

理化学研究所の データ管理システムと連携

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>

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
 - 利活用データとそれ以外のデータの定義と取扱
 - 利活用データの理研内統一リポジトリ（若しくは公共リポジトリ）への保存
- 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

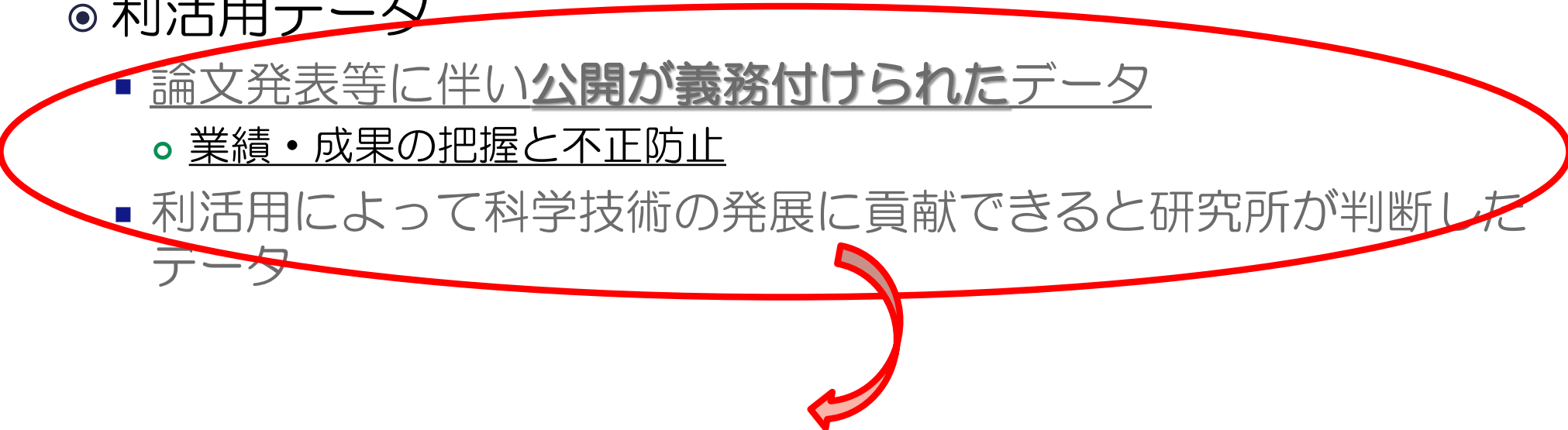
理研の研究データを適切に管理・公開する各種機能を備えたりポジトリの運用

理研データポリシーの策定

◎ 研究データ

- 研究所の研究活動を通じて得られたすべてのデータ

◎ 利活用データ

- 論文発表等に伴い公開が義務付けられたデータ
 - 業績・成果の把握と不正防止
 - 利活用によって科学技術の発展に貢献できると研究所が判断したデータ
- 
- A red oval highlights the first two items of the '利活用データ' section. A red arrow points from the bottom of the oval down towards the '研究データリポジトリの運用' section.

◎ 研究データリポジトリの運用

- メタデータやデータアクセス権の適切な付与
- 研究データの改ざん、消失、不正取得を防止する機能を備える

理研データポリシーと研究分野毎のデータ管理

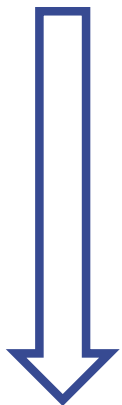
- ◎ 研究ドメイン毎にデータの取り扱い方法には差がある
 - 理研データポリシーは最低限の内容しか定められていない
 - 利活用データは原則リポジトリに登録し研究所が管理する
 - 利活用データを取り扱う基準は研究所の規定及び研究契約に従う
 - 原則10年保持する



- ◎ 研究センタ毎にデータガイドラインを策定
 - どのリポジトリにデータを保管するか
 - Ex. 公的レポジトリにデータを保管し、そのリンクを研究所が取得できるようにする
 - データの取り扱い基準はどうするか
 - Ex. 個人情報の含まれるデータは匿名化して復号情報は機密レベル3として非公開、匿名化情報は特定のリポジトリで公開

研究データリポジトリの要件と変遷

- ◎ 利活用データの内、公的リポジトリがないものに関して適切なオープンクローズ戦略を設定し、公開が可能なリポジトリ
 - 公的リポジトリに収められたデータへのリンクも管理可能
 - 研究所内業績管理システムと連携可能
 - メタデータを利用した高度な検索が可能
 - 公開データの利用傾向をデータ公開者にフィードバック可能（インセンティブ）



- ◎ 一方.....研究室毎がデータサーバを抱えることに対するセキュリティ／コストのリスクが提起
 - 多数の研究所が多数のサーバをそれぞれのクオリティで管理
 - プロジェクト終了時や縮小時の保守費用不足の懸念

- ◎ 利活用データ以外の研究データ管理も行えるのでは？
 - 結果的にはこちらの顧客の方が多い

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、バージョンング、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
 - 研究管理と公開を高度に連携したい
 - GakuNinRDM<->Jairo Cloud の連携機構が未完成だった（計画はある）
 - 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

最終的に合流の可能性も視野に入れたシステム設計を開始

柔軟なデータ保存・送出機能の確保

- ◎ GakuNinRDM はデータの保存や転送方法に関する拡張は最重要課題ではない
 - データの見せ方やプロジェクト間の関係、共有範囲の設定
 - データの公開や、研究促進のための共有事項の整理及び不正防止に用いることを期待
 - もちろん拡張はされて続けていますが...
- ◎ スパコンやデータプロバイダーとのシームレスな連携をつかさどるレイヤがあった方が便利
 - Online Storage
 - 計算機へのマウント機能や同期機能（バックアップ）
 - マウントしたディレクトリにセンサーが書き込むだけで、リポジトリにデータを送れる
 - GakuNinRDM がストレージとしてすでに利用可能な設計
 - いくつかの実装では変更履歴等も追跡してくれる

