NII RCOS
Research Center for Open Science and Data Platform

NII RDCの データガバナンス機能について

2022年12月7日

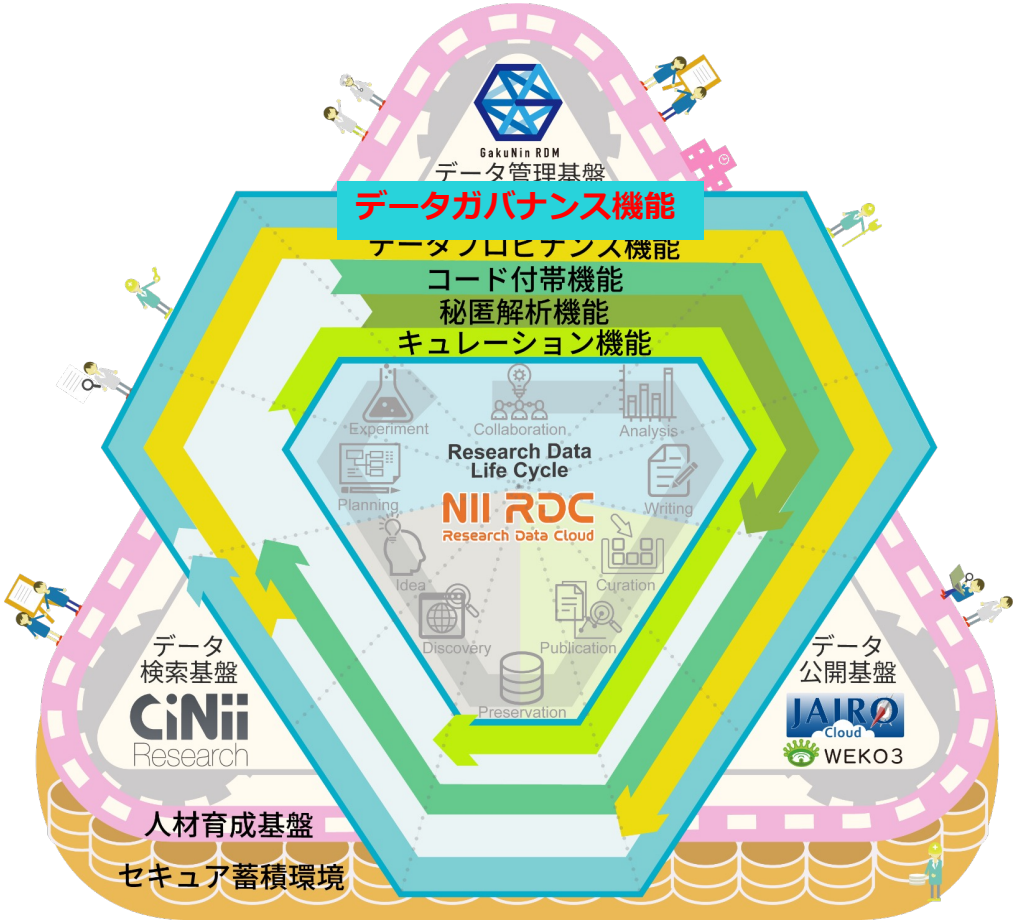
国立情報学研究所
オープンサイエンス基盤研究センター

横山重俊

1. データガバナンス機能とは

データガバナンス機能の位置付け (1/2)

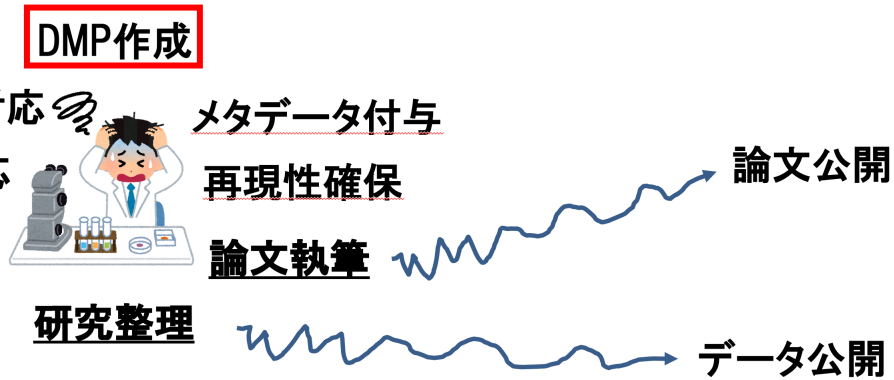
研究データ管理品質を向上させるために
ワークフロー、モニタリング、研究記録機能を提供



データガバナンス機能の位置付け (1/2)

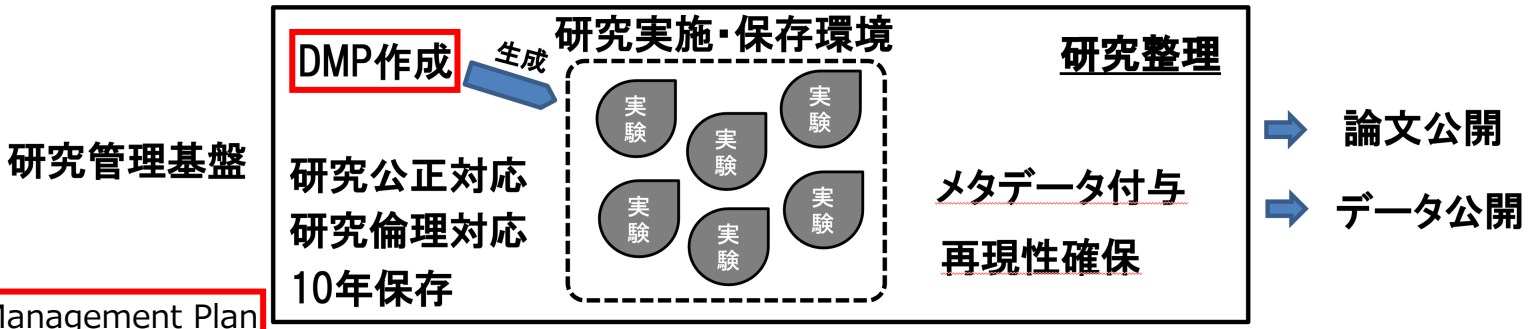
AS IS

研究者の
努力に依存



TO BE

研究管理
基盤による
支援



DMP: Data Management Plan

データガバナンス機能の位置付け (2/2)

2022年度 NII オープンフォーラムにおける
九州大学 附属図書館副館長・理系図書館長, データ駆動イノベーション推進本部 研究データ管理支援部門長,
大学院システム情報科学研究院　冨浦 洋一教授
ご講演資料「九州大学における研究データガバナンス構築に向けて―九大のRDMとNIIデータガバナンス機能への期待」より

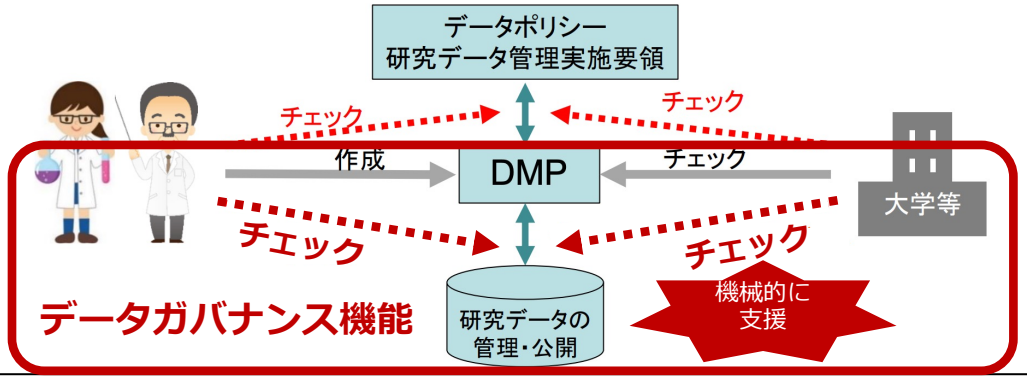
https://www.nii.ac.jp/openforum/upload/f2c67eb99b1013be0fc9a007440442edcd5bf06c.pdf (p.5)

KYUSHU UNIVERSITY

データガバナンス

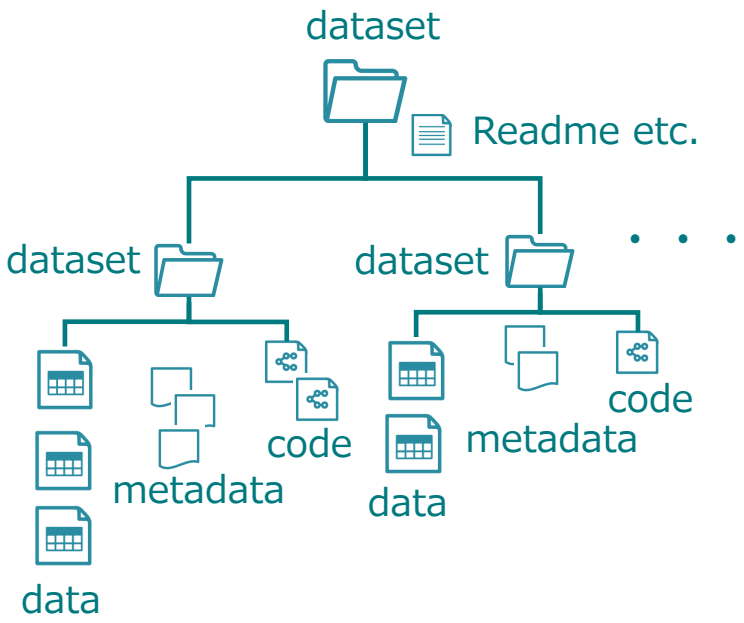
5

- 研究データの管理・公開等の取り扱いに関する組織の考え方を記述したものが**データポリシー**.
  そして，研究データの具体的な取り扱い方法を記述したものが**研究データ管理実施要領**.
- 研究プロジェクトXの**DMP**は，そのプロジェクトにおける研究データの生成・収集および研究データの管理・公開等の取り扱いの計画を研究者自身が記述したもの．したがって，DMPがデータポリシーや研究データ管理実施要領に準拠しており，研究者がDMPに記述したとおりのデータの取り扱いを行うならば，データガバナンスが保証される.

