NII RCOS
Research Center for Open Science and Data Platform

# GakuNin RDMと連携するデータ解析機能

藤原一毅
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

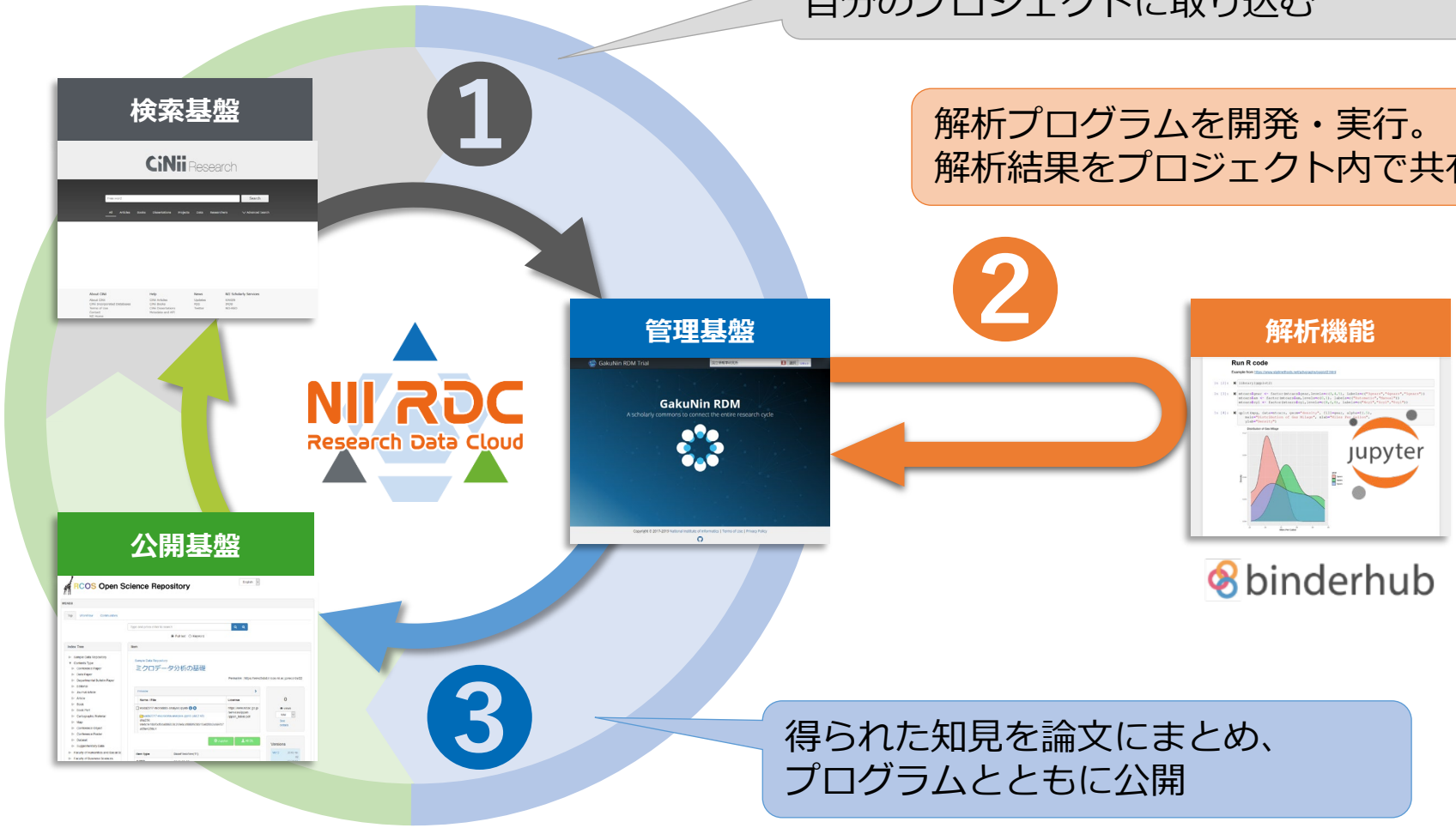
2022/12/07
研究データ管理（RDM）説明会2022 in 大阪

# データとコードが循環する世界

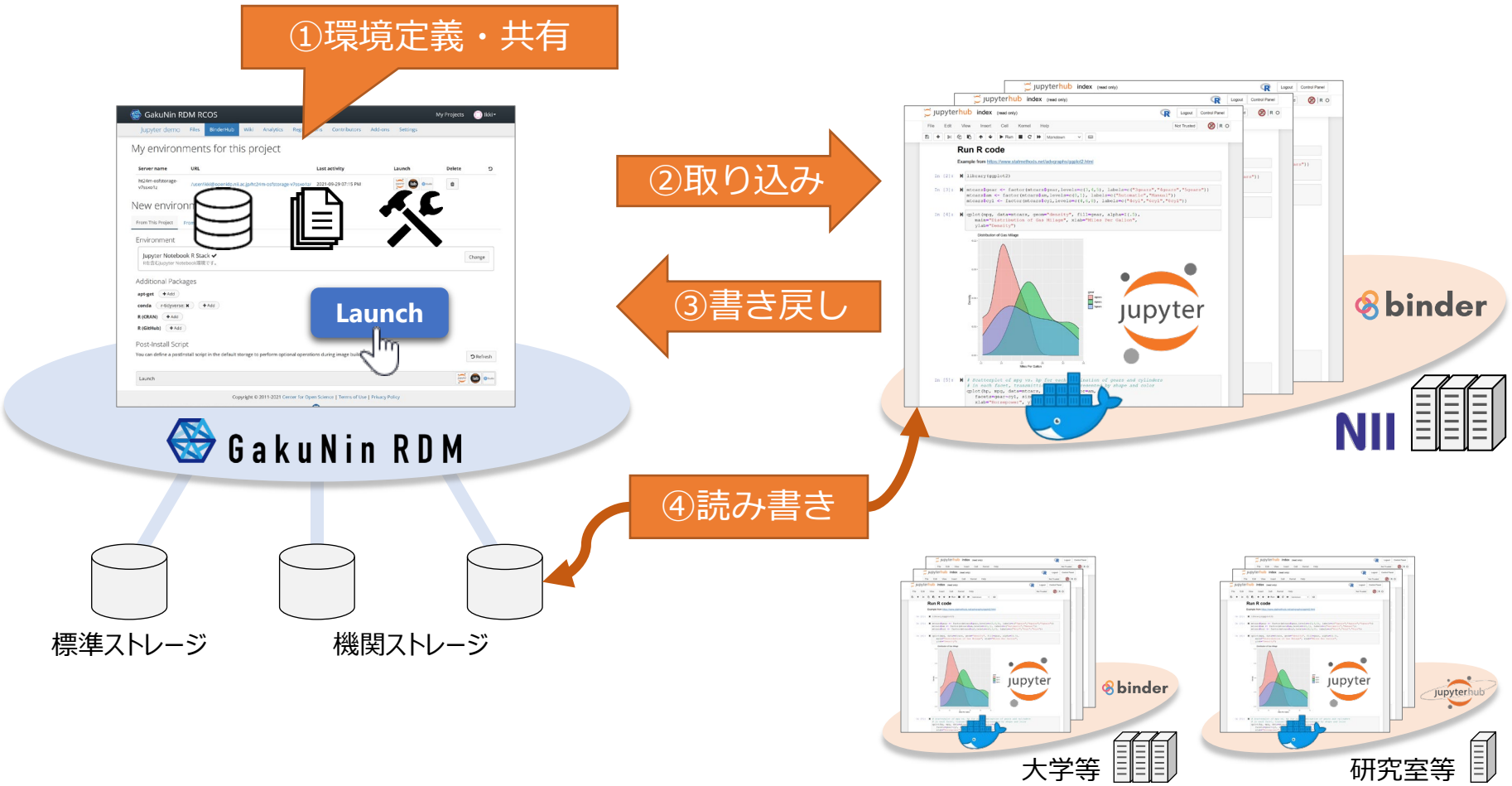
先行研究のデータを発見。GakuNin RDM の
自分のプロジェクトに取り込む

