

RIKENのデータ管理状況

情報統合本部 データ管理システム開発ユニット
實本英之

R2DMS: 研究情報管理システムとは

- ◎ 研究データを概要やメタデータ等と合わせ、保存・管理・共有・公開するシステム
 - 研究データ管理
 - 日々の研究データを保管する
 - 研究データを限られた範囲で共有し、共同研究を促進する。
 - データ生成機器から計算処理、データの保管・公開までをシームレスに接続する
 - 研究データ公開・利活用
 - 研究データに適切な説明・メタデータを付け公開する
 - 業績発表に関して査読時のデータ提出および、客観的証明データを公開する
 - 様々なカタログサービス・検索サービスに研究ページのリンクを提供する

R2DMS の仕組み

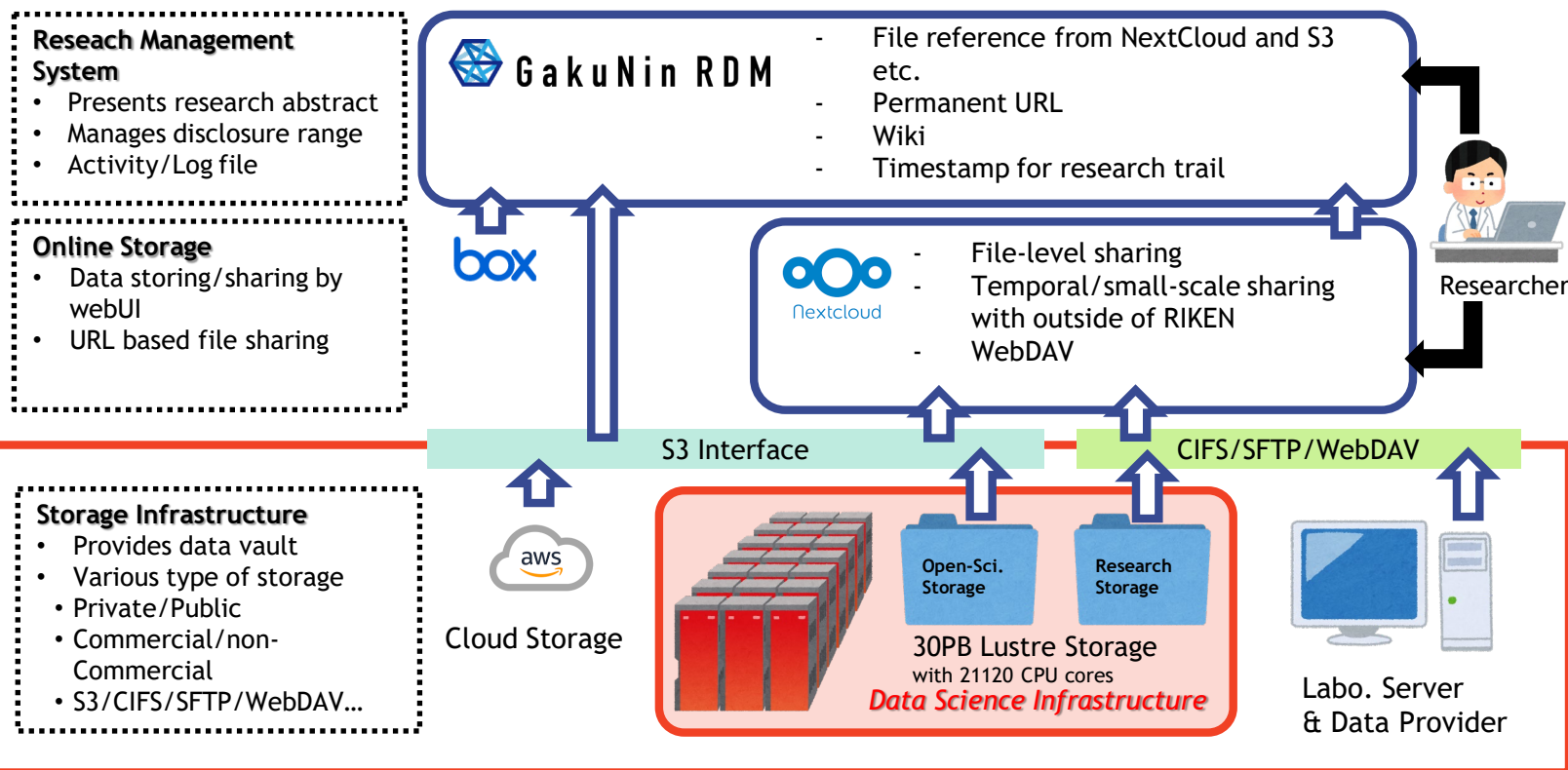
◎ プロジェクト管理サービスとオンラインストレージの複合システム

- GakuNinRDM プロジェクト管理サービス (NII RDCサービスのクローン)
- NextCloud オンラインストレージサービス

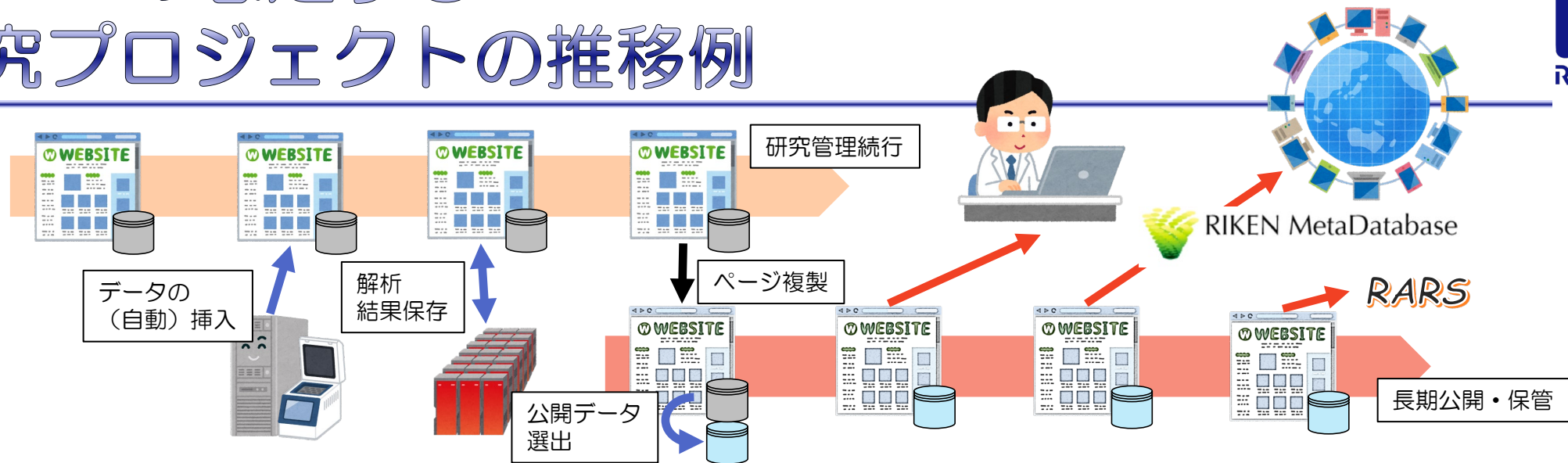
- GakuNinRDM のデータ保管場所として、NextCloud や S3 を用いる

