

理研における GAKUNINRDMを用いた 研究データ管理

理化学研究所 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 <Hideyuki.Jitsumoto@riken.jp>

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

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

科学データ基盤：HOKSUAI 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



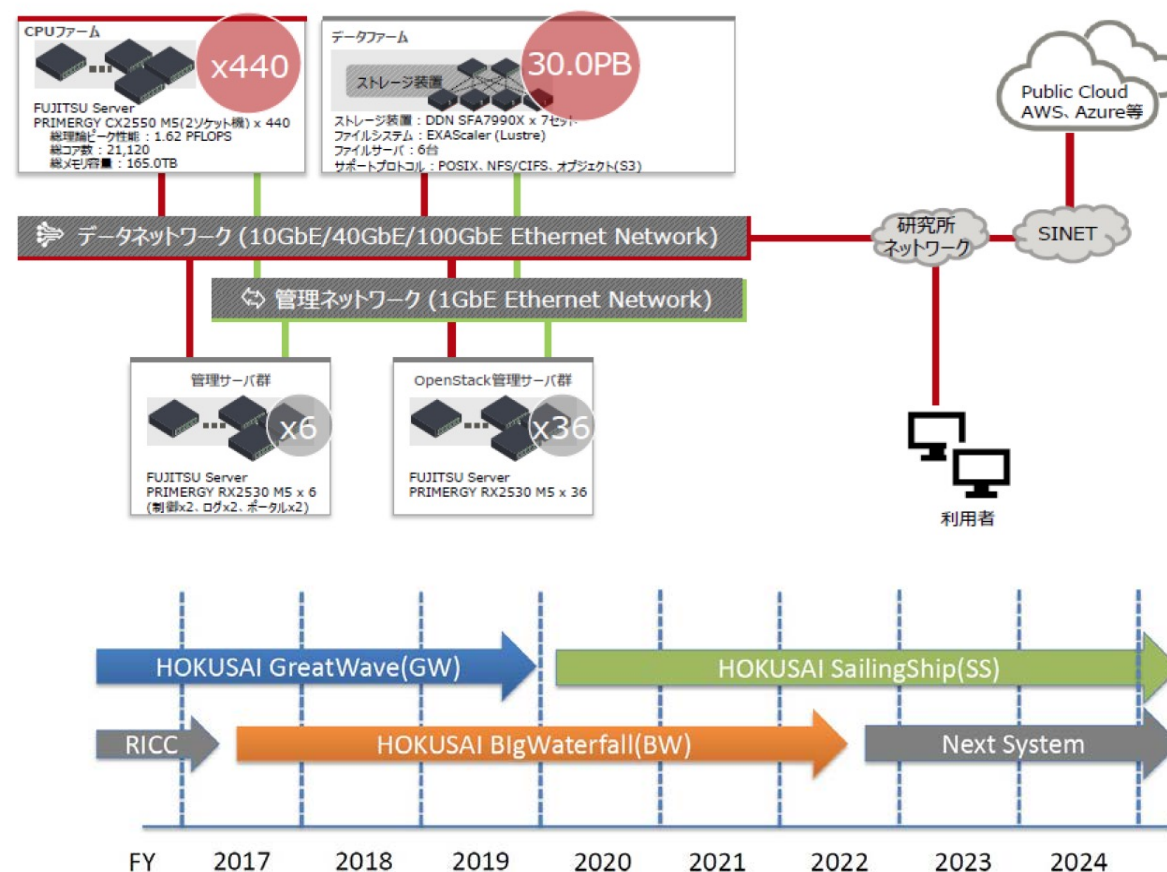
- ◎ Hokusai Sailing Ship (HSS)

