理研オープンサイエンス体系と研究管理

理化学研究所 情報統合本部 基盤研究開発部門
データ管理システム開発ユニット
實本　英之 <hideyuki.jitsumoto@riken.jp>

自己紹介

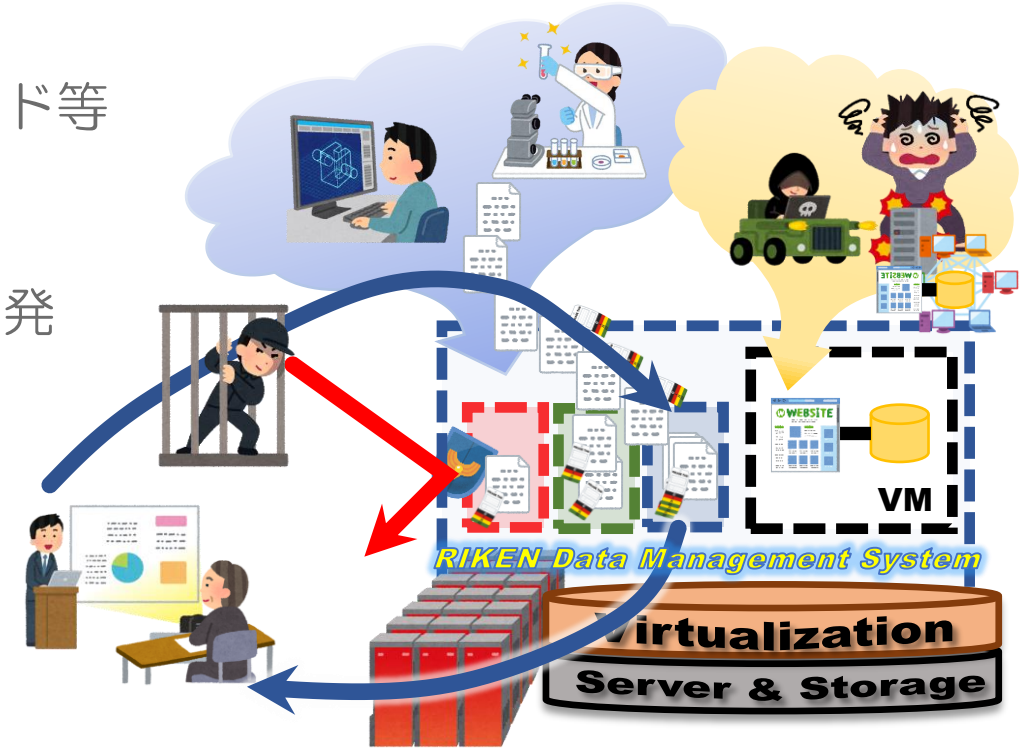
實本　英之

- 理化学研究所　情報統合本部　基盤研究開発部門
  データ管理システム開発ユニット
  - オープンサイエンスに向け、主にデータ管理するリポジトリサービスを検討・構築・（試験）運用する部署

- バックグラウンド：高性能計算分野
  - 高性能計算ミドルウェア(MPI), 並列耐故障性
  - 東京工業大学 TSUBAME1, 2, 3
  - 東大 Oakleaf-FX, HA8000(T2K), SR16000(Yayoi)
  - HPCI 東拠点ストレージ

ユニットの目的

- 適切なオープン・クローズ戦略に基づき、理研で産出される多種多様・膨大な研究データを効率よく蓄積、データ処理やデータの利活用を支援するデータ管理システムの開発・運用
  - 研究機関・研究者毎に特有のデータ管理ポリシーに沿った管理提供手法の検討・開発
  - 外部計算機、ストレージ、パブリッククラウド等との連携に関する検討・開発
  - 効率的かつ拡張性の高いストレージ、データベース、クラウド利用技術の検討・開発
  - 運用の効率化と利用者支援、またそのための利便性向上ツールの開発



