第3回データワークショップ

SPring-8データセンター構想の進捗報告

初井宇記

理化学研究所

アウトライン

- SPring-8データセンター構想の進捗報告
  - 類型化と課題、検討中の機能
- スケジュール
- まとめ

SPring-8データセンター構想: インフラ

新設

SACLA

SPring-8

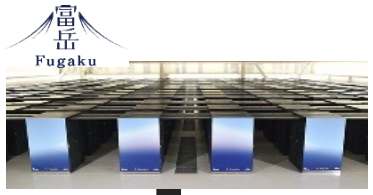
データ圧縮
専用計算機
BL５本

データ
アップロード端末

基幹
ネットワーク
100 G

データセンターBL5本分

富岳・東大等
HPCI計算リソース



7.6 Mcore

SINET6
100 G

2023夏ハンラ

SINET6
400 G

NIMS
マテリアルDX
プロジェクト*

大学・国研等

先端量子ビーム施設

J-PARC MLF
Next to Meet@ MLF

次世代放射光施設

検討中*

検討中

HPCI共用ストレージ

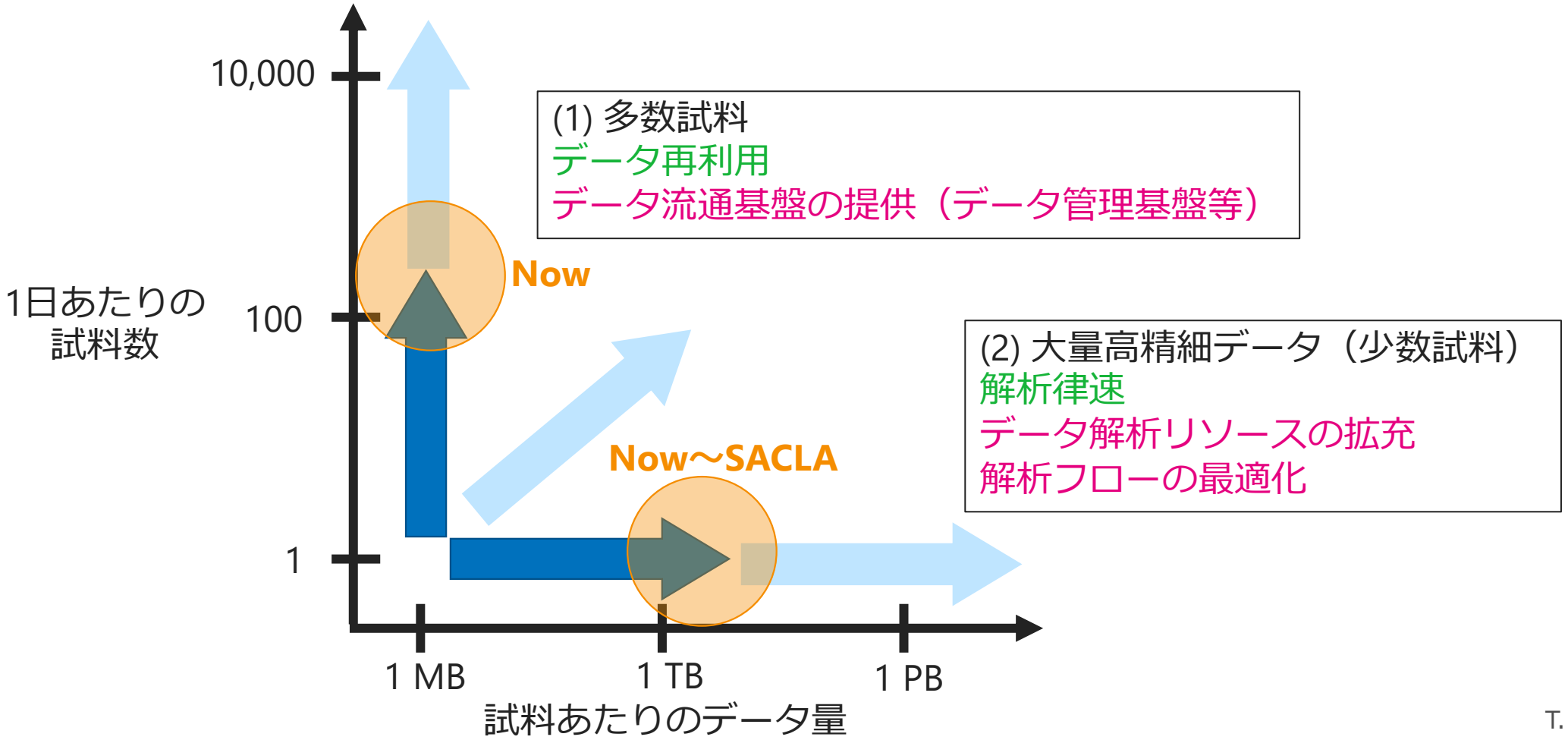
国立情報学研究所
GakuNin RDM

企業内
計算インフラ

商用クラウド

* クラウド上に設置を予定

データ増大に関する類型化と課題、検討中の機能



SPring-8データセンター構想の動機・現状および検討中のサービス

SPring-8・放射光技術の進歩

- 光源
- 要素技術
  - X線光学
  - 検出器
  - ロボット
- エンドステーション
  - 自動化

第2回WS
MDXを中心に

データの飛躍的な増大

1A) 多数試料

- 1日に100以上の試料測定が可能
- 自動化の高度化・ロボット導入によりさらに多数の試料測定が可能に
- 試料当たりのデータは小さい（＜1 TB)
- MDXでの利用が想定される

1B) 多数試料

- 複数手法・複数施設を活用した分析

第1回WS

2) 大量高精細データ・少数試料

- 高精細な3次元像の時間変化追跡が可能
- 元素・化学状態の区別が可能
- 少数試料について測定
- 試料あたりのデータは大きい（＞1 TB)