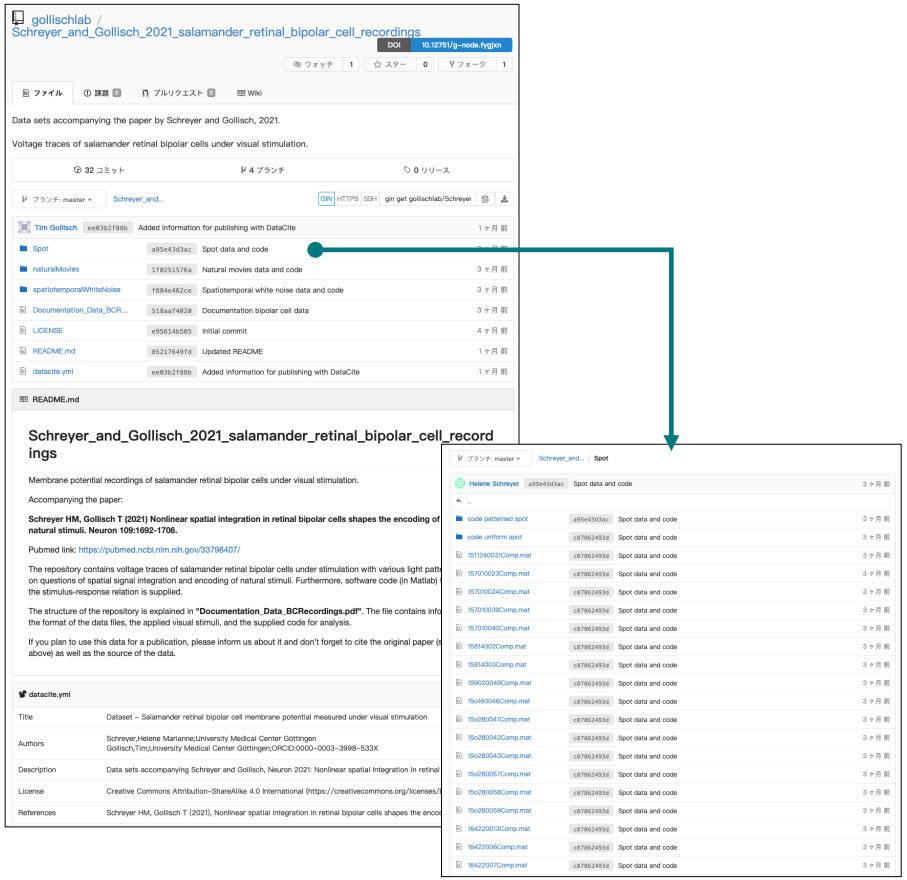


研究データ

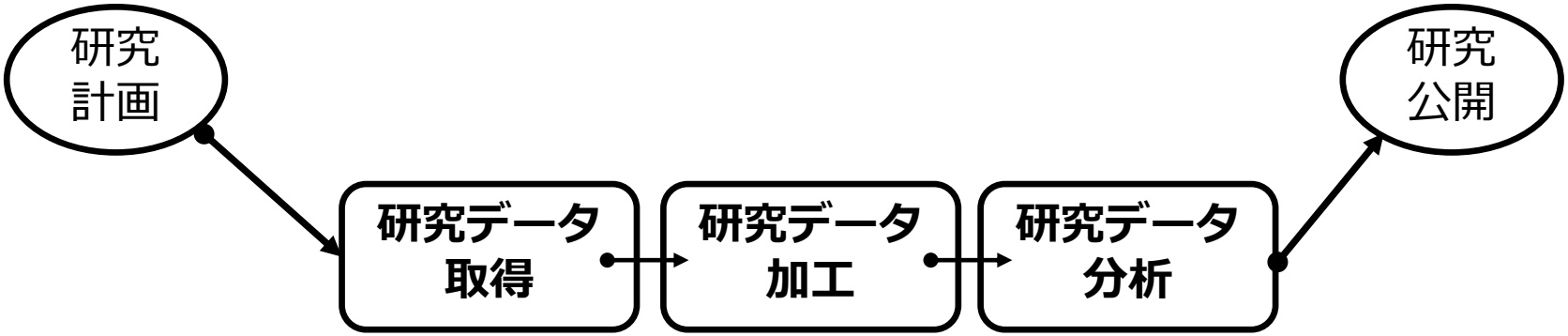
研究データの構成

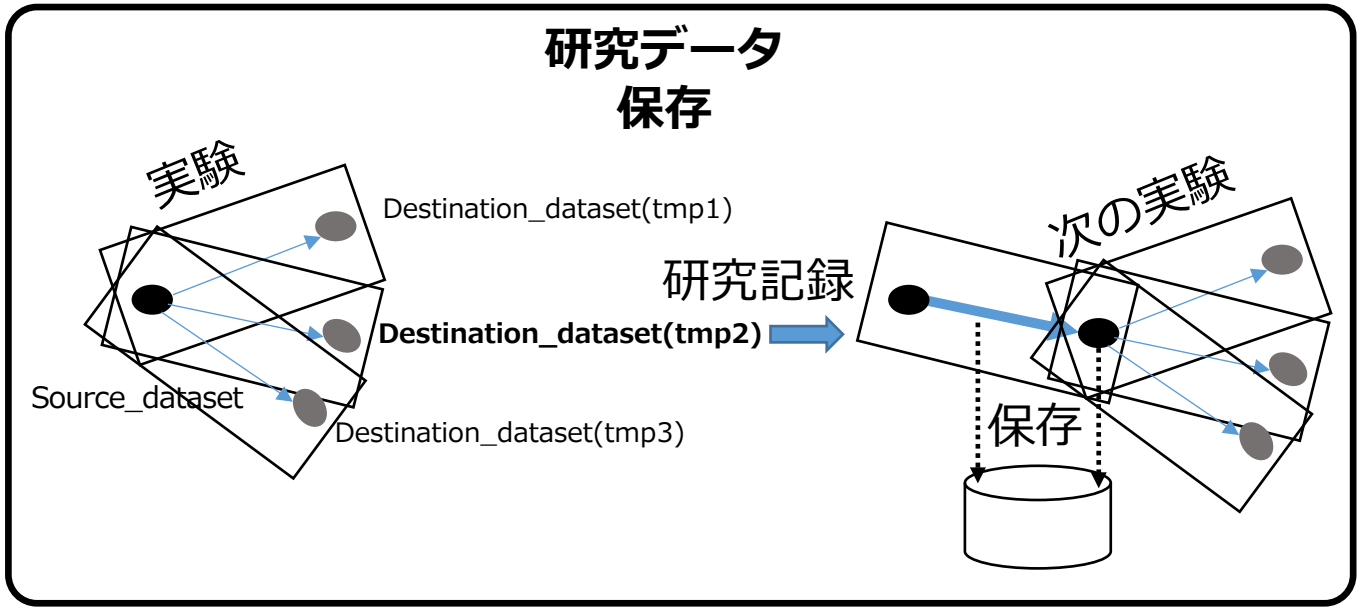


研究データの例



研究データ管理


研究
計画
研究データ
取得
研究データ
加工
研究データ
分析
研究
公開


研究データ
保存
実験
Destination_dataset(tmp1)
研究記録
Destination_dataset(tmp2)
Source_dataset
Destination_dataset(tmp3)
次の実験
保存

