NII RCOS
Research Center for Open Science and Data Platform

データ分析を再現・再利用可能にするパッケージ機能の構想

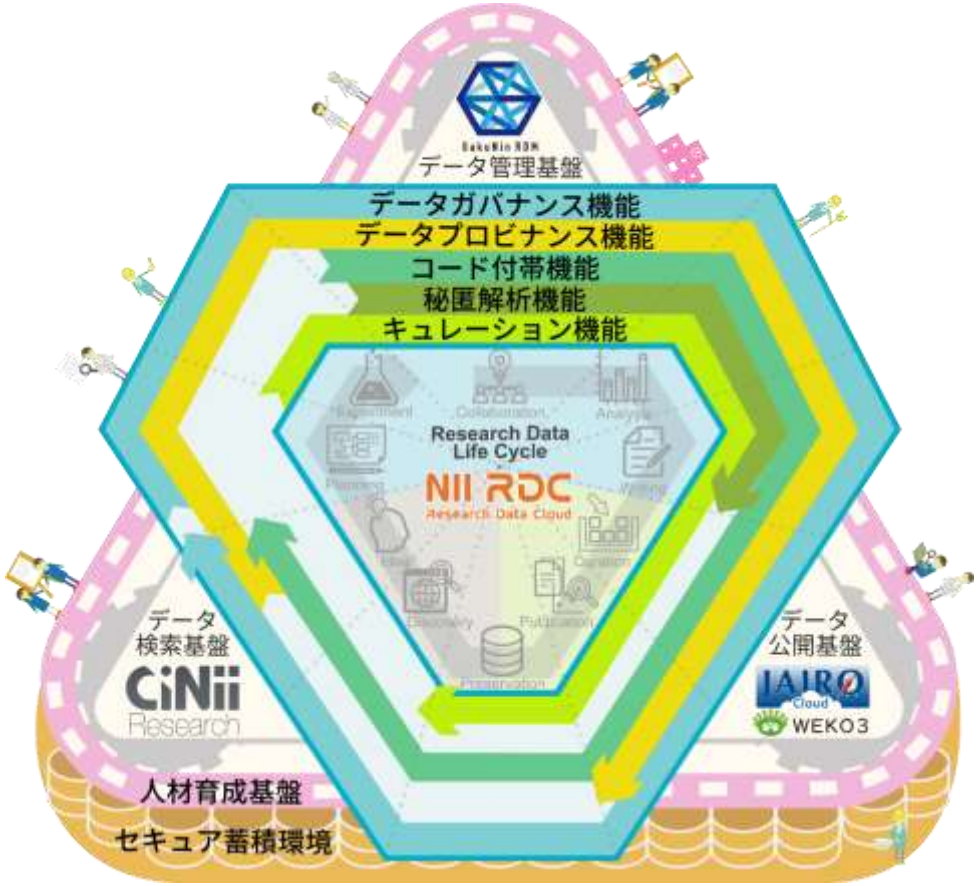
藤原一毅
国立情報学研究所 アーキテクチャ科学研究系／
オープンサイエンス基盤研究センター (RCOS)

2023-06-21
Japan Open Science Summit (JOSS 2023)

本日の話題

- **GakuNin RDM データ解析機能の紹介** いまここ
- 計算再現パッケージ機能の構想

次世代 NII Research Data Cloud



データガバナンス機能　管理

研究データ管理の計画と執行、モニタリングを組織的に実現する機能。研究データ管理計画（DMP）の作成を支援し、研究者が守るべきポリシーや計画に沿った環境を自動的にセットアップ。組織は研究データ管理状況をモニタリングすることで研究の効率化と研究公正を促進します。