第3回WS

データの創出基盤・管理基盤・活用基盤の整備

利用シーン

1A-1) NIMSのMDXプロジェクトへデータを登録
1A-2) SPring-8データセンターへデータを格納
1A-3) NIIのGakuNin RDMでデータを管理
1A-4) 企業内サーバーでデータ活用
1A-5) 商用クラウドでデータ活用
1A-6) データベース作成・公開へのサポート

第3回WS

1B) 複数量子ビーム施設のデータをアクセス

2)大量高精細データに対するトータルソリューション

2-1) 実験中に解析まで行い、実験成功率を向上
2-2) 実験後の解析迅速化により成果創出を加速
2-3) 大規模計算による高度な情報抽出
2-4) 富岳と連携した解析サービス

1A-2) SPring-8データセンターへデータを登録

ビームラインに設置している端末

- SPring-8データセンターのアップロード用のweb画面が表示される。
- 自動化されたビームラインは半自動でアップロードされる。

自動化ビームラインについて具体的に検討している例については
次の平木の講演をお聞きください。

SPring-8データセンター構想: インフラ

新設

SACLA

SPring-8

データ圧縮
専用計算機
BL５本

データ
アップロード端末

基幹
ネットワーク
100 G

データセンター

先端量子ビーム施設

J-PARC MLF
Next to Meet@ MLF

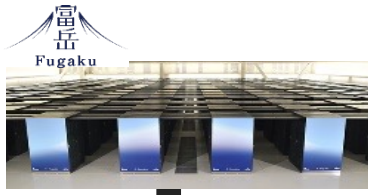
次世代放射光施設

検討中*

検討中

SINET6
100 G

富岳・東大等
HPCI計算リソース



7.6 Mcore

HPCI共用ストレージ

SINET6
400 G

NIMS
マテリアルDX
プロジェクト*

国立情報学研究所
GakuNin RDM

大学・国研等

企業内
計算インフラ

商用クラウド

* クラウド上に設置を予定

SPring-8データセンター構想: インフラ

新設

SACLA

