NII RCOS
Research Center for Open Science and Data Platform

ビッグデータの利活用に向けた研究データ管理・ガバナンス構想

国立情報学研究所（NII）
オープンサイエンス基盤研究センター（RCOS）
特任研究員
平木俊幸

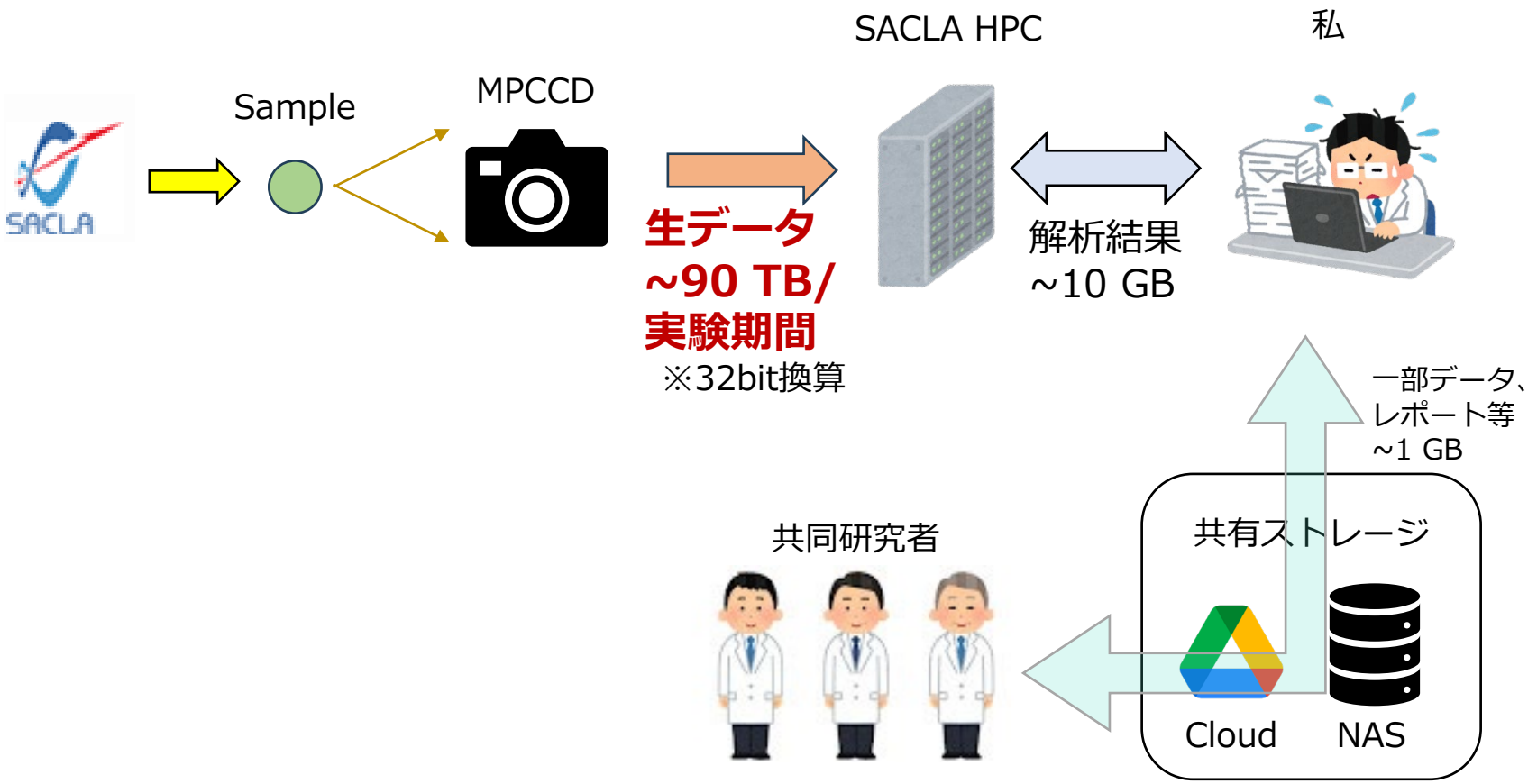
2023年6月21日

Table of contents

ビッグデータ管理上の課題

自身の経験から
– 博士課程在籍時 (2015/4~2018/3)