- NextCloud は各種実験機器やファイルサーバとのデータ転送に用いることが可能

データ実体はストレージインフラに保存される



R2DMSの想定する 研究プロジェクトの推移例



研究開始

- ・研究データ管理計画の設定
- ・R2DMS上にプロジェクト作成
- ・プロジェクトに関するメタデータ等の設定

研究データ 作成・蓄積

- ・プロジェクトにデータを保存
- ・データのメタデータも併せて保存

研究データ 解析

- ・プロジェクトデータを使って解析、結果をプロジェクトに保存

発表承認

- ・成果発表をPIに申告、データ公開承認を取得
- ・プロジェクトページをフォークし公開ページを作成
- ・適切に公開データを選出

査読

- ・公開ページを査読者に通知し、論拠データを提供

発表・公開

- ・PIの承認の元、公開ページを公開し論拠データを公知化
- ・適切なライセンス等定め利活用化

業績登録

- ・RARS内リポジトリ入力欄に公開ページを記載



研究管理・フロー用例(1/2)

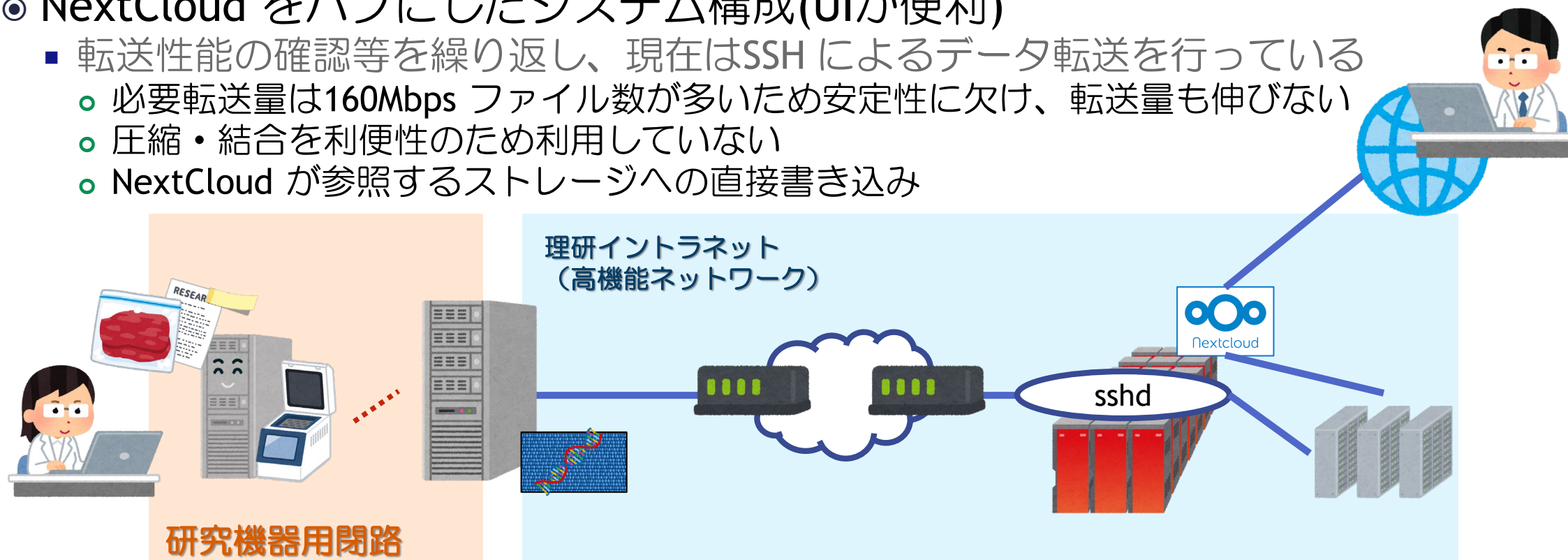
- ゲノムプラットフォーム (GPF)

◎ R2DMS 内サービスを利用したデータフロー設計のパイロットスタディ

- シーケンサにかけた人以外のゲノム解析結果について、R2DMS内に格納、必要に応じて共有・公開できる体制
 - 現時点でR2DMSは個人情報情報を格納できるシステムではない

◎ NextCloud をハブにしたシステム構成(UIが便利)