SPring-8

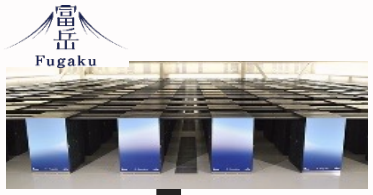
データ圧縮
専用計算機
BL５本

データ
アップロード端末

基幹
ネットワーク
100 G

データセンター

富岳・東大等
HPCI計算リソース



7.6 Mcore

SINET6
100 G

HPCI共用ストレージ

先端量子ビーム施設

J-PARC MLF
Next to Meet@ MLF

次世代放射光施設

検討中*

検討中

データ創出

NIMS
マテリアルDX
プロジェクト*

SINET6
400 G

国立情報学研究所
GakuNin RDM

データ管理

大学・国研等

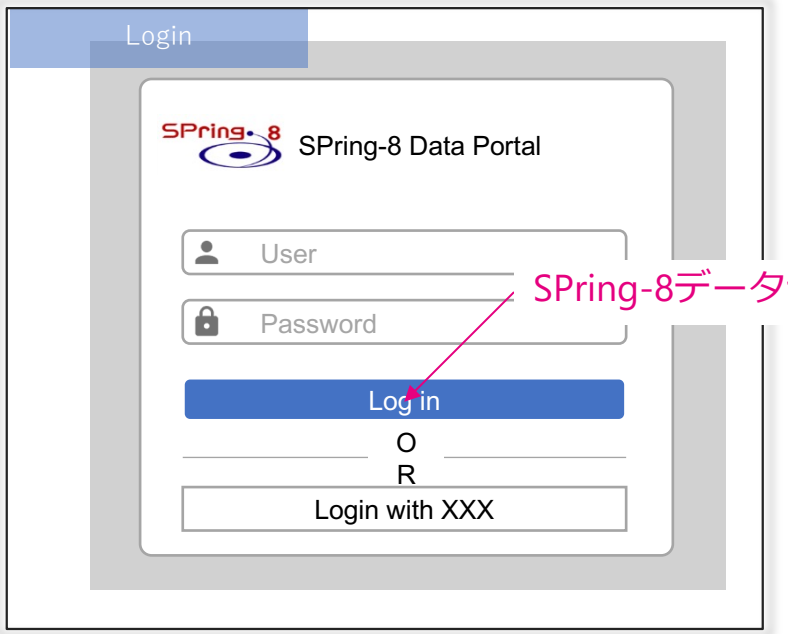
企業内
計算インフラ

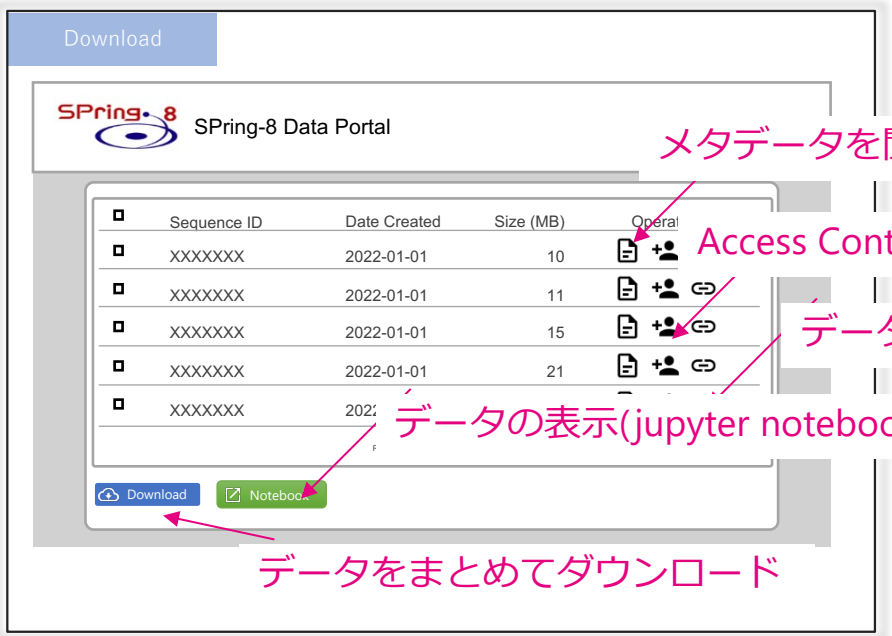
商用クラウド

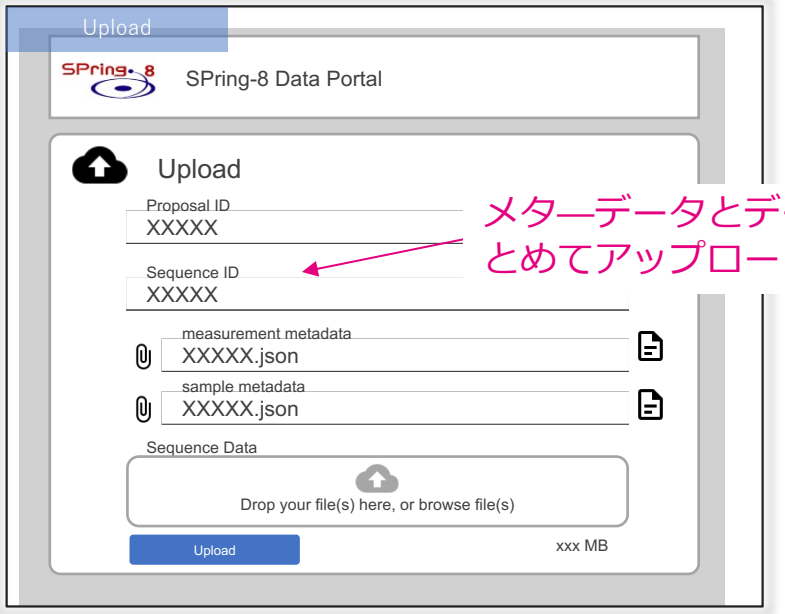
データ活用

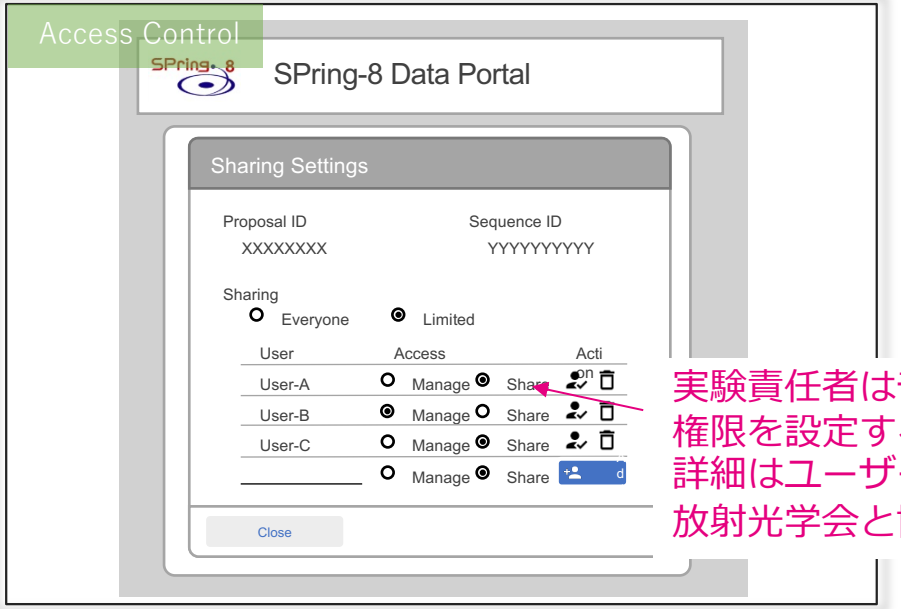
* クラウド上に設置を予定

SPring-8 Data-Flow Systemのユーザー画面（検討中の案）









データの管理について

課題番号　XXXXXXXXX

Sequence ID XXXXXX

Sequence ID XXXXXX

Sequence ID XXXXXX

研究者A

研究者B

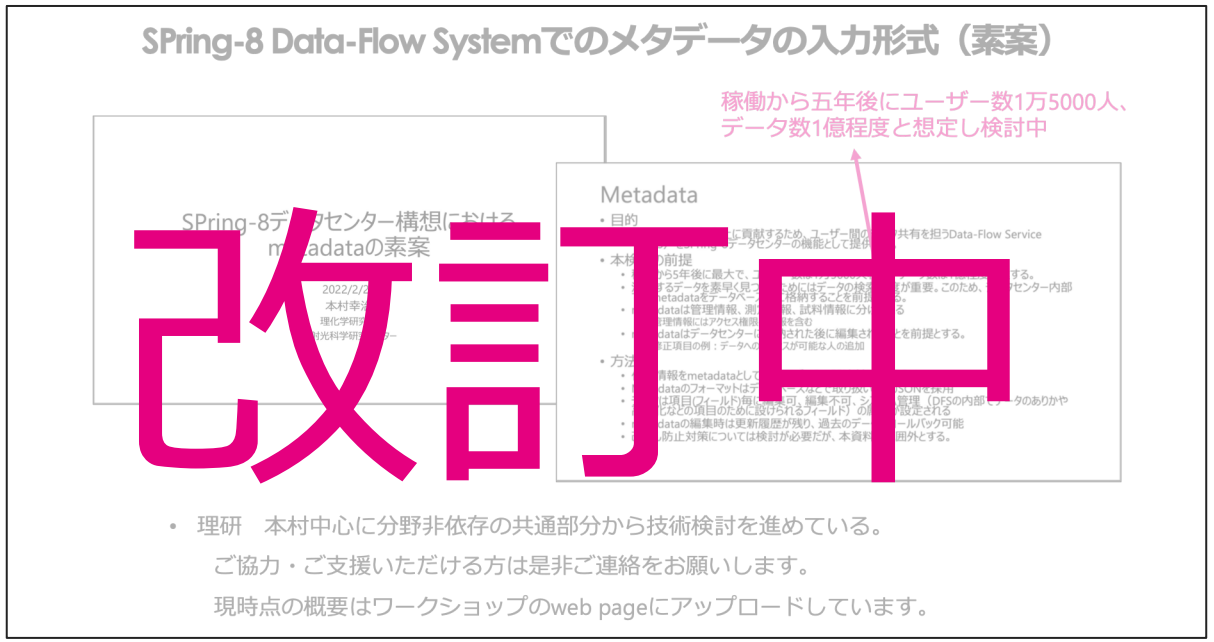
- 課題番号
  - （実装案）課題番号毎にアクセス権限を設定できるようにする
    - 皆様のご意見をお待ちしております。
      - Sequence毎のアクセス管理が必要の場合はお知らせください。
- Sequence
  - 一定の時間測定したデータをsequenceとよび、メタデータを付与する。
  - 以下のメタデータ付与は義務とする。
    - 管理用メタデータ（課題番号、実験日など）