解析プログラムを開発・実行。
解析結果をプロジェクト内で共有

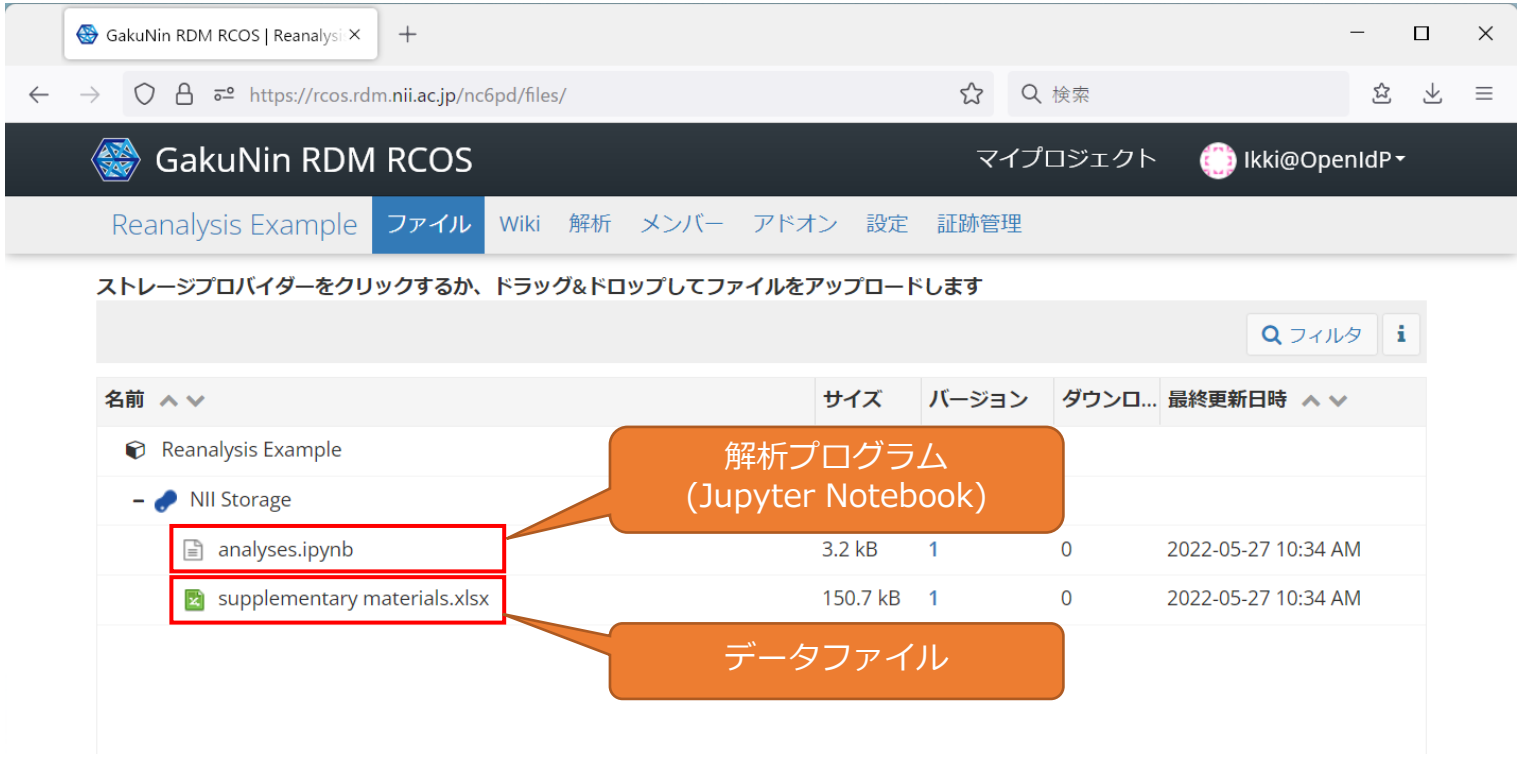
得られた知見を論文にまとめ、
プログラムとともに公開



# GakuNin RDM データ解析機能



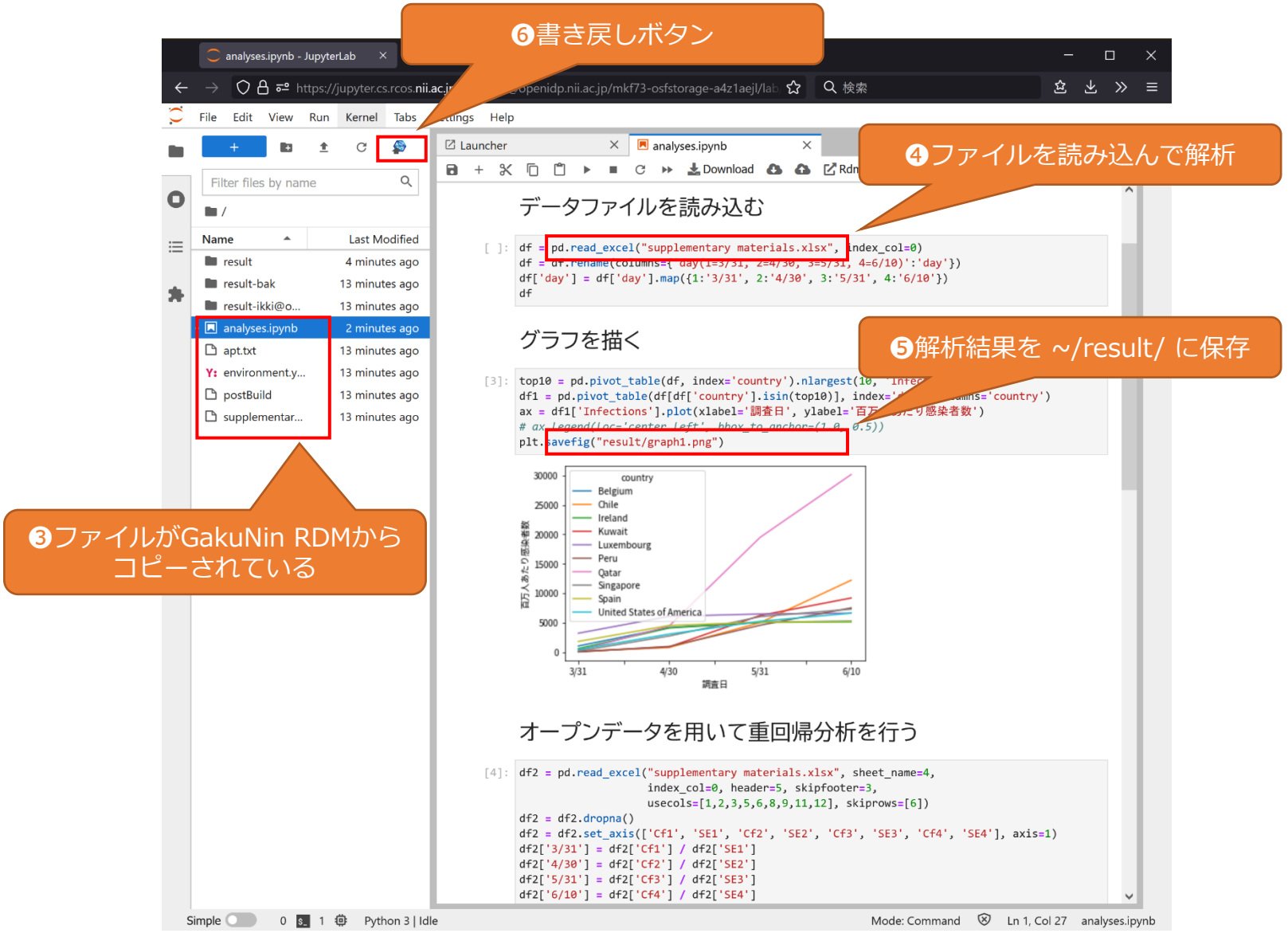
# デモ (1/4)


GakuNin RDM RCOS | Reanalysi
https://rcos.rdm.nii.ac.jp/nc6pd/files/
検索
GakuNin RDM RCOS
マイプロジェクト
Ikki@OpenIdP
Reanalysis Example
ファイル
Wiki
解析
メンバー
アドオン
設定
証跡管理
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
フィルタ
名前
サイズ
バージョン
ダウンロ...
最終更新日時
Reanalysis Example
NII Storage
analyses.ipynb
3.2 kB
1
0
2022-05-27 10:34 AM
supplementary materials.xlsx
150.7 kB
1
0
2022-05-27 10:34 AM
解析プログラム
(Jupyter Notebook)
データファイル