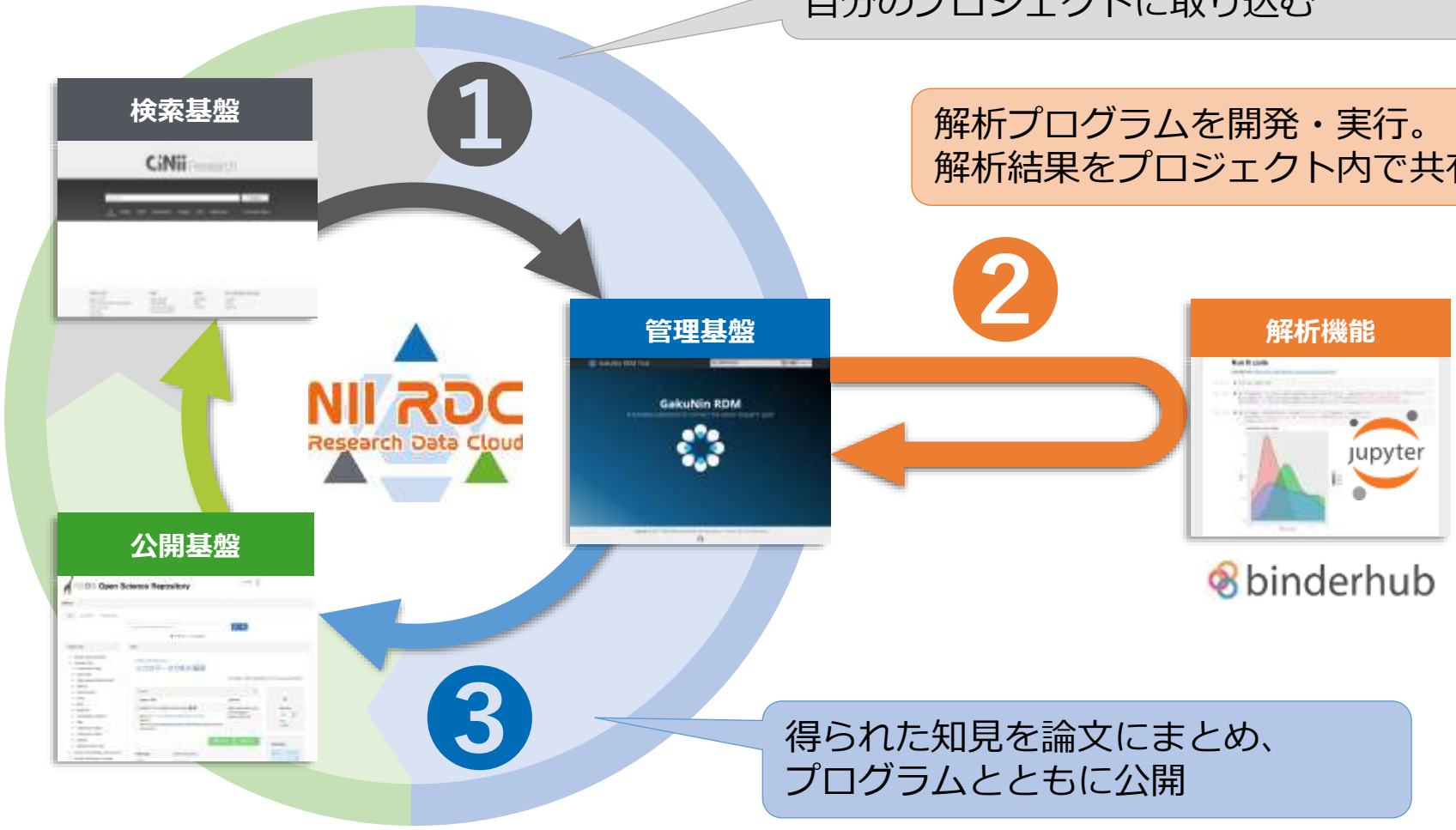
データプロビナンス機能　信頼

データを利用する研究者がデータの来歴を確認できるとともに、データを提供する研究者が自身のデータの利用状況を確認できる機能。研究不正の疑いから研究者と組織を守り、データ公開への取り組みを後押しします。

コード付帯機能　活用

研究者が用いたデータ・プログラム・実行環境定義をまとめて「計算再現パッケージ」として公開・再利用できる機能。先行研究のデータ解析を他の研究者が確実に再現し、発展的な研究を円滑に始められるようにします。