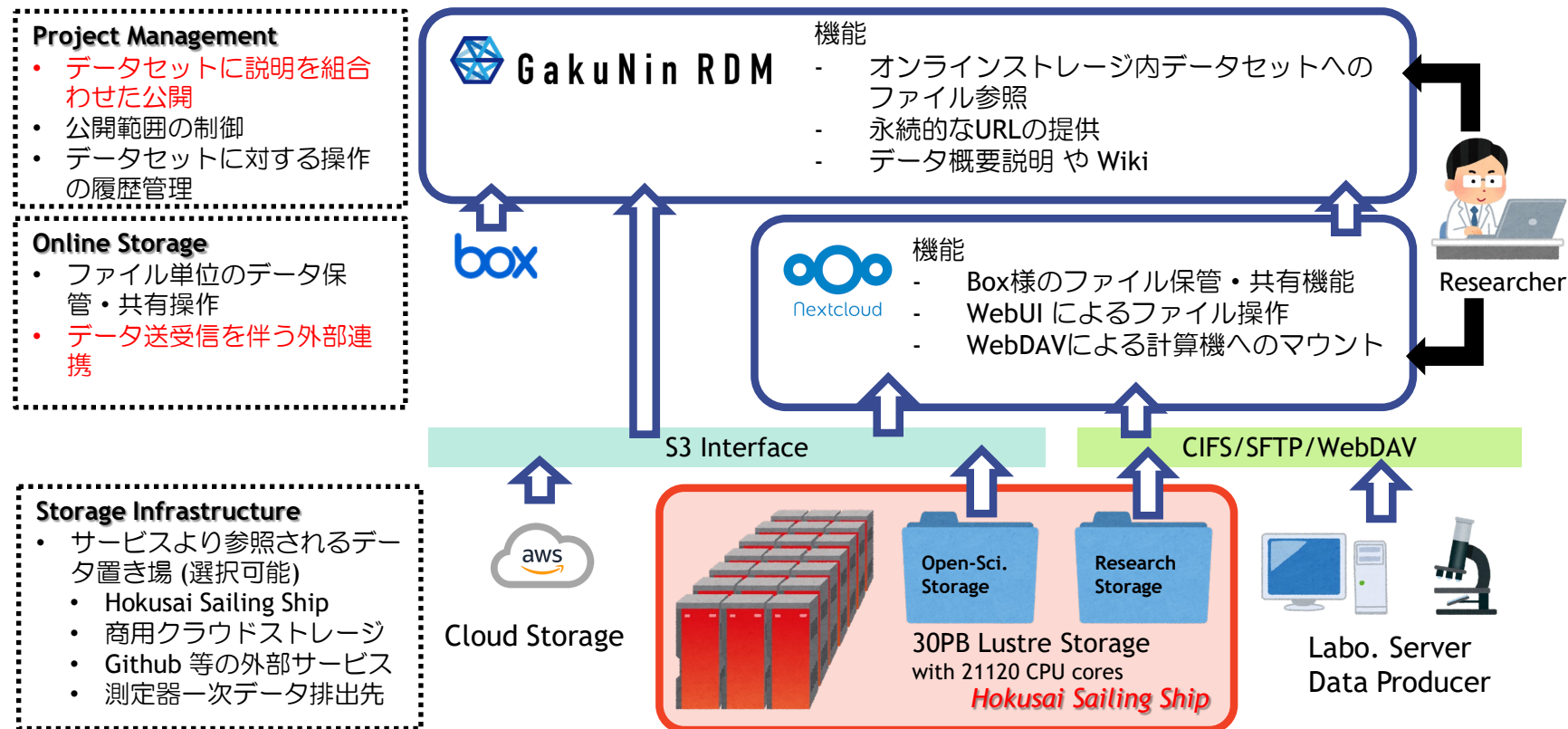
理研研究情報管理システム(R2DMS) システム構成図

◎ 理研データポリシーにおいて定められたリポジトリ

■ サービス構成

- プロジェクト管理サービス：GakuNinRDM（公開・管理基盤）
- オンラインストレージサービス：NextCloud（保存・共有基盤）

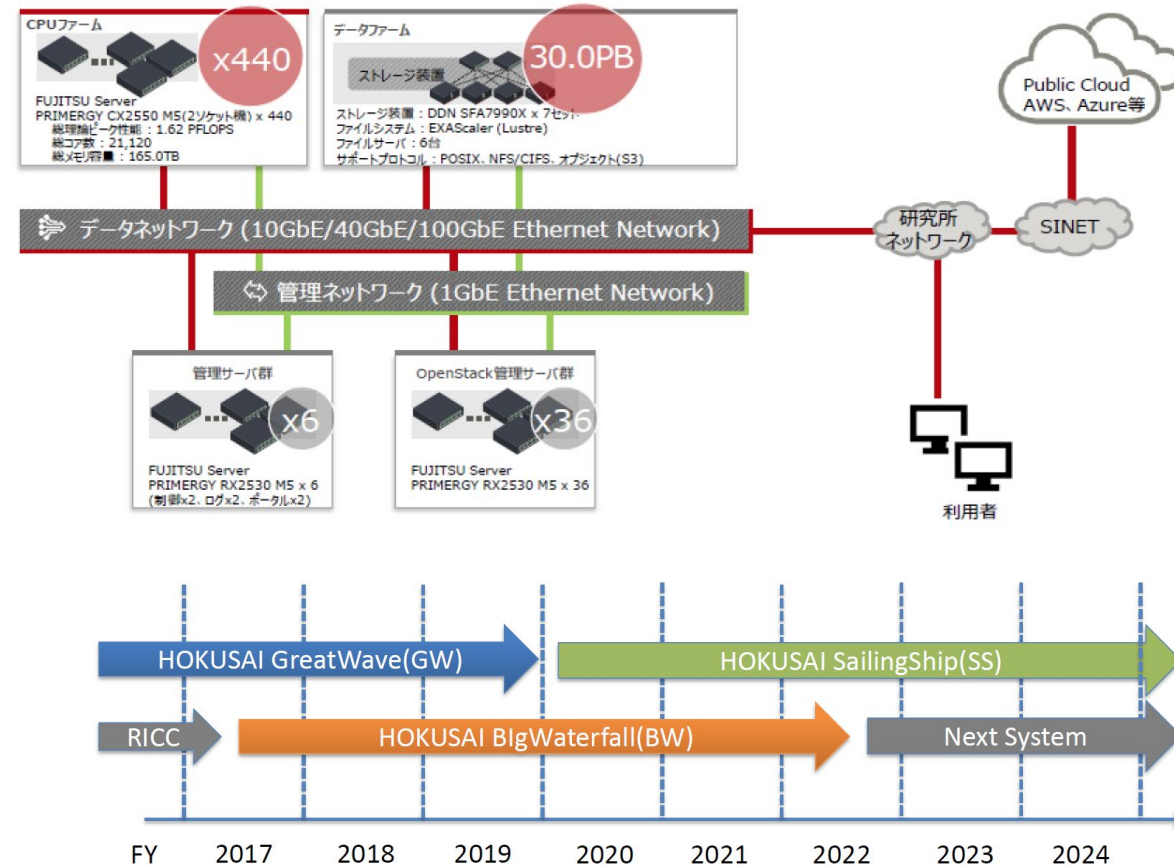