- 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

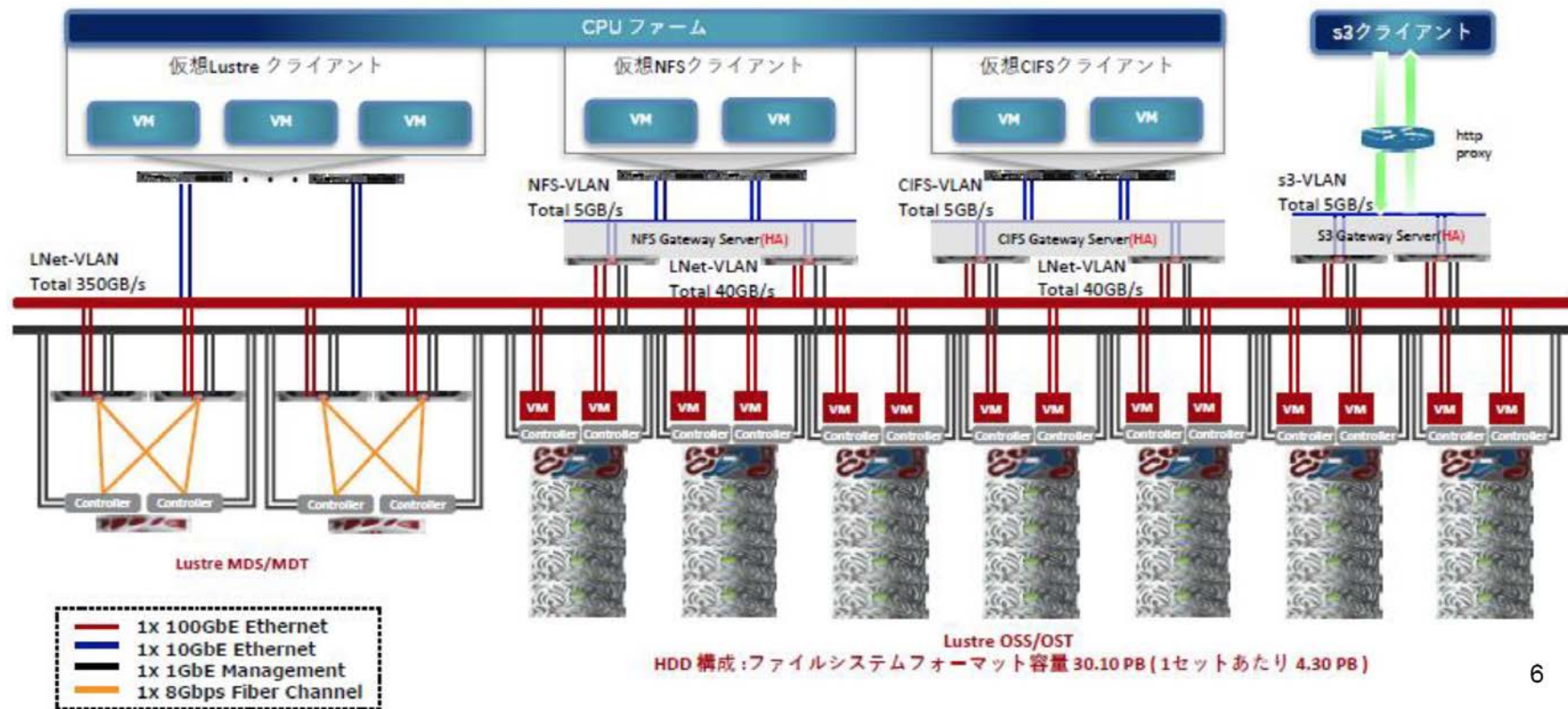
◎ 従来型スパコン(Hokusai-GW) から data-oriented な IaaS/CaaS システムへの転換