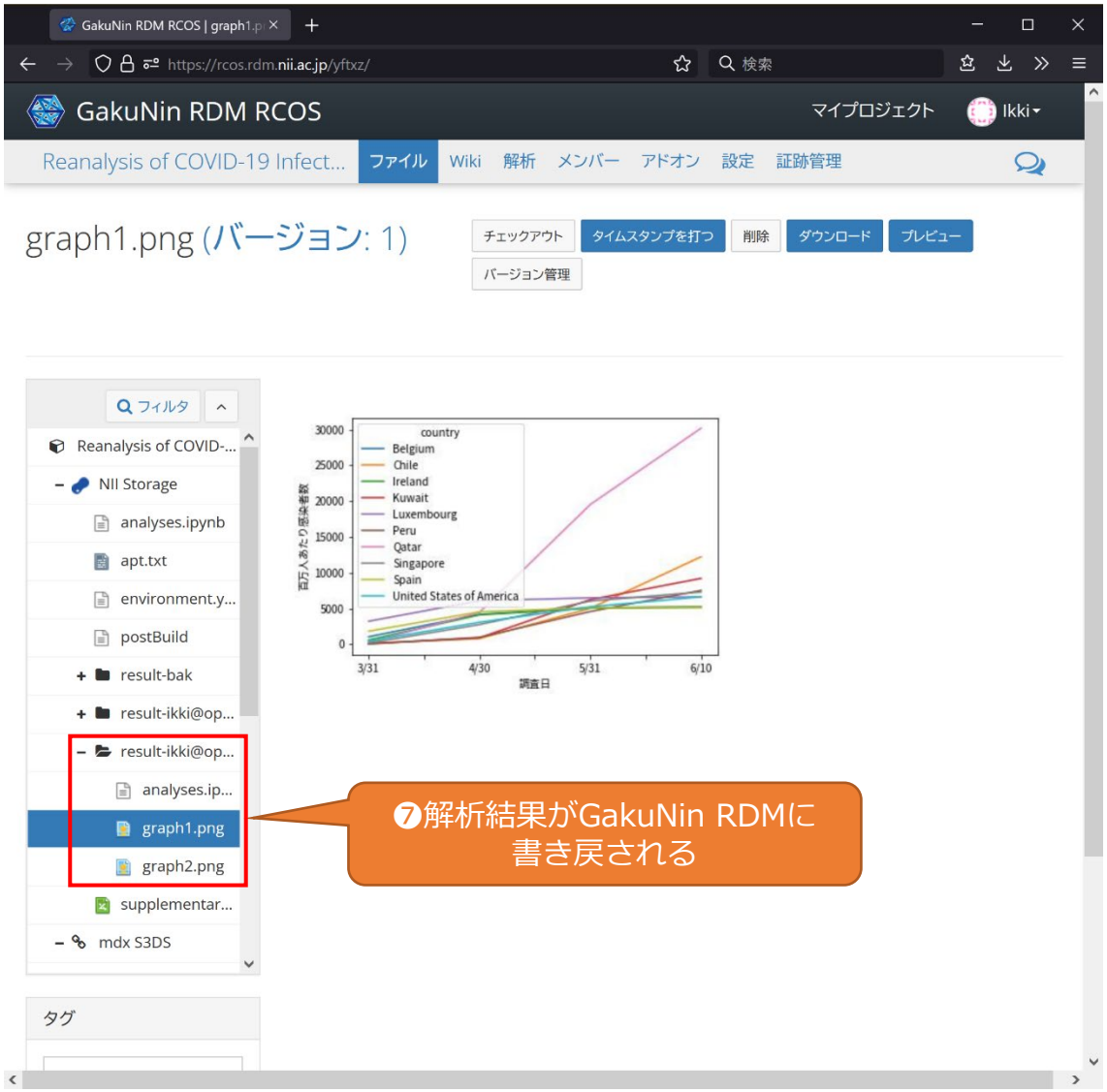
# デモ (2/4)


Reanalysis Example BidnerHub
https://rcos.rdm.nii.ac.jp/nc6pd/binderhub
検索
GakuNin RDM RCOS
マイプロジェクト
Ikki@OpenIdP
Reanalysis Example
ファイル
Wiki
解析
メンバー
アドオン
設定
新しい解析環境
❶解析環境を構成
基本イメージ
Python 3.9 + R 4.1.3 ✔
Jupyter Notebook, JupyterLab, RStudio, Shinyが使えます。
変更
追加パッケージ
apt-get
fonts-noto-cjk: ✖
+追加
conda
seaborn: ✖
openpyxl: ✖
+追加
pip
+追加
R (MRAN)
+追加
自動実行スクリプト
#!/bin/bash
set -x
...
保存
❷計算機を選択して起動！
環境作成
このプロジェクトのデフォルトストレージの内容がコピーされます。
新しい解析環境を作成: https://binder.cs.rcos.nii.ac.jp