メタデータ

第2回データワークショップで提示させていただいた素案

SPring-8 Data-Flow System
メタデータは以下の3種とする

- 管理情報
- 測定関係情報
- 試料関係情報



前回からの進捗

背景

- 理研では統合情報本部においてデータ構造化を進めている

現状

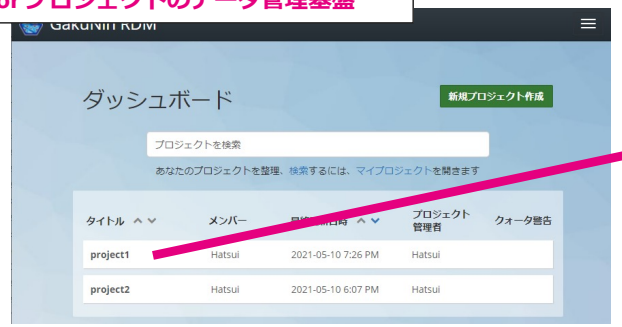
- SPring-8についても連携してメタデータ構造を改訂中
- 世界標準の規格をなるべく参照
- NIMSのデータ構造と連携できる構造化とする
- 放射光学会でのデータ構造化の議論とも整合性のある形にする

理化学研究所
情報統合本部
小林 紀郎（D.Eng.）

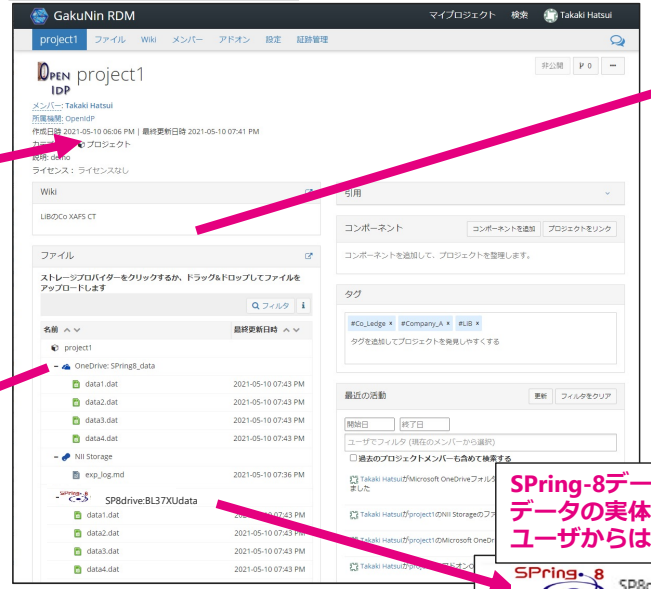
是非インプットをお願いします。

1A-3) NIIのGakuNin RDMでデータを管理

NII上に構築されているGakunin RDM
or プロジェクトのデータ管理基盤

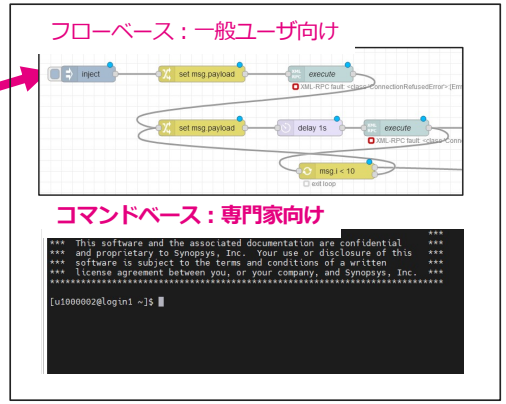


研究テーマごとのページ



所属機関サーバ
プロジェクトサーバ
商用クラウド

SPring-8データセンター or
HPCI共用ストレージ上（新規）
実験中の解析を最先端のアルゴリズムで
実施できる（ユーザのブラウザ上）

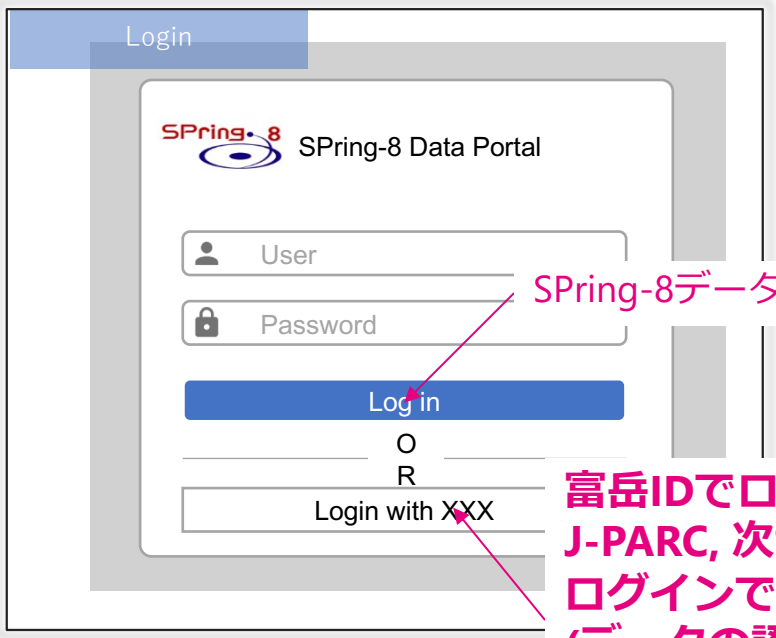


SPring-8データセンター or　HPCI共用ストレージ上（新規）
データの実体はSPring-8が管理するサーバにあるが
ユーザからは手元にあるように見える