- 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

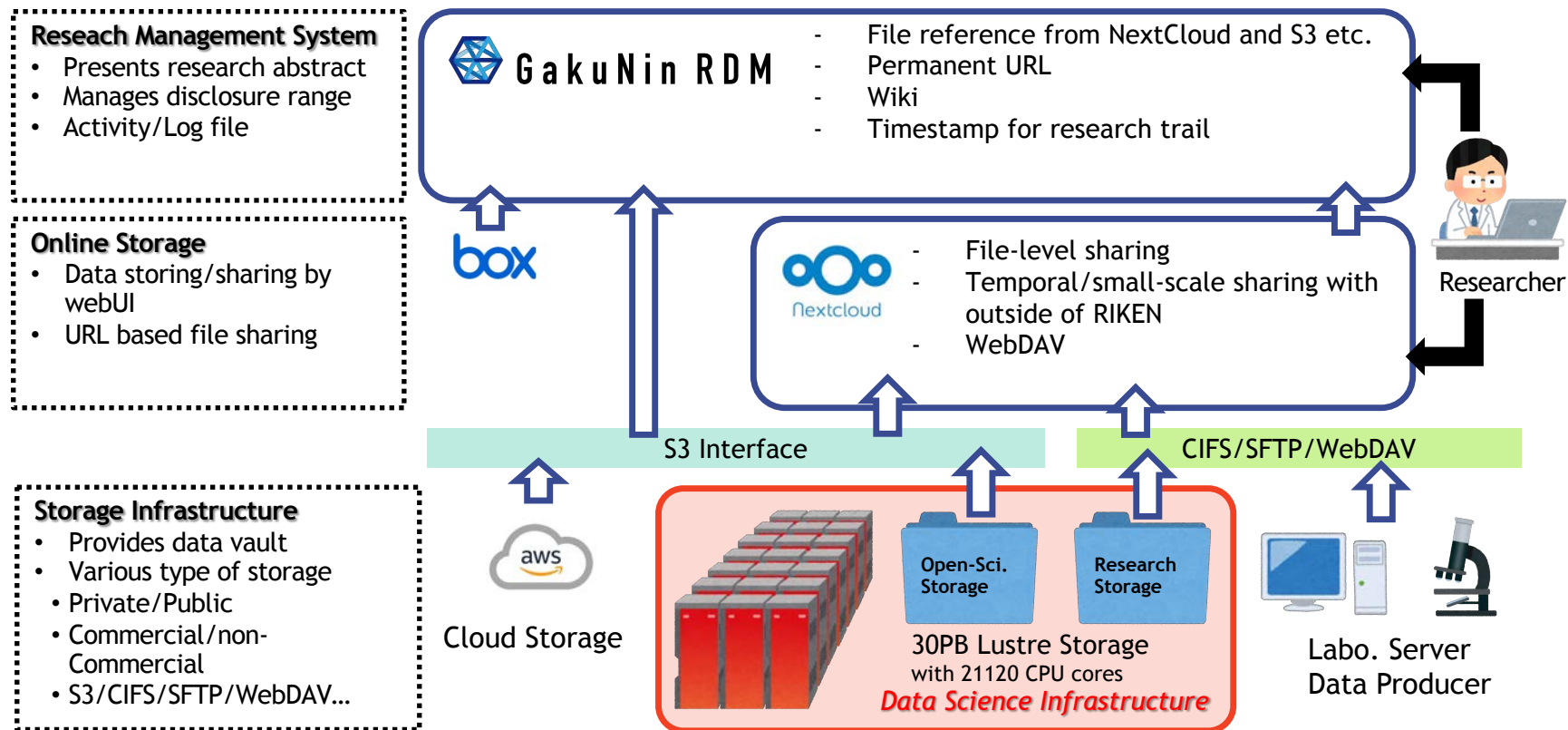
- ◎ データの保管・共有
 - 適切なアクセスコントロールに基づく公開範囲の限定
 - データ操作履歴の保持
 - 改ざん・不正利用の検出
- ◎ 研究データとしての管理・公開
 - データの概要の記述
 - 自由記述, タグ
 - 研究方針やバックグラウンド等の共有
 - データ間関係も考慮した検索・参照
- ◎ 高性能計算機や研究室サーバとの連携
 - 容易で柔軟なデータ取得手法
 - あるいは汎用的な.....