研究データ管理 と データガバナンスの関係

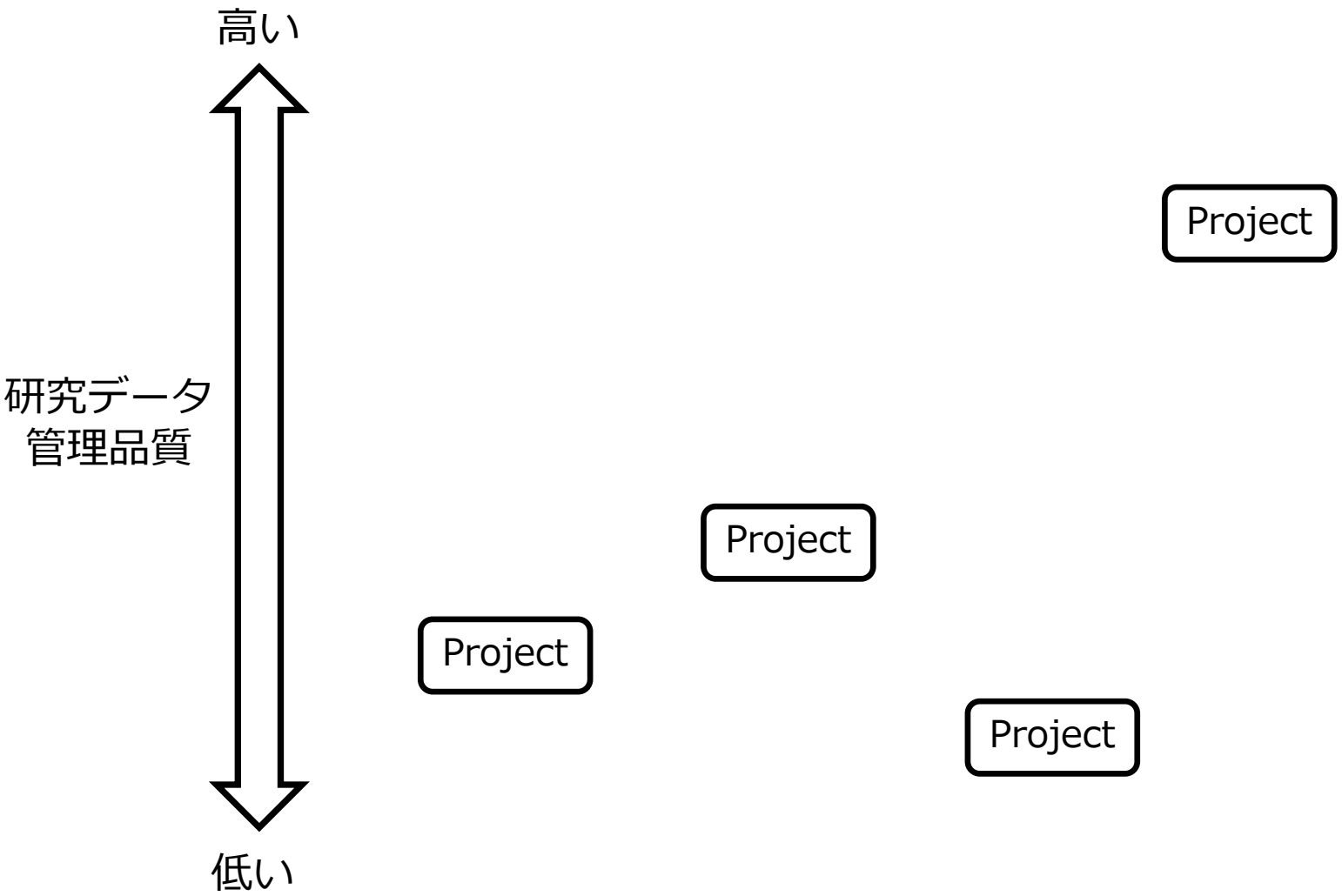
データガバナンス　：　研究データ管理品質を守る規律に従う行為

研究データ
管理品質

研究データ管理　：　研究データの取得・加工・分析・保存・公開などの行為

仮説：
研究者や研究グループ自身が，データガバナンスを，
日頃から行うことが，研究データ管理品質の向上に大きく寄与する.
（ガバナンスは監視ではない）

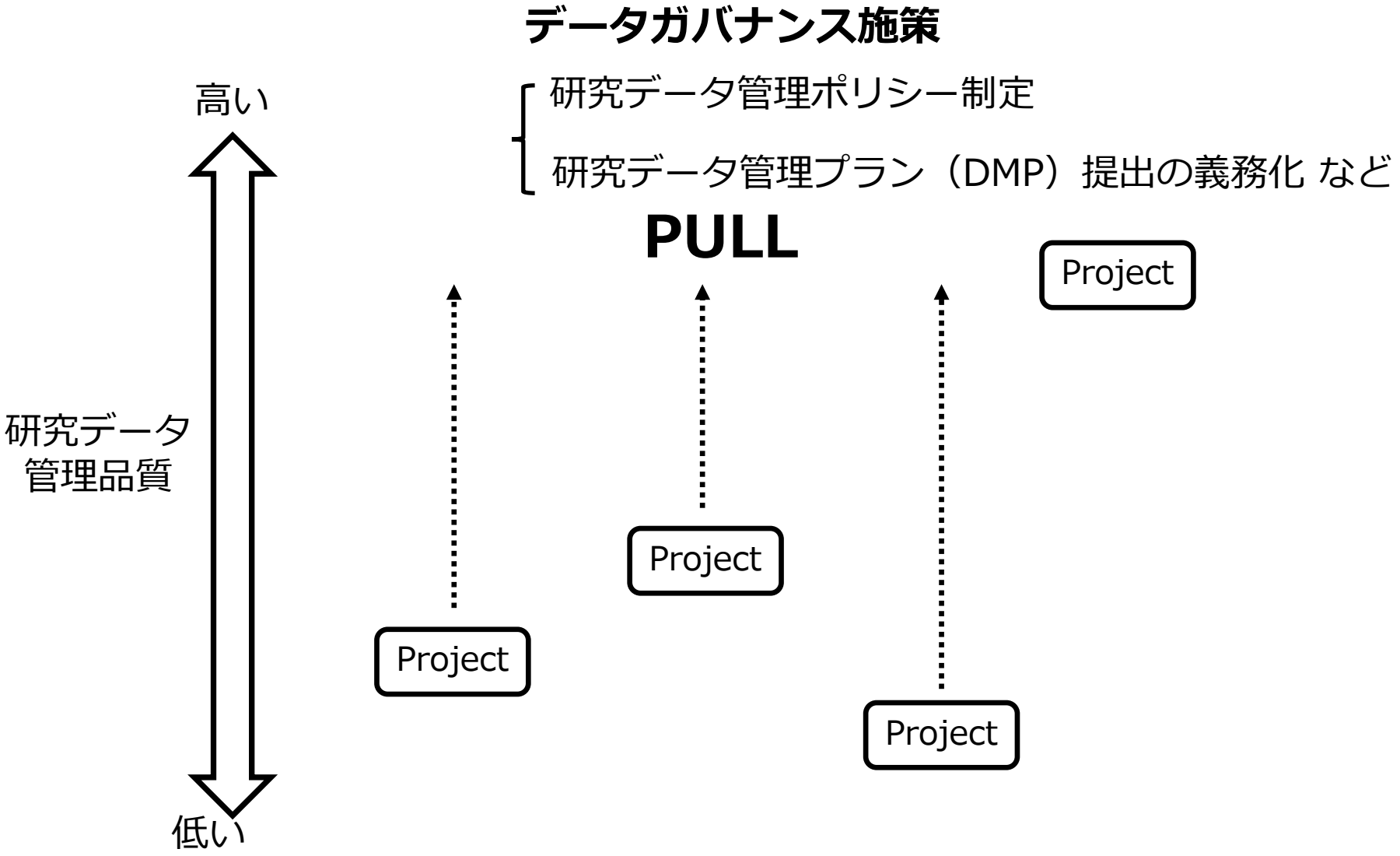
研究データ管理品質のばらつき



研究データ管理の品質が悪いと

- そもそも必要な時に，どこにあるのか分からない.
- 本当にあるのかすら，分からない.
- ある場所が分かっても ， 実際には手に入らない.
- 手に入っても，それをどうやって扱うのか分からない.
- データの由来が不明なので信用できない.
- 保護しなければならないデータを ，漏洩させてしまう .
- 整理が悪くて，とても公開までに手間がかかる.

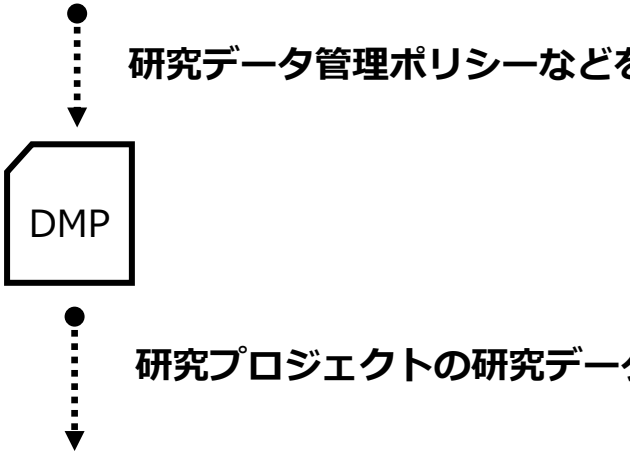
施策の展開による研究データ管理品質向上



データマネジメントプラン (DMP)

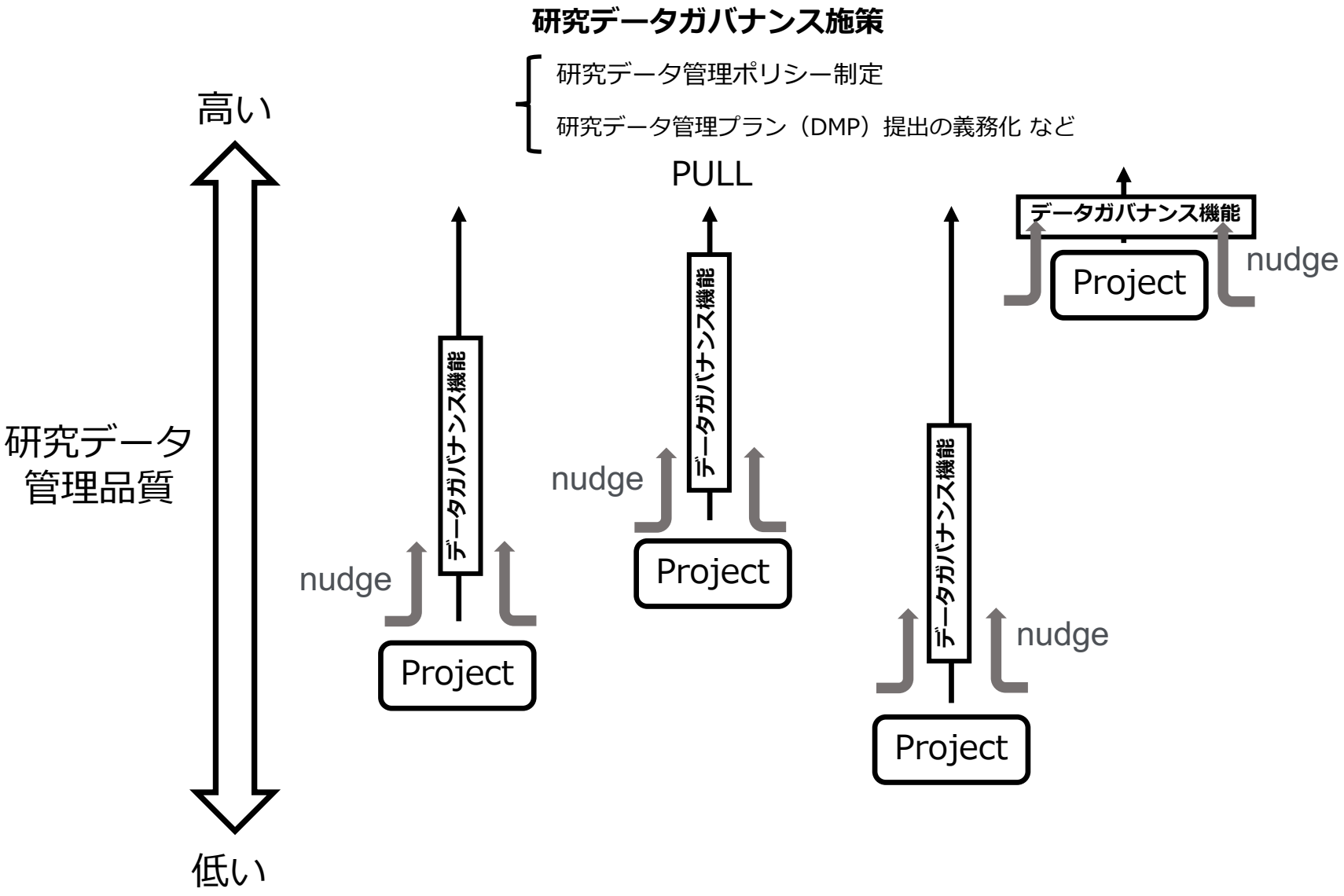
データマネージメントプラン(DMP)は，研究のために収集・作成する研究データの取扱いや整備・保存・公開についての計画を定めた**文書**である.
DMPの記述内容は提出先の助成機関によって異なるけれど，データ管理者，データの説明，データの種類，データ量，公開レベルなどが記述される.