データとコードが循環する世界

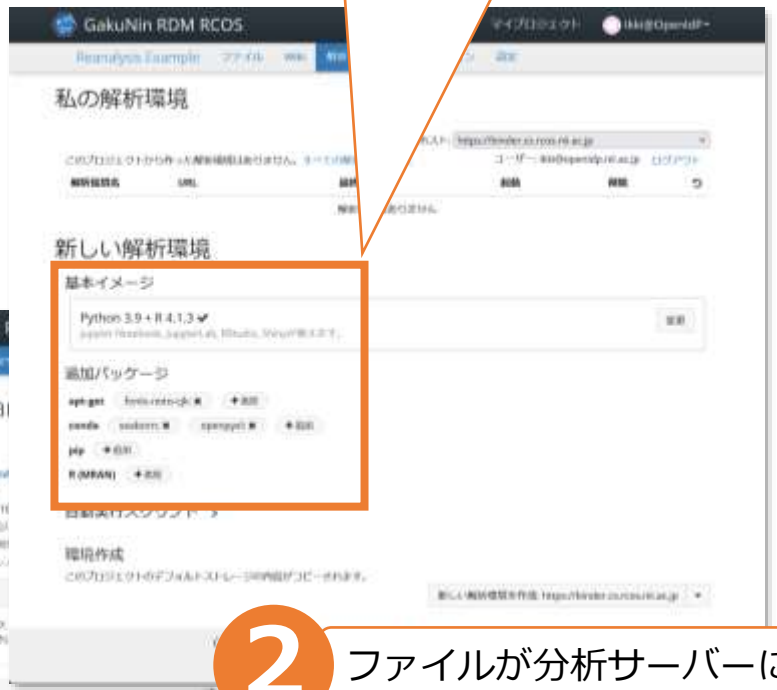


コード付帯機能の一部:
GakuNin RDM データ解析機能

- JupyterHub がインストールされた計算機と連携し、データ解析環境をGakuNin RDMから1クリックで構築
- NII所有の計算機のほか、クラウド上のVMなど外部計算機とも連携可能