X 線自由電子レーザー SACLA と X 線イメージング検出器 MPCCD を利用した広角 X 線回折実験を実施。



生データは SACLA HPC 上で管理できた（というよりそこから動かすことが困難であった）。

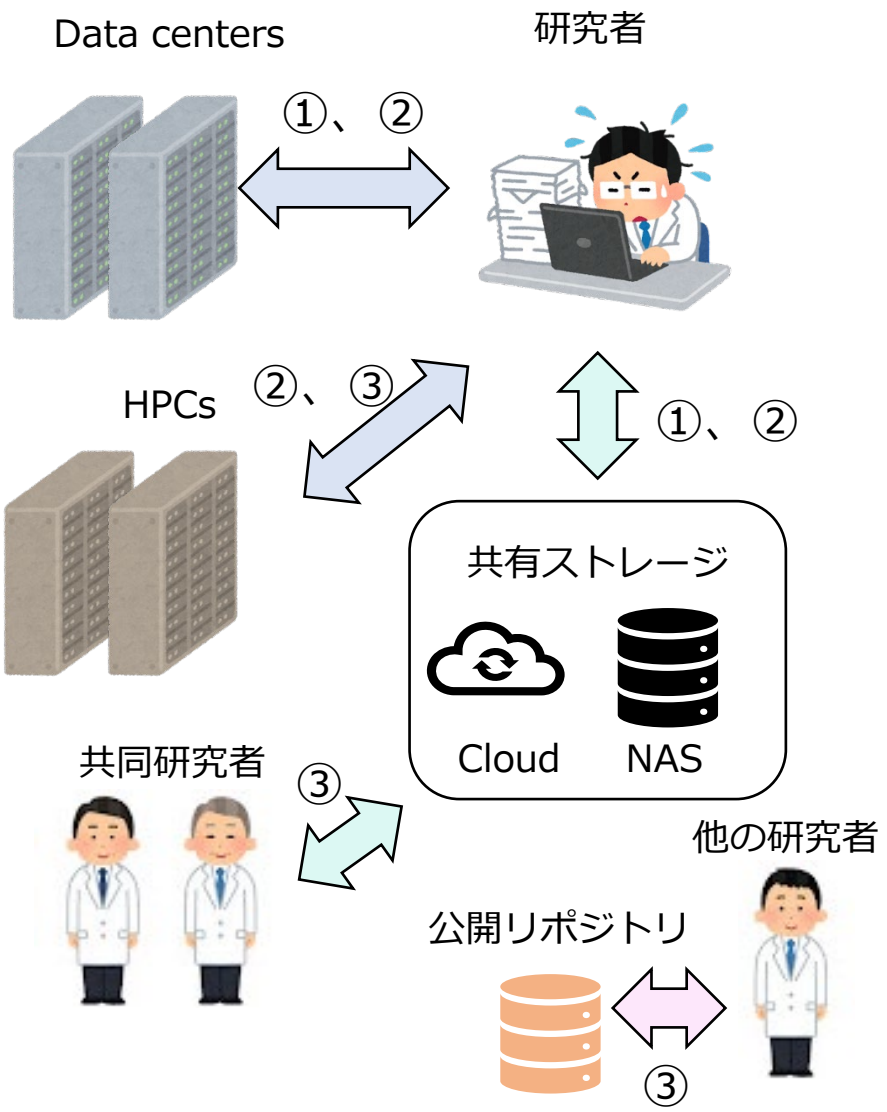
ビッグデータ管理上の課題

あえて分けるなら三つに分類できそう。

① データの置き場
- 誰が提供するのか？機関が調達？
- 共同利用施設やスパコンを利用？
- クラウドを借りるのが良いのか？
- **そもそもデータを移動できるのか？**
- **どのくらいの期間置けるのか？**
  - 例えば SACLA HPC ではポリシー上、最初の 3 年以内は hot access 可能なストレージに保存、その後 5 年間は tape storage に保存[1]。
- 増え続けるデータをどう管理するのか？

② データ管理方法
- データを一元的に管理できるのか？

③ データ利活用
- チーム内で共有し、使えるようにするにはどうすればよい？
- 公開したデータをほかの研究者が参照・利用しやすくするためにはどうすればよい？



[1] http://xfel.riken.jp/users/pdf/20200716_SACLA_data_retention_e.pdf

研究データ管理・ガバナンス構想

※あくまで個人的なアイデアである。
※必ずしも前述の課題を解決するものではない。

①置き場所について

※実現のためのコストは度外視。

ビッグデータは生成場所から繰り返し移動することが（特に経済的・時間的コストの観点で）困難。

↓

- **HPC またはクラウド環境のすぐ近くにビッグデータを置く。**
- **（研究活動上）現実的な時間内でビッグデータを他の HPC の近くにマイグレーションできるようなネットワークを構築する。**
  - まるで手元にデータあるような感覚で移動できるとベスト。