研究情報管理システム概観

◎ 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（公開基盤）
- オンラインストレージによるファイルの管理（保管基盤）



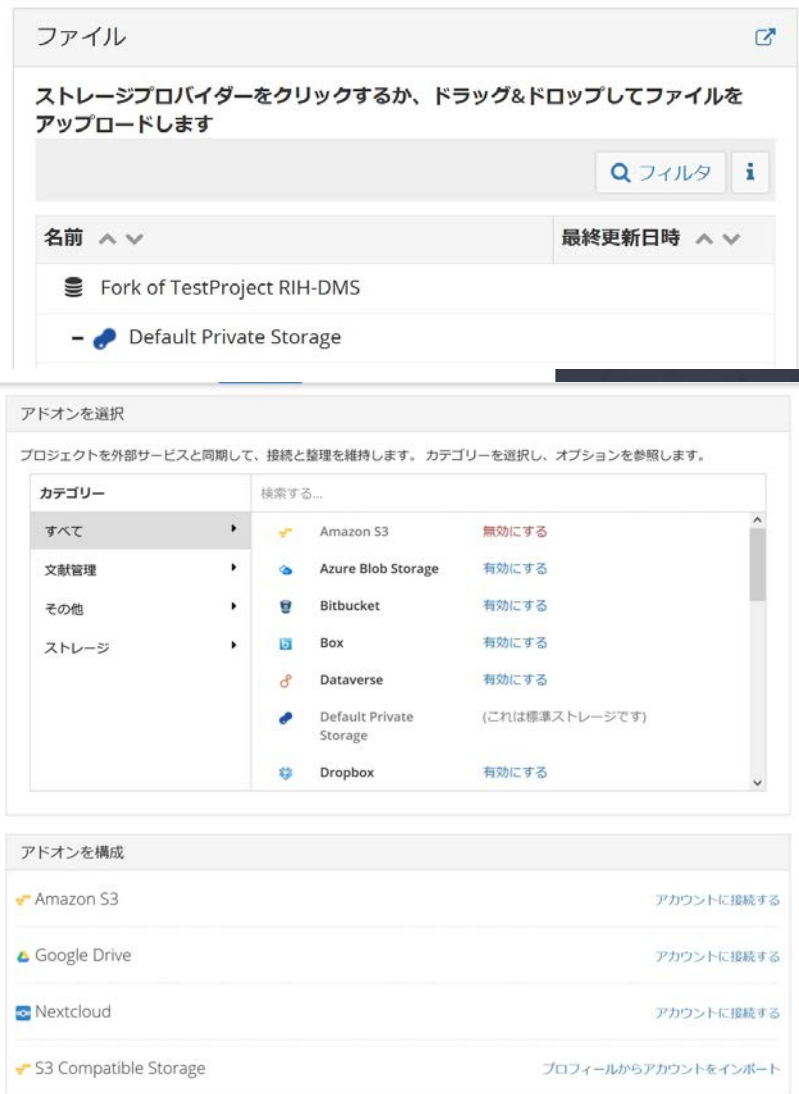
GAKUNINRDM : プロジェクト管理基盤

- ◎ 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、バージョンing、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

利活用情報領域と一般情報領域

- ◎ 利活用情報（情報＝データ＋著作）
 - 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研究所（各センター）が判断したもの

- ◎ 外部ストレージアドオンによってGakuNinRDMに適切な情報領域（ストレージセット）を接続して利用する
 - 利活用情報領域：S3互換ストレージ／無料提供→公開に向けたもの
 - 一般情報領域：NextCloud／有料提供→提携・共有に向けたもの



The screenshot displays the GakuNinRDM interface, divided into two main sections: 'ファイル' (Files) and 'アドオンを選択' (Select Add-on).