■ ユーザは HSS 等のストレージを GakuNinRDM/NextCloud に結び付けて利用する



参考：科学データ基盤：HOKUSAI SAILING SHIP

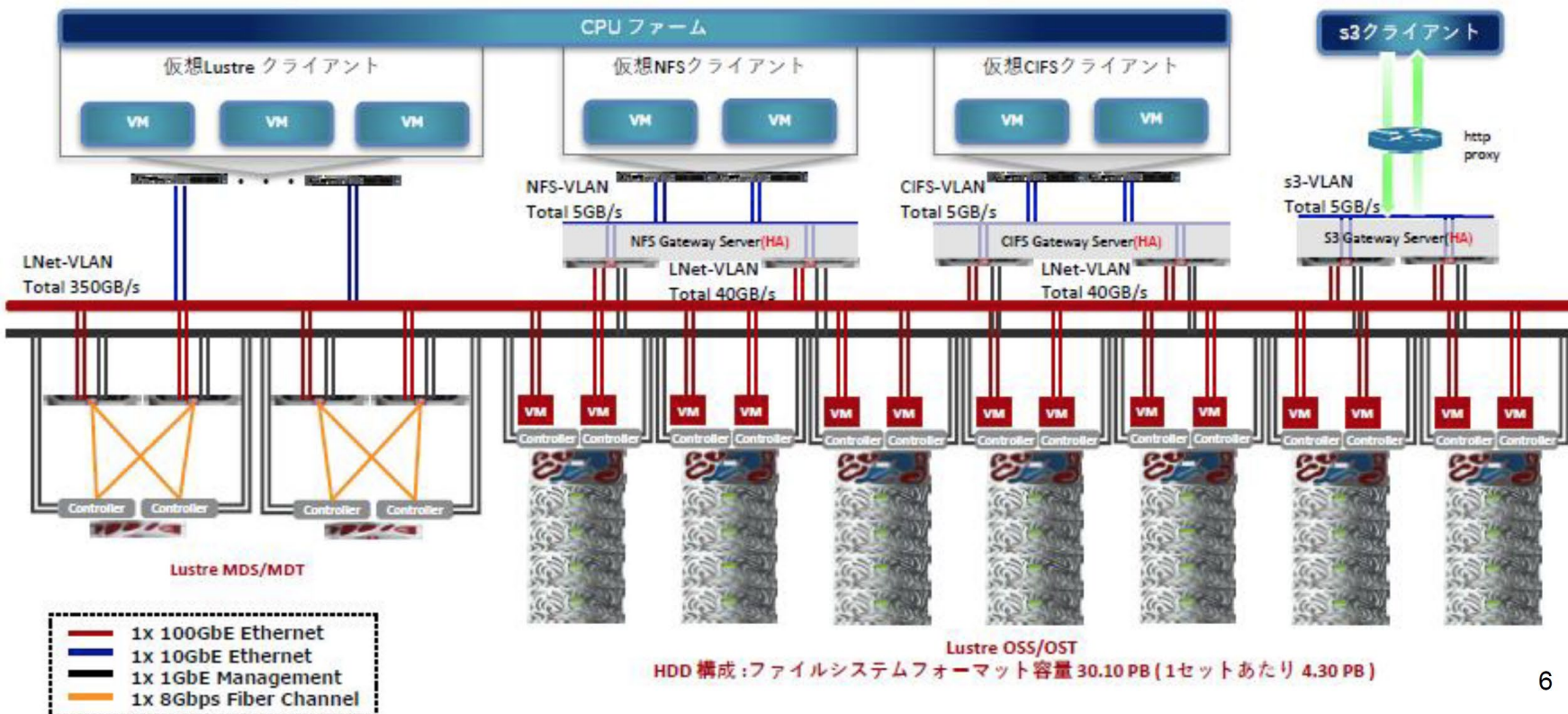
◎ 従来型スパコン(Hokusai-GW) から data-oriented な IaaS/CaaS システムへの転換