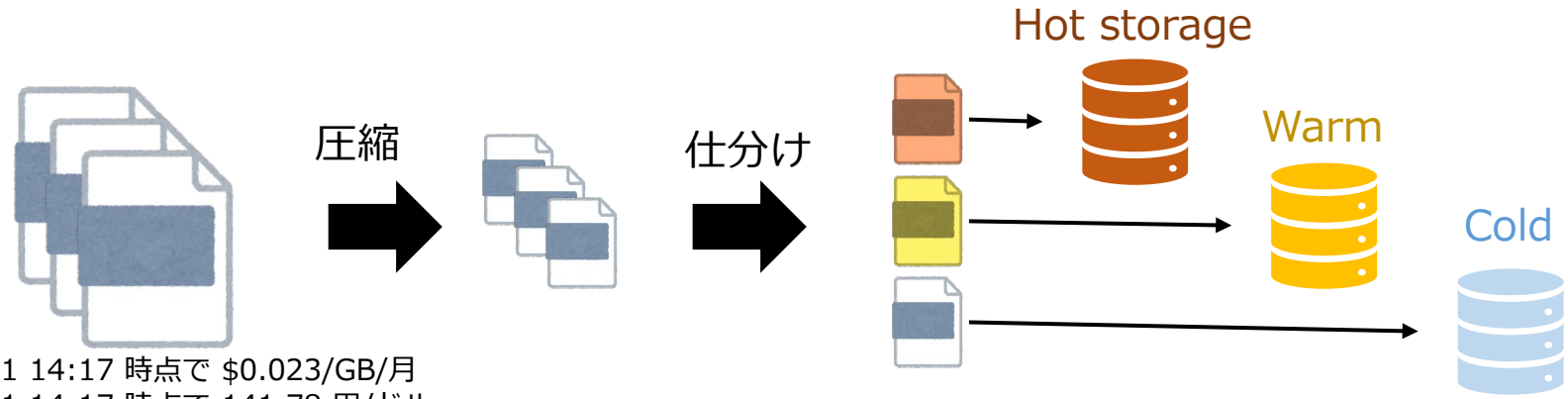
①置き場所について

※実現のためのコストは度外視。

ビッグデータは生成場所から繰り返し移動することが（特に経済的・時間的コストの観点で）困難。

↓

- **HPC またはクラウド環境のすぐ近くにビッグデータを置く。**
- **（研究活動上）現実的な時間内でビッグデータを他の HPC の近くにマイグレーションできるようなネットワークを構築する。**
  - まるで手元にデータあるような感覚で移動できるとベスト。
- **可能な限りデータを圧縮する。**
  - 例えば Google Cloud Storage[1] の standard storage を利用すると、100 TB → 80 TB で 326,094 円/月 → 260,875.2 円/月 (-65,218.8 円/月)[2]
- **データを仕分けする。**
  - Hot（高頻度でアクセス）, warm（低頻度）, cold（アーカイブ）の三段階？
  - アクセス頻度に応じて料金が変わるクラウドストレージサービスが多いので、仕けが大事。



