01-07-2020).
[8] Docker: Docker, https://www.docker.com/ (accessed on 01-07-2020).
[9] Jupyter: repo2docker, https://github.com/jupyter/repo2docker (accessed on 01-07-2020).
[10] CERN: zenodo, https://zenodo.org/ (accessed on 01-07-2020).
[11] 竹房あつ子，佐賀一繁，丹生智也，横山重俊，合田憲人: 学認クラウドオンデマンド構築サービスの推進，AXIES 2018 年度年次大会 (2018).
[12] 国立情報学研究所: 学認クラウドオンデマンド構築サービス，https://cloud.gakunin.jp/ocs (accessed on 01-07-2020).
[13] 長久　勝，政谷好伸，谷沢智史，中川晋吾，合田憲人: Notebook を介した作業ノウハウの継承・移転を分析するための基盤，インターネットと運用技術研究会, 情報処理学会，pp. 1–6 (2019).
[14] 国立情報学研究所: GakuNin RDM（研究データ管理基盤），https://rcos.nii.ac.jp/service/rdm/ (accessed on 01-07-2020).