研究室の他の装置のデータと一括管理が可能
SPring-8のデータは実体はSPring-8にあるが参照先が分かる。
解析もここからアクセス可能とする予定
データ公開機能もあり

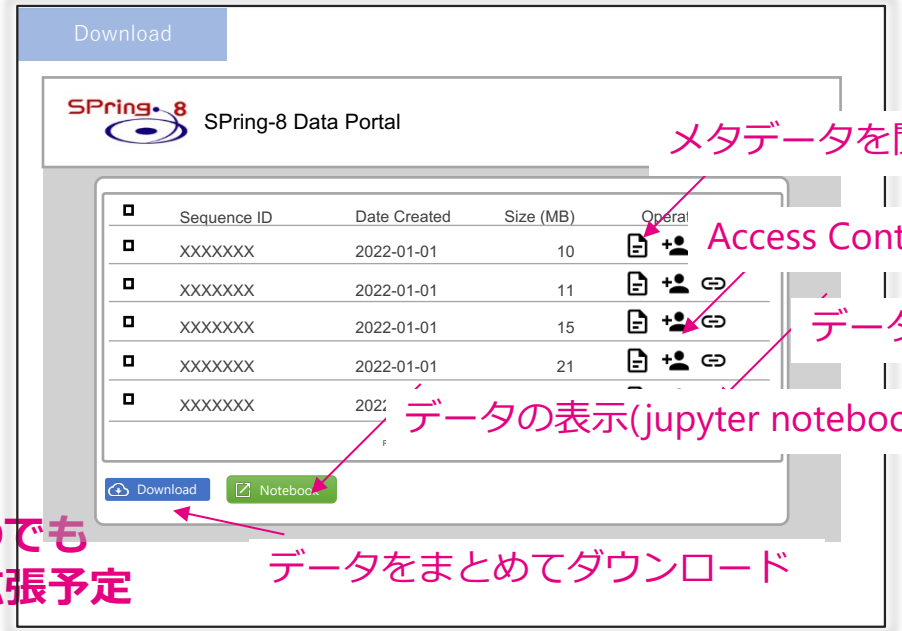
国立情報学研究所(NII)には**WEKO3**というサービスがあり、この機能を使えば
GakuninRDMで管理されているデータを公開することも可能。

1B-1) 複数量子ビーム施設のデータをアクセス



SPring-8データセンターのID

富岳IDでログイン
J-PARC, 次世代放射光IDでも
ログインできるように拡張予定
(データの認証・認可
先端大型施設間で連携)



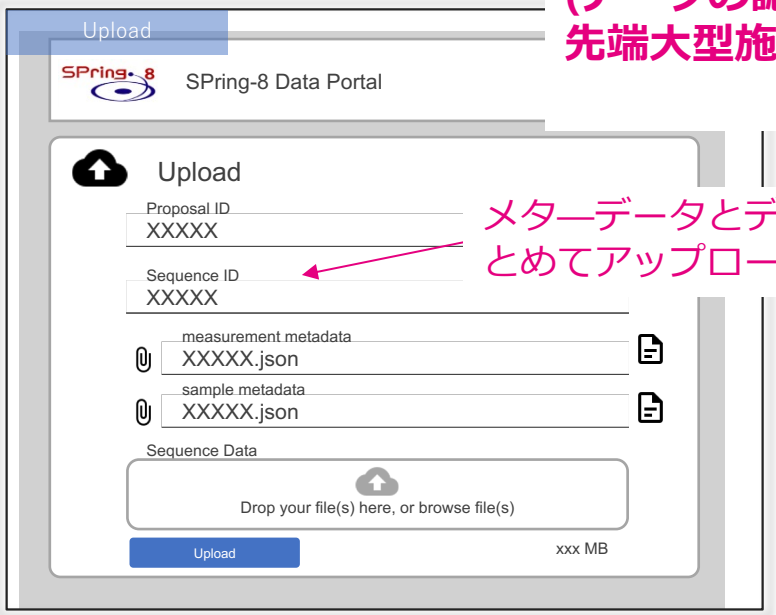
メタデータを閲覧

Access Controlへのリンク

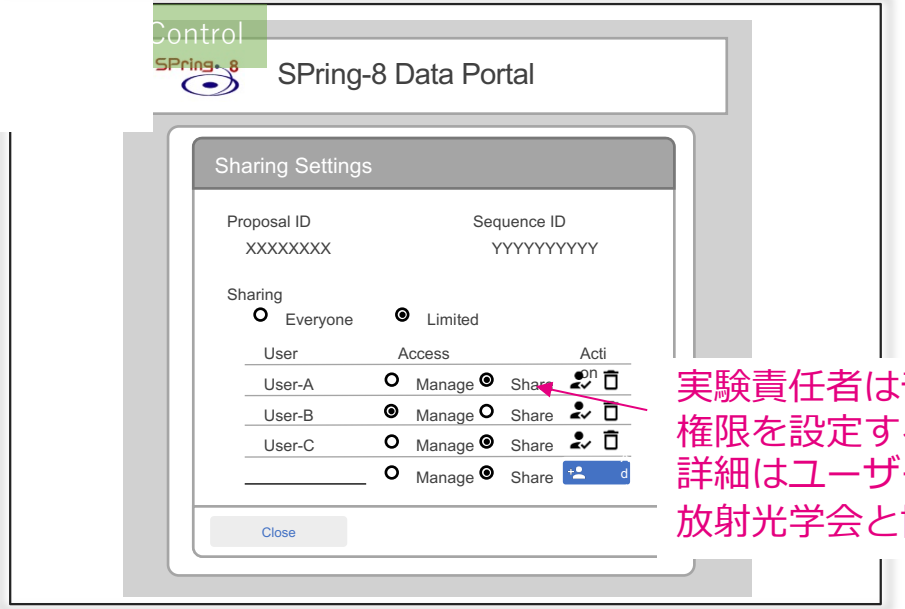
データへのリンク(REST API)