# デモ (3/4)


❻書き戻しボタン
❹ファイルを読み込んで解析
❺解析結果を ~/result/ に保存
❸ファイルがGakuNin RDMからコピーされている
analyses.ipynb - JupyterLab
File Edit View Run Kernel Tabs Settings Help
Launcher
analyses.ipynb
Download
Filter files by name
Name
Last Modified
result 4 minutes ago
result-bak 13 minutes ago
result-ikki@o... 13 minutes ago
analyses.ipynb 2 minutes ago
apt.txt 13 minutes ago
environment.y... 13 minutes ago
postBuild 13 minutes ago
supplementar... 13 minutes ago
データファイルを読み込む
df = pd.read_excel("supplementary materials.xlsx", index_col=0)
df['day'] = df['day'].map({1:'3/31', 2:'4/30', 3:'5/31', 4:'6/10'})
df
グラフを描く
plt.savefig("result/graph1.png")
country
Belgium
Chile
Ireland
Kuwait
Luxembourg
Peru
Qatar
Singapore
Spain
United States of America
3/31
4/30
5/31
6/10
オープンデータを用いて重回帰分析を行う
df2 = pd.read_excel("supplementary materials.xlsx", sheet_name=4,
index_col=0, header=5, skipfooter=3,
usecols=[1,2,3,5,6,8,9,11,12], skiprows=[6])
df2 = df2.dropna()
df2 = df2.set_axis(['Cf1', 'SE1', 'Cf2', 'SE2', 'Cf3', 'SE3', 'Cf4', 'SE4'], axis=1)
df2['3/31'] = df2['Cf1'] / df2['SE1']
df2['4/30'] = df2['Cf2'] / df2['SE2']
df2['5/31'] = df2['Cf3'] / df2['SE3']
df2['6/10'] = df2['Cf4'] / df2['SE4']
Simple
Python 3 | Idle
Mode: Command
Ln 1, Col 27
analyses.ipynb

# デモ (4/4)


GakuNin RDM RCOS | graph1.p
https://rcos.rdm.nii.ac.jp/yftxz/
検索
GakuNin RDM RCOS
マイプロジェクト
Ikki
Reanalysis of COVID-19 Infect...
ファイル
Wiki
解析
メンバー
アドオン
設定
証跡管理
graph1.png (バージョン: 1)
チェックアウト
タイムスタンプを打つ
削除
ダウンロード
プレビュー
バージョン管理
フィルタ
Reanalysis of COVID-...
NII Storage
analyses.ipynb
apt.txt
environment.y...
postBuild
result-bak
result-ikki@op...
result-ikki@op...
analyses.ip...
graph1.png
graph2.png
supplementar...
mdx S3DS
タグ
country
Belgium
Chile
Ireland
Kuwait
Luxembourg
Peru
Qatar
Singapore
Spain
United States of America
30000
25000
20000
15000
10000
5000
0
3/31
4/30
5/31
6/10
調査日
❼解析結果がGakuNin RDMに
書き戻される

# 外部計算機連携のバリエーション




| | システム | 運用主体 | 認証方法 | 同時起動数 | ドメイン名＋サーバ証明書 | バックエンド |
|---|---|---|---|---|---|---|
| A | Binder | NII | 学認 | 10個／ユーザ | ○ | Kubernetes |
| B | Binder | 機関 | 学認, OAuth, LDAP, etc. | 任意設定 | 必要 | Kubernetes |
| C | JupyterHub | 研究室等 | OAuth, LDAP, ローカル | 1個／ユーザ | 不要 | Linux VM |

- ワークフローエンジンとの連携機能を開発中。Sapporo を介して各種エンジンに対応
- スパコン等との連携も設計中。Open OnDemandを介して各種スケジューラに対応