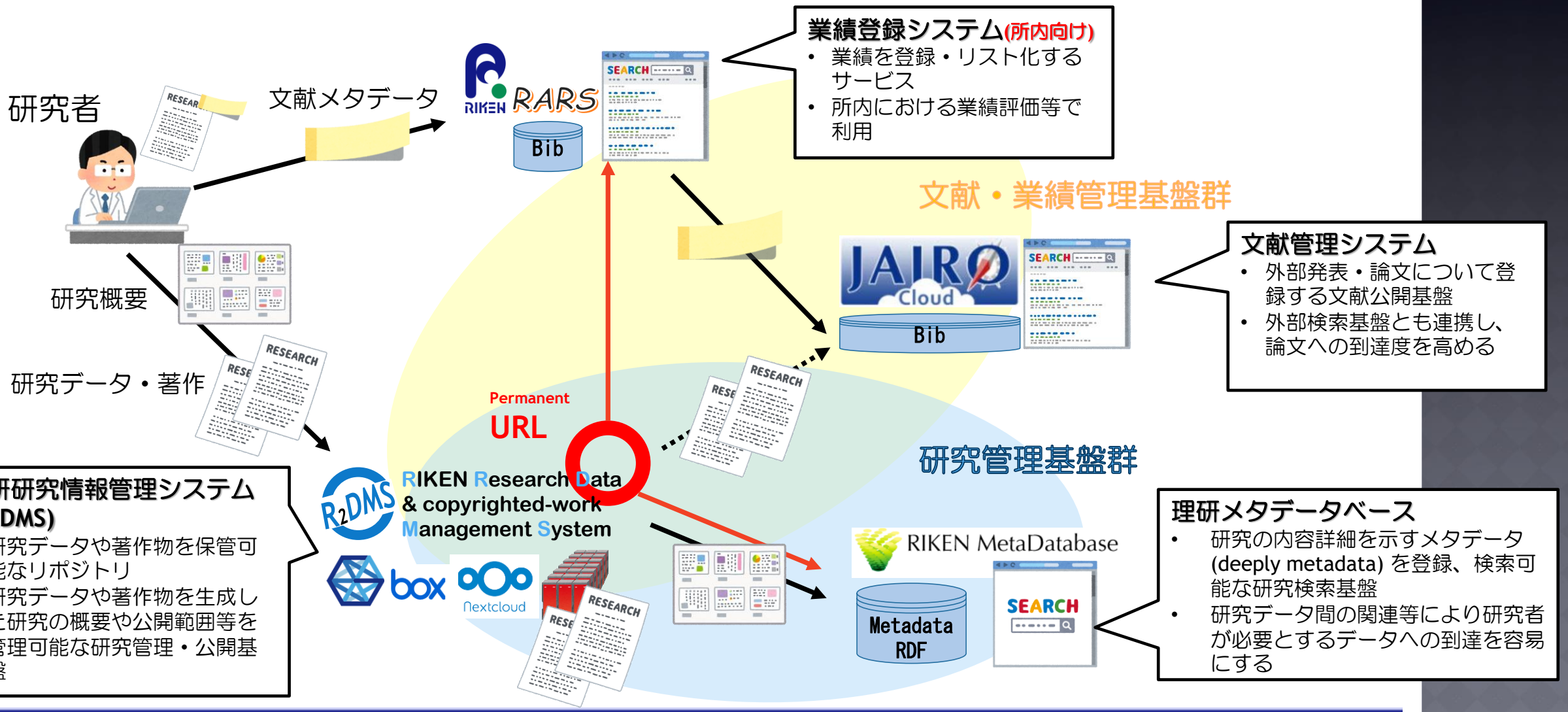
オープンサイエンスのためのデータ基盤

社会インフラとして、オープンサイエンスのためのデータ基盤構築が必要である - *統合イノベーション戦略:内閣府(2018/6/15)*

- 理研データポリシー(2019)
  - 研究データの取り扱いに対する基本方針
  - 研究データは研究者の魂であるとともに、「公的資金で運営される研究所で生成された国民の知の共有財産」
    - 原則、最終的に公開する方針→利活用データ・研究証跡

理研発の研究データを適切に管理・公開する各種機能を備えたリポジトリの運用

理研オープンサイエンス体系 (※最終形)



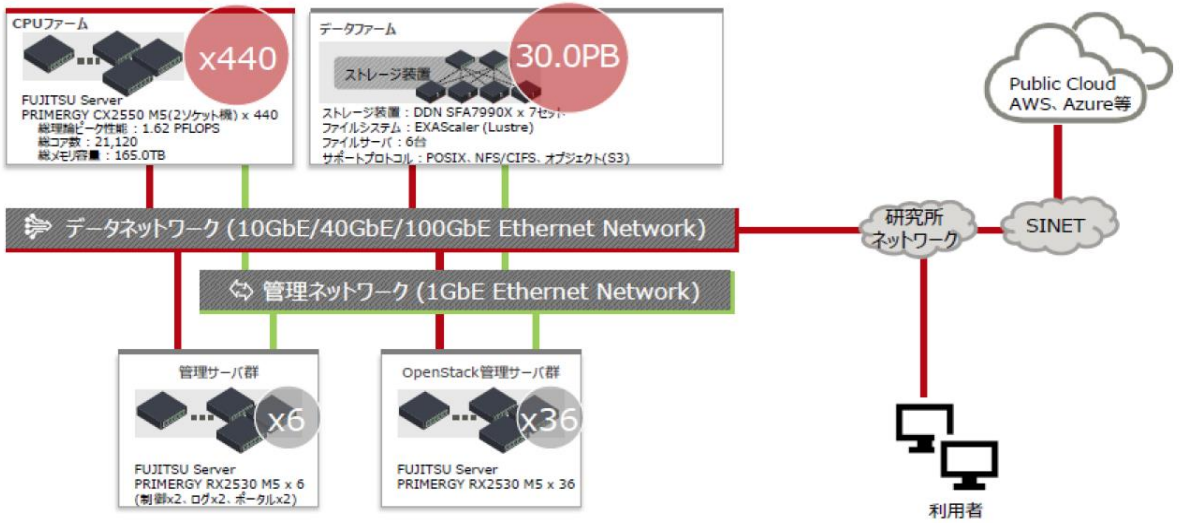
科学データ基盤：HOKUSAI SAILING SHIP

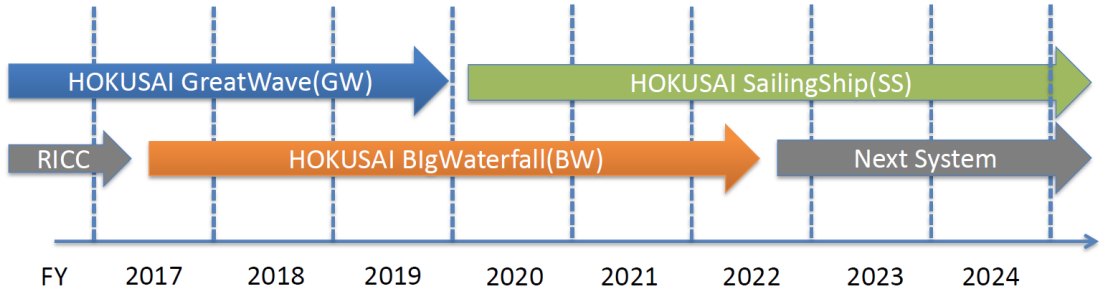
- 各研究分野における様々な研究スタイルを視野に入れたパラダイムシフト
  - データ処理の必要性増大
  - アプリケーション実行環境への柔軟性
  - 研究環境の変化への柔軟な対応

- Hokusai Sailing Ship (HSS)
  - 様々な資源へのデマンド処理
  - セキュアなデータ保存・運用・管理
  - クラウド技術の利用と連携