研究データガバナンス　：　研究データ管理品質を守る規律に従う行為



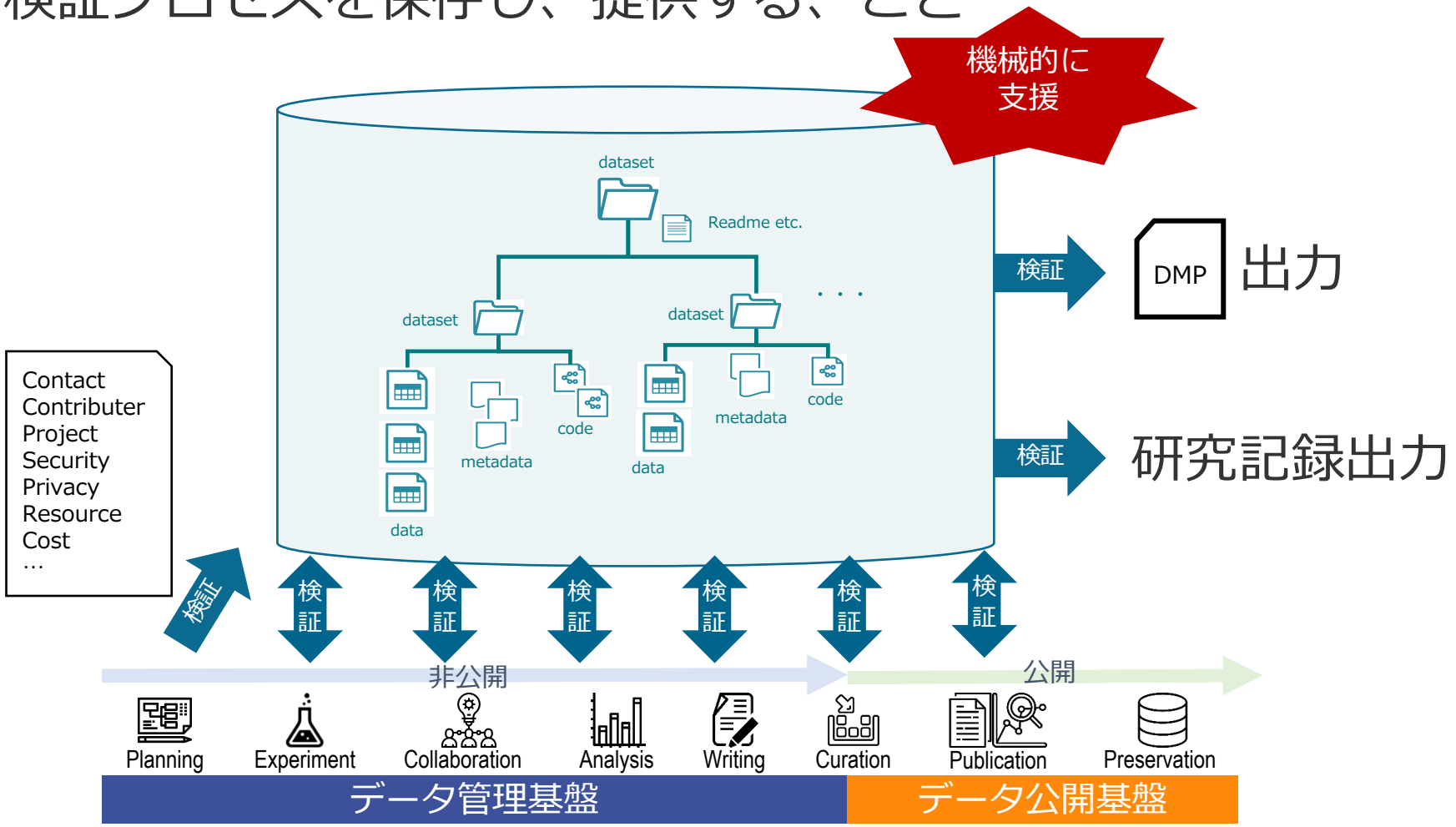
研究データ管理　：　研究データの取得・加工・分析・保存・公開などの行為

データガバナンス機能の役割



「データガバナンスする。」とは

ステート（＝メタデータ・データセット）の変化を検証し、
検証プロセスを保存し、提供する、こと



DMPとデータガバナンス機能の関係

ポリシー ↓ プラン ↓ データ管理実施

研究者

DMPで管理側とコミュニケーション
（ データガバナンス機能を使って
DMPに沿ったデータ管理を実施）

DMP

研究管理者　　FA

DMPで研究者とコミュニケーション
（研究データ管理の実態との整合性は
データガバナンス機能が保証していると期待）

データガバナンス機能

生成

プランの実行に必要な制約

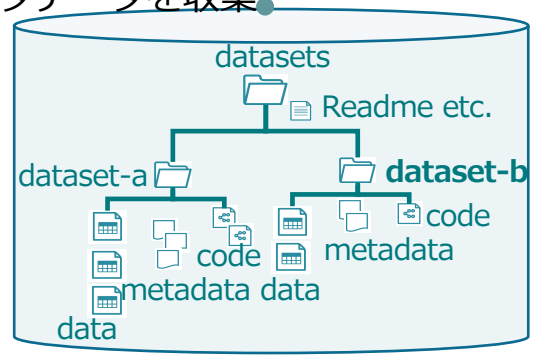
検証

メタデータのコレクション(状態)