# こんな用途に使えます

## 研究

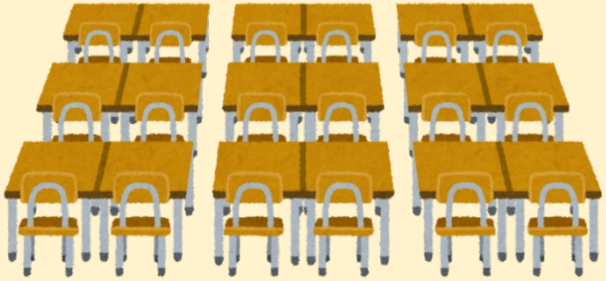
- ご自身の研究のためのデータ分析

## 公開・共有

- 他の研究者の二次分析に資するデータとプログラムの公開

## 教育・学習

- 学生たちにデータ分析をさせるゼミ・講義・演習など



## 引き継ぎ

- 先輩の研究環境を後輩が再現し、研究を継続する

# 詳しい情報

## 導入手続き

- データ解析機能は、GakuNin RDM のオプション機能として、機関単位で提供されます。
- 利用機関の情報基盤センター等で初期設定を行うと、その機関に所属するユーザーが利用可能となります。
- 現在 GakuNin RDM を正式利用されている機関の担当者様に、解析機能の追加についてご案内してあります。