ファイル (Files) Section:

- Instruction: 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 (Click on a storage provider or drag&drop to upload files).
- Search bar: フィルタ (Filter).
- Table headers: 名前 (Name) and 最終更新日時 (Last updated time).
- Table content:

名前	最終更新日時
Fork of TestProject RIH-DMS	
- Default Private Storage	

アドオンを選択 (Select Add-on) Section:

- Instruction: 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。 (Maintain connection and organization by synchronizing projects with external services. Select a category and refer to options).
- Table with categories and storage providers:

カテゴリー	検索する...																
すべて	<table border="1"> <thead> <tr> <th>Storage Provider</th> <th>Status</th> </tr> </thead> <tbody> <tr> <td>Amazon S3</td> <td>無効にする (Disable)</td> </tr> <tr> <td>Azure Blob Storage</td> <td>有効にする (Enable)</td> </tr> <tr> <td>Bitbucket</td> <td>有効にする (Enable)</td> </tr> <tr> <td>Box</td> <td>有効にする (Enable)</td> </tr> <tr> <td>Dataverse</td> <td>有効にする (Enable)</td> </tr> <tr> <td>Default Private Storage</td> <td>(これは標準ストレージです) (This is the default storage)</td> </tr> <tr> <td>Dropbox</td> <td>有効にする (Enable)</td> </tr> </tbody> </table>	Storage Provider	Status	Amazon S3	無効にする (Disable)	Azure Blob Storage	有効にする (Enable)	Bitbucket	有効にする (Enable)	Box	有効にする (Enable)	Dataverse	有効にする (Enable)	Default Private Storage	(これは標準ストレージです) (This is the default storage)	Dropbox	有効にする (Enable)
Storage Provider	Status																
Amazon S3	無効にする (Disable)																
Azure Blob Storage	有効にする (Enable)																
Bitbucket	有効にする (Enable)																
Box	有効にする (Enable)																
Dataverse	有効にする (Enable)																
Default Private Storage	(これは標準ストレージです) (This is the default storage)																
Dropbox	有効にする (Enable)																
文獻管理																	
その他																	
ストレージ																	

アドオンを構成 (Configure Add-on) Section:

- Table showing configured add-ons:

Add-on	Action
Amazon S3	アカウントに接続する (Connect to account)
Google Drive	アカウントに接続する (Connect to account)
Nextcloud	アカウントに接続する (Connect to account)
S3 Compatible Storage	プロフィールからアカウントをインポート (Import account from profile)