共通スキーマで表現

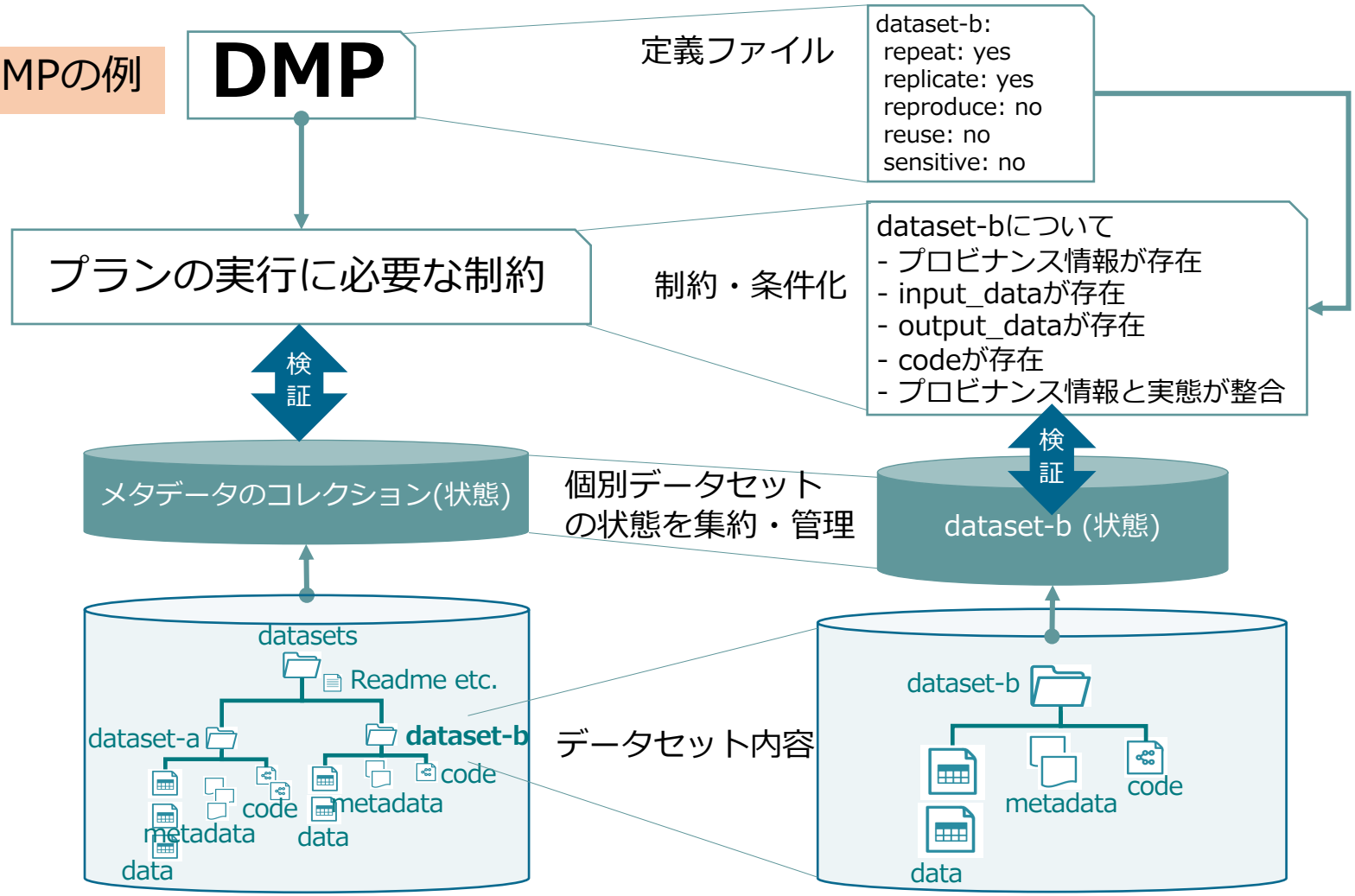
システムから
メタデータを収集

研究者



データガバナンス機能の利用例

実際のDMPの例



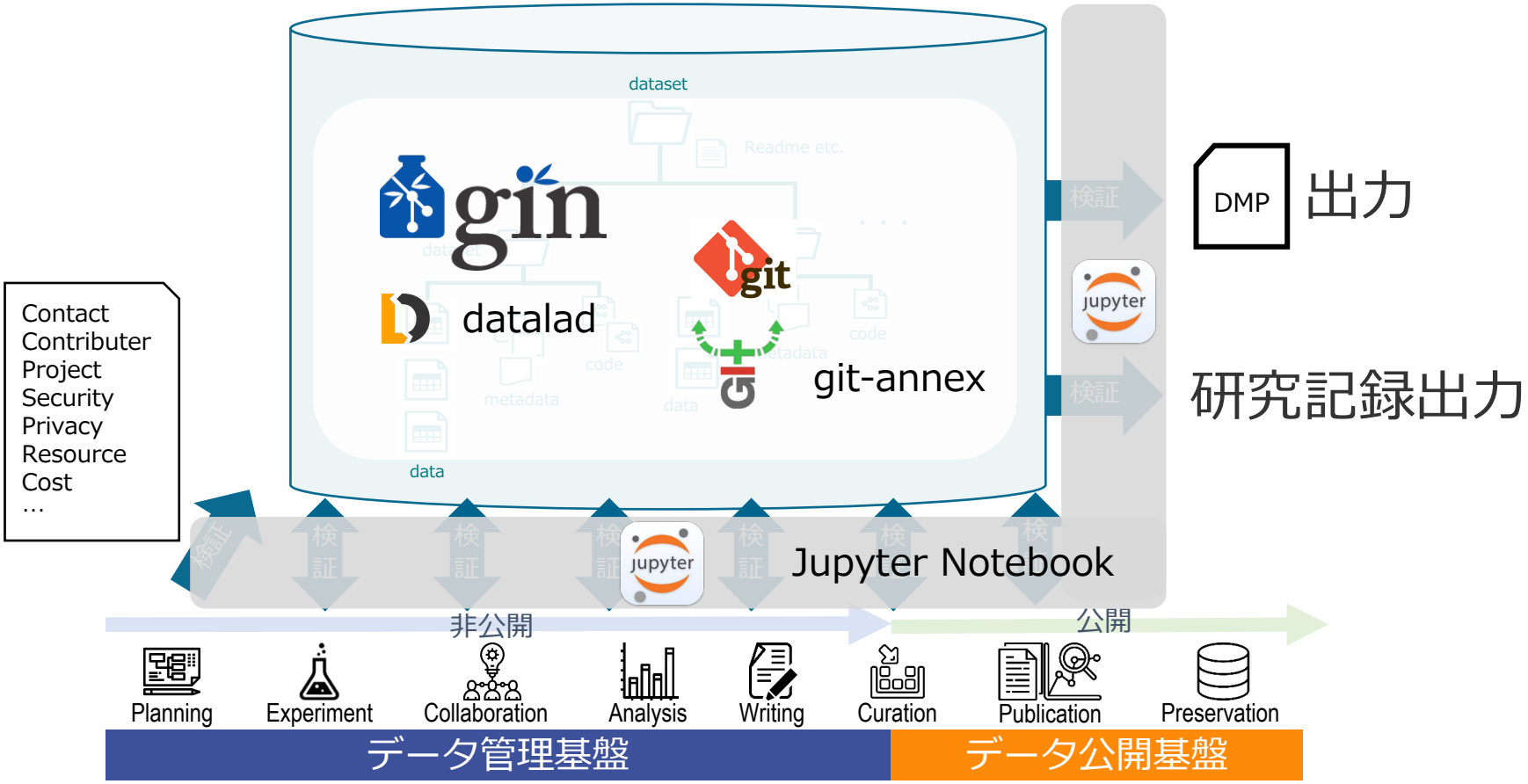
DMPに基づく確実なデータ管理の実行

プロビナンス機能からも情報を取得してメタデータの一つとして保持

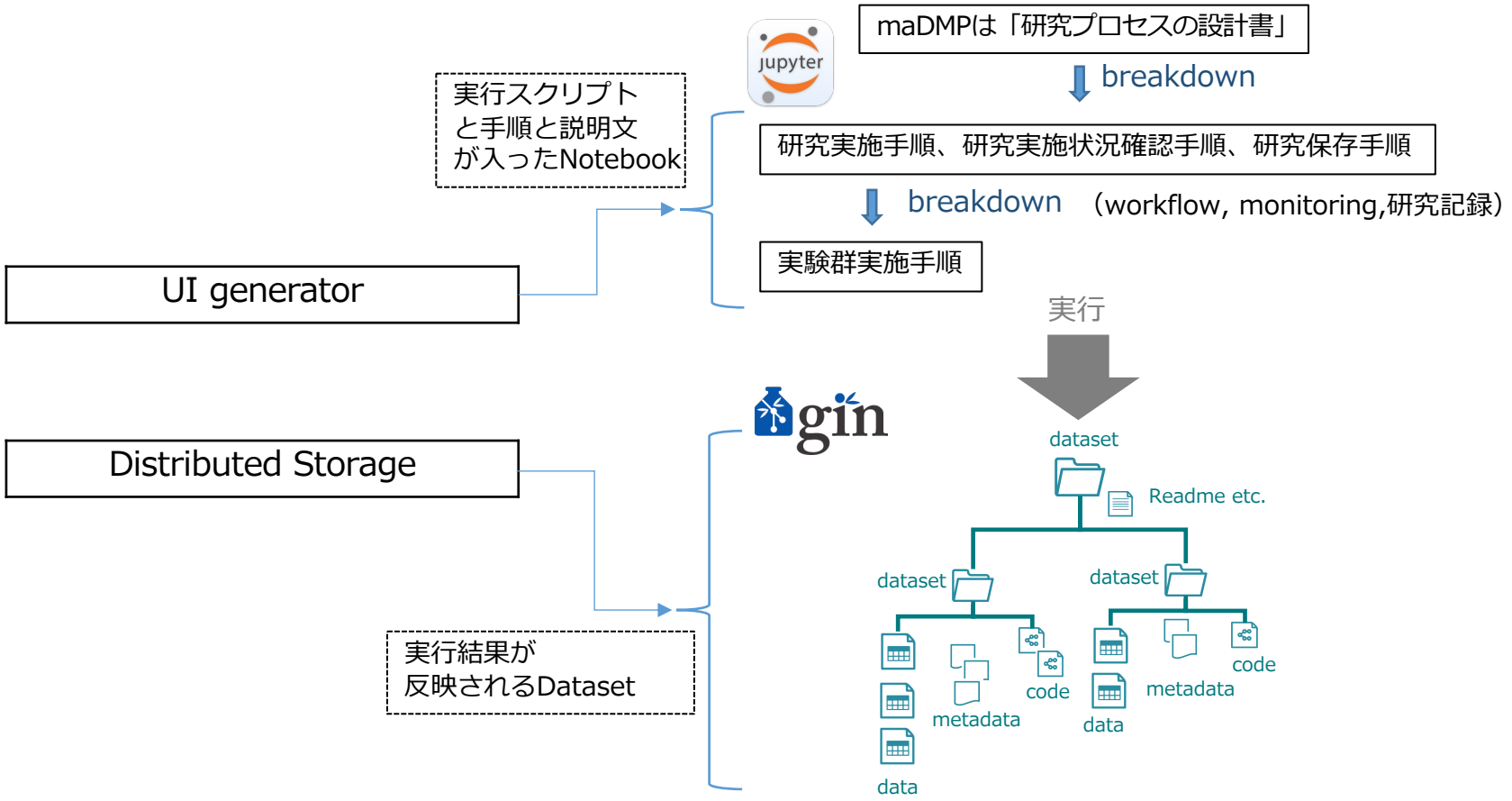
2. データガバナンス機能プロトタイプシステム

プロトタイプシステムのアーキテクチャ (1/2)

ステート（＝メタデータ・データセット）の保存をgit, git-annexを使ったgin/dataladというオープンソースプロダクトで実装、検証プロセスをJupyter Notebookで実現

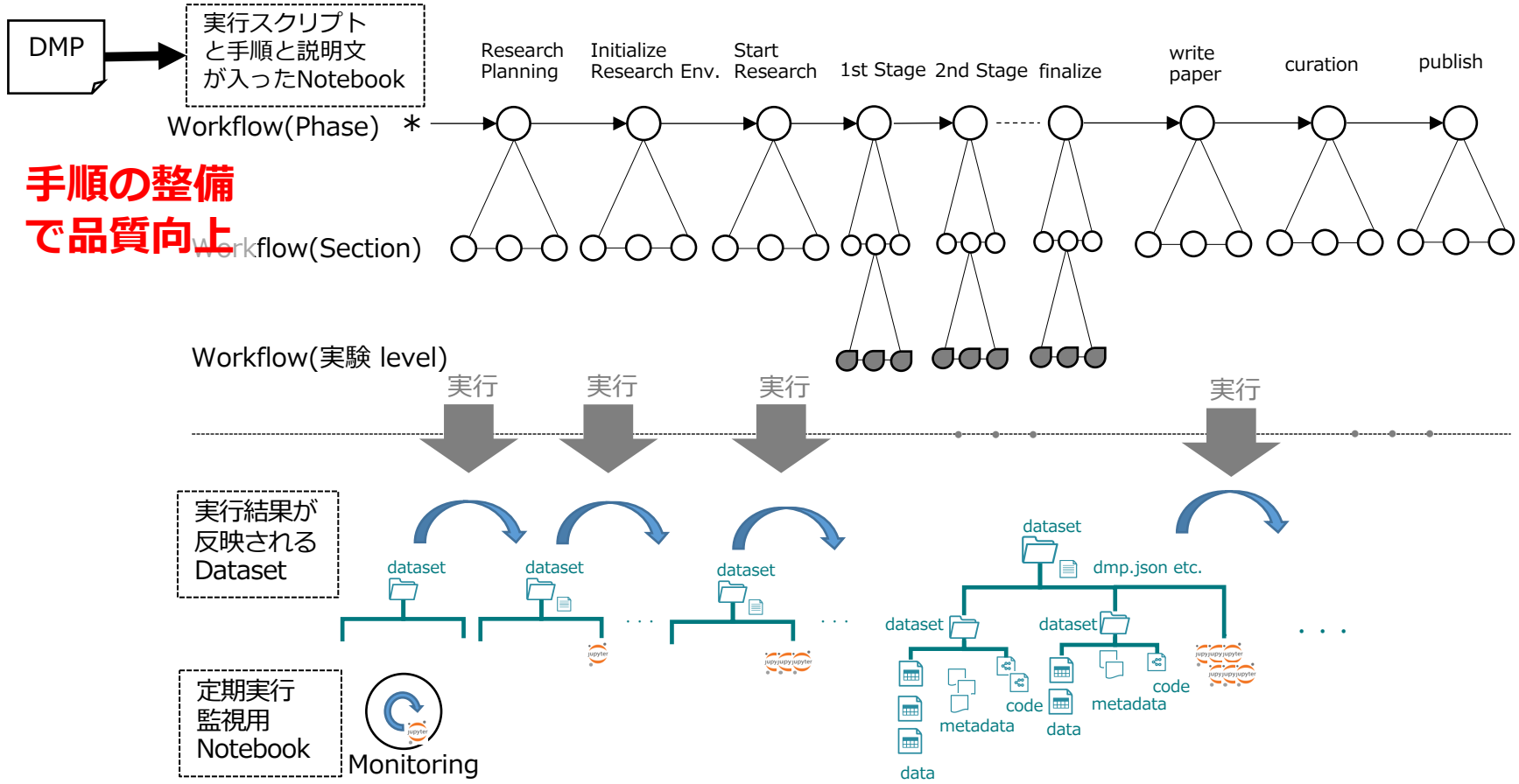


プロトタイプシステムのアーキテクチャ (2/2)