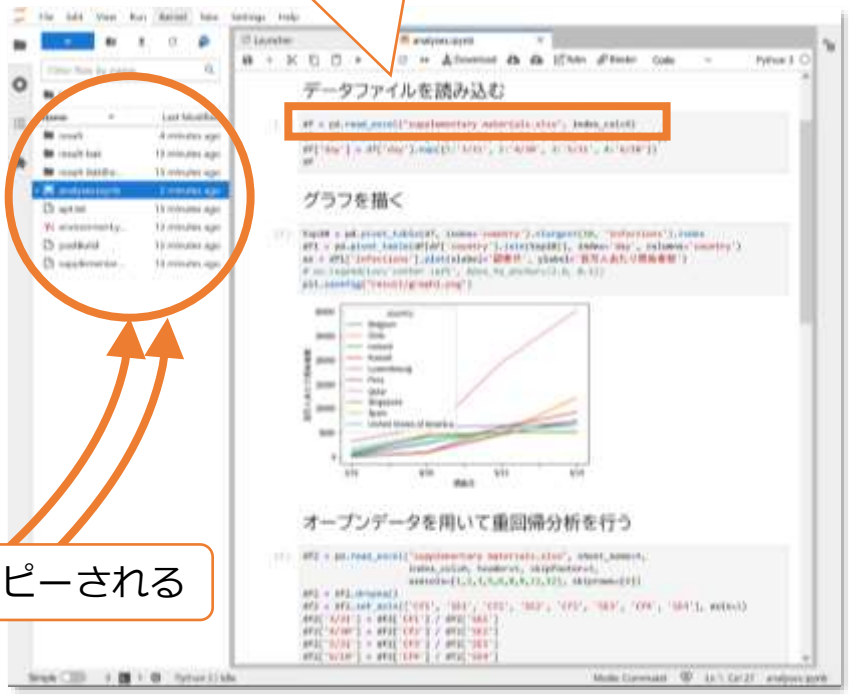
基本的な使い方

1 分析環境を選んで作成ボタンを押す



2 ファイルが分析サーバーにコピーされる

3 分析サーバー上で、そのファイルを読み込むプログラムを書いて実行する



4 分析結果を管理基盤に書き戻す

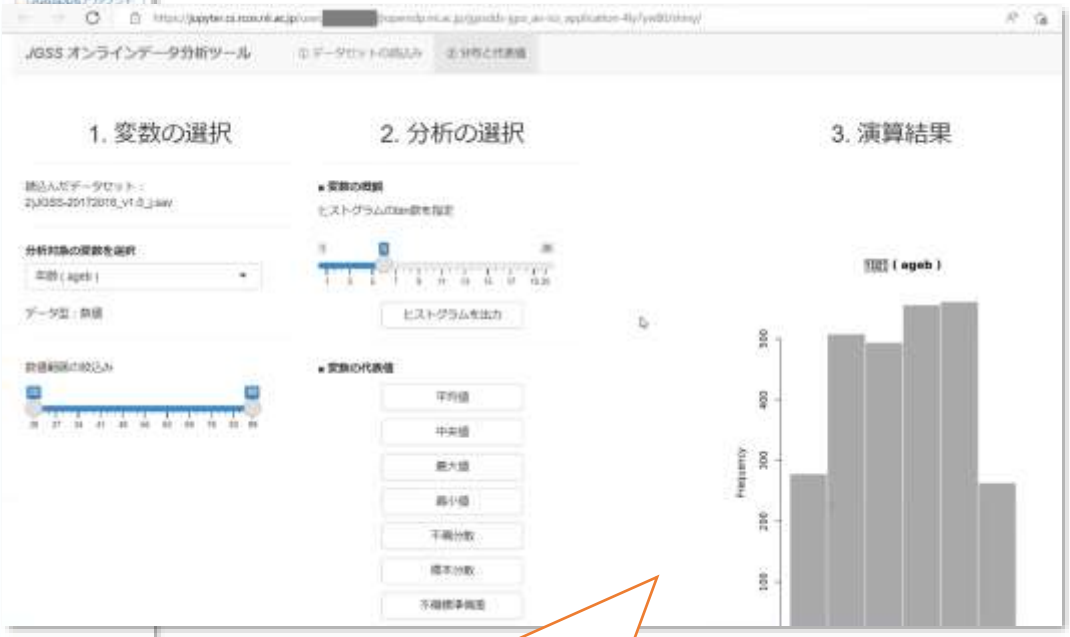
応用例1: 深層学習による動画解析

動画データ・解析プログラム提供: 海洋研究開発機構 研究プラットフォーム運用開発部門

応用例2: 社会学データの簡易分析アプリ

1 リポジトリで起動リンクをクリック

2 NIIのサーバー上で簡易分析アプリが起動



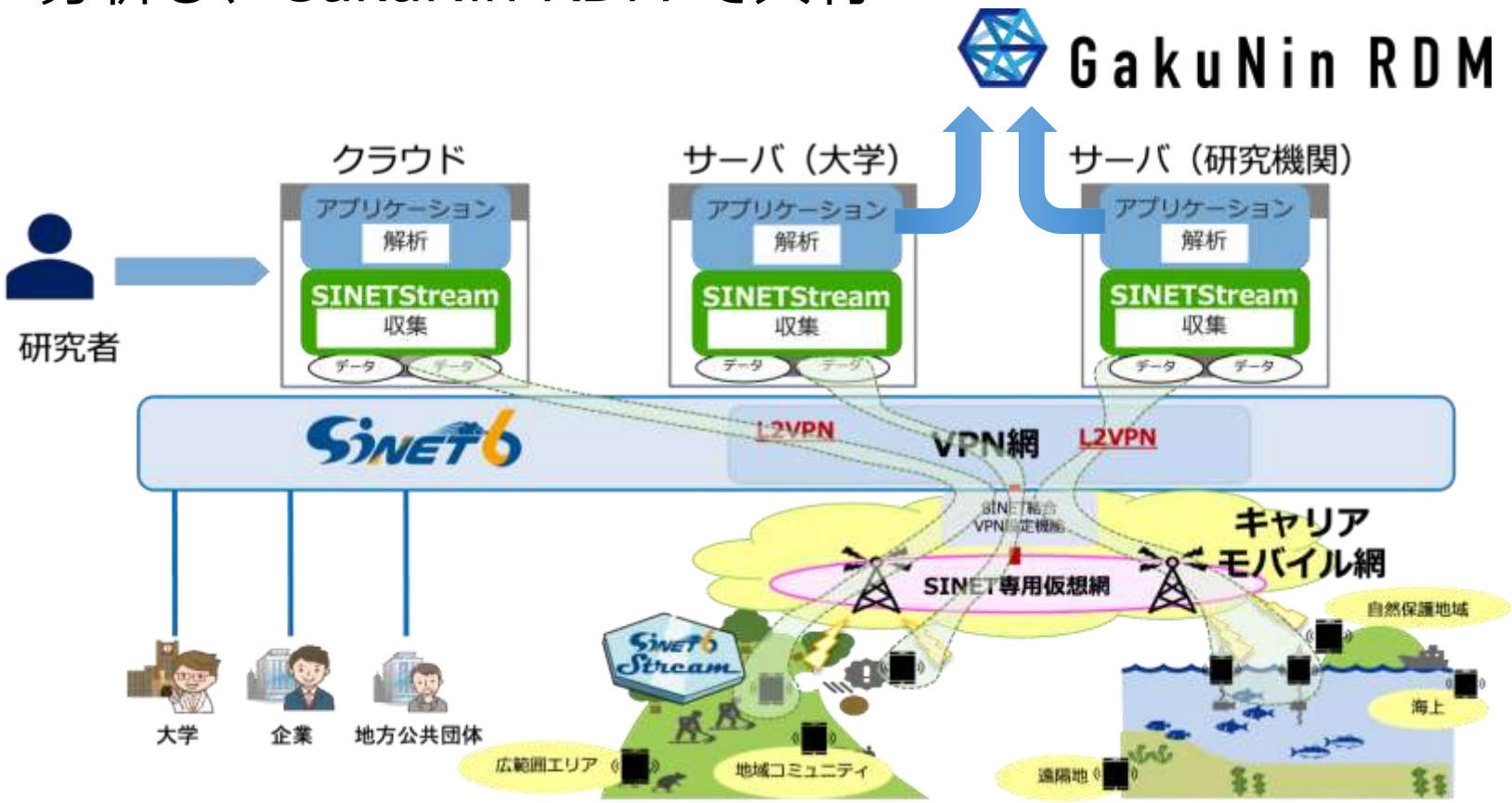