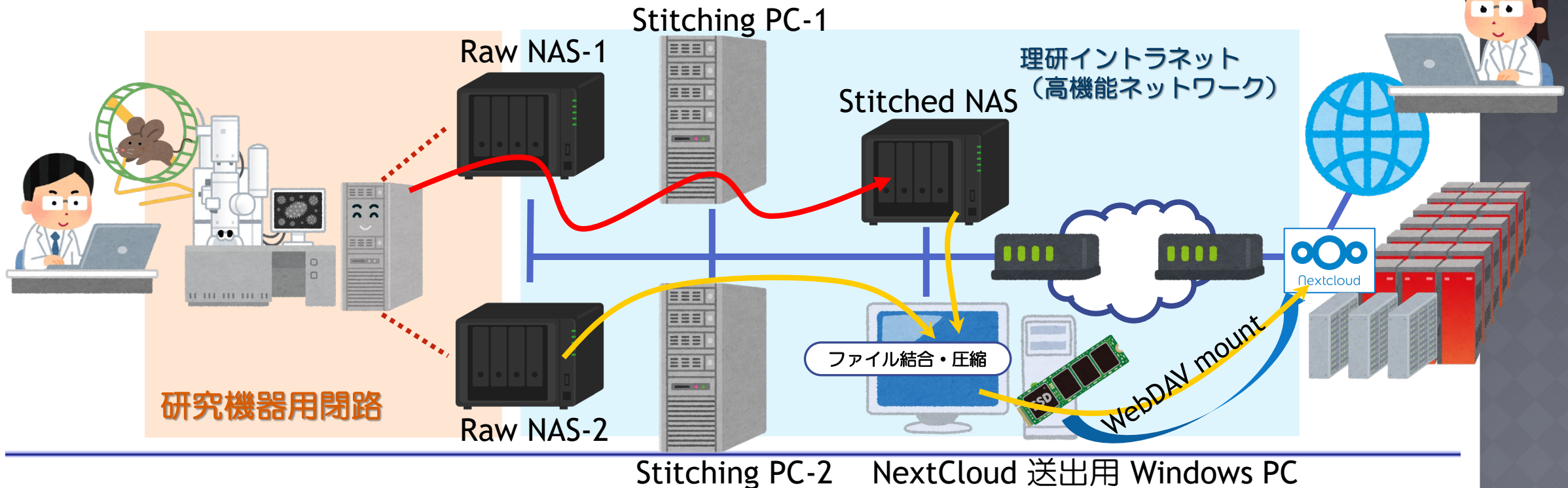
- 転送性能の確認等を繰り返し、現在はSSH によるデータ転送を行っている
 - 必要転送量は160Mbps ファイル数が多いため安定性に欠け、転送量も伸びない
 - 圧縮・結合を利便性のため利用していない
 - NextCloud が参照するストレージへの直接書き込み



研究管理・フロー用例(2/2)

- 脳顕微鏡データの蓄積と共有

- ◎ 依頼された試料に対してスライス画像である Raw Dataおよびそれを処理した3D Stitched Data を NextCloud を経由し蓄積、共有する
 - 1TB/2day: Raw 800GB(圧縮後400GB), Stitched 400GB
 - Raw Data 結合・圧縮 4.5時間 + 転送7時間 => 半日程度で収まる (40-50MB/s)
 - 研究室側全パスが1Gbps なので今後 10Gbps へ拡張予定
 - NextCloud の性能ボトルネックが出そう (2.5Gbps くらいまでは確認したが.....)