科学データ基盤：HOKUSAI SAILING SHIP

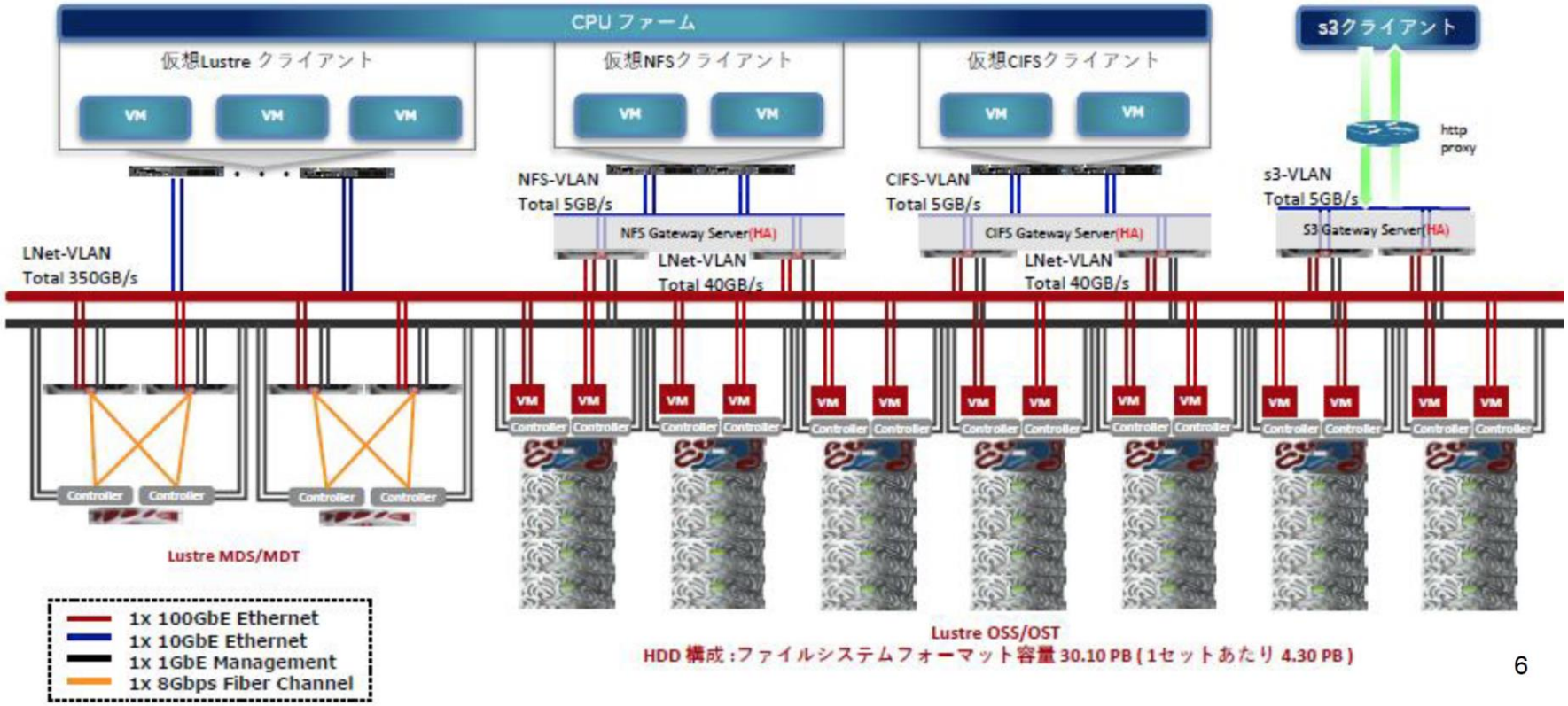
- 従来型スパコン(Hokusai-GW) から data-oriented な IaaS/CaaS システムへの転換
  - 理研内部用設備
  - Mirantis Cloud Platform (商用OpenStack dist.)
    - PlaaS (private Infrastructure as a Service)
  - CPU Farm (440 node)
    - FUJITSU PRIMERGY CX2550 M5
    - 1.62 PFlops / 21120 core
    - 165 TB Total Memory
  - Data Farm
    - DDN SFA7990X x 7sets
    - 30 PB Lustre





データファーム

- 30PB DDN EXAscaler Lustre ファイルシステム, 350GB/s
- NFS/CIFS GW, S3DS(S3ゲートウェイ), 5GB/s



リポジトリの要件

- データの保管・共有
  - 適切なアクセスコントロールに基づく公開範囲の限定
  - データ操作履歴の保持
    - 改ざん・不正利用の検出
- 研究データとしての管理・公開
  - データの概要の記述
    - 自由記述, タグ
  - 研究方針やバックグラウンド等の共有
  - データ間関係も考慮した検索・参照
- 高性能計算機や研究室サーバとの連携
  - 容易で柔軟なデータ取得手法
  - 課金体系

対象とする情報の種類
- 利活用情報領域と一般情報領域

- 情報（情報＝データ＋著作）の管理方法や課金ポリシーを決定するために研究管理と研究公開を区別する
  - 利活用情報領域（公開）
    - 論文発表等により公開が義務付けられたもの
    - 利活用によって科学技術の発展に貢献できると理研（各センター）が判断したもの

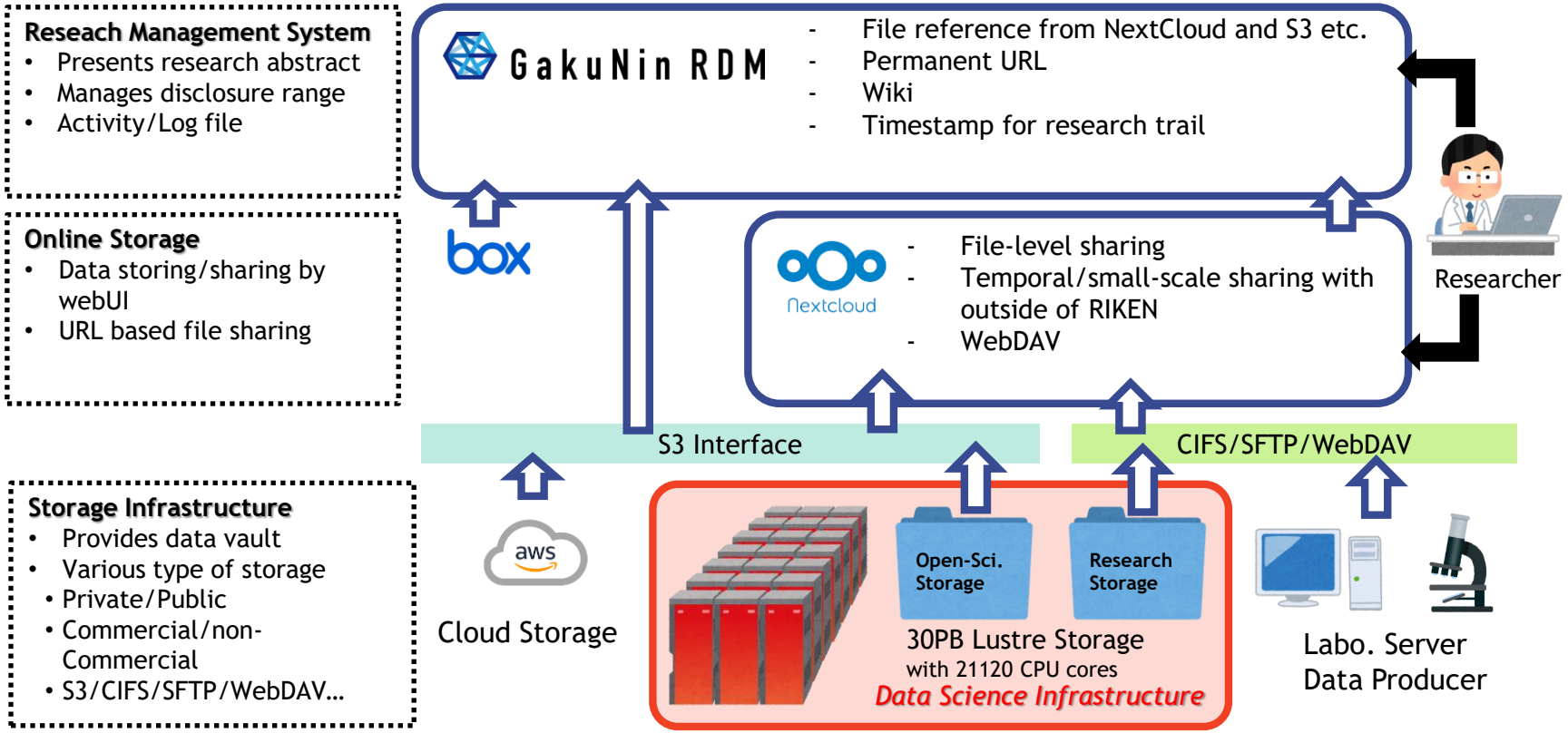
理研の規定により公開を求めているもの→基本無料
データの利用法・拡張性を考慮し S3 プロトコルで用いるオブジェクトストレージを利用