- 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



参考：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



◎ 研究データ・著作の保管・共有・再利用

- 単純保管や受け渡し
 - 研究データについてファイル・ディレクトリ単位での保管および、限定的な受け渡し
- 研究管理
 - データセットとしてあるテーマに関わる情報を一元的に管理し、限定したメンバーに共有、連携研究を駆動する
- 研究公開
 - 利活用データとして、広く世界に研究概要・データを公開し利用してもらう。共同研究等の起点とする

◎ 業績・論文根拠・証跡の保管と公開

- 査読データの提出
 - 査読時におけるデータ公開、限定的な公開や一般公開からの差異
- 業績論拠データの公開
 - 業績登録に関わる論拠データの長期保管と公開

日本における流れと理研データ基盤 - 研究DXプラットフォームの構築

- ◎ 研究データの創出・統合・利活用を一気通貫するDXプラットフォーム
- ◎ マテリアル分野をユースケース→他の分野への拡張

データ創出

- マテリアルデータ創出
 - 共用先端リサーチインフラ（大学）
 - グリーン社会実現データ創出（NIMS）
- 大量創出
 - SPring-8 データ創出基盤整備



- 各研究室の計測器等



データ統合・管理

- NIMS データ中核拠点
 - データ共有
 - データの蓄積とAI解析

- NII 研究データ流通基盤
管理公開基盤**



- RIKEN 研究データ流通基盤
管理公開基盤

データ利活用

データ駆動型研究

- 富岳



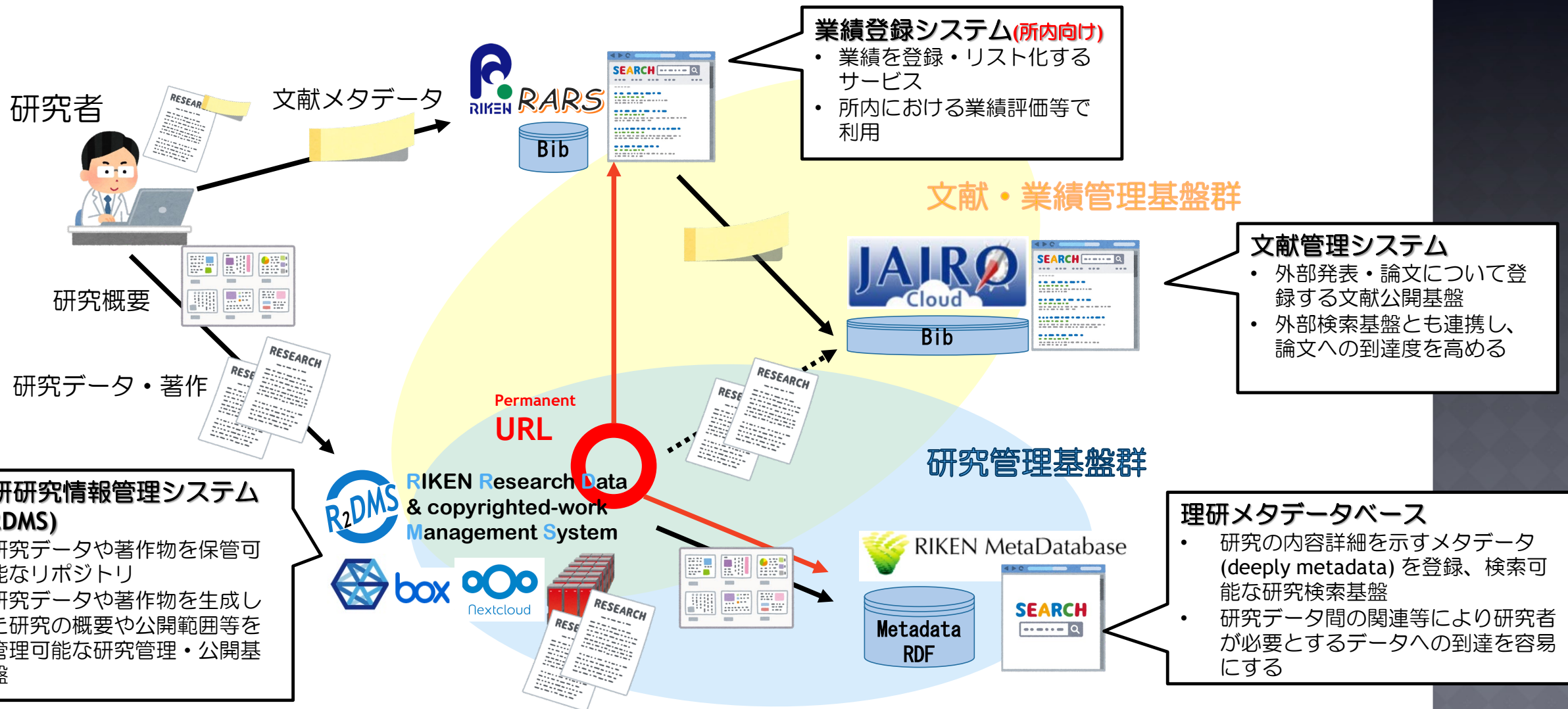
- RAIDEN (AIP)