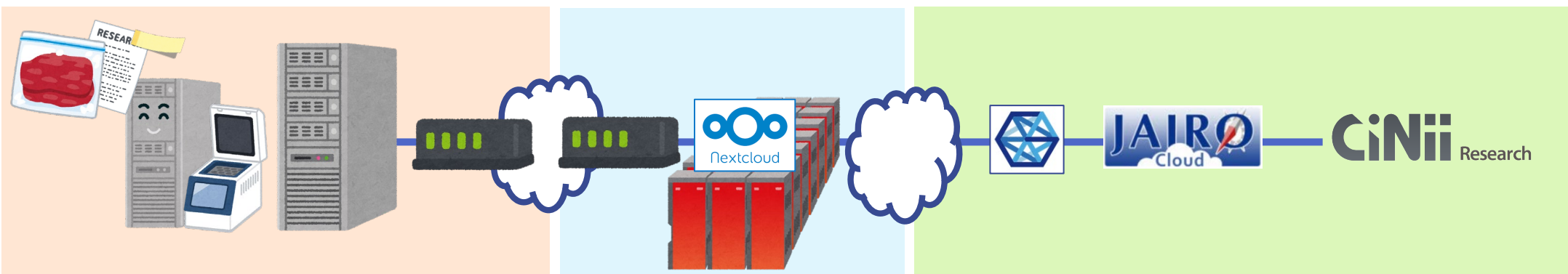
データ追跡・メタデータ管理草案

いつメタデータをつけるか

◎ 言いかえると「いつデータの証跡管理を始めるか」

- 研究機器
 - 生成と生成条件の情報
- 研究室のファイルサーバ
 - データのキュレーション
- 拠点のファイルサーバ
 - 拠点内共有に必要な情報
- NII RDC
 - 利活用するために必要な情報

保存すべきデータ化
 ・ただし生成情報は事前に
 保管する必要がある



データ採番とメタデータの成長

◎ メタデータ付与の是非

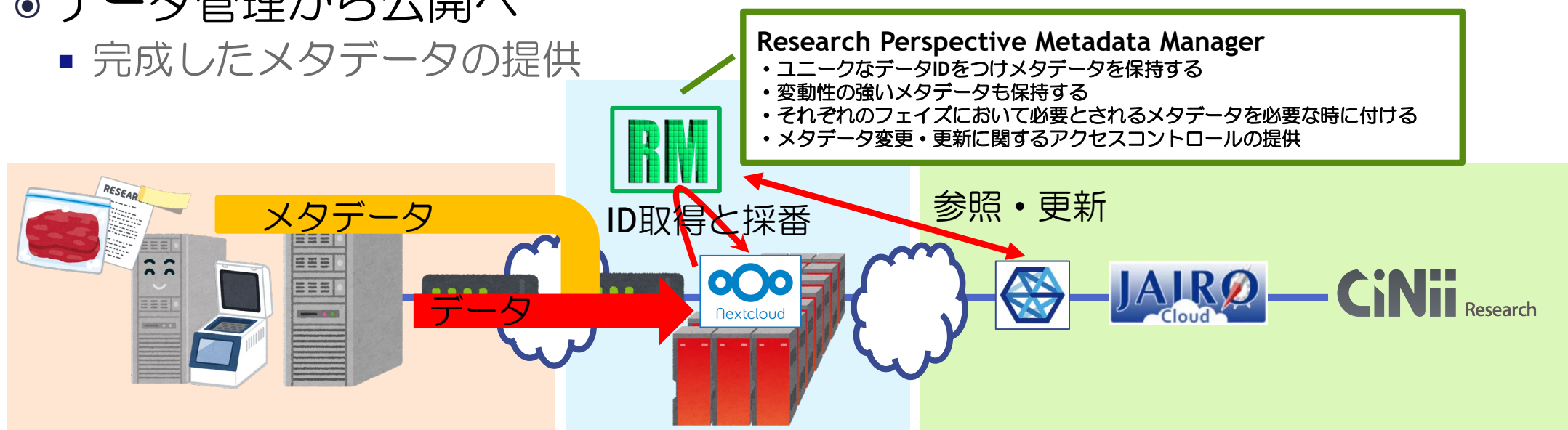
- 必要ないデータにメタデータをつけるのは無駄
→メタデータはデータのフェイズごとに補完されていく