プロトタイプシステムの機能

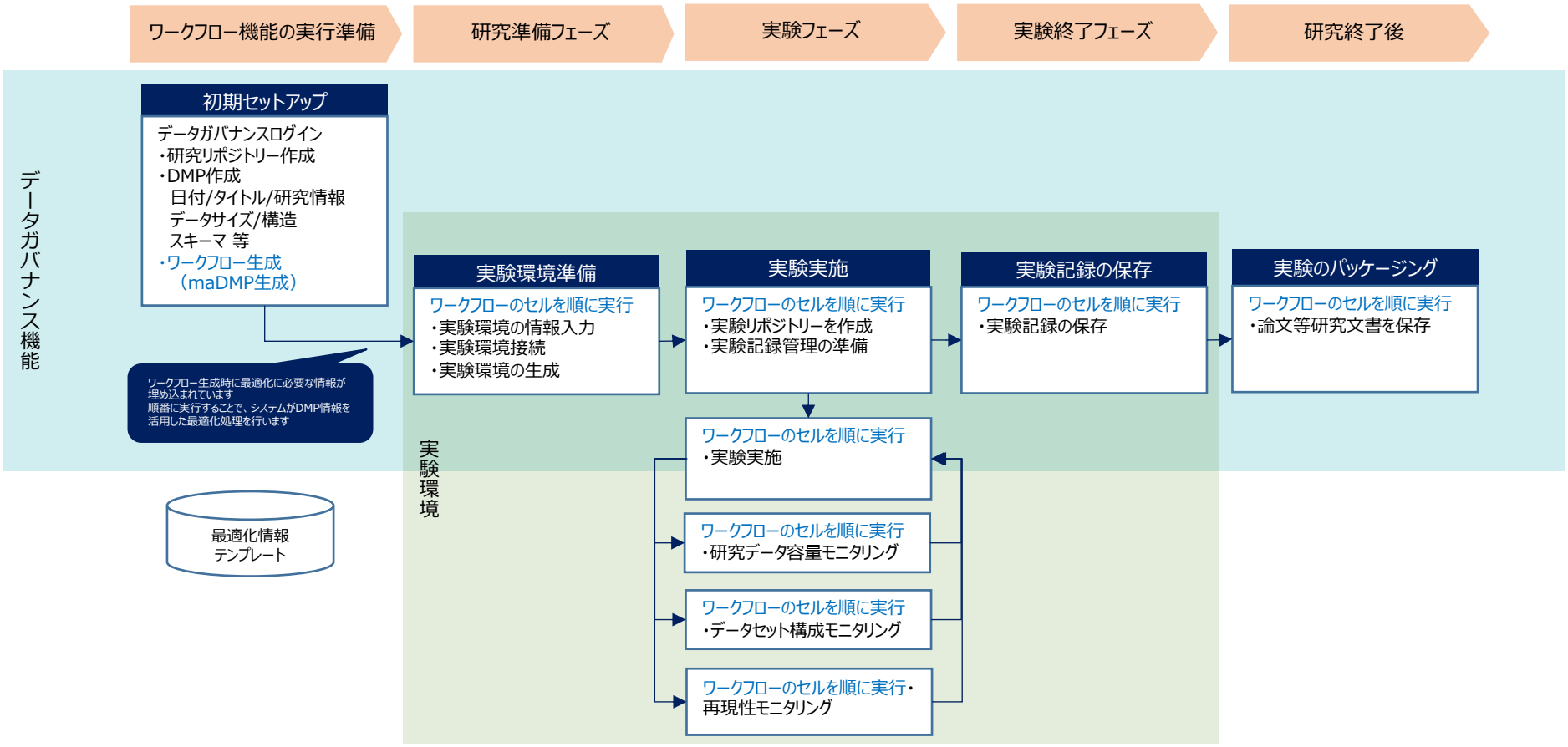
データガバナンスを実施する実行可能な手順書(Notebook)をDMPから生成



手順の整備で品質向上

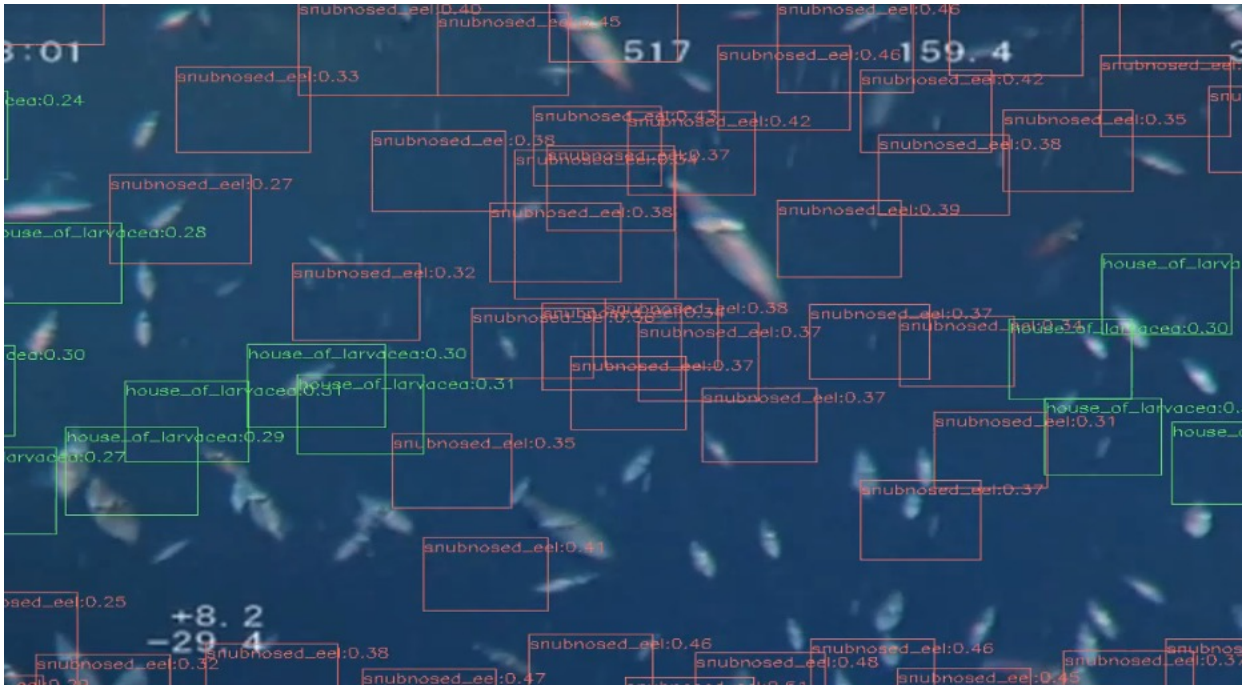
継続監視で品質向上

研究活動支援範囲


ワークフロー機能の実行準備
研究準備フェーズ
実験フェーズ
実験終了フェーズ
研究終了後
データガバナンス機能
初期セットアップ
データガバナンスログイン
・研究リポジトリー作成
・DMP作成
日付/タイトル/研究情報
データサイズ/構造
スキーマ 等
・ワークフロー生成
（maDMP生成）
ワークフロー生成時に最適化に必要な情報が埋め込まれています
順番に実行することで、システムがDMP情報を活用した最適化処理を行います
最適化情報
テンプレート
実験環境
実験環境準備
ワークフローのセルを順に実行
・実験環境の情報入力
・実験環境接続
・実験環境の生成
実験実施
ワークフローのセルを順に実行
・実験リポジトリーを作成
・実験記録管理の準備
実験記録の保存
ワークフローのセルを順に実行
・実験記録の保存
実験のパッケージング
ワークフローのセルを順に実行
・論文等研究文書を保存
ワークフローのセルを順に実行
・実験実施
ワークフローのセルを順に実行
・研究データ容量モニタリング
ワークフローのセルを順に実行
・データセット構成モニタリング
ワークフローのセルを順に実行・
再現性モニタリング

実験適用例

底生生物の検知／追跡

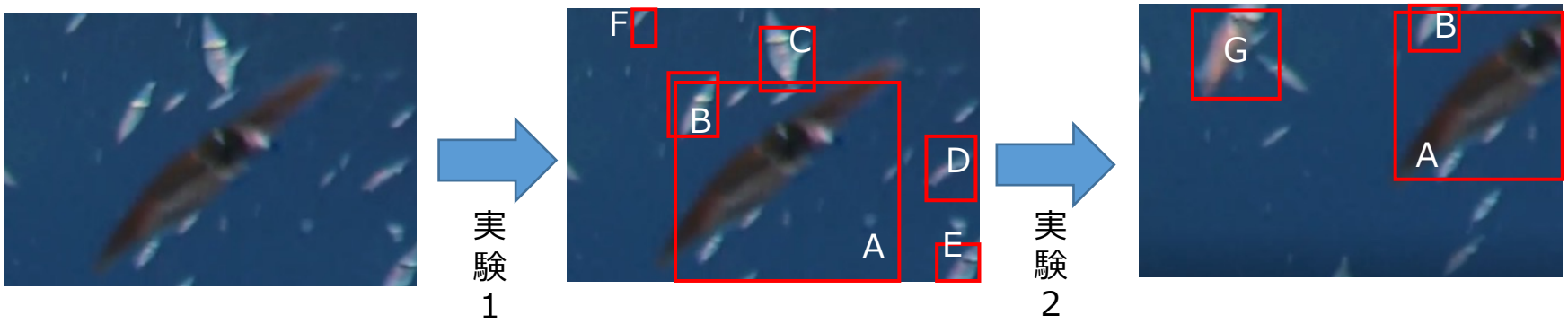


データ提供：海洋研究開発機構(JAMSTEC)

実験適用例の概要

海中を映した動画中に含まれる、底生生物を検知／追跡する。

- 実験1 物体検出
  動画の各フレームで生物を検知する検知器を作成・評価する。
- 実験2 物体検出＋追跡
  動画のあるフレームで検出した生物が、
  次フレームでどこに移動したかを追跡するAIを作成し、
  追跡性能を評価する。



データ提供：海洋研究開発機構(JAMSTEC)

実験１「detr(物体検出の学習と評価)」の説明

実験概要、目的

概要：機械学習による物体検出の学習とその評価の出力。
目的：良い物体検出モデルを作成する。

実験構成

以下を構築

input_data
- configs　設定ファイル（実験のパラメータ）
- data　機械学習の教師データと評価用データ（実験の入力データ）
- weight　機械学習の初期重み（実験の入力データ）

source
- train.py　学習用プログラム
- eval.py　評価用プログラム

実験の実行

main.ipynbを実行（環境構築、学習用プログラム実行）

↓

約10分後、以下のフォルダに結果が格納される。

output_data
- hogehoge_20220309183628　モデルと学習曲線

評価値を取得して、以下を確認する

・モデルと学習曲線（モデルの性能が良くなってるか（数値））

↓

モデルと学習曲線のフォルダパスをconfigに適用して、
eval.pyを実行し、推論結果の評価値と推論動画を得る。
そして、モデルの性能が良くなっているかを目視確認する。

実験リポジトリ

```
├─input_data
│  ├─configs
│  │      hogehoge.yaml
│  │
│  ├─data
│  │  └─11175505
│  │      ├─annotations
│  │      │    instances_train.json
│  │      │    instances_val.json
│  │      │
│  │      ├─images
│  │      │    frame_000000.png
│  │      │    frame_000010.png
│  │      │    ・・・
│  │      │    frame_001440.png
│  │      │
│  │      └─video_and_score_text
│  │          │  11195506.mov
│  │          └─11195506
│  │              frame_000000.txt
│  │              frame_000001.txt
│  │              ・・・
│  │              frame_001754.txt
│  │
│  └─weight
│      detr-r50-e632da11.pth
│
├─output_data
│  │
│  ├─hogehoge_20220309183628
│  │      │  hogehoge_valloss_min.pth
│  │      │
│  │      └─2022_0309_1836
│  │          [train.pyによる評価データ（学習曲線）]
│  │
│  └─inference
│      └─hogehoge_20220309183628
│              [eval.pyによる評価データ（推論の評価値と推論動画）]
│
└─source
   │  eval.py
   │  train.py
   │  main.ipynb
   ├─detr
   └─Object-Detection-Metrics
```