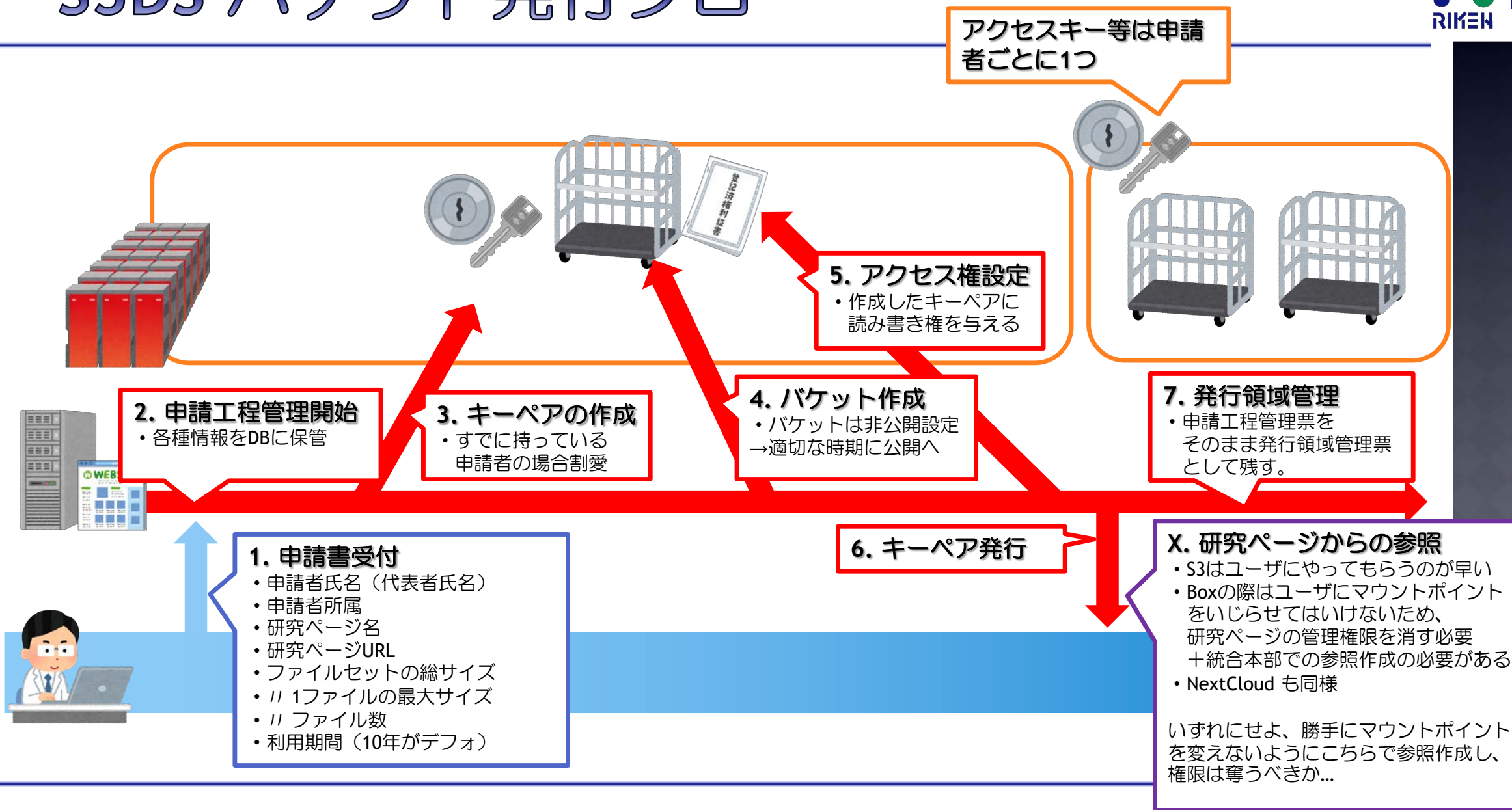
◎ DDNにより Lustre の上に実装された S3 互換インターフェース

- Lustre を利用しているので：
 - オブジェクトストレージ特有のスケーラビリティを持たない
 - Lustre ストレージとしてアクセスすることができる
 - バケット = ディレクトリ
 - バケットに対してクォタをかけられる
 - コールドストレージとの連携なども可能なのでは???
 - 内部でほかのファイルシステムもマウントできそう？<非推奨だそうで。
- 公開に向けた S3 インターフェースの提供
 - s3cmd を直接使うアプリケーションもあるらしい
 - 他サービスのアカウントティングとは独立かつ、ユーザ名・パスワードより、推測しづらい形のアクセスキーペアを用いる

- ◎ 論文・発表は年間10000件ほどある
 - 利活用情報として S3 を手動で 10000件発行するのは無理がある

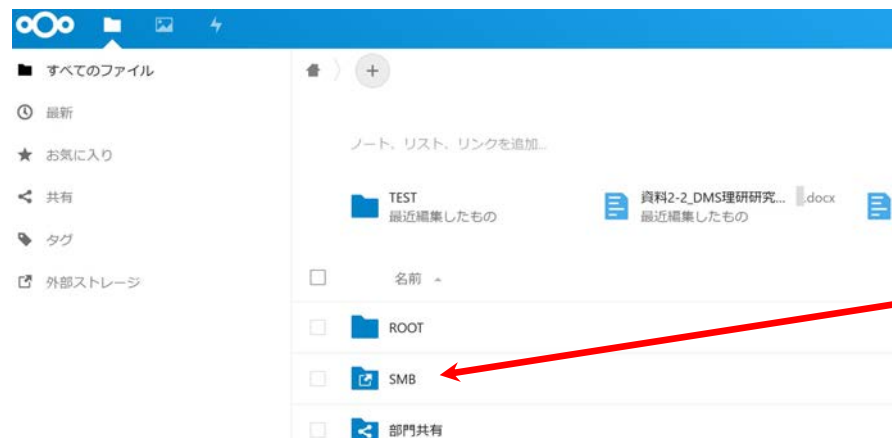
- ◎ S3外部管理サーバ
 - ユーザからの申請をWebより受け取り、S3 キーペアとバケット（非公開）を発行する
 - Lustre レベルでクォタをかける？
 - 図書館業務やユーザとの連携により、正しくバケットのアクセスコントロールを定める
 - バケットを管理し、利用者やキーペアのデータベースを持つ