3 プログラムを書かずにデータの概要を把握

https://jgssdds.repo.nii.ac.jp/records/2000721

アプリ開発: 大阪商業大学JGSS研究センター

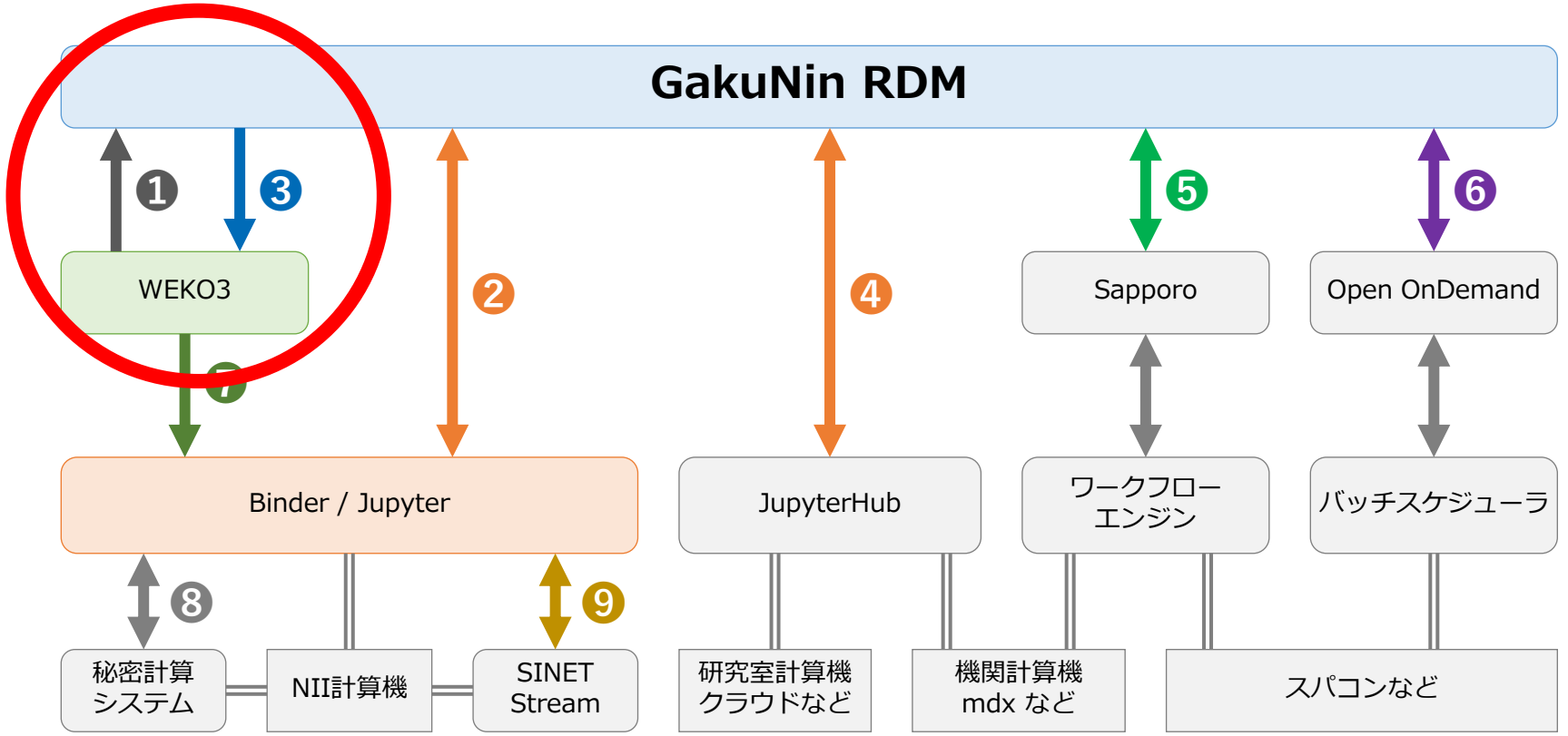
応用例3: IoTデータのリアルタイム分析

- 広域データ収集基盤 SINETStream と連携
- IoT機器などから流れてくるデータをリアルタイムに分析し、GakuNin RDM で共有



https://www.sinetstream.net/

将来展望: コード付帯機能群の開発

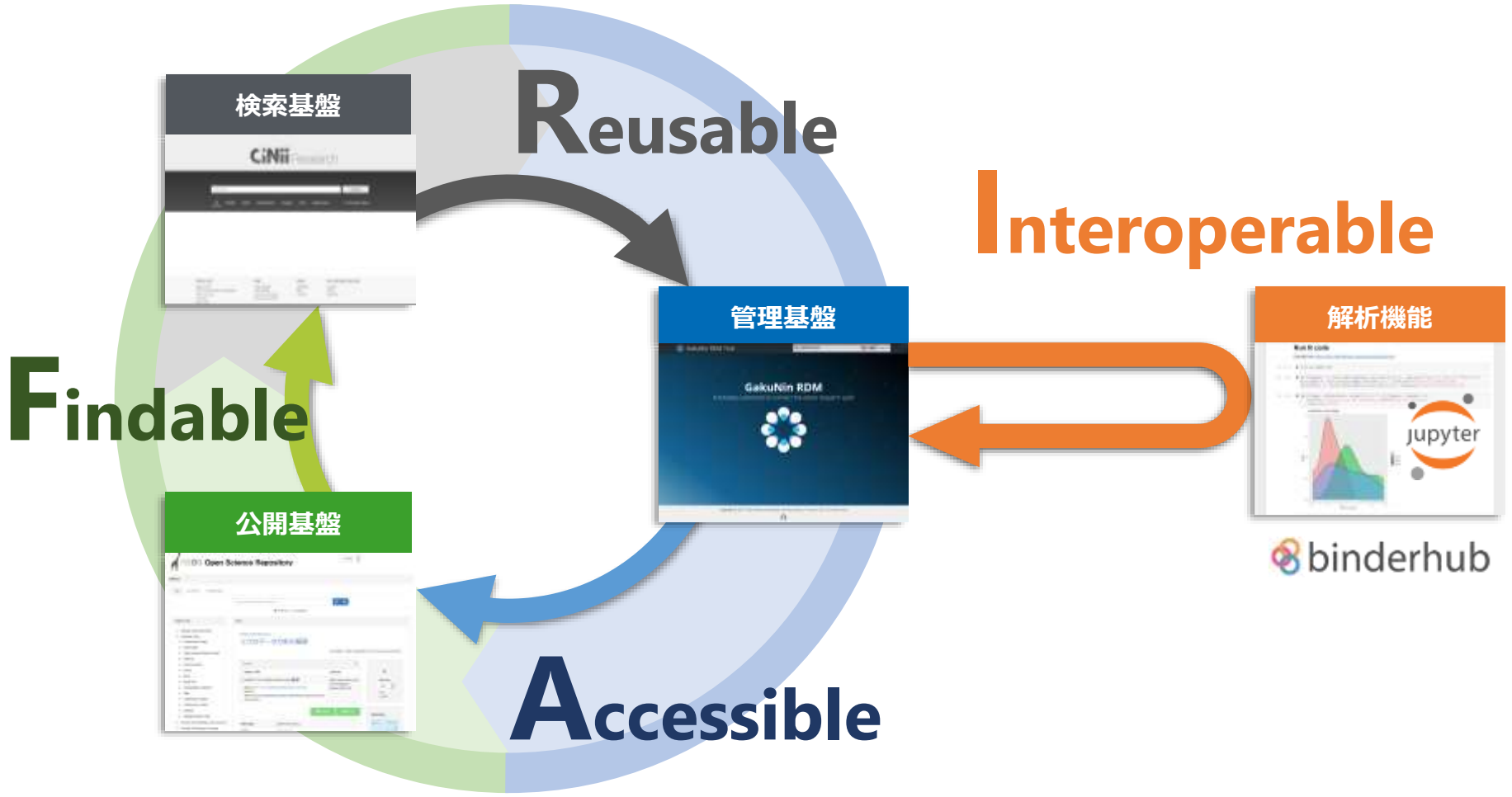


❶❸ 計算再現パッケージ機能	GRDMプロジェクトをWEKOで公開、他者がGRDMに取り込み再利用	開発中
❷❹ GakuNin RDMデータ解析機能	Jupyterによるデータ解析環境をGRDMから構築	運用中
❺ 外部ワークフローエンジン連携機能	ワークフローエンジンをGRDMから起動、結果をGRDMに回収	開発中
❻ 外部バッチスケジューラ連携機能	バッチスケジューラにジョブをGRDMから投入	開発中
❼ WEKOオンライン分析機能	NIIのBinderを使ってWEKOから解析環境を構築	運用中
❽ 秘密計算システム統合機能	秘密分散によるセキュアな解析環境をJupyterから利用	設計中
❾ SINETStream連携検討	SINETStreamによるリアルタイムデータ収集環境を構築	開発中