[1] 2023/6/21 14:17 時点で $0.023/GB/月
[2] 2023/6/21 14:17 時点で 141.78 円/ドル

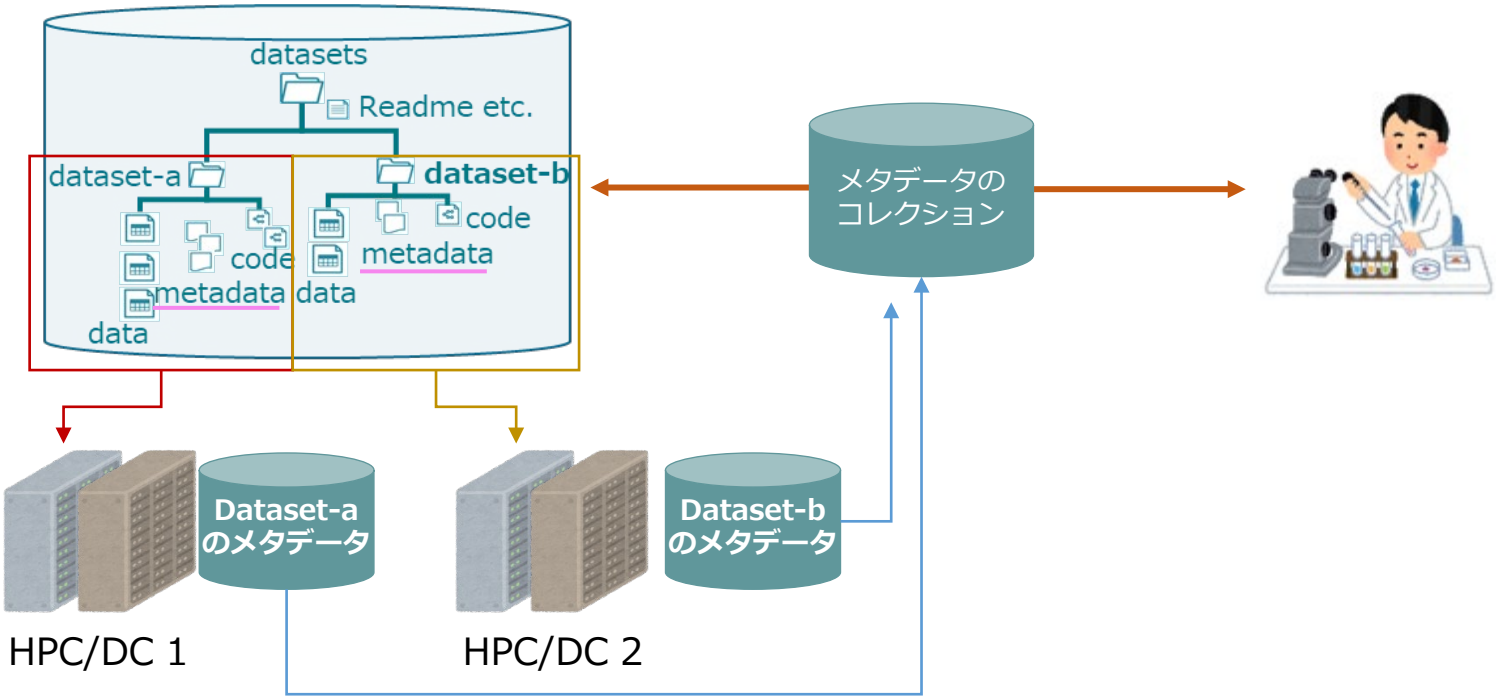
②データ管理方法について

※実現のためのコストは度外視。

プロジェクト内の（ビッグ）データが複数の置き場所に分かれることがよくある。これらのデータを一元管理できるとよいが、、

↓

- **データに適切なメタデータ（データを説明するデータ）を付ける。**
- **メタデータを用いて対象データの位置を把握できるような機能・サービスを各 HPC/DC に設ける。**
- **上記機能・サービスすべてにアクセスできるサービスを設ける。**



③データの利活用について

※実現のためのコストは度外視。

チーム内外でのデータの利活用を促進するには？
→ 日頃からデータを FAIR[1] な状態にする。

To be Findable（見つけられるために）
- （メタ）データが、グローバルに一意で永続的な識別子（ID）を有すること。
- データがメタデータによって十分に記述されていること。
- （メタ）データが検索可能なリソースとして、登録もしくはインデックス化されていること。
- メタデータが、データの識別子（ID）を明記していること。

To be Accessible（アクセスできるために）
- 標準化された通信プロトコルを使って、（メタ）データを識別子（ID）により入手できること。
  - そのプロトコルは公開されており、無料で、実装に制限が無いこと。
  - そのプロトコルは必要な場合は、認証や権限付与の方法を提供できること。
- データが利用不可能となったとしても、メタデータにはアクセスできること。

To be Interoperable（相互運用できるために）
- （メタ）データの知識表現のため、形式が定まっていて、到達可能であり、共有されていて、広く適用可能な記述言語を使うこと。
- （メタ）データがFAIR原則に従う語彙を使っていること。
- （メタ）データは、他の（メタ）データへの特定可能な参照情報を含んでいること。

To be Re-usable（再利用できるために）
- （メタ）データが、正確な関連属性を豊富に持つこと。
  - （メタ）データが、明確でアクセス可能なデータ利用ライセンスと共に公開されていること。
  - （メタ）データが、その来歴と繋がっていること。
  - （メタ）データが、分野ごとのコミュニティの標準を満たすこと。

[1] https://biosciencedbc.jp/about-us/report/fair-data-principle/

データガバナンス機能の位置づけ

① データの置き場
- 誰が提供するのか？機関が調達？
- 共同利用施設やスパコンを利用？
- クラウドを借りるのが良いのか？
- そもそもデータを移動できるのか？
- どのくらいの期間置けるのか？