データの表示(jupyter notebook)

データをまとめてダウンロード



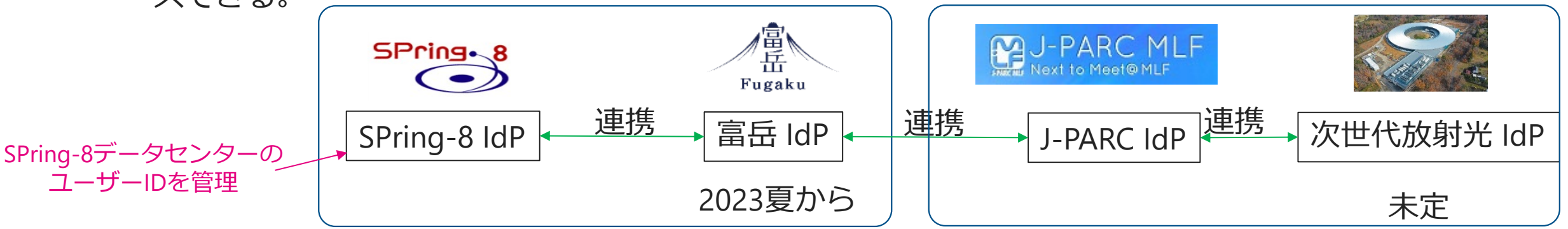
メターデータとデータをまとめてアップロード



実験責任者はデータのアクセス権限を設定することができる。詳細はユーザーコミュニティ・放射光学会と協議

1B-1) 複数量子ビーム施設のデータをアクセス

- 連携が必要な機能
  - データの認証・認可：共通化することで一つのプログラムから複数施設のデータを簡単にアクセスできる。



将来的にNIIの認証基盤とも連携

- 共通化が必要な機能
  - データアクセス方法の共通化
    例：データアクセスの基幹部分のソフトウェアがHOST名を変えるだけでどの施設でも使える

検討中の実装
　Amazon社のオブジェクトデータサービスS3と互換性のあるサービス
ユーザーから見た使い方
　pythonライブラリ boto3によりSPring-8データセンター内のデータをアクセスできるようになる。

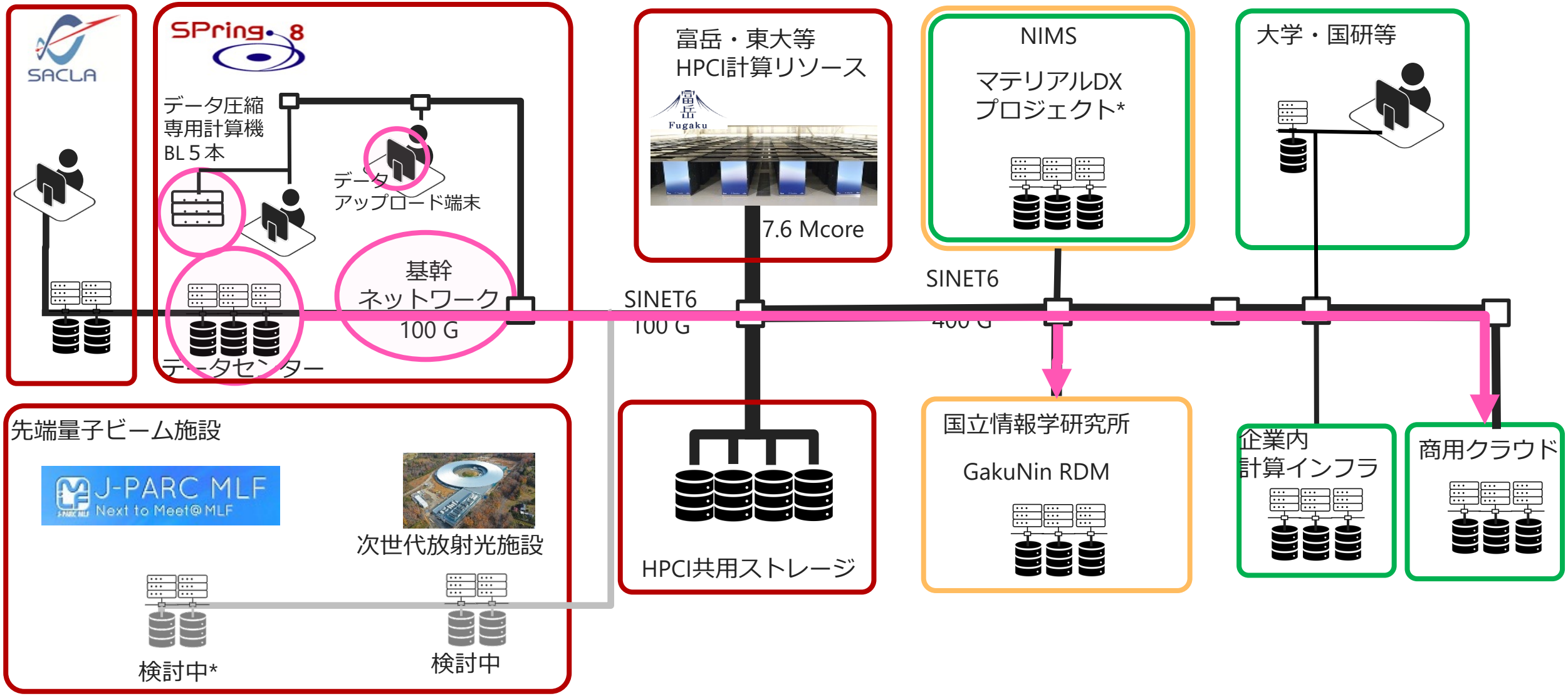
1A-4) 企業内サーバーでデータ活用
1A-5) 商用クラウドでデータ活用

- ヒアリング状況
  - 企業内サーバへのデータ移動
    - VPNなどを介すため帯域が狭く、実務上ダウンロードできない場合が多い。
  - 商用クラウド（Amazon AWS, Microsoft Azure他)
    - 利用できる企業が増えている。