- HSS, HGW (RIH)



理研オープンサイエンス体系 (※最終形)



対象とする情報の種類と情報領域

◎ 情報領域

- 研究情報（情報＝データ＋著作）が格納できるストレージ領域
 - システム側で利用を管理する最小単位

◎ 情報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している

■ 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

◎ 情報領域管理責任者

- 情報領域の利用・管理に責任を持つ者
- 所属長（研究室のリーダー）
 - 研究データは理研が所有し、その管理責任は所属長にある

◎ 情報領域管理担当者

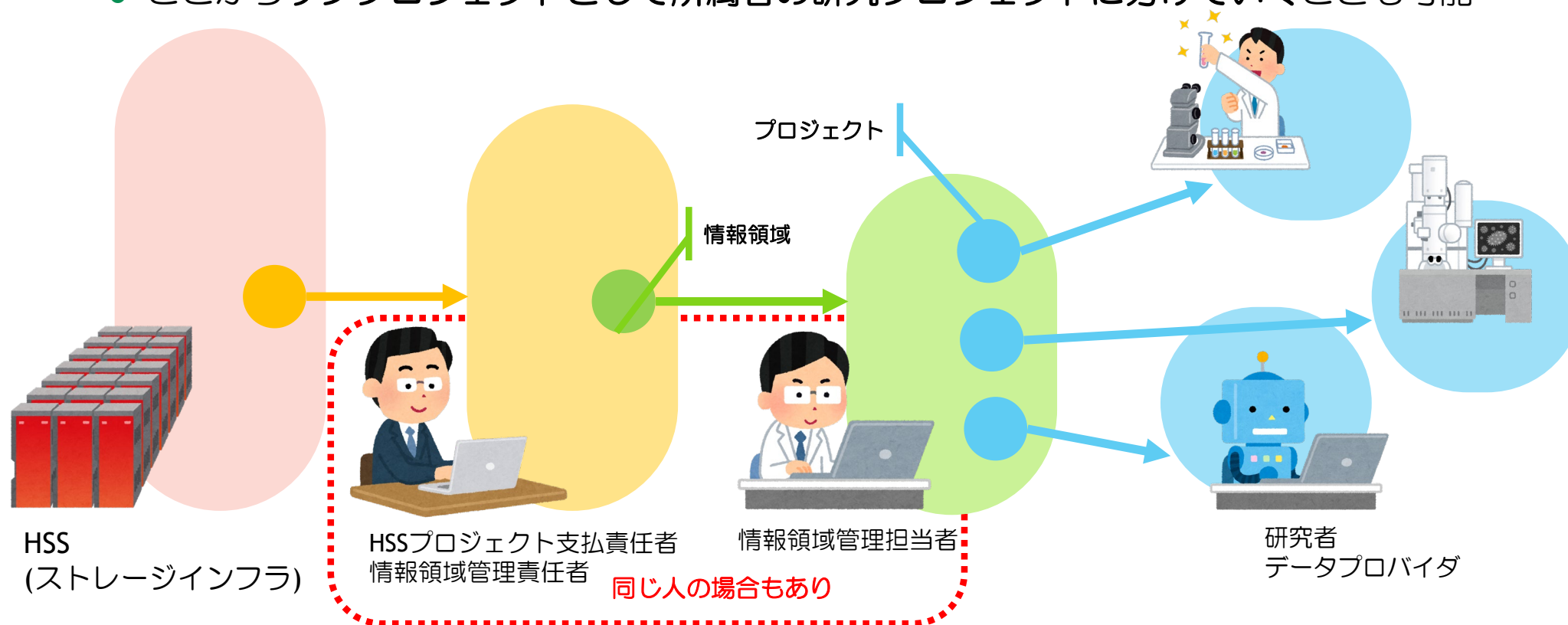
- 情報領域に存在する研究情報を主体的に管理するもの
 - 情報管理時
 - 利用設定・DMP等の設定や部門・部署のガイドラインの適用
 - 情報公開時（管理の内容も行う）
 - 情報の品質の保全に「努める」
 - 著作権違犯や公開可否、利用条件の確認・設定等
 - 場合によってはキュレーション等の協力を求めるかもしれない
 - **情報内容に関して**安全保障輸出管理制度に抵触しないか確認する

◎ 責任者と担当者は同一人物でもよい

プロジェクト

◎ プロジェクト

- 情報領域に研究情報を保管するあらゆる研究の構成単位（規模は問わない）
 - 個人データ管理、論文データ管理、チーム研究管理、センター研究管理
 - 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



対象ユーザとアカウント

◎ 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（領域ベースアカウント）
 - データ基盤連携アカウント
 - NextCloud と HSS を結び付けるのに利用し領域毎に発行
 - 個人に紐づけられていないので、引継ぎが必要な設定等に利用
 - アクセスキーペア
 - S3バケット毎に発行

◎ ゲスト

- NextCloud, GakuNinRDM
- 理研職員が本人の責任の下、既定数まで発行可能
 - NextCloud に外部システムを追加し、ゲストアカウントの発行とグループの設定をユーザができるようにしてある。
- GakuNinRDM: 管理部局による生成が必要（フロー未定）
 - サービス開発中