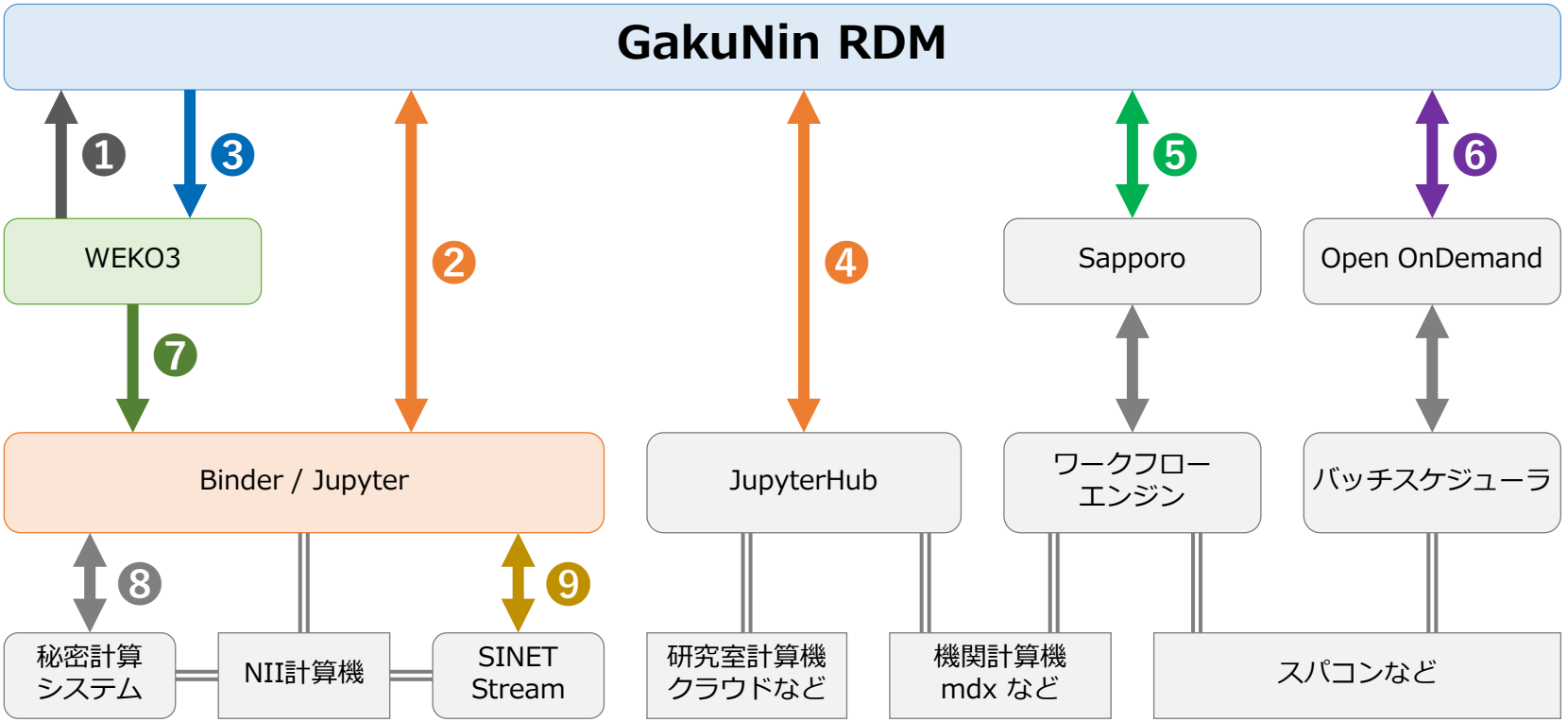
## 資料

- マニュアル https://support.rdm.nii.ac.jp/
- 解説動画 https://youtu.be/_FzOpDTQrBQ
- オープンフォーラム https://www.nii.ac.jp/openforum/2022/day3_nii-rdc4.html

## お問い合わせ

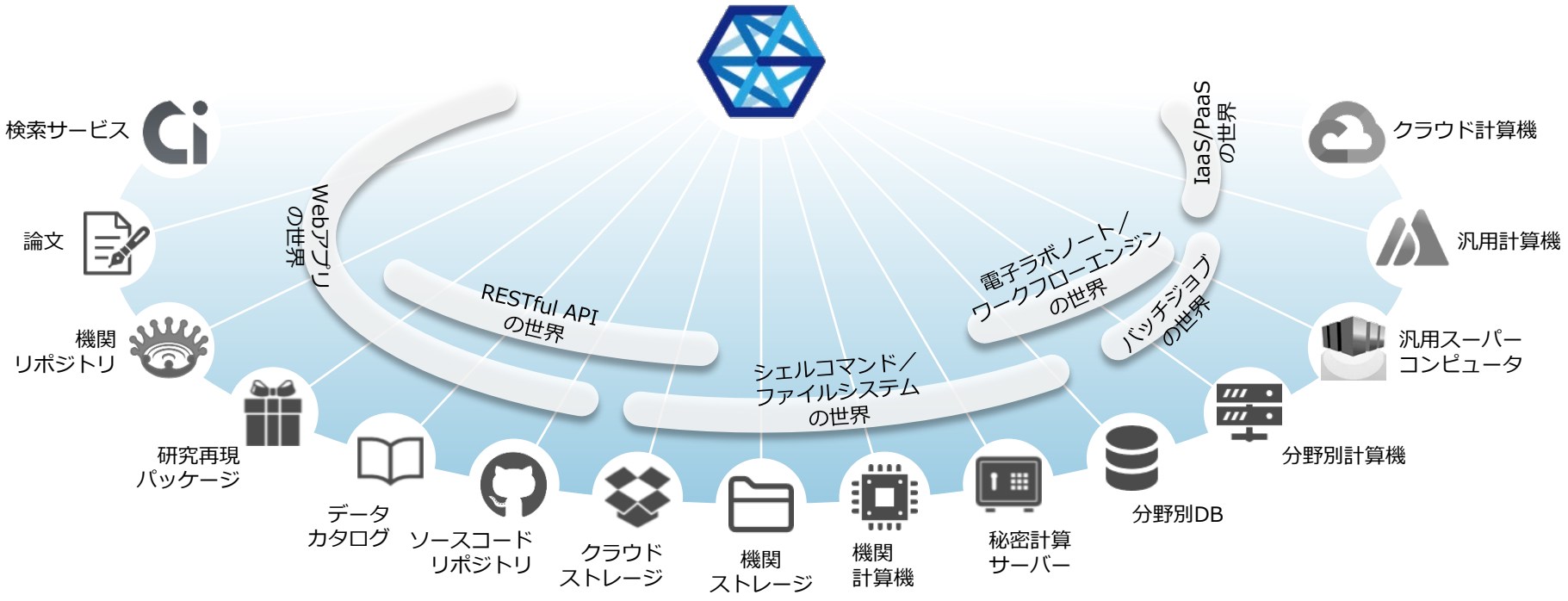
- 開発・活用に関すること→ cs-support@nii.ac.jp
- 導入手続きに関すること→ rdm_support@nii.ac.jp

# コード付帯機能群




| | | |
|---|---|---|
| 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 | 設計中 |
| 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 | 設計中 |
| 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 | 開発中 |
| 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 | 設計中 |

# 目指したい将来像




**GakuNin RDM を核として**
**データの世界と計算機の世界を結ぶ**