  - 一般情報領域（管理）
    - その他の研究データ管理に用いる領域
    - ユーザ間で限定的にデータを共有し、研究活動促進のために用いる

常のデータ置き場としてユーザが選択し利用する→基本有料
ユーザの用意する様々なストレージを利用可能としそちらに支払をする

研究情報管理システム(R2DMS)概観

- GakuNinRDM/NextCloud から成るデータ・著作保管公開基盤
  - プロジェクト管理システムによる研究の管理（管理・公開基盤）
  - オンラインストレージによるファイルの管理（保存・連携基盤）



GAKUNINRDM：プロジェクト管理基盤

- NII が OSF を日本学術機関に合わせてカスタマイズした Webアプリケーション
  - プロジェクトを管理するための各種機能を持つ
    - データセットの参照、プロジェクト概要説明、ライセンス表示、ブランディング、バージョニング、データ追跡、データスタンプ……
  - NIIが Shibboleth SP として GakuNin を用いてサービス展開している
    - GakuNin フェデレーションに入っていれば利用可能

- RIKEN GakuNinRDM：NII GakuNinRDM Clone
  - 理研内部の研究フローとの融合のために、特別な開発・運用・改造の可能性
    - NIIと合流するにも事前運用テストは必要
  - 理研と共同研究をしている中小企業との連携（GakuNin外アカウント）
    - 経産省系統一認証基盤や OpenIdP で利用可能になる見込みはある
  - 内部基盤サービスとの連携
    - 高度なFindable を提供するメタデータサーバ等との連携

安定的な運用法が確立したら合流する予定

プロジェクトとは

- R2DMS でデータセットを共有する単位（狭義）
  - 部門・部署等が遂行するあらゆる課題をいい、これに対して従事する人員及び資源の集合（広義）

- R2DMS では（業績に関する情報を除き）特にプロジェクトの規模・単位・目的を限らない
  - 個人データを管理するためだけのプロジェクト
  - 論文データを管理するためだけのプロジェクト
  - チームの研究を管理するプロジェクト
  - センター全体のすべての研究を管理するプロジェクト
    - ここからサブプロジェクトとして所属者の研究プロジェクトに分けていくことも可能

NEXTCLOUD： オンラインストレージ

- 外部ストレージアドオンによってストレージインフラを接続
  - CIFS, SFTP, WebDAV, S3
  - 有料ストレージとして HSS の共有ストレージ(Lustre領域) を推奨
    - 180 yen/TB/month
- WebDAV インターフェースにより外部連携が可能
- NextCloud 側でのファイル変更・削除は上位の GakuNinRDM から認識される (予定)

アカウント

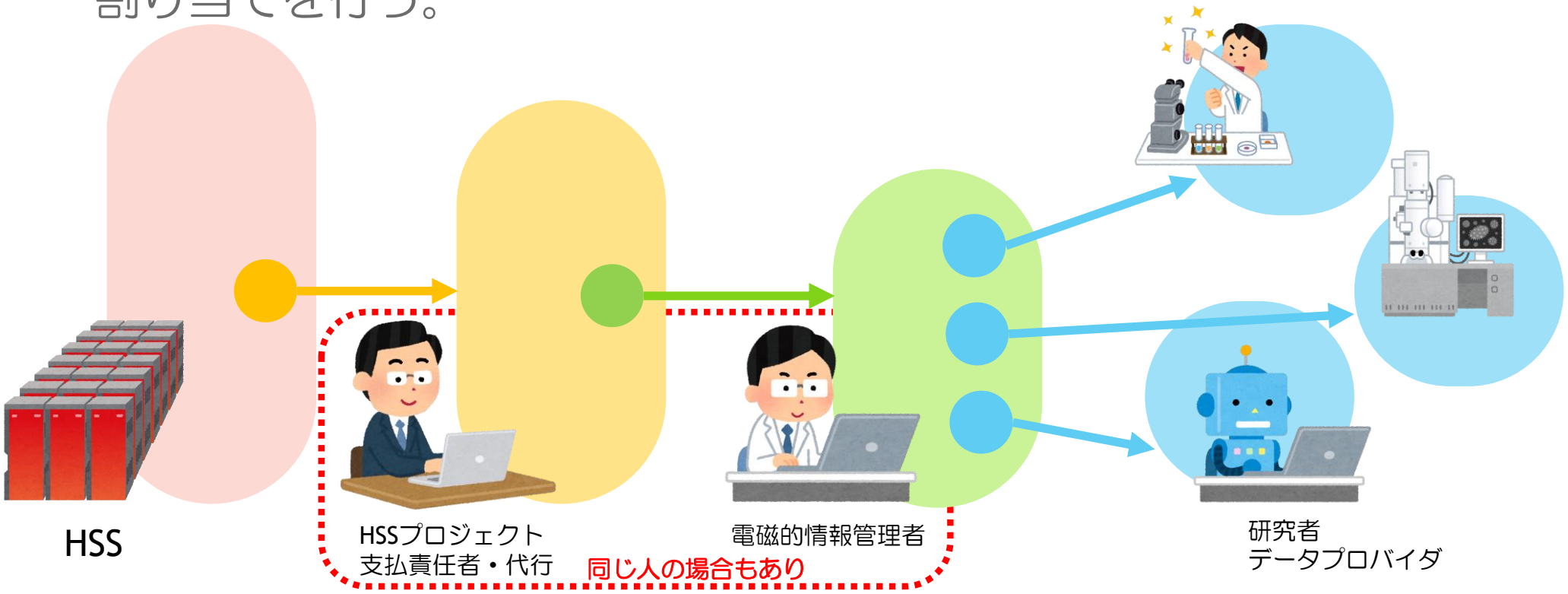
- 理研所属員
  - 基本的には理研統合認証基盤 (Shibboleth IdP) or Okta 認証
    - 双方とも人事DBが元帳
    - HSS Tenant, GakuNinRDM, NextCloud, Box
  - 一部非連携アカウント（プロジェクトベースアカウント）
    - CIFS 連携アカウント
      - NextCloud と HSS を結び付けるのに利用しプロジェクト毎に発行
      - 電磁的情報管理者が管理し、NextCloud 設定にも利用できる
    - アクセスキーペア
      - S3バケット毎に発行
- ゲスト
  - NextCloud: 理研職員が本人の責任の下、既定数まで発行可能
  - GakuNinRDM: 管理部局による生成が必要（フロー未定）