実験２「DeepSORT」＋「ByteTrack」の説明

実験概要、目的

概要：物体の検出＋追跡と、その評価の出力。
　　実験１の出力（検出量の重み）を使って行う。
目的：良い追跡パラメータを見つけ、追跡する。

実験構成

以下を構築

input_data

config	設定ファイル（実験のパラメータ）
video	処理対象の動画と評価データ（実験の入力データ）
weight	検出用の重み（実験の入力データ）

source

inference_detr+tracker.py	追跡用プログラム
main.ipynb	実行用プログラム

実験の実行

main.ipynbを実行（環境構築、検出＋追跡用プログラム実行）

↓

約１時間後、以下のフォルダに結果が格納される。

output_data

result.txt	計算ログ
result_metric.txt	評価値
video.mp4	検出枠＋追跡枠を動画に書き込んだもの

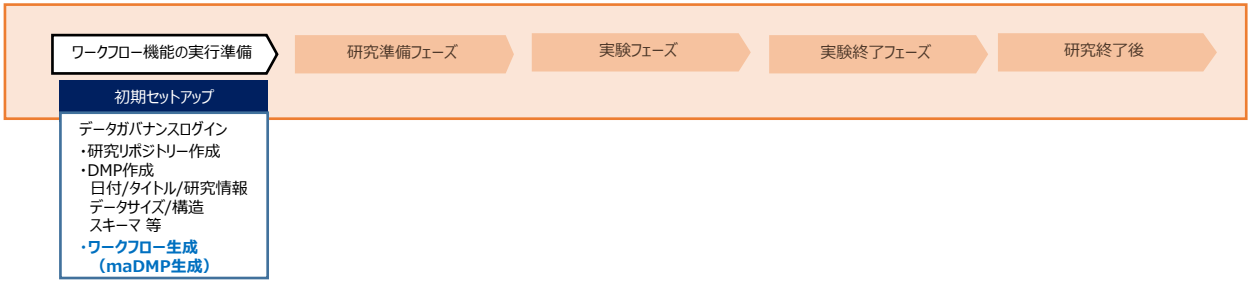
評価値を取得して、以下を確認する

・推論結果の評価値と推論動画（追跡の性能がよくなってるか）

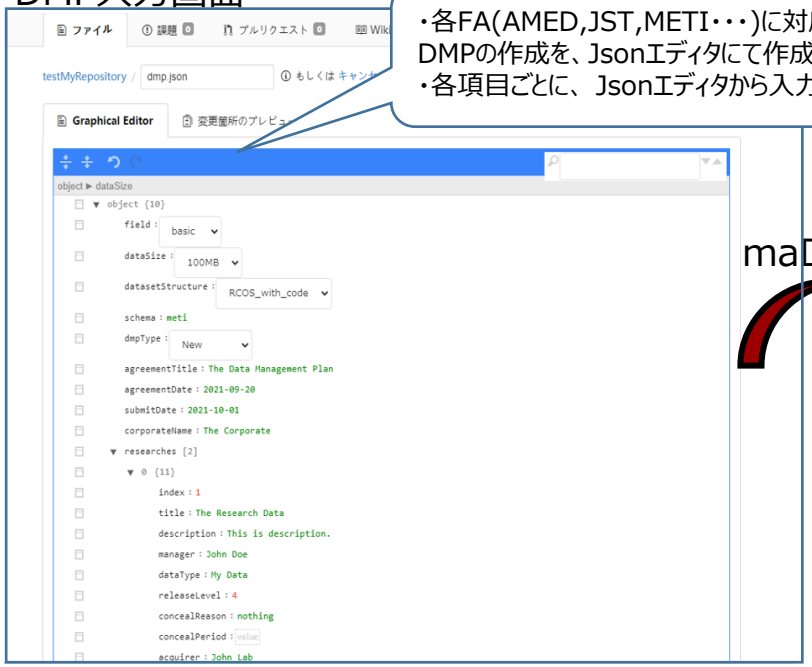
実験リポジトリ

```
├─input_data
│  ├─config
│  │    colors.txt
│  │    hogehoge.yaml
│  │    labels.txt
│  │
│  ├─video
│  │  │  11195506.mov
│  │  └─11195506
│  │        frame_000000.txt
│  │          ・・・
│  │
│  └─weight
│       ckpt.t7
│       hogehoge_valloss_min.pth（実験1の出力）
│
├─output_data
│    result.txt
│    result_metric.txt
│    video.mp4
│
└─source
   │  inference_detr+tracker.py
   │  main.ipynb
   ├─detection
   └─tracking
```

①DMPの作成および、maDMPの生成

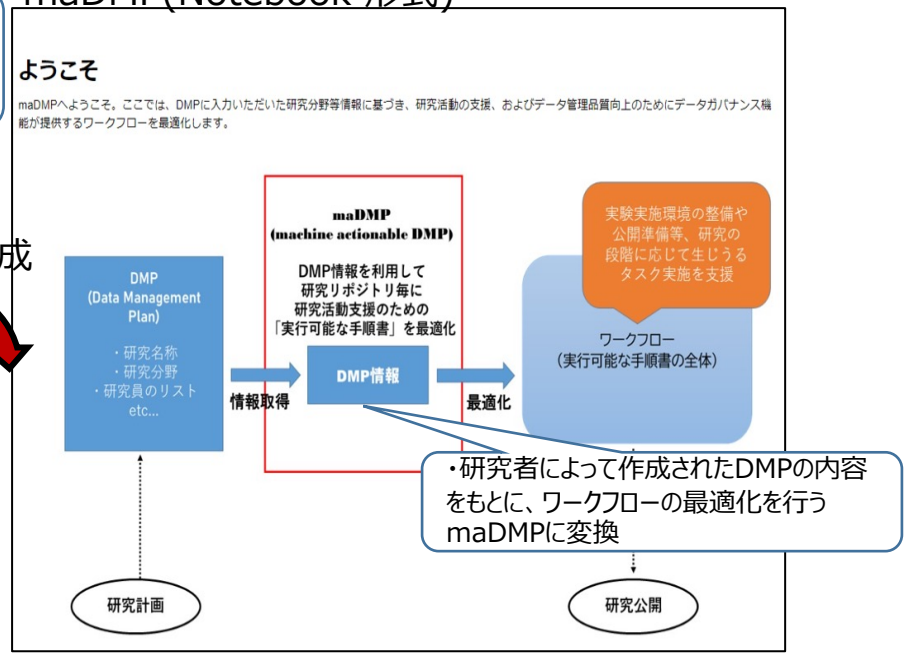


DMP入力画面

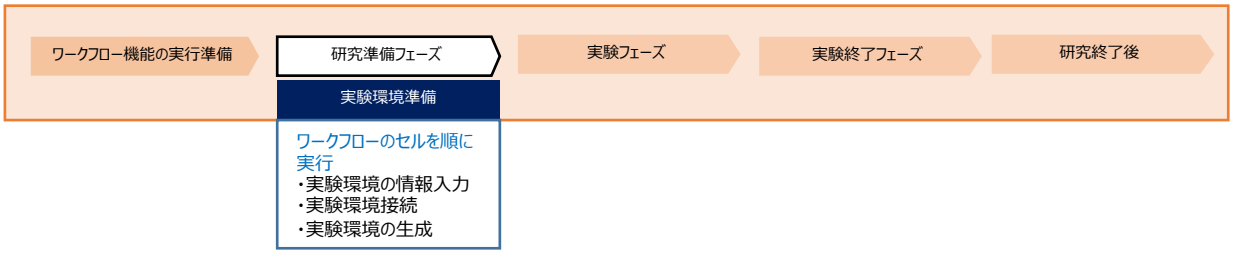


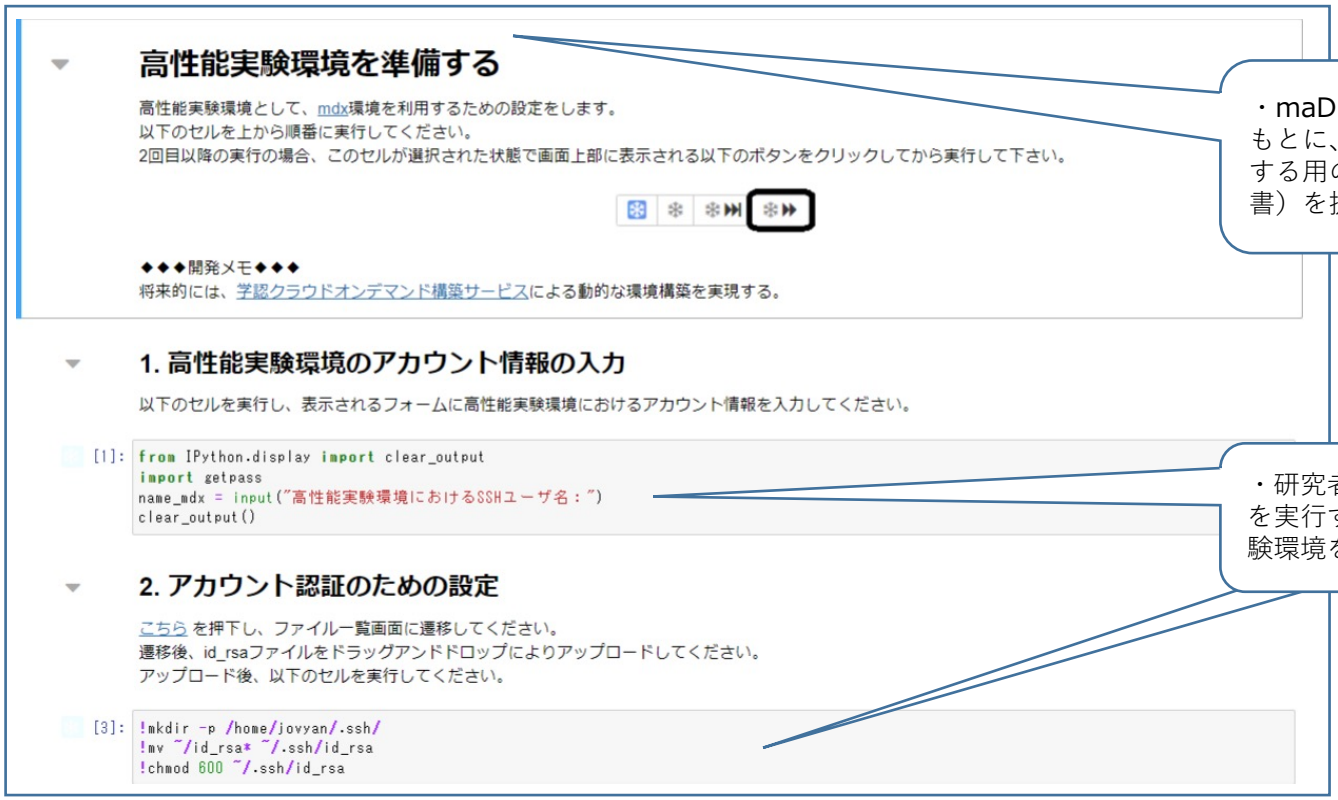
maDMP生成

maDMP(Notebook 形式)



②高性能実験環境準備用のワークフロー





・maDMPに記載された内容をもとに、高性能実験環境を準備する用のワークフロー（手順書）を提供。

・研究者は、NoteBookのセルを実行するのみで、高性能実験環境を準備できる。

③実験の実施　研究サンプル(底生生物の検知／追跡)

ワークフロー機能の実行準備 → 研究準備フェーズ → 実験フェーズ → 実験終了フェーズ → 研究終了後

実験フェーズ
- 実験実施
  - ワークフローのセルを順に実行
  - ・実験実施

```
[1]: !pip install opencv-python cython-bbox motmetrics
     !pip install torch==1.11.0+cu113 torchvision==0.12.0+cu113 -f https://download.pytorch.org/whl/cu113/torch_stable.html
     !conda install -c conda-forge lap -y
     !python3 inference_detr+tracker.py --config ../input_data/config/hogehoge.yaml

     Collecting opencv-python
       Downloading opencv_python-4.5.5.64-cp36-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (60.5 MB)
          ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 60.5/60.5 MB 46.4 MB/s eta 0:00:0000:0100:01
     Collecting cython-bbox
       Downloading cython_bbox-0.1.3.tar.gz (41 kB)
          ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 41.3/41.3 KB 12.7 MB/s eta 0:00:00
       Preparing metadata (setup.py) ... done
     Collecting motmetrics
       Downloading motmetrics-1.2.0-py3-none-any.whl (151 kB)
          ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 151.6/151.6 KB 41.9 MB/s eta 0:00:00
     Requirement already satisfied: numpy>=1.14.5 in /opt/conda/lib/python3.9/site-packages (from opencv-python) (1.19.5)
     Collecting pytest
       Downloading pytest-7.1.1-py3-none-any.whl (297 kB)
          ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 297.0/297.0 KB 61.7 MB/s eta 0:00:00
     Collecting pytest-benchmark
       Downloading pytest_benchmark-3.4.1-py2.py3-none-any.whl (50 kB)
          ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 50.1/50.1 KB 16.4 MB/s eta 0:00:00
     Collecting flake8-import-order
       Downloading flake8_import_order-0.18.1-py2.py3-none-any.whl (15 kB)
```

・実験の開始方法について、自由な実行をサポート
（再現性担保のため、一部制限あり）

③実験の実施　研究サンプル(底生生物の検知／追跡)