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

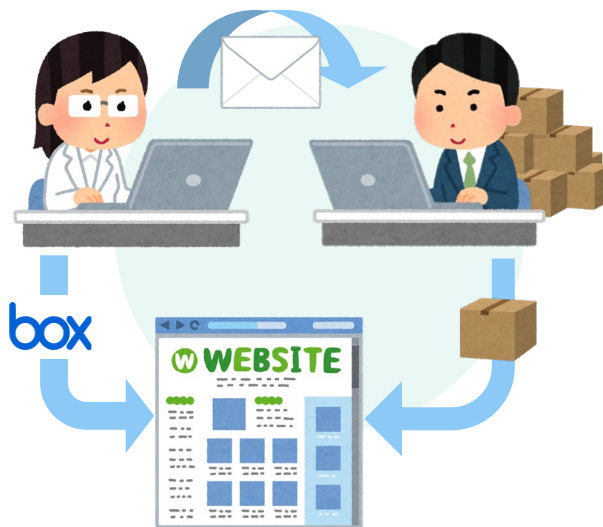
- 各センターの DMP/ガイドラインに従った決定
- 情報領域管理者の管理する領域 (NextCloud: 有料)
- 個人アカウントが管理する領域 (Box: 無料)

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



- ◎ 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
 - 機関ストレージとしてNextCloudを使うと、GakuNinに連動した個人アカウントが紐づけられる
- ◎ 基本情報領域
 - NextCloud
 - 主にHokusai Sailing Ship 上に有料で領域を作り利用
 - 情報領域ごとにデータ連携アカウント等で管理し、所属員の退所等に関わらず存続していく
 - Box... (を予定していたのですが.....)
 - 理研が無料で所属者に割り当てるクラウド領域
 - Box の利用ポリシー（ディレクトリ命名規則や ACL など）を無視した設定が可能になってしまう。(ex. 所内ユーザしか利用できない場所を公開リポジトリにつけて所内ユーザ権限で使ってしまう) → 何らかの抑止機構を実装するまでは利用できない

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

DMS データ管理

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



名前 ^ v 最終更新日時 ^ v

 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して、プロジェクトを整理します。

タグ

DMS x データ管理 x NextCloud x Box x 機密度 2 x タグを追!

最近の活動

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	 Amazon S3 有効にする
その他	 Azure Blob Storage 有効にする
ストレージ	 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

 Nextcloud authorized by Hideyuki Jitsumoto [アカウントを切断](#)

Current Folder: /UsefulData/

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

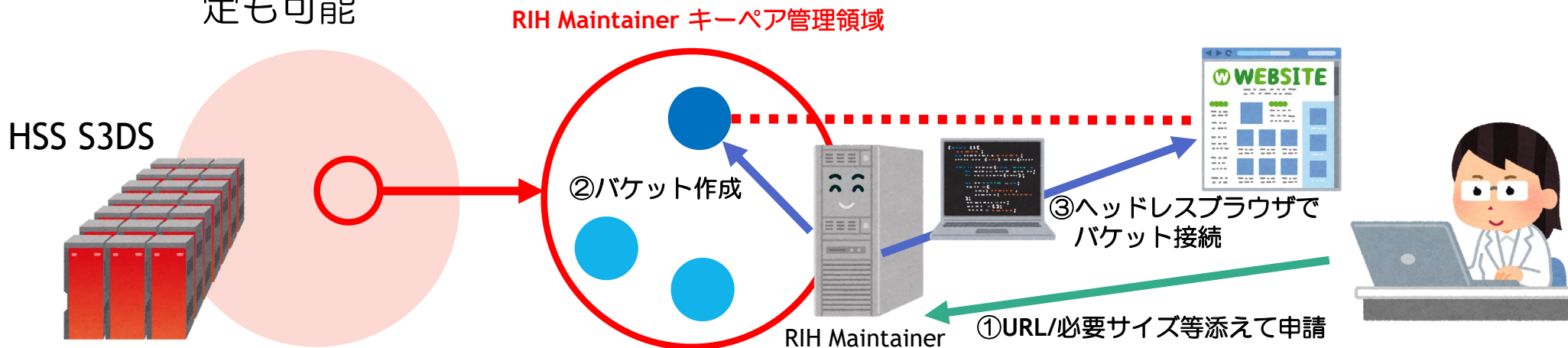
理研へ利活用情報領域（ストレージ）の申請

- ・ 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



- ◎ 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



◎ 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

利活用データ公開申し込み

申し込み
申し込み確認/変更
ログアウト

管理者用検索
管理者用DB確認

申し込み

● 依頼者情報

依頼者: Hideyuki Jitsumoto(191880)

● 代表者情報

氏名: and ID:

● 申し込み内容

研究ページ(URL):
保存データ総量(GB):
保存データ最大数:
保存データ1ファイル最大量(GB): ☐ 15GB未満 ☒ 15GB以上
利用ストレージ: ☐ Box ☒ S3
備考:

※申し込み処理は、少々時間が掛かります。そのままお待ちください。

RIKEN

GakuNinRDMによる研究情報公開

ファイル

Wiki

メンバー

GakuNinRDMによる研究情報公開

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 07:25 PM

カテゴリー: データ

説明:
〇〇学会 第〇回△△研究会発表 成果データ
ライセンス: ライセンスなし

Wiki

メタデータ
ファイルリスト内 metadata.json を参照のこと

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

名前

最終更新日時

GakuNinRDMによる研究情報公開

Nextcloud: GakuNinRDMによる研究情報公開

操作画面スナップショット.JPG 2020-06-12 11:53 AM

Default Private Storage

S3 Compatible Storage: testbucket2 (sailings...

