

GakuNin RDMと連携する データ解析機能

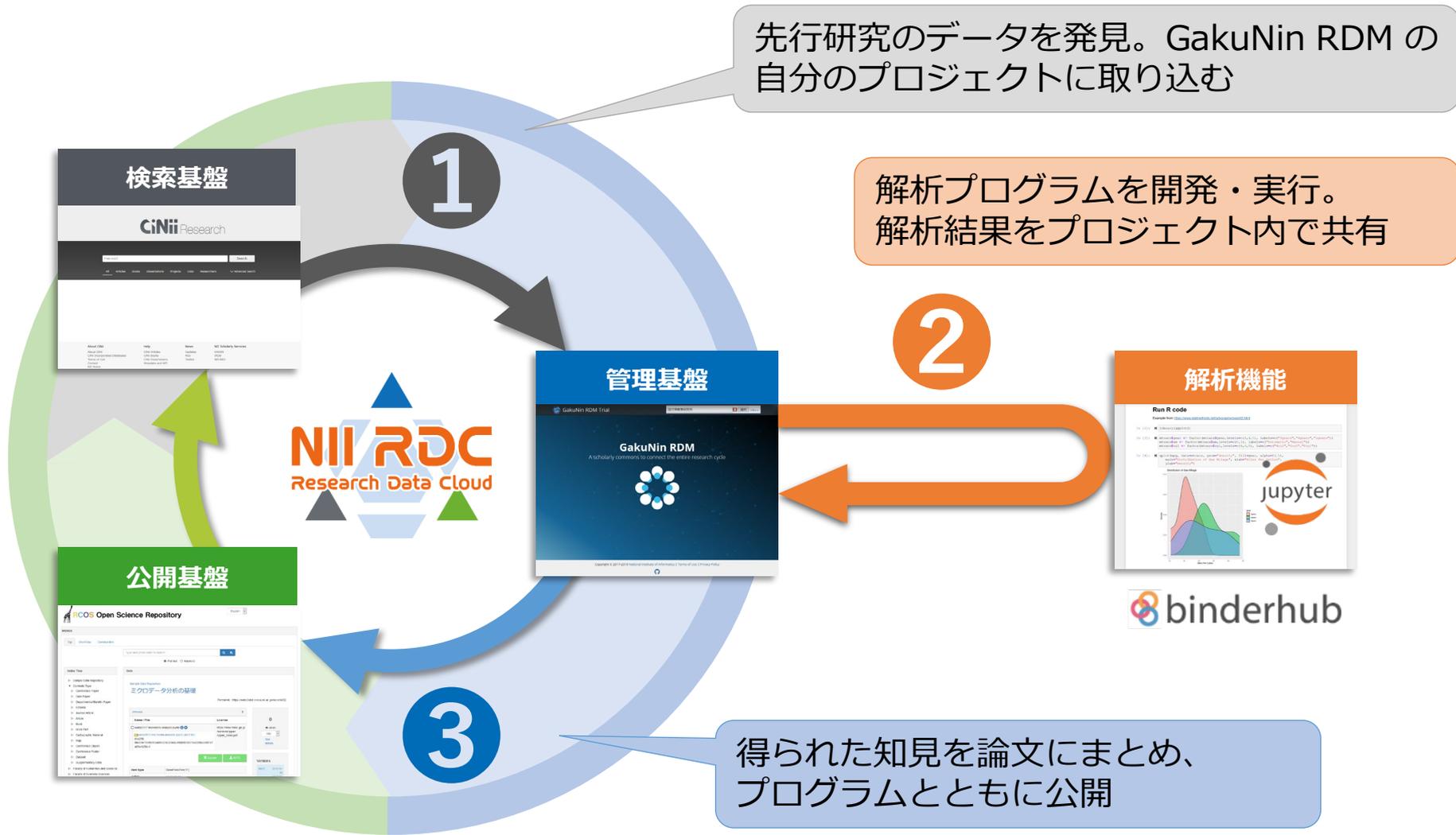
藤原一毅

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

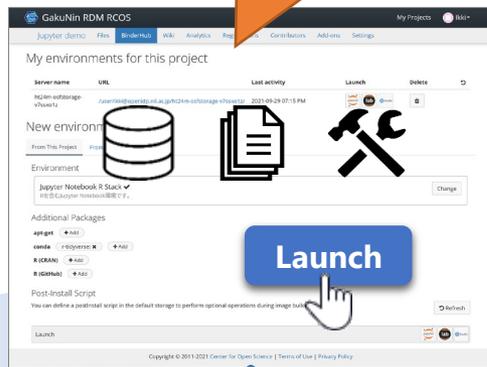
研究データ管理（RDM）説明会2022 in 大阪

データとコードが循環する世界



GakuNin RDM データ解析機能

①環境定義・共有



GakuNin RDM

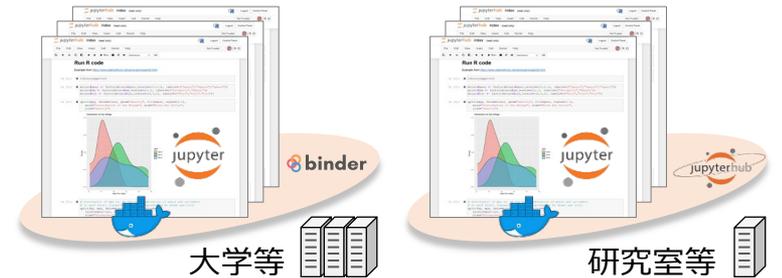
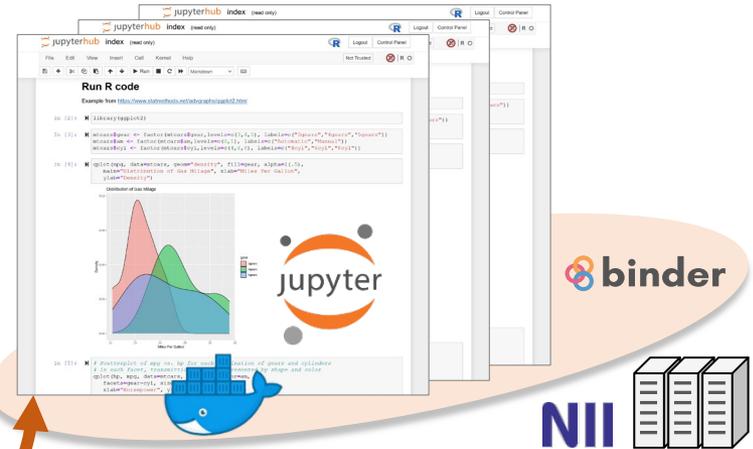
標準ストレージ

機関ストレージ

②取り込み

③書き戻し

④読み書き



デモ (1/4)

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

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)



新しい解析環境

① 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

② 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)

③ファイルがGakuNin RDMからコピーされている

④ファイルを読み込んで解析

⑤解析結果を ~/result/ に保存

⑥書き戻しボタン

```
[1]: df = pd.read_excel("supplementary_materials.xlsx", index_col=0)
df = df.rename(columns={'day(1=3/31, 2=4/30, 3=5/31, 4=6/10)': 'day'})
df['day'] = df['day'].map({'1': '3/31', 2: '4/30', 3: '5/31', 4: '6/10'})
df

[3]: top10 = pd.pivot_table(df, index='country').nlargest(10, 'Infections')
df1 = pd.pivot_table(df[df['country'].isin(top10)], index='country', columns='day')
ax = df1['Infections'].plot(xlabel='調査日', ylabel='百万人あたり感染者数')
# ax.legend(loc='center left', bbox_to_anchor=(1, 0, 0.5))
plt.savefig("result/graph1.png")

[4]: 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']
```

デモ (4/4)