S3DS バケット発行フロー



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する
 - 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)



外部ストレージ i

外部ストレージを使用すると、外部ストレージサービスおよびデバイスをセカンダリのNextcloudストレージデバイスとしてマウントできます。また、ユーザーが独自の外部ストレージサービスをマウントできるようにすることもできます。

フォルダー名	外部ストレージ	認証	設定
			XXXXXXXX
			/lustre1
			/nextcloud/CXXXXXX
			ドメイン
		ユーザー名とパス	...
☒	SMB	SMB / CIFS	☐ 隠しファイルを表示
☐	ROOT		☐ ファイルをリストアップする際に ACL アクセスを確認する
☐	SMB		CXXXXXX
☐	部門共有		

NEXTCLOUD 外部拡張サーバ

- ◎ ゲストアカウントおよびグループ定義を行うための外部サーバを用意
 - ゲストアカウントや適切なグループ定義は、アクセスコントロールに関するヒューマンエラーの防止に必須
 - NextCloud は管理者（唯一のマスター）のみしかこれらが不能
→ 外部サーバから occ コマンドで DB をいじることにより
権限のデリゲーションを可能とした。

Nextcloud Portal

Hideyuki Jitsumoto (191880@riken.jp) ▾

管理グループ一覧

作成できるグループの最大数になりました。作成できるグループ数を増やしたい場合は情報システム部門までご連絡ください。

#	グループ名	グループの説明	
1	G191880_Test		メンバー 編集 削除
2	G191880_test_exp	説明用テストグループ	メンバー 編集 削除

苦労しているところ

- ◎ 全ての利活用データは個人に紐づくかに、残ってほしい
 - 個人が退職等でいなくなる際、GakuNinRDMのリンクをすべて付け替えるのは非現実的
 - S3 ストレージ と NextCloud ストレージ

しかし...

- ◎ GakuNinRDM は個人ごとに外部ストレージアドオンを管理する
 - 1人が同じストレージプロバイダに対して複数のアカウントを設定できない = ストレージの引継ぎ利用ができない
- ◎ GakuNinRDM の機関ストレージは個人ごとにスペースが作られる
 - S3はマスターとなる空間に全員がバケットを作っていけるらしい
 - 自由に作られてしまうと困る（実はアドオン版も同じ挙動）
（公開データは図書館によるキュレーション等をする）

◎ NextCloud にプロジェクト用アカウントを作成

■ CIFS 連携アカウント

- HSS の Lustre 共有ディスクを NextCloud に接続するための CIFS アカウントが NextCloud のアカウントとしても機能する
→ 引継ぎが発生するようなストレージはすべて CIFS アカウントを使い GakuNinRDM から接続する