※Hokusai Sailing Ship VM 内アカウントと外部アカウント群の関連性はなし

- 結局ユーザが勝手に指定できてしまうため、制限を設けなかった（輸出管理規制除く）
- Lustre 上の UUID もノードマッピングされているだけのため、外部連携には難がある

データ管理体制

- 電磁的情報管理者
  - プロジェクトにおいて研究情報の管理を積極的に行うもの
  - データ管理計画(DMP)の検討や、プロジェクトメンバへの情報領域の割り当てを行う。



一般情報の管理概要

R2DMSにプロジェクトページを作成
- プロジェクトの設定
  - タイトル
  - メンバー
  - DMP

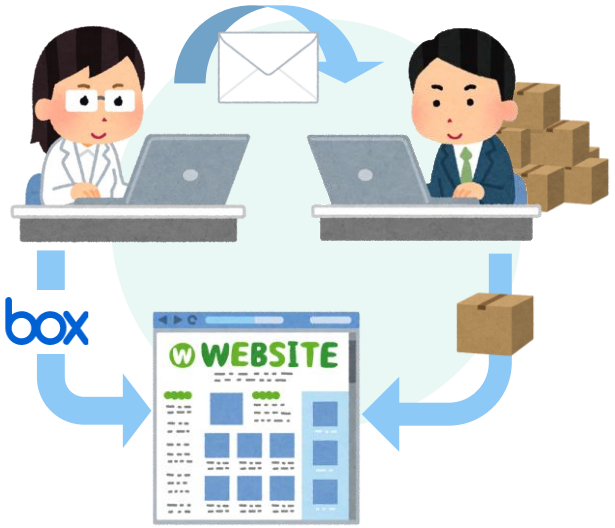
プロジェクト自体はタイトルのみで作成可能

適切な一般情報領域の接続
- 各センターの DMP/ガイドラインに従った決定
  - 電磁的情報管理者の管理する領域 (NextCloud: 有料)
  - 個人アカウントが管理する領域 (Box: 無料)

ファイルを入れて共有開始
- メタデータの付与等

R2DMS RIKEN Research Data & copyrighted-work Management System

box Nextcloud



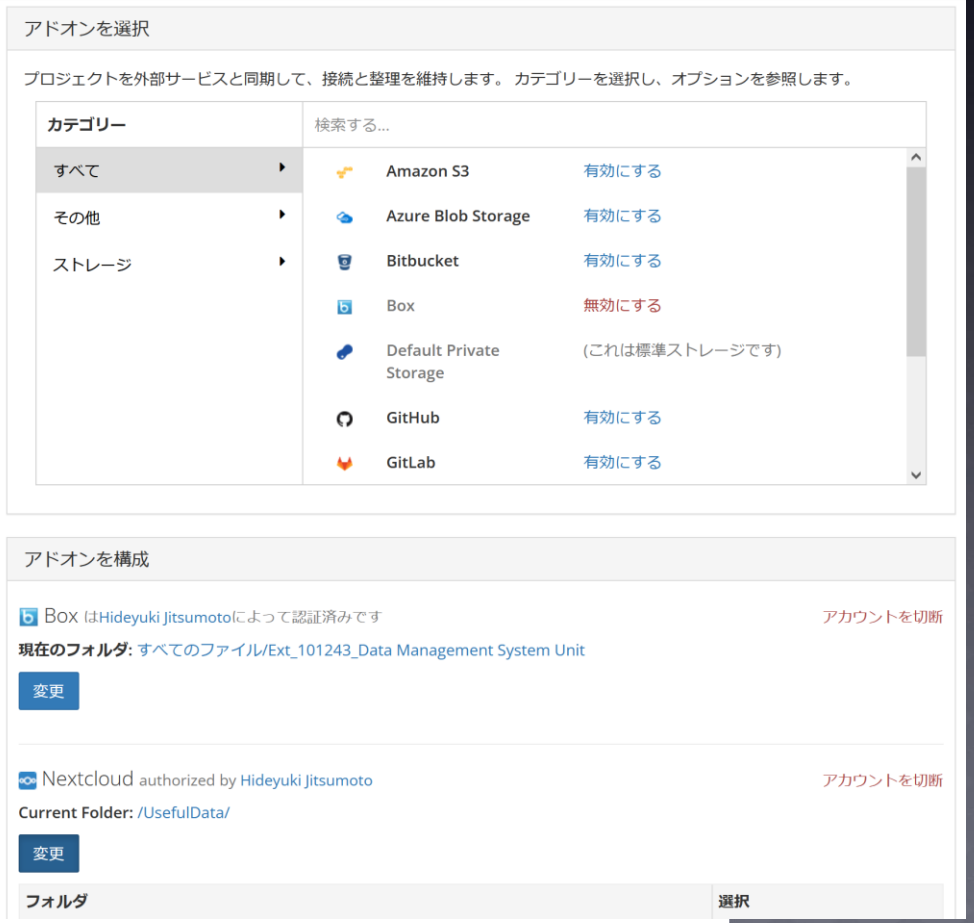
研究管理

- GakuNinRDM のプロジェクトページに情報領域を外部アドオンで結び付けて用いる
  - 退職する可能性のある個人の権限による操作をなるべく避ける
  - 2つの基本情報領域
    - NextCloud
      - 主にHokusai Sailing Ship 上に有料で領域を作り利用
      - センターや研究室のプロジェクトで利用し、個人に紐づけない仕組みを提供する→電磁的情報管理者専用アカウントによるアサイン
    - Box
      - 理研が無料で所属者に割り当てるクラウド領域
      - 主に個人のアカウントで提供し一次的なものに用いる
        - 結び付けるにあたり個人アカウントを使うことになるが、アクセスコントロール自体は組織主体のものになっているため、引継ぎ手続きを明確化するなら利用可能(DMP)