◎ データとメタデータの一致性

- データに採番を行いメタデータを対応させる
- データIDによるデータ追跡

◎ データ管理から公開へ

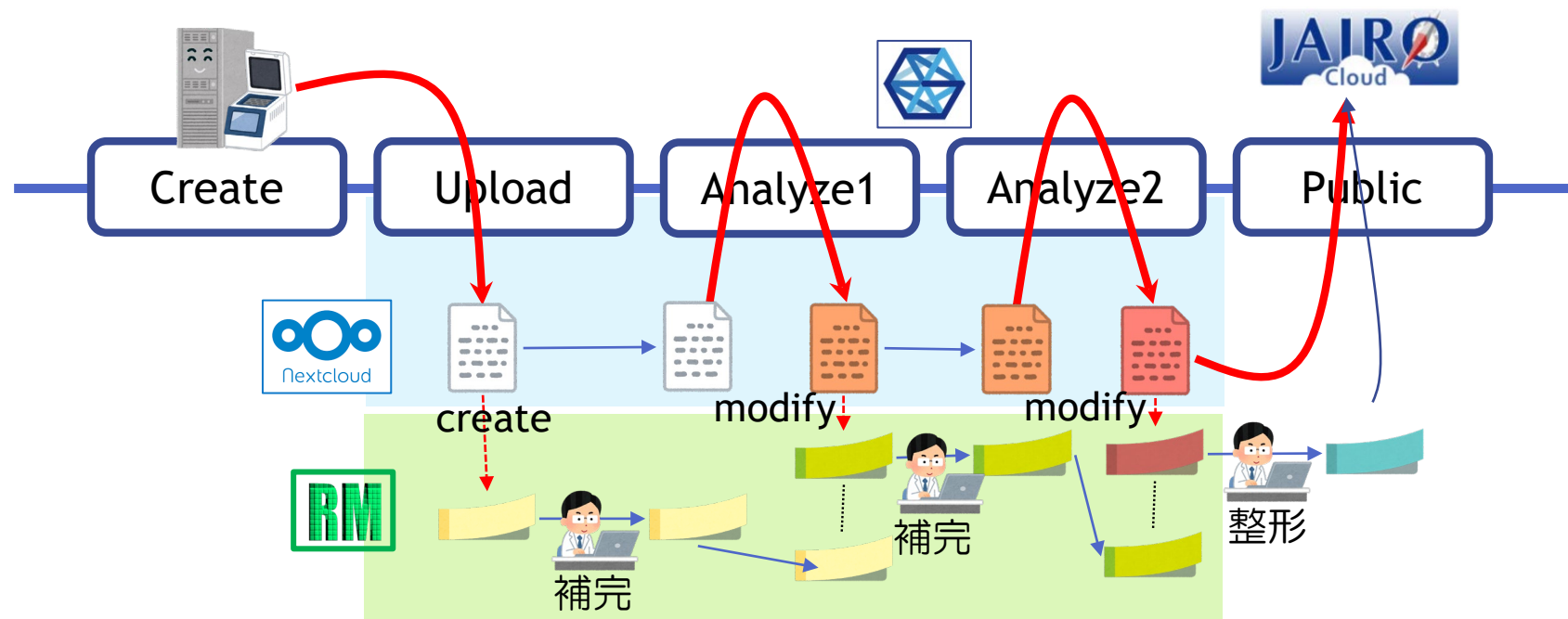
- 完成したメタデータの提供



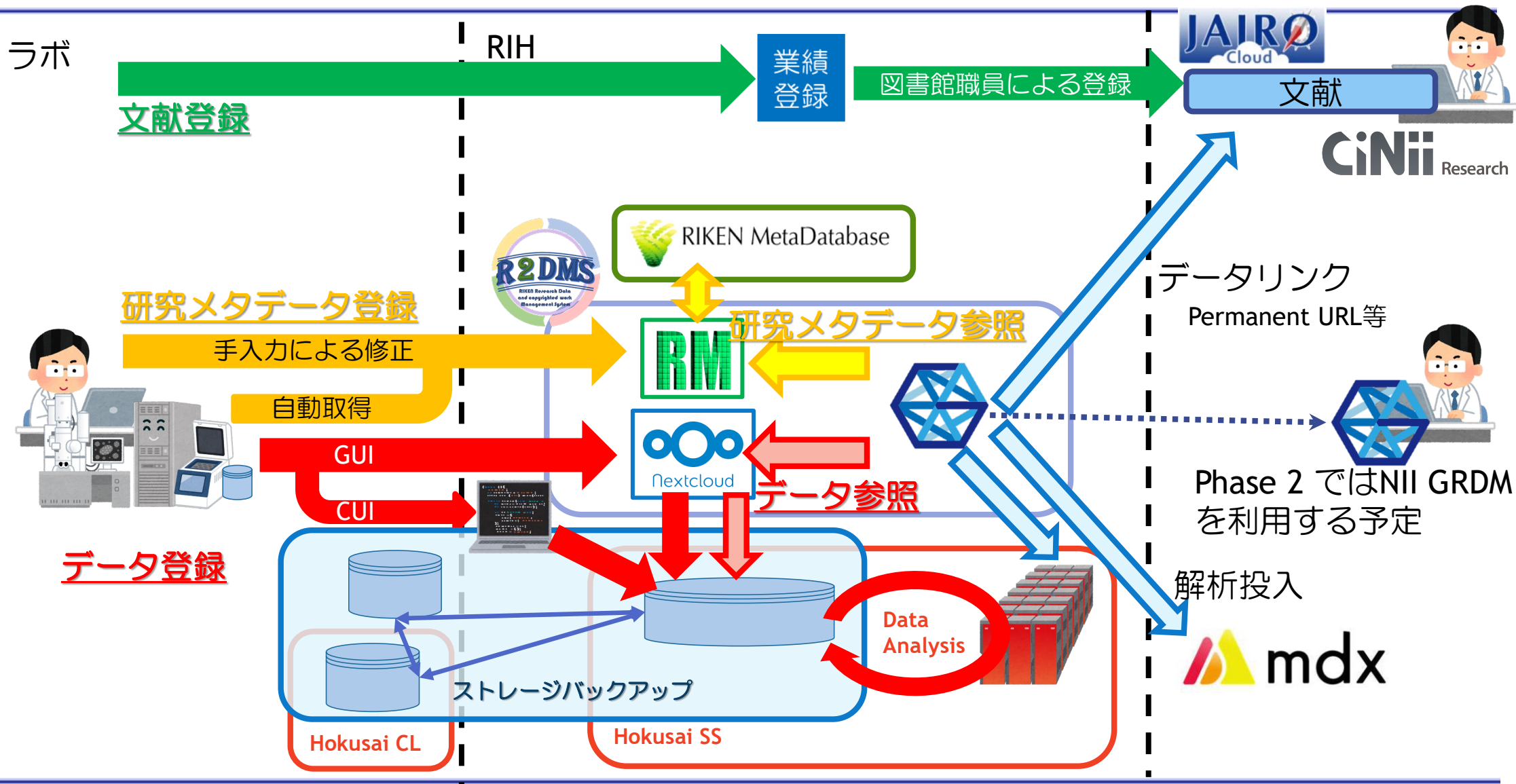
NEXTCLOUD の機能と追跡(検討中)

◎ NextCloud flow

- ファイルの操作に応じてスクリプトを実行できる
 - 作成・更新・削除・閲覧・・・・
- 処理されるたびにメタデータにログを追加
- ファイル間の依存関係を自動生成



まとめ代わりに RIKEN オープンサイエンスPF PHASE1



RIKEN オープンサイエンスPF PHASE3くらい