- サービス方針案
  - 1A-4) 企業内サーバーでデータ活用
    - 小さなデータ（<< 1 GB)の場合ダウンロードできる
    - SPring-8 Data-Flow Systemで対応
  - 1A-5) 商用クラウドでデータ活用
    - SINET6に直接接続しているデータセンターを対象に、中程度のデータでもダウンロード可能なサービスを検討する。
      - 対応できるデータサイズはこれから性能確認を行う
    - SPring-8 Data-Flow Systemで対応
    - プログラムから直接読み出すユーザにはboto3によるデータアクセスを提供

SPring-8データセンター構想: インフラ

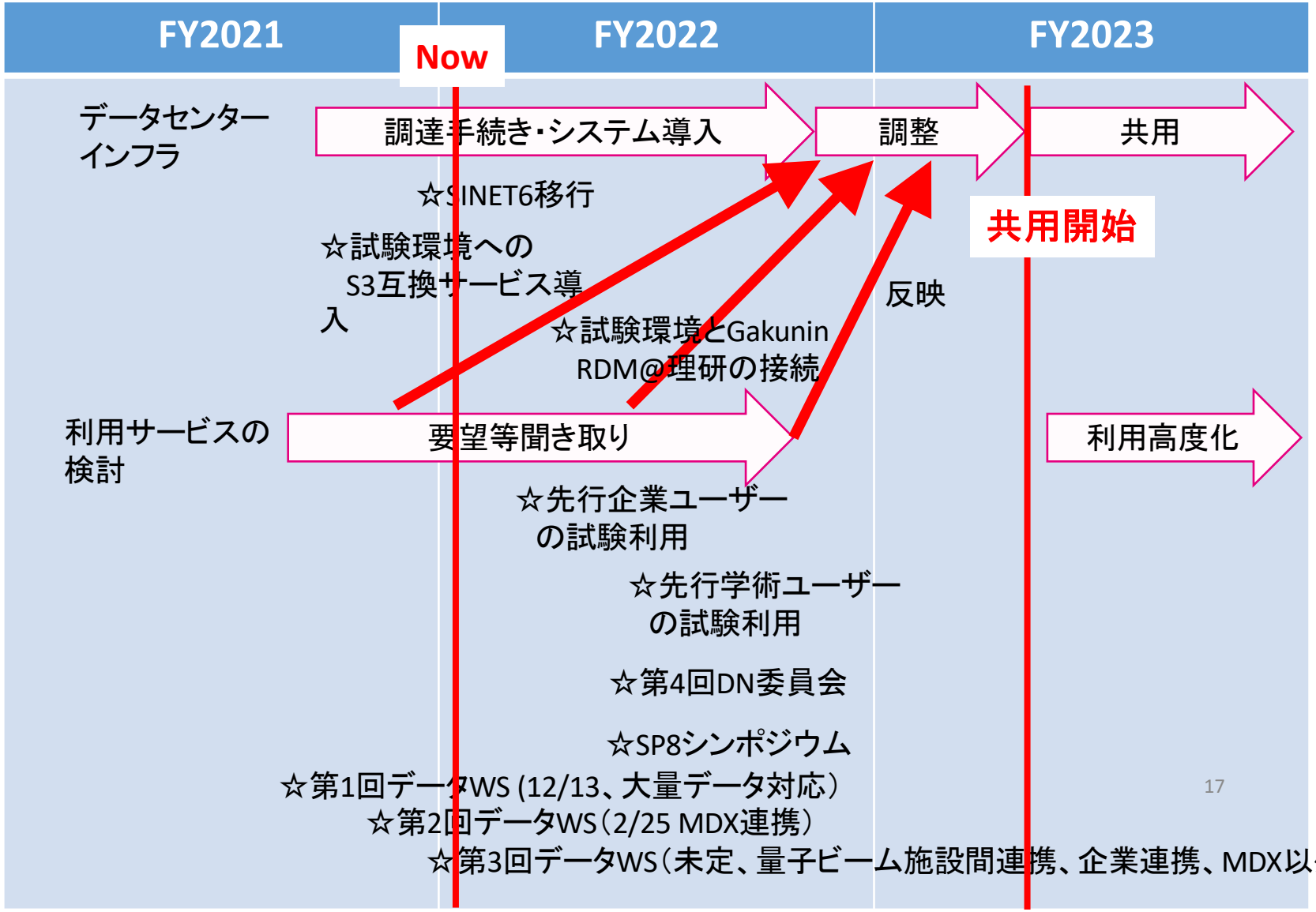
新設



* クラウド上に設置を予定

新設
2021年度補正予算

スケジュール



- MDXプロジェクト参画代表者
- 専用ビームライン2機関
- 企業　2法人

などからフィードバック・ご要望をいただいています。

まとめ

- SPring-8データセンター構想の進捗報告
  - 類型化と課題、検討中の機能
  - データ管理の考え方
    - 課題番号：アクセス管理するレイヤー
    - Sequence: SPring-8データセンターでデータを管理するに当たり、新たに導入するデータの粒度
- スケジュール
- SPring-8データセンターに関する情報
  - https://dncom.spring8.or.jp/
- ご意見のある方は、是非ご連絡をお願いします。
  - dncs@spring8.or.jp