    他にもGitlab や S3 などが DMP 次第で考えられる
  - プロジェクトページをブランチすることで容易な研究公開ページの作成手段を提供

研究管理


RIKEN
マイプロジェクト
検索
Hideyuki Jitsumoto
DMS データ管理
ファイル
Wiki
メンバー
アドオン
設定
証跡管理
DMS データ管理
非公開
公開にする
0
メンバー: Hideyuki Jitsumoto
所属機関: 無し
作成日時 2021-10-26 06:53 PM | 最終更新日時 2021-10-26 07:00 PM
カテゴリー：データ
説明:
DMS で利用するあらゆるデータについての親プロジェクト
ライセンス：ライセンスなし
Wiki
機密度２データを含むため、本階層でのユニット外共有は不可
ファイル
ストレージプロバイダーをクリックするか、ドラッグ&ドロップしてファイルをアップロードします
フィルタ
名前
最終更新日時
DMS データ管理
Box: Ext_101243_Data Management Syste...
2021年度開発検討項目
DMS低機密度研究室管理データ
DMS外部機関アップロード
GakuNinRDM 環境資料
NextCloud 簡易利用マニュアル個別申請...
引用
コンポーネント
コンポーネントを追加
プロジェクトをリンク
コンポーネントを追加して、プロジェクトを整理します。
タグ
DMS
データ管理
NextCloud
Box
機密度２
タグを追加
最近の活動
更新
フィルタをクリア
開始日
終了日
ユーザでフィルタ (現在のメンバーから選択)
過去のプロジェクトメンバーも含めて検索する


アドオンを選択
プロジェクトを外部サービスと同期して、接続と整理を維持します。 カテゴリーを選択し、オプションを参照します。
カテゴリー
検索する...
すべて
その他
ストレージ
Amazon S3
有効にする
Azure Blob Storage
有効にする
Bitbucket
有効にする
Box
無効にする
Default Private Storage
(これは標準ストレージです)
GitHub
有効にする
GitLab
有効にする
アドオンを構成
Box はHideyuki Jitsumotoによって認証済みです
アカウントを切断
現在のフォルダ: すべてのファイル/Ext_101243_Data Management System Unit
変更
Nextcloud authorized by Hideyuki Jitsumoto
アカウントを切断
Current Folder: /UsefulData/
変更
フォルダ
選択

利活用情報の公開概要

各センターに該当データが利活用情報か等に関する検討をする
- センターガイドライン・理研データポリシーに従った判断
- 利用条件・範囲の設定

R2DMSに研究ページを作成
- 既存研究ページの利用も可能
  - 限定公開→完全公開
  - 研究共有用ページのブランチ

R2DMS RIKEN Research Data & copyrighted-work Management System

box Nextcloud

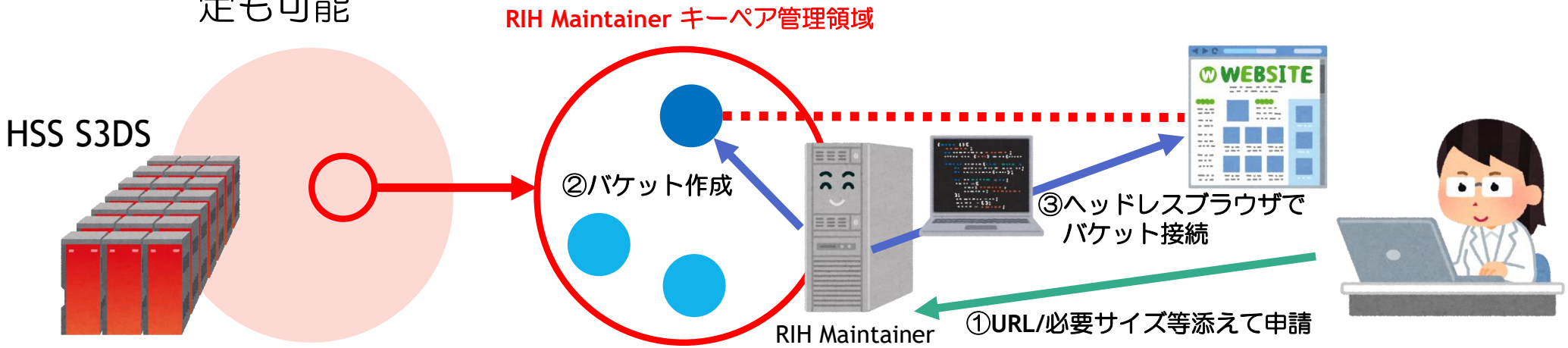
理研へ利活用情報領域（ストレージ）の申請
- 申請代表者氏名
- 連絡先
- 研究ページ名

→研究ページに理研がストレージを接続

ファイルを入れて公開

研究公開

- GakuNinRDM に公開用のプロジェクトページを作成し、理研側から利活用情報領域を割り振って利用する
  - プロジェクトページは新規作成、既存研究管理ページをブランチ双方が可能。
  - ユーザに RIH Maintainer アカウントを招待させ、情報領域申請をWeb サービスで行うと、プロジェクトページに利活用情報領域が結び付けられる。
    - RIH Maintainer 権限で、外部アドオンで情報領域を結び付け
      - 主としてS3バケット、データの規模によっては Box
      - 情報領域の確保は外部サーバ、外部アドオンの設定はヘッドレスブラウザを用いることにより自動化されている
    - 申請サービスでは S3 バケットのアクセスエンドポイントの確認や全公開・非公開設定も可能



研究公開

- 利活用情報領域にユーザがデータをアップロードする
  - 研究管理ページのブランチの場合 GakuNinRDM 上でドラッグ＆ドロップが可能
  - アップロード後は一般情報領域は切断する必要がある