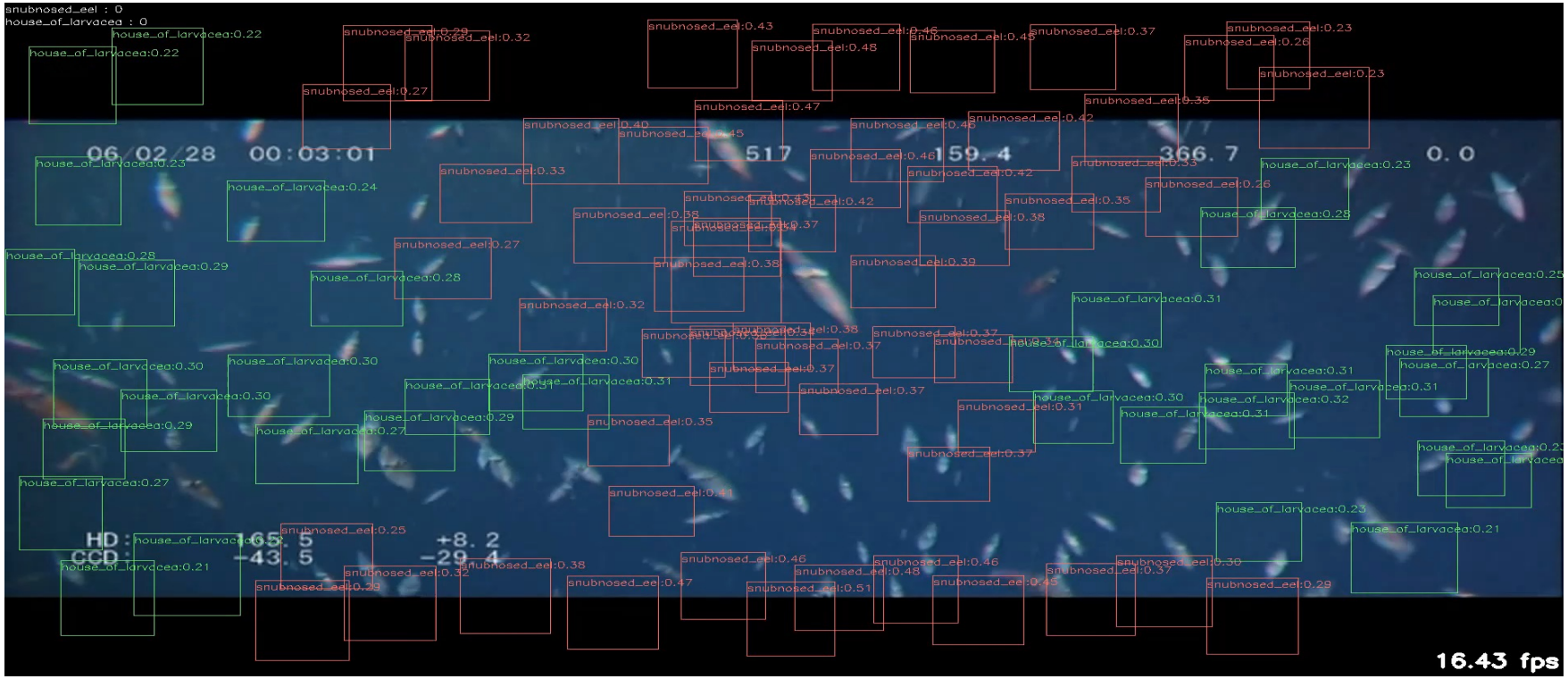
ワークフロー機能の実行準備 | 研究準備フェーズ | 実験フェーズ | 実験終了フェーズ | 研究終了後

実験実施

ワークフローのセルを順に実行
・実験実施

イカ　アナゴ

※動画は開発中のものです



データ提供：海洋研究開発機構(JAMSTEC)

④実験の記録

ワークフロー機能の実行準備 | 研究準備フェーズ | 実験フェーズ | 実験終了フェーズ | 研究終了後

実験記録の保存

ワークフローのセルを順に実行
・実験記録の保存

2. 実験をデータガバナンス機能に途中保存する

※データの保存先としてAWS S3準拠のオブジェクトストレージを利用する場合は、
こちらを実行してください。

```
from IPython.display import display, Javascript
display(Javascript('IPython.notebook.save_checkpoint();'))
```

```
import os
import glob

# Git管理のパスのリストを作成する
%cd ~/
files = os.listdir()
# ディレクトリ一覧からGit-annex管理するディレクトリ(input_dataとoutput_data)を排除する
dirs = [f for f in files if os.path.isdir(f)]
dirs.remove('input_data')
dirs.remove('output_data')
# HOME直下のファイルを取得
files = [f for f in files if os.path.isfile(f)]
# Git管理するパスの配列を作成する
files.extend(dirs)
save_path = files

# Git-annex管理するパスの配列を作成する
annexed_save_path = ['input_data', 'output_data']
```

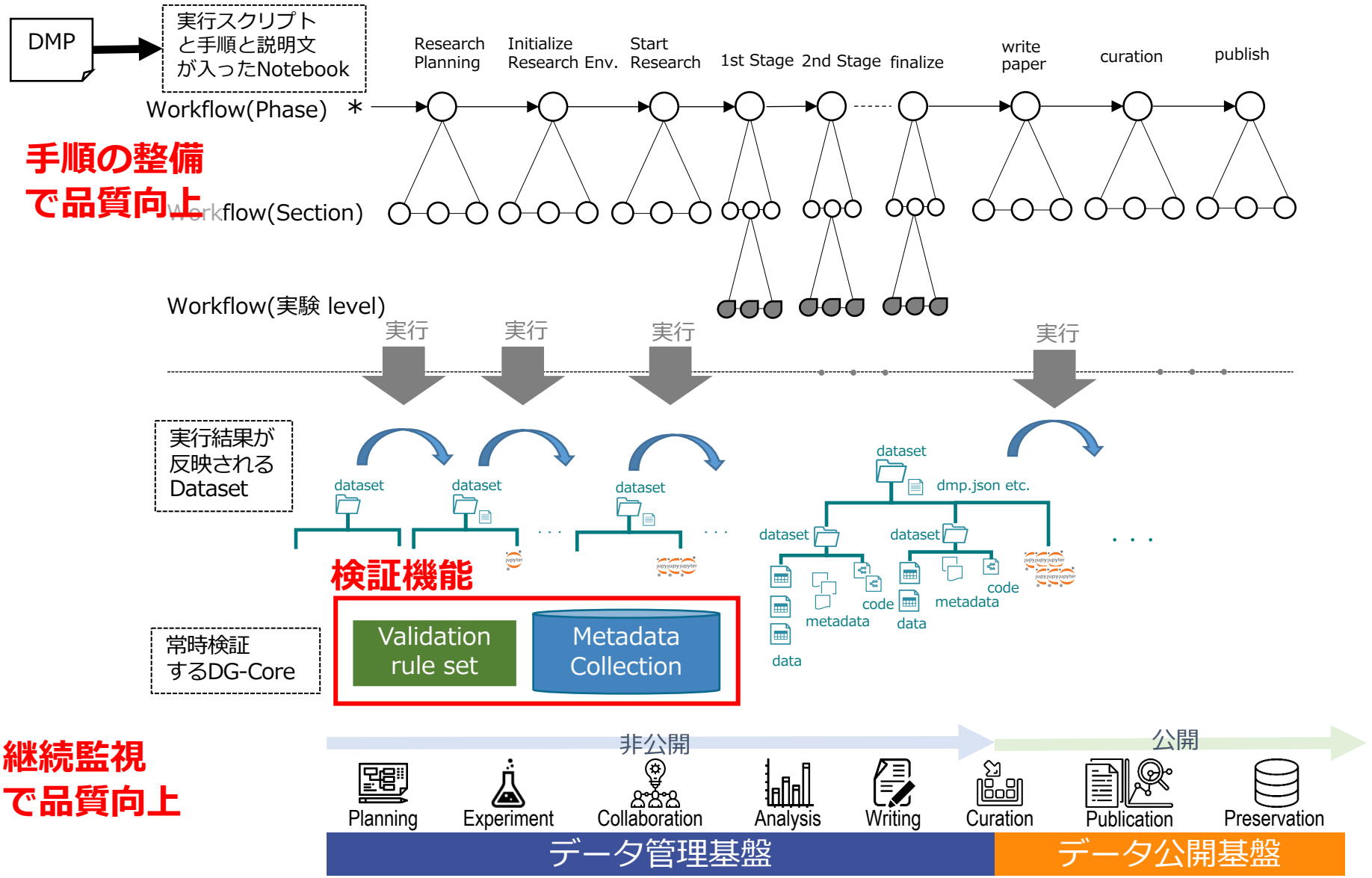
```
import papermill as pm

%cd ~/
# Git-annex管理ファイルを保存
pm.execute_notebook(
    'EX-WORKFLOW/util/base_datalad_save_push.ipynb',
    '/home/jovyan/.local/push_log.ipynb',
    parameters = dict(SAVE_MESSAGE = message + ' (1/2)', PATH = annexed_save_path, IS_RECURSIVE = False)
)
```

・実験の記録について、NoteBookのセルによる手順書を提供
セルの実行のみで、実験の記録をGitによるバージョン管理で行うことができるようにする

3. 今後の進め方

検証機能の拡充


DMP
実行スクリプト
と手順と説明文
が入ったNotebook
Research Planning
Initialize Research Env.
Start Research
1st Stage
2nd Stage
finalize
write paper
curation
publish
Workflow(Phase) ＊
手順の整備
で品質向上
Workflow(Section)
Workflow(実験 level)
実行
実行
実行
実行
実行結果が
反映される
Dataset
dataset
dataset
dataset
dataset
dmp.json etc.
dataset
dataset
code
metadata
data
code
metadata
data
検証機能
常時検証
するDG-Core
Validation
rule set
Metadata
Collection
継続監視
で品質向上
非公開
公開
Planning
Experiment
Collaboration
Analysis
Writing
Curation
Publication
Preservation
データ管理基盤
データ公開基盤

今後の進め方

1. 検証機能の拡充
   - 共通スキーマ策定
   - DMPと共通スキーマの関係定義
   - 共通スキーマに基づく研究データ管理・検証機能実現方式決定
   - 検証機能の実装
2. 手順書のブラッシュアップ
   - プロトタイプ利用ユーザからのフィードバックへの対応など
3. データガバナンス機能 次期プロトタイプ の実現
   - 検証機能拡充及び手順書のブラッシュアップを統合
4. 他基盤との連携実施

4. まとめ

データガバナンス機能の目的

「研究プロジェクトのデータ管理を
機械的に支援すること」

具体的には以下を実施

- 研究データ管理のための共通スキーマの策定
- 共通スキーマに基づく研究データ管理・検証機能の提供
- 共通スキーマに基づくインタフェースの提供
  - 研究者にとって管理しやすいインタフェース
  - 他基盤との相互運用のためのインタフェース

データガバナンス機能導入による効果

- 機械的に検証されたデータ管理状態が保証される
  → 研究効率・品質が向上
  → 研究管理部門とのコミュニケーション品質向上
- 研究データ管理の他基盤との連携が促進される
  → オープンサイエンス推進

問合せ先： yoko@nii.ac.jp