  - 例えば SACLA HPC ではポリシー上、最初の 3 年以内は hot access 可能なストレージに保存、その後 5 年間は tape storage に保存[1]。
- 増え続けるデータをどう管理するのか？

② データ管理方法
- データを一元的に管理できるのか？

③ データ利活用
- チーム内で共有し、使えるようにするにはどうすればよい？
- 公開したデータをほかの研究者が参照・利用しやすくするためにはどうすればよい？

- HPC またはクラウド環境のすぐ近くにビッグデータを置く。
- HPC 同士を接続するネットワークを構築・増強する。
- データを圧縮・仕分けする。
- データに適切なメタデータを付ける。
- そのメタデータに基づきデータの位置を把握できる機能を設ける。
- データを FAIR にする。

適切にメタデータがついているかどうか、データが FAIR であるかどうか不安…

→ **データガバナンス機能がこの検証（と環境利用）をサポートする。**

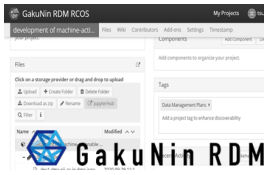
NII RDC におけるデータガバナンス機能の構想と現状

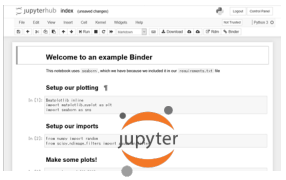
NII RDC におけるデータガバナンス機能の構想

所属機関のポリシーの下で DMP に沿って研究者が望む形で研究データ管理（RDM）を実践（データガバナンス）することを機械的に支援

プラン作成支援 → プロジェクト自動構築 → 解析環境自動連携 → データ管理支援 → データ公開支援 → 組織のDMP活用支援

DMP作成機能





研究機関の戦略立案支援
- データのインパクトを評価
- 共同研究の可視化
- データ人材の発掘

データガバナンス機能

非公開 → 公開

Planning / Experiment / Collaboration / Analysis / Writing / Curation / Publication / Preservation / Discovery / Idea

データ管理基盤 | データ公開基盤 | データ検索基盤

研究データの状態を自動検証する（モニタリング）
＋ DMP に基づき研究データ基盤を orchestrate する（リサーチフロー）

ガバナンスシート（構想）

研究者がデータガバナンスを実践するためのルールセット。

ルール

FAIR principles[1]

To be Findable:

F1. (meta)data are assigned a globally unique and eternally persistent identifier.
F2. data are described with rich metadata.
F3. (meta)data are registered or indexed in a searchable resource.
F4. metadata specify the data identifier.