- プロジェクトページのユニークURLを公開に利用
  - ページ内公開ボタンを押す
  - 現在は公開のためのキュレーションフローがまだない

→釦押し即公開

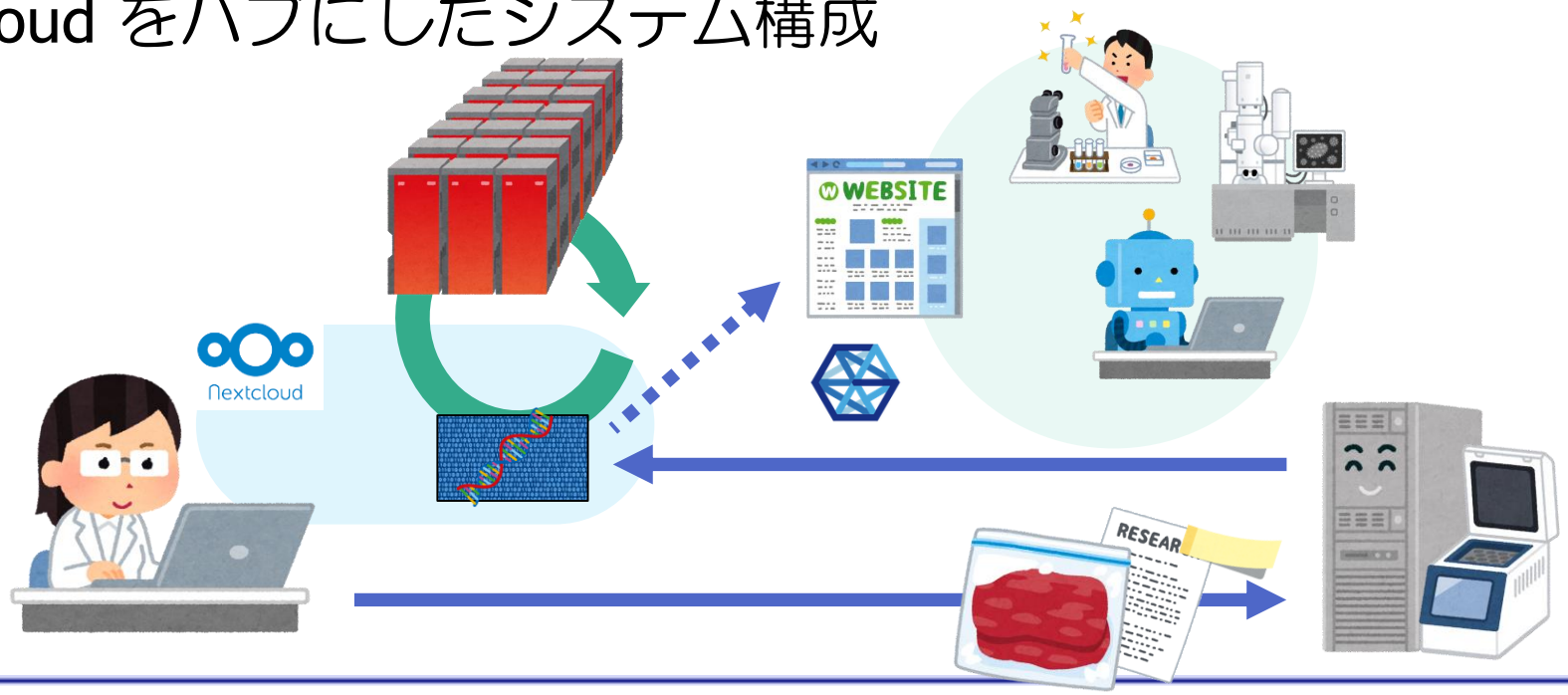
業績管理と研究情報公開

- 理研の業績に当たる内容は R2DMS 若しくは外部の公共リポジトリへの情報登録が必要となる予定
  - 業績管理システム RARS から該当情報へのリンクがたどれるよう登録が必要になる
    - 現時点ではまだ強制性はない
  - R2DMS では利活用情報としての取り扱いになる

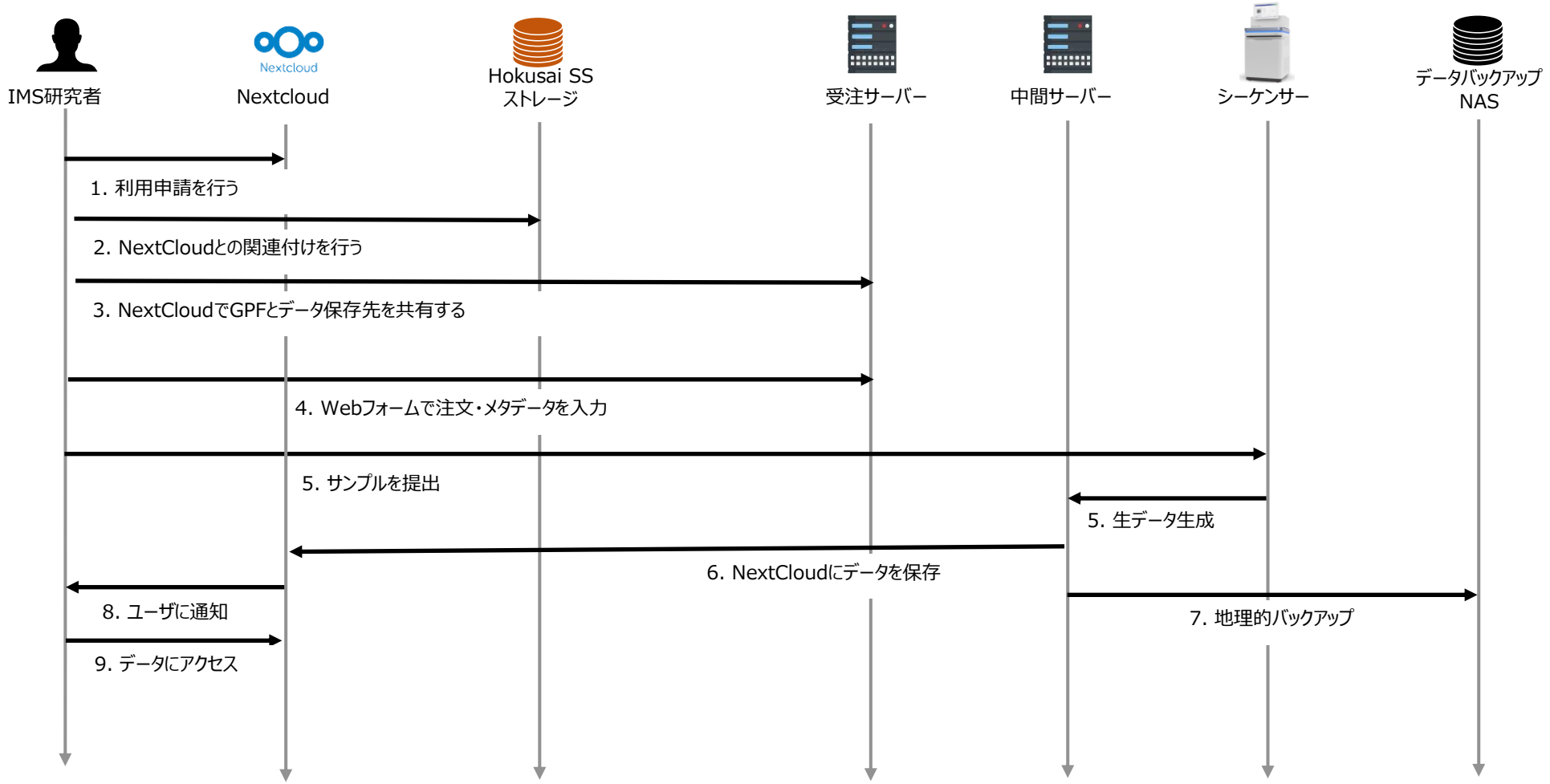
- 業績管理システムの持つ業績タイトル名でGakuNinRDMに自動的にプロジェクトページが作成されるような連携を計画中

研究フローとR2DMS
- IMS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
  - RIKEN IMS(生命医科学研究センター) とのテストプロジェクト
  - シーケンサにかけた人以外のすべてのゲノム解析結果について、R2DMS内に格納、必要に応じて公開できる体制
    - 現時点でR2DMSは個人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