謝辞

- [SPring-8データセンター構想]
  - 城地保昌(JASRI,理研),中嶋享(JASRI), 本村幸治(理研), 平木俊幸(理研),杉本崇(JASRI),
  - 山鹿光裕(JASRI), 矢橋牧名(理研)
  - 中町将貴（理研）、渡邊一輝（理研）

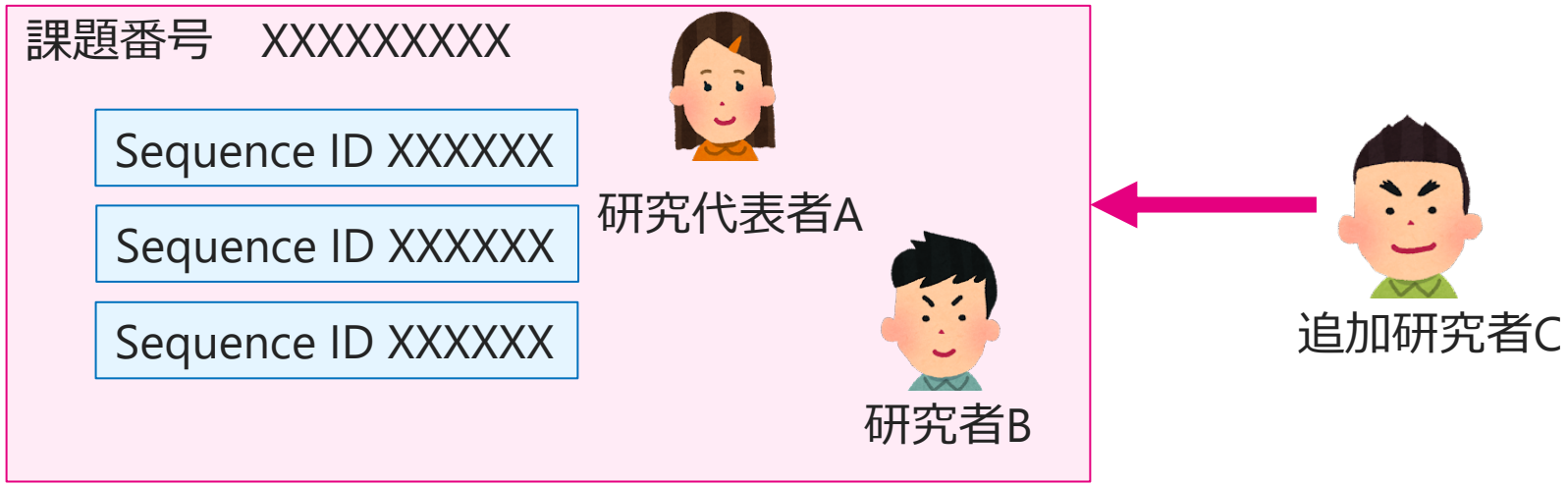
- [HPCI連携、データ圧縮、Workflowツール、認証]
  - 理研計算科学研究センター(R-CCS, 富岳)
    - 松田元彦, 原田浩、金山秀智, 山本啓二, 佐藤賢斗,
    - 佐野健太郎, 庄司文由, 松岡聡

- [研究データ管理]
  - 理研情報統合本部(R-IH)
    - 實本英之, 冨士健太郎,小林紀郎, 美濃導彦

- 国立情報学研究所(NII)
  - 込山悠介, 山地一禎, 喜連川優

SPARE SLIDE

データの管理について



- 課題番号
  - （実装案）課題番号毎にアクセス権限を設定できるようにする
    - 皆様のご意見をお待ちしております。
      - Sequence毎のアクセス管理が必要の場合はお知らせください。

- Sequence
  - 一定の時間測定したデータをsequenceとよび、メタデータを付与する。
  - 以下のメタデータ付与は義務とする。
    - 管理用メタデータ（課題番号、実験日など）

データのオープン化と
SPring-8データセンターの関係

1A-6)データベース作成・公開へのサポート

Big Data Value Chain[1]

[1] Freitas A., Curry E. (2016) Big Data Curation. In: Cavanillas J., Curry E., Wahlster W. (eds) New Horizons for a Data-Driven Economy. Springer, Cham. https://doi.org/10.1007/978-3-319-21569-3_6
[2] digital object identifier

SPring-8の場合

• メタデータを付与したデータの格納（３年） • 認証・認可基盤によりアクセス権限を課題責任者が設定・変更可能	• 標準試料データを用いたデータ校正	• Data Quality • Trust/Provenance • Annotation • Data Validation • …	• データの格納 • データアクセスの提供 • DOI[2]の付与	• データの活用のための機能提供

SPring-8データセンター構想
- 解析律速の解消
  SPring-8内データセンターおよび富岳等HPCIスパコンによる計算リソース
- データ共有の簡便化
  SPring-8 data-flow system

ユーザーコミュニティやプロジェクト（ARIM、MDX等）が主導

データ公開のために必要なインフラ：永続的なデータ保存、検索・アクセス基盤の整備

データ公開基盤（各研究機関、国立情報学研究所、NIMS等）を担当される機関と連携し、データ公開化に協力して参ります。

データのオープン化に関するSPring-8データセンターの役割

考え方

SPring-8データセンター構想
連携機関と協力してSPring-8ユーザーにデータに関する課題を解決していく
SPring-8に設置するデータインフラは以下を担う
データ創出
実験中のデータ処理
一時保管(論文出版に必要な期間を目標としています。)

データの所有権、使用権

データネットワーク委員会にて詳細を決定して参ります。
ご意見がある方は、データネットワーク委員会の委員、もしくは理研側へご連絡ください。

前提としては、以下で進めるものと考えています。
【所有権】データの所有権はユーザが持つ
【使用権】施設者は保守･高性能化などに資する情報を得るためにデータを使用する場合がある。
ただし、収集したデータは第3者へは開示しない
【運用】データはバックアップをとらない。ユーザーは期限が来たら消去する必要がある。