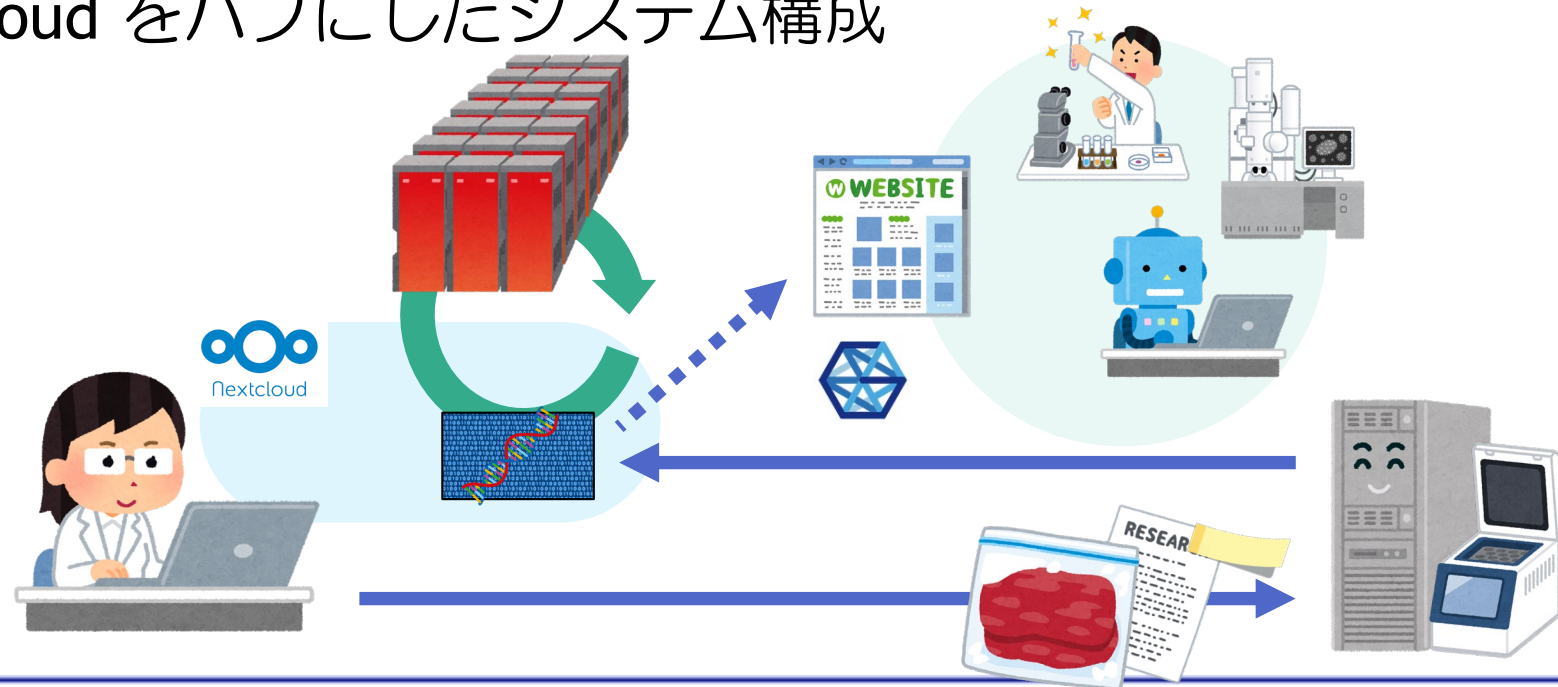
操作画面スナップショット.JPG

- ◎ 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
- ◎ 業績管理システム登録時には公開中の研究管理ページタイトルの候補から選択するだけですむ
 - OSF API を利用
 - 非公開のものは扱えない
 - デバッグモードを除いて管理者権限の API は発行できない仕組み
 - 公開に当たって招待されている RIH Maintainer 権限での操作を行う

研究フローとR2DMS

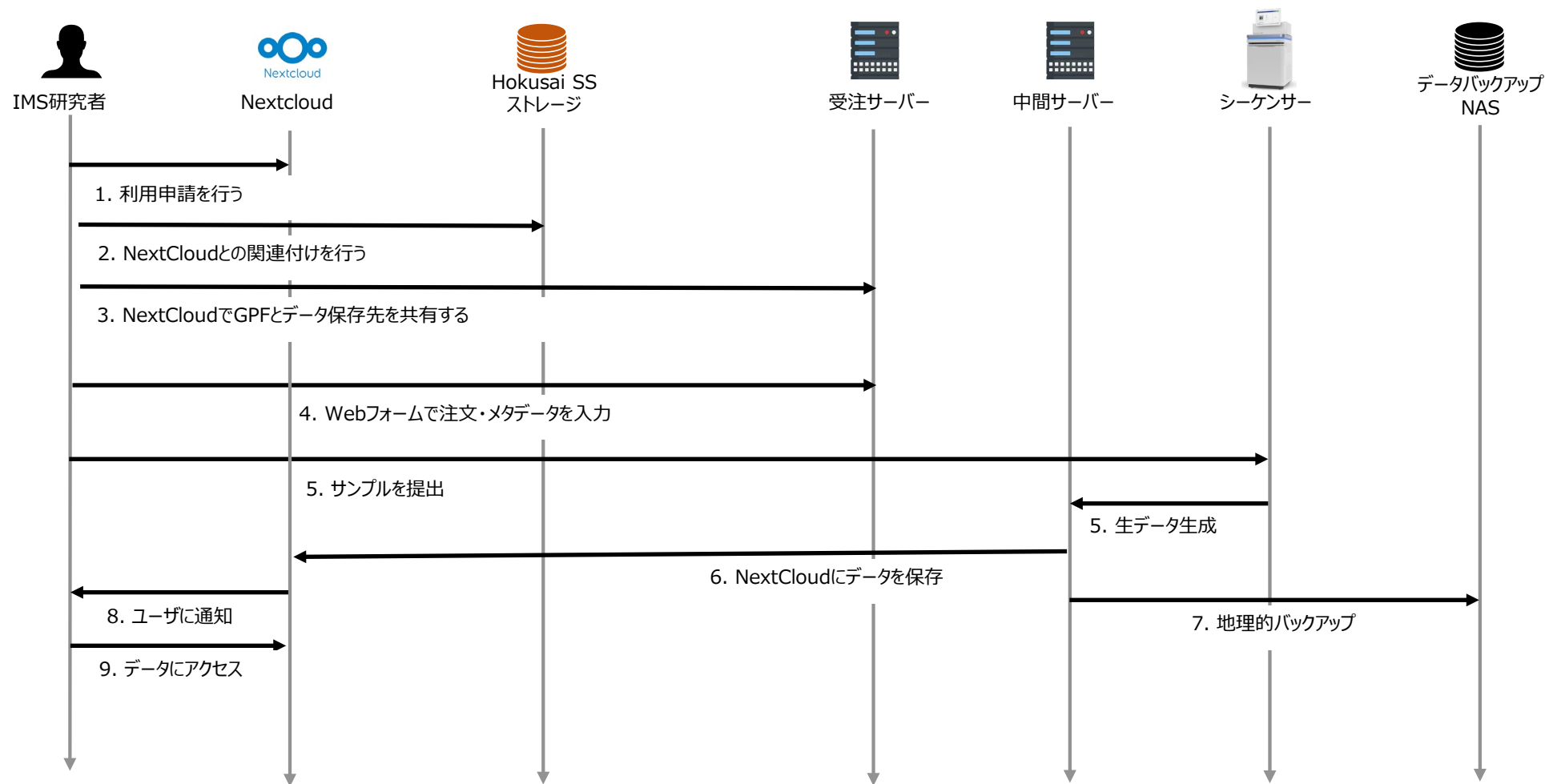
- IMSゲノムプラットフォーム(GPF)