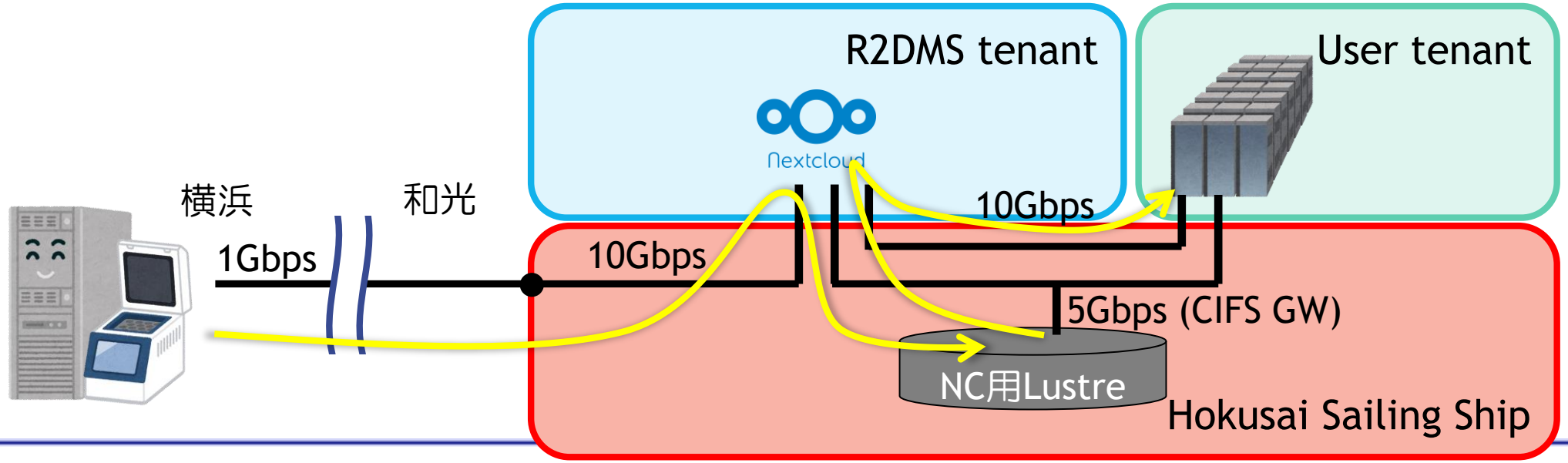


研究フローとR2DMS
- IMSゲノムプラットフォーム(GPF)



データ転送

- IMS GPF(横浜) - NextCloud(和光) - VM on Hokusai Sailing Ship (和光)
  - データ生成の仮定最大値： 50MB/s, 100GB dd転送
  - ① IMS GPF -> NextCloud (-> Lustre): 150MB/s (WebDAV)
  - ② NextCloud -> User VM: 360MB/s (WebDAV)
  - Lustre-UserVM 間はアカウント制御がNextCloudとHSSで一致しておらず直接転送ができない



実データ転送における問題

- 数万ファイル／数百KBのデータセット
  - 150MB/s -> 20MB/s 程度まで転送量が下がる
  - 1ファイル化、若しくはバースト転送が必要
- ファイル単位・データセットの意味
  - ファイルの保管形態には各研究ドメイン・研究者毎の意味がある
    - 1ファイル化するということ
      ＝雑多なファイルからあるポリシによってグルーピングする
  - アプリケーションはデータセットすべてを読み込みたいのではなく、データセットから一部を読み込みたいことが多い

- 転送時圧縮して転送先で解凍して保管する手法
  - 何らかのタスクスケジューラが必要か？
    （またHSSとNextCloudのアカウント問題が……）

NII GAKUNINRDM への合流

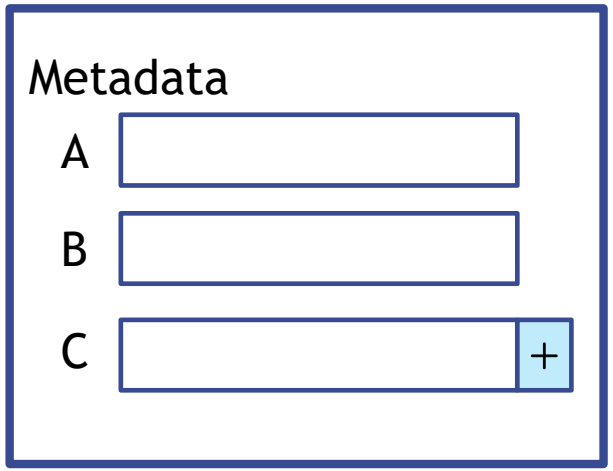
- RIKEN での運用・研究連携手法が確立し、必要な機能がそろった段階で NII GRDM へ合流を予定

  → GRDM の保守コストはとても高い
  - 所内センターの持つ計算資源・ストレージとの連携
    - アドオン開発
  - NII GRDM(管理基盤) - JAIRO Cloud(公開基盤) の連携手法が確立
    - ユーザの二度手間の排除
  - JAIRO Cloud におけるデータ提出者評価につながるデータの取得
    - 閲覧数・データダウンロード数→研究・データの利用率
    - データDL時のDL者情報の任意な取得→データ提出者へ共同研究の機会提供
  - ユーザが削除されたときのデータ管理継続性
  - GakuNin外ユーザの利用 (OpenIdP)

1,2年をめどに、NII とトランジションプランを検討中

今後について（年度末めど）

- GakuNinRDM による MetaData の管理
  - プロジェクトに対してメタデータを付与するようなGakuNinRDMアドオンを開発中
  - あらかじめ登録されているか、ユーザが用意するテンプレートに対してデータを入力しGakuNinRDM のファイルリスト下に保存する
  - メタデータはjson形式で保存され他サービスとの連携に用いる
    - RIKEN Metadatabase におけるカタログ自動生成等




```
Metadata :{
    A: value,
    B: value,
    C:{
        0: value,
    }
}
```

今後について（未定）

- ファイル展開
  - ファイルリストに特定のファイル名のファイルを含む場合、そのファイルが指定する圧縮ファイルを解凍格納する
    - NextCloud 領域全体を Lustre プロトコルでマウントする
    - 日毎程度でマウント領域をサーチし、特定のファイルを見つけた場合に処理をする。
      - アカウントが制御できないので、管理者権限で一様に行う
- 公開キュレーションフローの実装
  - ユーザの公開釦から直接公開に行かず、一度図書館部署等でのキュレーションを行う（著作権等の解決）
  - GakuNinRDM のアドオンによる実装を目指す
    - GakuNinRDM - JAIRO Cloud 連携次第によっては JAIROのキュレーティングが使える可能性がある