本日の話題

- GakuNin RDM データ解析機能の紹介
- **計算再現パッケージ機能の構想** いまここ

データとコードが循環する世界



FAIRな計算再現パッケージとは

- **F**indable
  - 公開されたパッケージを、論文や研究データと関連づけて発見できる
- **A**ccessible
  - 解析プログラムを含む GakuNin RDM プロジェクトをパッケージ化してリポジトリに公開できる
- **I**nteroperable
  - 異なる計算機上に同一の解析環境（Dockerコンテナ）を再構築し、他者の解析プログラムを再実行できる
- **R**eusable
  - 公開されたパッケージを、他の研究者が自身のGakuNin RDM プロジェクトとして再利用できる
    - プロジェクトに含まれる解析プログラムを自身の環境で再実行し、先行研究のデータ解析結果を再現できる

どんなメタデータが必要か

- **F**indableであるために…
  - 論文との関連性
    - DOI（で十分か？）
  - データとの関連性
    - プロジェクト内のデータ
    - プロジェクト外のデータ
      - 外部ストレージ、外部データリポジトリ、etc.
- **I**nteroperableであるために…
  - プログラム同士の関連性
    - プロジェクト内のプログラム
      - 実行ファイル、実行順序、実行環境、実行時パラメータ、etc.
    - プロジェクト外のプログラム
      - 言語別ライブラリ、パッケージリポジトリ、etc.
      - 外部ソースコードリポジトリ