- ◎ 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)



アカウント制御問題

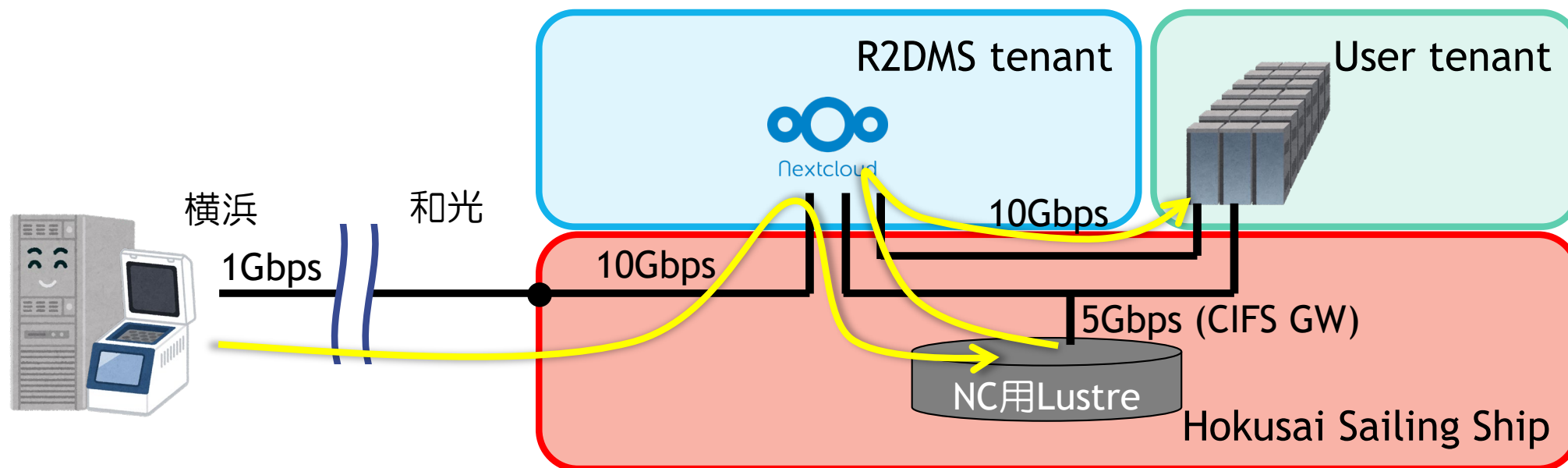
- ◎ リポジトリがスパコンストレージ上にあっても、アカウント管理ポリシーが一致していないと、直接利用はできない
 - リポジトリのアクセスコントロールを無視した提供になってしまう

 - ◎ HSSで起きていること
 - Lustreストレージの提供は VM へのノードマッピングで行う
 - NextCloud は全てのデータを httpd の権限で管理する
 - 利用者ごとに固有の UID 管理とはなっていない
- VMユーザがアクセス可能な NextCloud 内のデータは NextCloud 認証を通さないと制限されない。
- 情報領域管理者は CIFS アカウントでVMへのマウントを行うことは可能だが、権限設定を個別にやらねばならない。
(みんなができるとは限らない)

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/davfs2)
- ② NextCloud -> User VM: 360MB/s (WebDAV /davfs2)
 - Davfs2 のキャッシングが止められてない？



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
 - 200kB x 200並列の転送 (fio)
 - HSS VM内: 230MB/s → IMS GPF: 660kB/s
 - レイテンシが100倍
 - 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ファイル化すること
= 雑多なファイルからあるポリシー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

◎ 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 (json schema)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-ld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

B

C

+

Metadata :{

A: value,

B: value,

C:{

0: value,

}

}

今後について

◎ 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

◎ DOIの付与と管理