◎ S3 外部サーバとバケットの作成場所

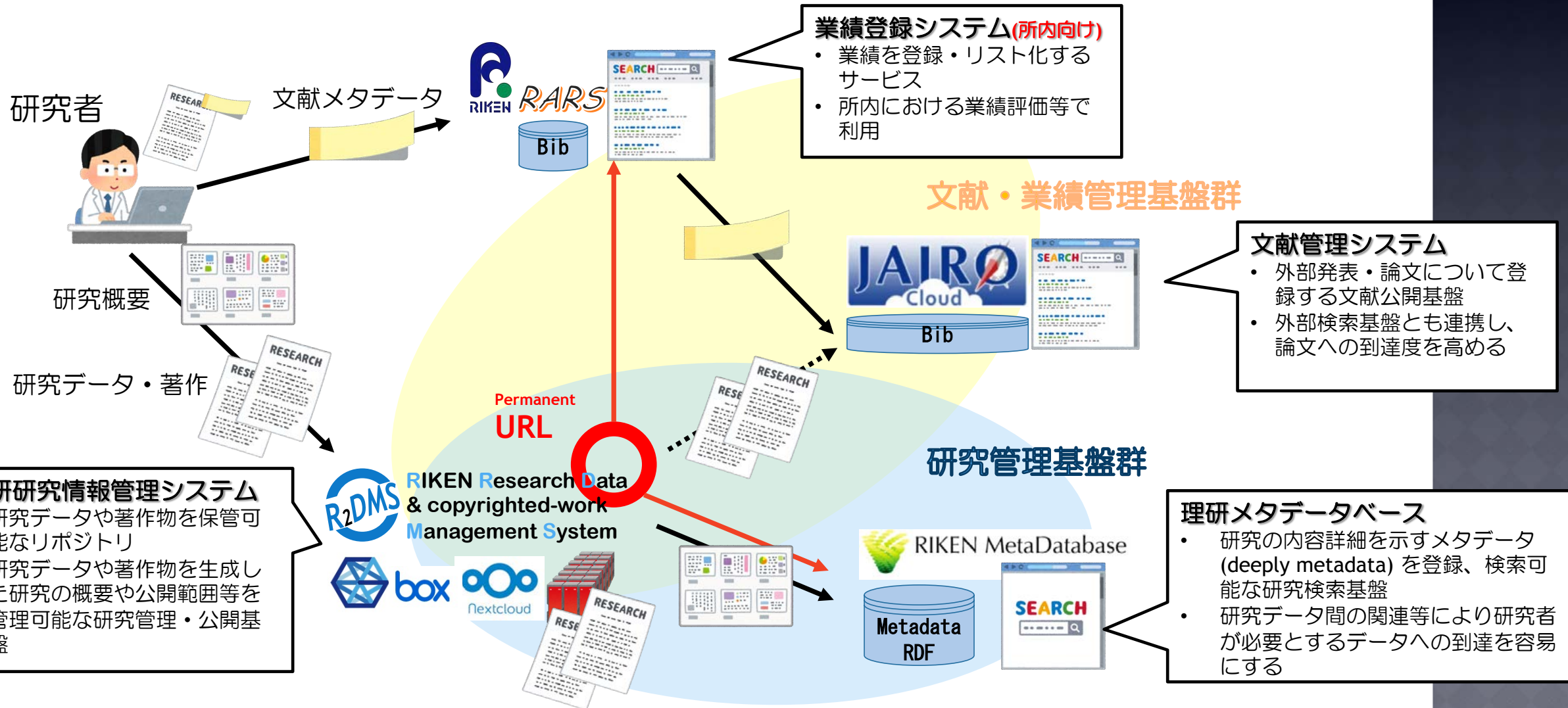
■ 外部管理サーバとの連携

- 申請者ごとに1つキーペアを発行し、バケットはそのキーペアの権限で作成する
- 申請者がいなくなった際にもキーペアは削除せず、申請管理を行っている外部サーバ上のステータスを変更する。

■ 代表申請者

- 引継ぎが発生するようなキーペアは同じ申請者が管理し、利用者に渡す運用をしてもらう

理研オープンサイエンス体系 (※最終形)



今後に向けて：研究情報管理システムと研究フロー

- ◎ データプロバイダからデータ駆動型処理、データ公開までを研究情報管理システムを中心にフロー構築する
 - データプロバイダからのデータを WebDAV-NextCloud経由, SSH-Lustre経由などでHSSのストレージに保存
 - データプロバイダの持つメタデータも自動的にファイル化し、HSS ストレージに保存される
 - GakuNinRDM のプロジェクトページを自動作成し HSS ストレージと結び付ける
 - HSSで解析を行い、結果もストレージに保存
 - GakuNinRDM のプロジェクトページで共有、研究を促進する
 - 必要があれば GakuNinRDM プロジェクトページを公開する