セッション
6/19（月）D1, D2
6/23（金）E3

RO-Crate 形式の「NII RDC アプリケーションプロファイル」を開発中

事例: Materials Cloud

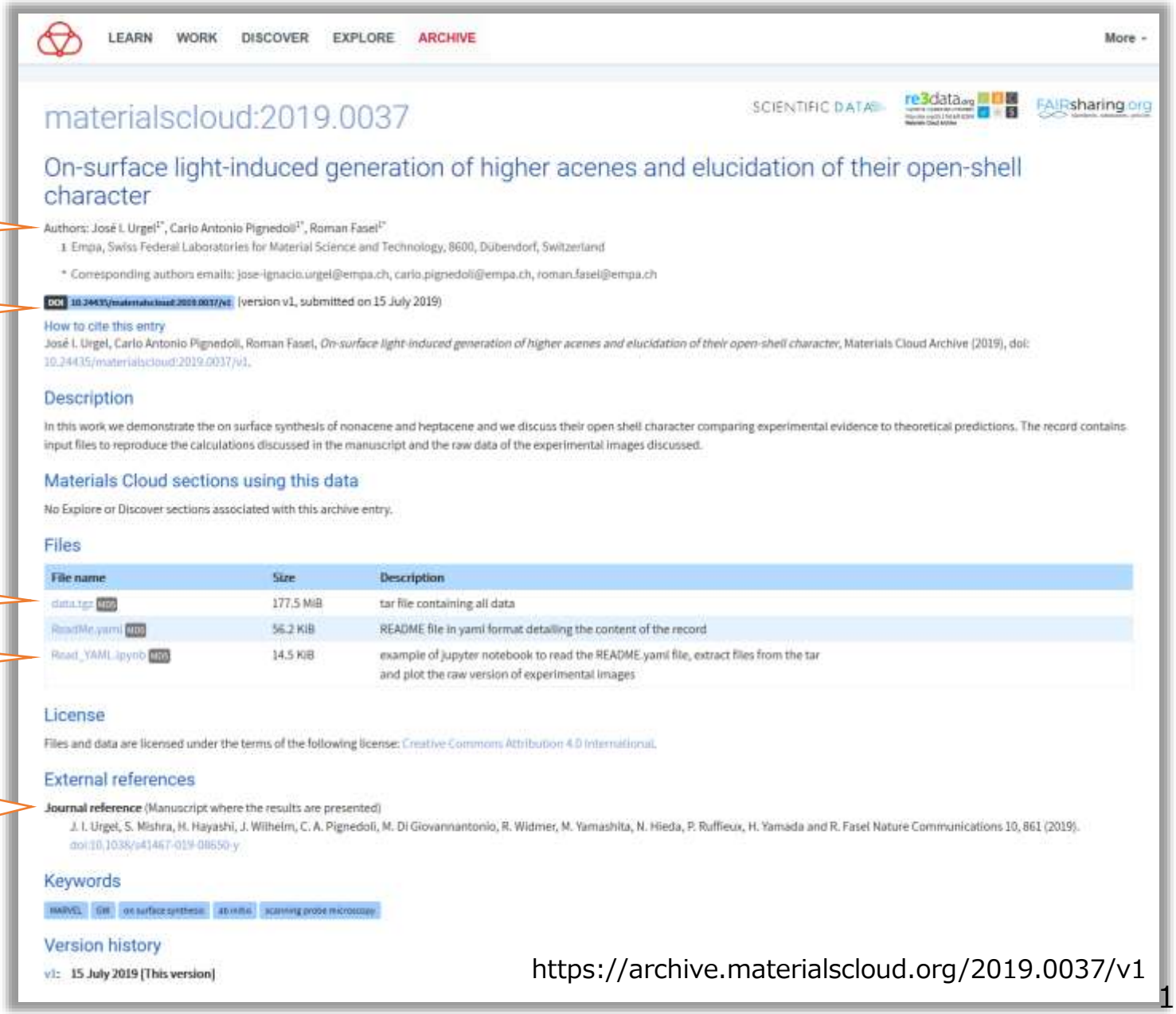
作成者

DOI

データ

分析ノート

このデータを使った論文



LEARN WORK DISCOVER EXPLORE ARCHIVE More

materialscloud:2019.0037

SCIENTIFIC DATA re3data.org FAIRsharing.org

On-surface light-induced generation of higher acenes and elucidation of their open-shell character

Authors: José I. Urgel[1*], Carlo Antonio Pignedoli[1*], Roman Fasel[1*]

1 Empa, Swiss Federal Laboratories for Material Science and Technology, 8600, Dübendorf, Switzerland

* Corresponding authors emails: jose-ignacio.urgel@empa.ch, carlo.pignedoli@empa.ch, roman.fasel@empa.ch

DOI 10.24435/materialscloud:2019.0037/v1 (version v1, submitted on 15 July 2019)

How to cite this entry

José I. Urgel, Carlo Antonio Pignedoli, Roman Fasel, *On-surface light-induced generation of higher acenes and elucidation of their open-shell character*, Materials Cloud Archive (2019), doi: 10.24435/materialscloud:2019.0037/v1.

Description

In this work we demonstrate the on surface synthesis of nonacene and heptacene and we discuss their open shell character comparing experimental evidence to theoretical predictions. The record contains input files to reproduce the calculations discussed in the manuscript and the raw data of the experimental images discussed.

Materials Cloud sections using this data

No Explore or Discover sections associated with this archive entry.

Files

File name	Size	Description
data.tgz MD5	177.5 MiB	tar file containing all data
ReadMe.yaml MD5	56.2 KiB	README file in yaml format detailing the content of the record
Read_YAML.ipynb MD5	14.5 KiB	example of jupyter notebook to read the README.yaml file, extract files from the tar and plot the raw version of experimental images

License

Files and data are licensed under the terms of the following license: Creative Commons Attribution 4.0 International.

External references

Journal reference (Manuscript where the results are presented)

J. I. Urgel, S. Mishra, H. Hayashi, J. Wilhelm, C. A. Pignedoli, M. Di Giovannantonio, R. Widmer, M. Yamashita, N. Hieda, P. Ruffieux, H. Yamada and R. Fasel Nature Communications 10, 861 (2019). doi:10.1038/s41467-019-08650-y

Keywords

MARVEL GW on surface synthesis ab initio scanning probe microscopy

Version history

v1: 15 July 2019 [This version]

https://archive.materialscloud.org/2019.0037/v1

NII RDC の一部として 計算再現パッケージ機能はどうあるべき？

- データと解析プログラムをまとめて Findable にするサービスは研究分野ごとに存在
  - 分野ごとに切り口や粒度が異なる
- NII RDC は分野横断的な汎用サービス
  - 幅広い研究分野をどのようにサポートするか？
  - 汎用的なメタデータ規格が存在するか？ 独自開発するか？
  - 既存の分野別メタデータ規格を利用すべきか？

有識者と議論しながら設計を進めていく

RCOS

rcos@nii.ac.jp