To be Accessible:

A1 (meta)data are retrievable by their identifier using a standardized communications protocol.
A1.1 the protocol is open, free, and universally implementable.
A1.2 the protocol allows for an authentication and authorization procedure, where necessary.
A2 metadata are accessible, even when the data are no longer available.

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles.
I3. (meta)data include qualified references to other (meta)data.

To be Re-usable:

R1. (meta)data have a plurality of accurate and relevant attributes.
R1.1. (meta)data are released with a clear and accessible data usage license.
R1.2. (meta)data are associated with their provenance.
R1.3. (meta)data meet domain-relevant community standards.

データ管理計画（DMP）

研究データポリシー

研究データの metadata schema と検証ルール※

Governance rule list

FAIR-F
FAIR-A
FAIR-I
FAIR-R
License
DMP-x
Policy-x
⋮

各種ルールを機械処理可能な形で定義。

Select

Governance sheet

Rules:
~~FAIR-F~~
~~FAIR-A~~
FAIR-I
FAIR-R
License
DMP-x
Policy-x
⋮

研究者が選択したガバナンスルールのリスト

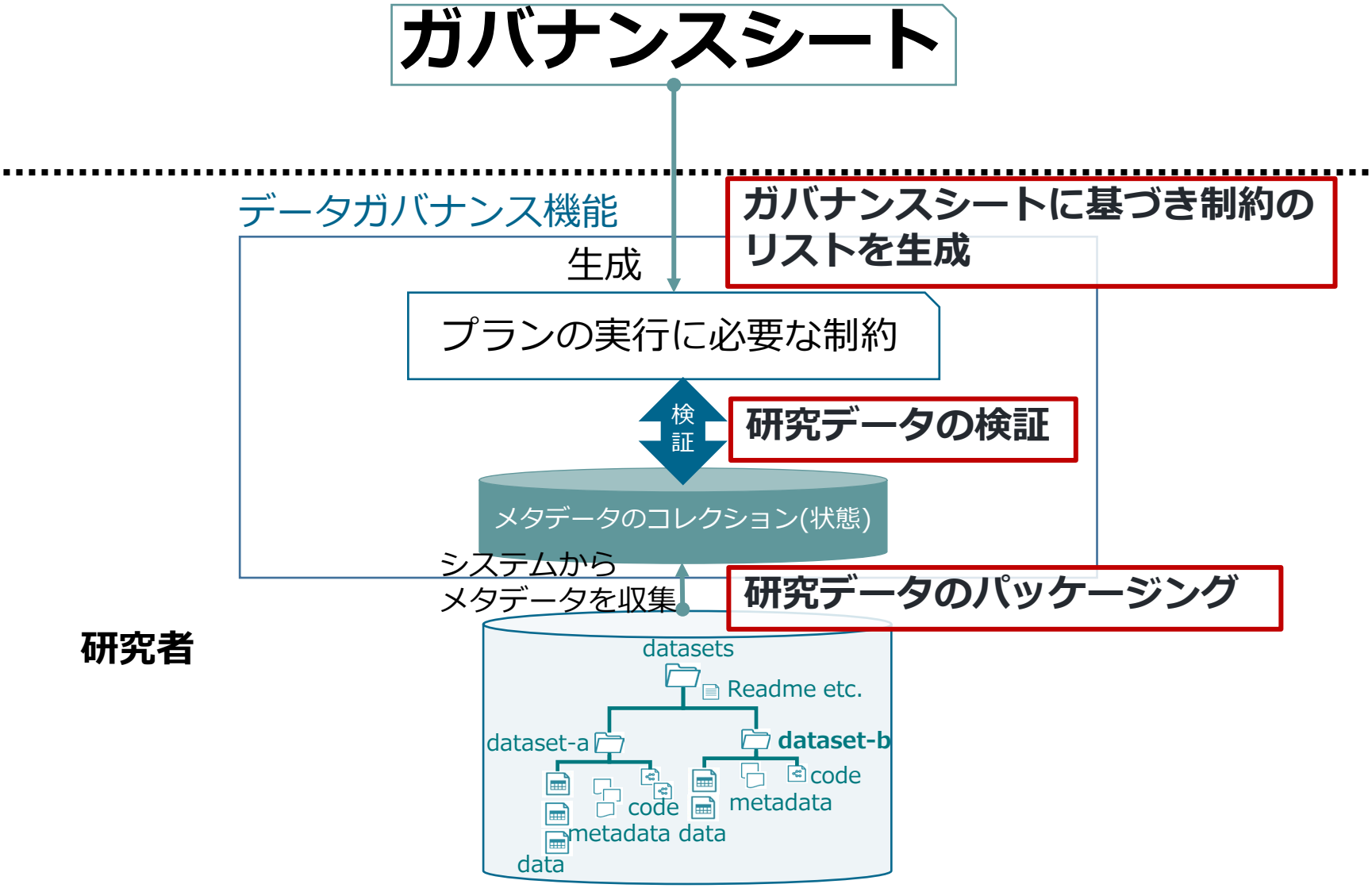
- **大学・研究機関が自身の研究データポリシーに基づくガバナンスルールを自由に定義可能とする。**
- **研究者（場合によっては大学・研究機関も）がプロジェクトごとにガバナンスルールを自由に定義可能とする。**

※分野ごとの metadata schema と検証ルールのユニーク性は、それぞれのコミュニティで拡張する形で実現可能。

[1] https://force11.org/info/the-fair-data-principles/

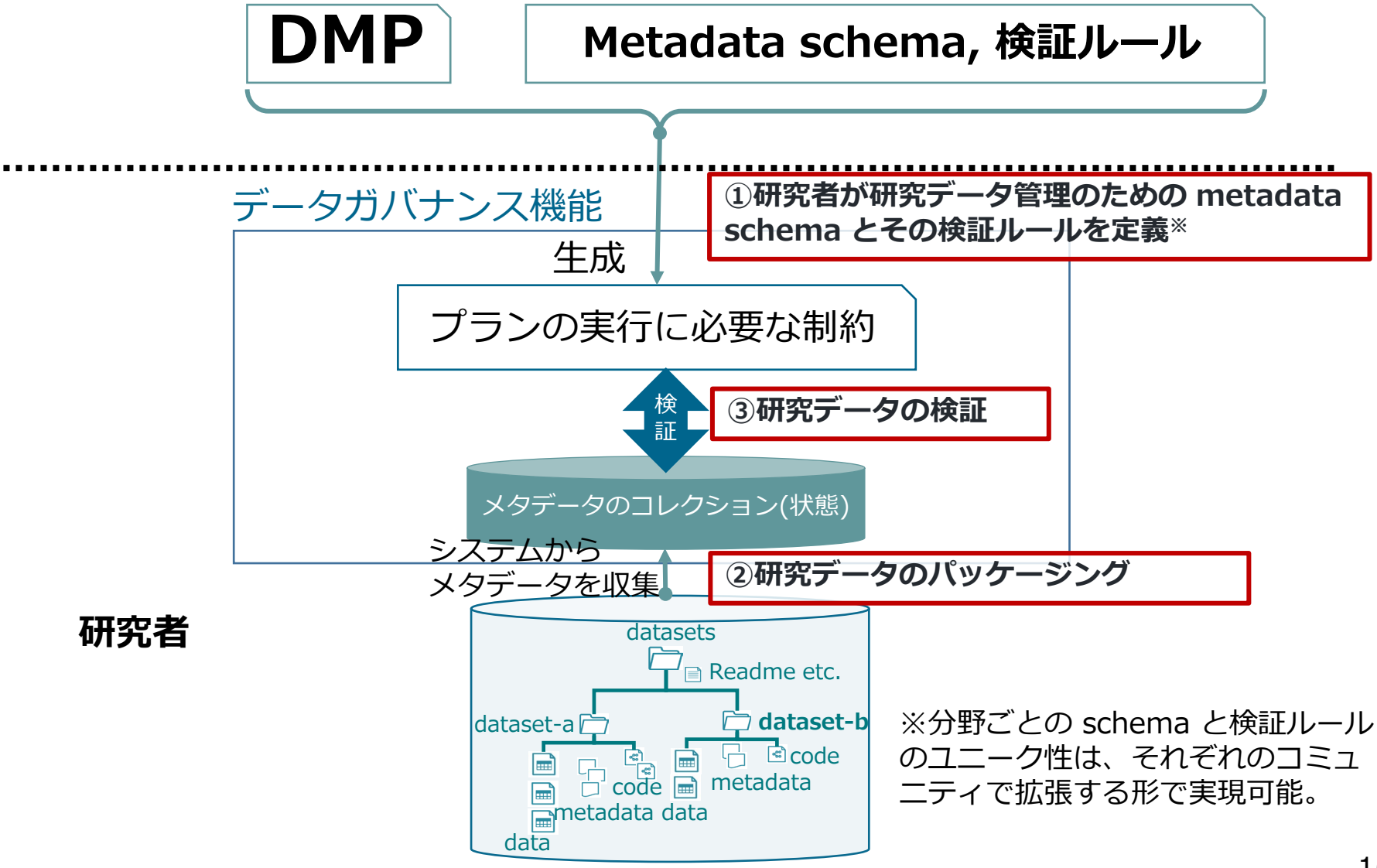
モニタリング（構想）

研究データの状態がガバナンスシートで指定された制約を満たしているかどうか検証する



モニタリング（現状）

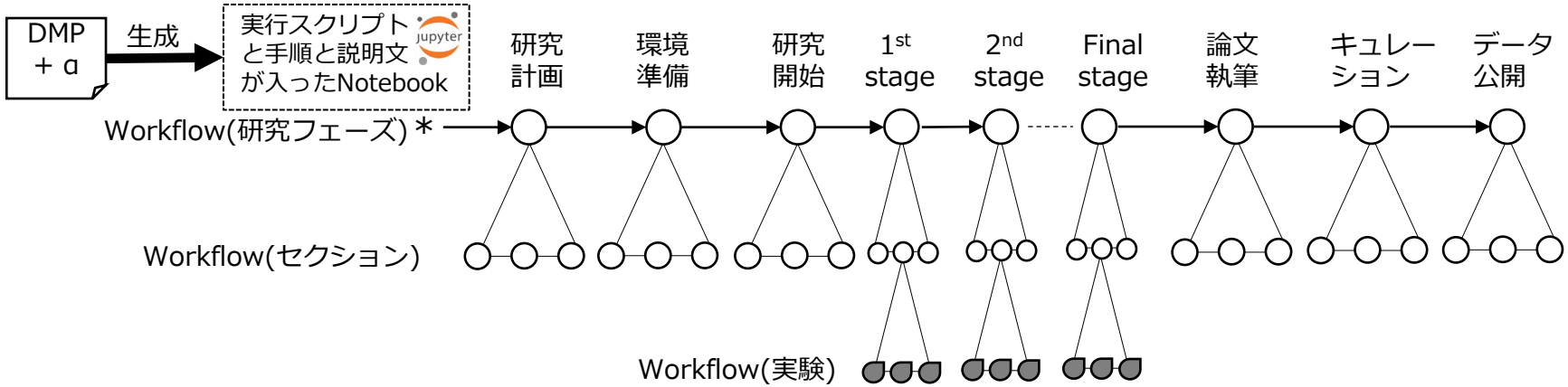
研究データの状態が DMP 等による制約を満たしているかどうか検証する



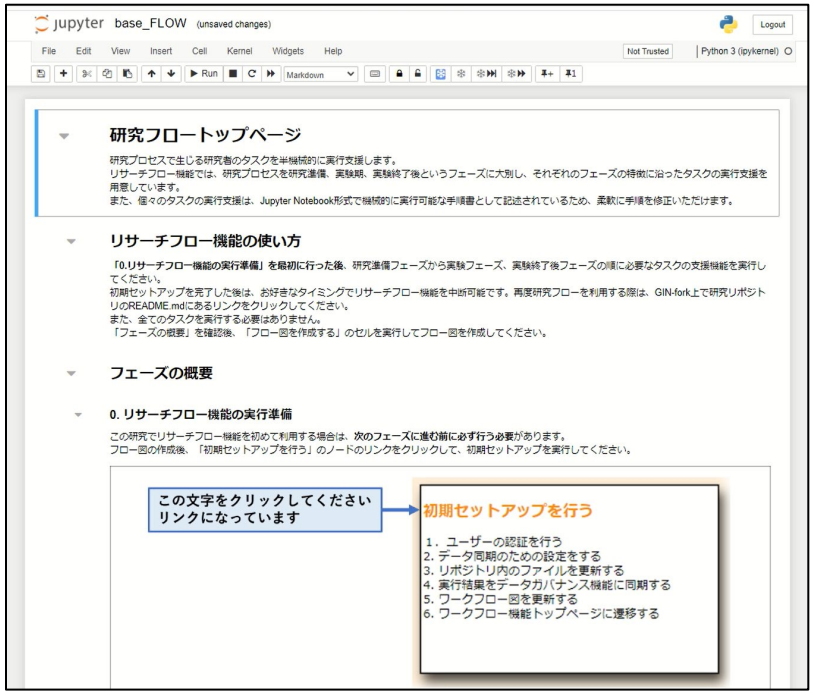
※分野ごとの schema と検証ルールのユニーク性は、それぞれのコミュニティで拡張する形で実現可能。

リサーチフロー（現状）

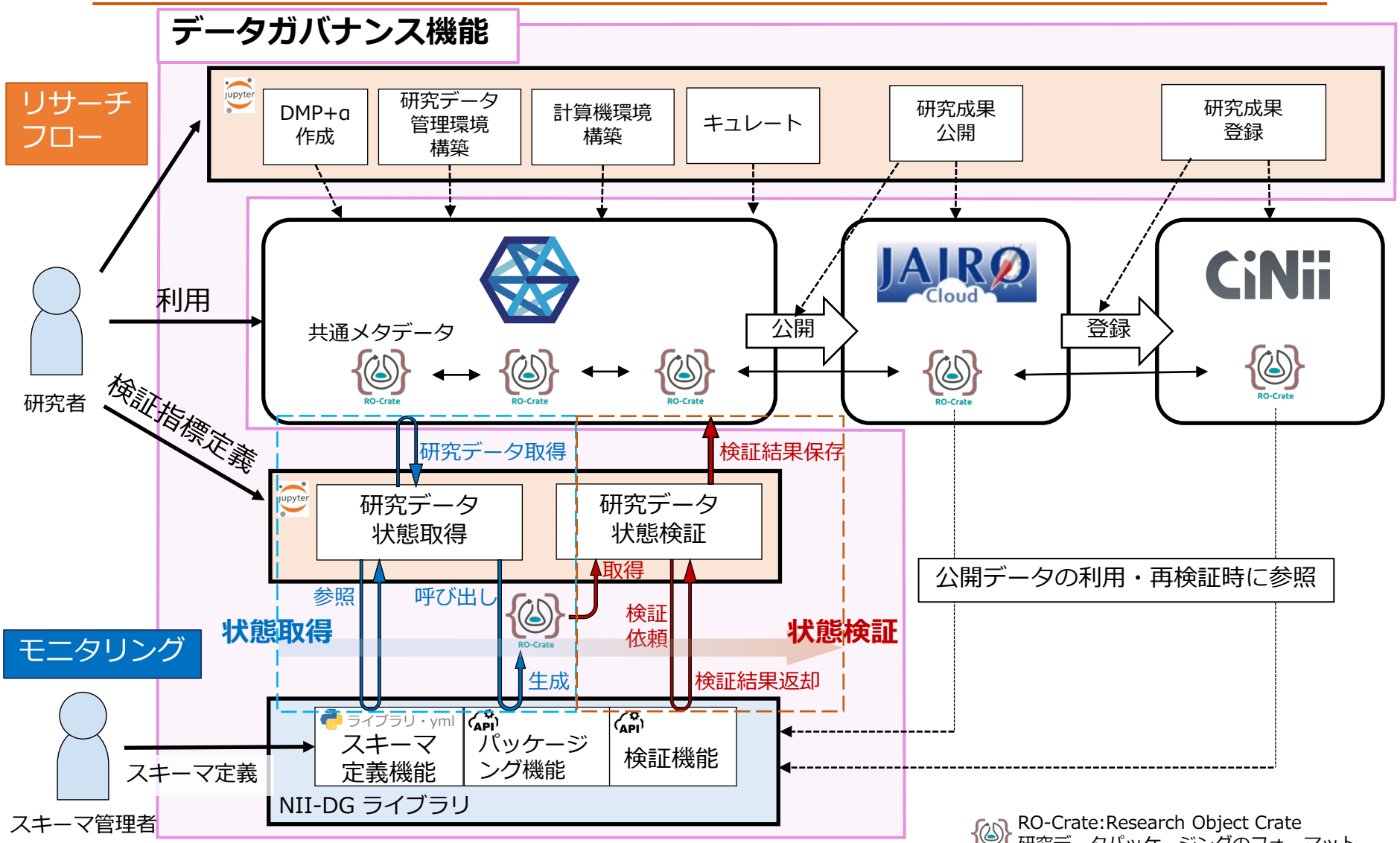
DMP 等によって表現された計画に基づき生成された、実行可能な手順書



リサーチフロー
（Notebook）の例



データガバナンス機能の動作イメージ



RO-Crate:Research Object Crate
研究データパッケージングのフォーマット規格のひとつ

まとめ

アイデアとデータガバナンス機能の方向性

(JUST my opinion)

課題	解決に向けたアイデア
① **データの置き場** • 誰が提供するのか？機関が調達？ • 共同利用施設やスパコンを利用？ • クラウドを借りるのが良いのか？ • そもそもデータを移動できるのか？ • どのくらいの期間置けるのか？ • 例えば SACLA HPC ではポリシー上、最初の 3 年以内は hot access 可能なストレージに保存、その後 5 年間は tape storage に保存[1]。 • 増え続けるデータをどう管理するのか？	• HPC またはクラウド環境のすぐ近くにビッグデータを置く。 • HPC 同士を接続するネットワークを構築・増強する。 • データを圧縮・仕分けする。
② **データ管理方法** • データを一元的に管理できるのか？	• データに適切なメタデータを付ける。 • そのメタデータに基づきデータの位置を把握できる機能を設ける。
③ **データ利活用** • チーム内で共有し、使えるようにするにはどうすればよい？ • 公開したデータをほかの研究者が参照・利用しやすくするためにはどうすればよい？	• データを FAIR にする。

データガバナンス機能がメタデータの付与状況やデータの FAIRness の検証、研究環境の構築を支援する（将来的にはメタデータ付与関連も？）。また、研究者に作成・提出が求められている DMP に沿った研究データ管理を実践することも支援する。

データガバナンス機能の機能評価試験版サービスの利用案内

NII RCOS
Research Center for Open Science and Data Platform

GakuNin RDM におけるデータガバナンス機能の機能評価試験版サービスの提供を 2023/6/19 より開始しました。

先行ユーザーからのフィードバックを受けてデータガバナンス機能の改善を実施し、実証実験レベルへのブラッシュアップを計画しております。

詳しくは GakuNin RDM のサポートポータル（https://support.rdm.nii.ac.jp/）の「お知らせ」をご確認いただくか、以下の「問い合わせ先」までご連絡ください。

- 提供予定期間：　2023/6/19～2024/3/31
- 問い合わせ先：　データガバナンス機能サポート
  dg_support@nii.ac.jp
- 問い合わせ時に必要な情報：　氏名、所属、連絡用メールアドレス、利用希望者リスト、参加希望理由