The screenshot shows the GakuNin RDM RCOS web interface. The browser address bar displays `https://rcos.rdm.nii.ac.jp/yftxz/`. The page title is "Reanalysis of COVID-19 Infect...". The navigation menu includes "ファイル" (Files), "Wiki", "解析" (Analysis), "メンバー" (Members), "アドオン" (Add-ons), "設定" (Settings), and "証跡管理" (Provenance Management). The user is logged in as "Ikki".

The main content area shows a file named "graph1.png (バージョン: 1)". Below the filename are buttons for "チェックアウト" (Check out), "タイムスタンプを打つ" (Stamp timestamp), "削除" (Delete), "ダウンロード" (Download), and "プレビュー" (Preview). A "バージョン管理" (Version management) button is also present.

On the left, a file explorer sidebar shows a tree view of the workspace. The "result-ikki@op..." folder is expanded, and the file "graph1.png" is highlighted with a blue selection bar. A red box highlights the "result-ikki@op..." folder and its contents.

On the right, a line graph titled "country" displays the number of cases per 100,000 people over time. The x-axis is labeled "調査日" (Survey date) with markers for 3/31, 4/30, 5/31, and 6/10. The y-axis is labeled "百万人あたり感染数" (Number of cases per 100,000 people) with a scale from 0 to 30,000. The legend includes Belgium, Chile, Ireland, Kuwait, Luxembourg, Peru, Qatar, Singapore, Spain, and United States of America. Qatar shows the highest and most rapidly increasing number of cases.

An orange callout bubble with a white number 7 and a checkmark points to the "graph1.png" file in the sidebar. The text inside the bubble reads: "⑦ 解析結果がGakuNin RDMに書き戻される" (⑦ Analysis results are written back to GakuNin RDM).

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。

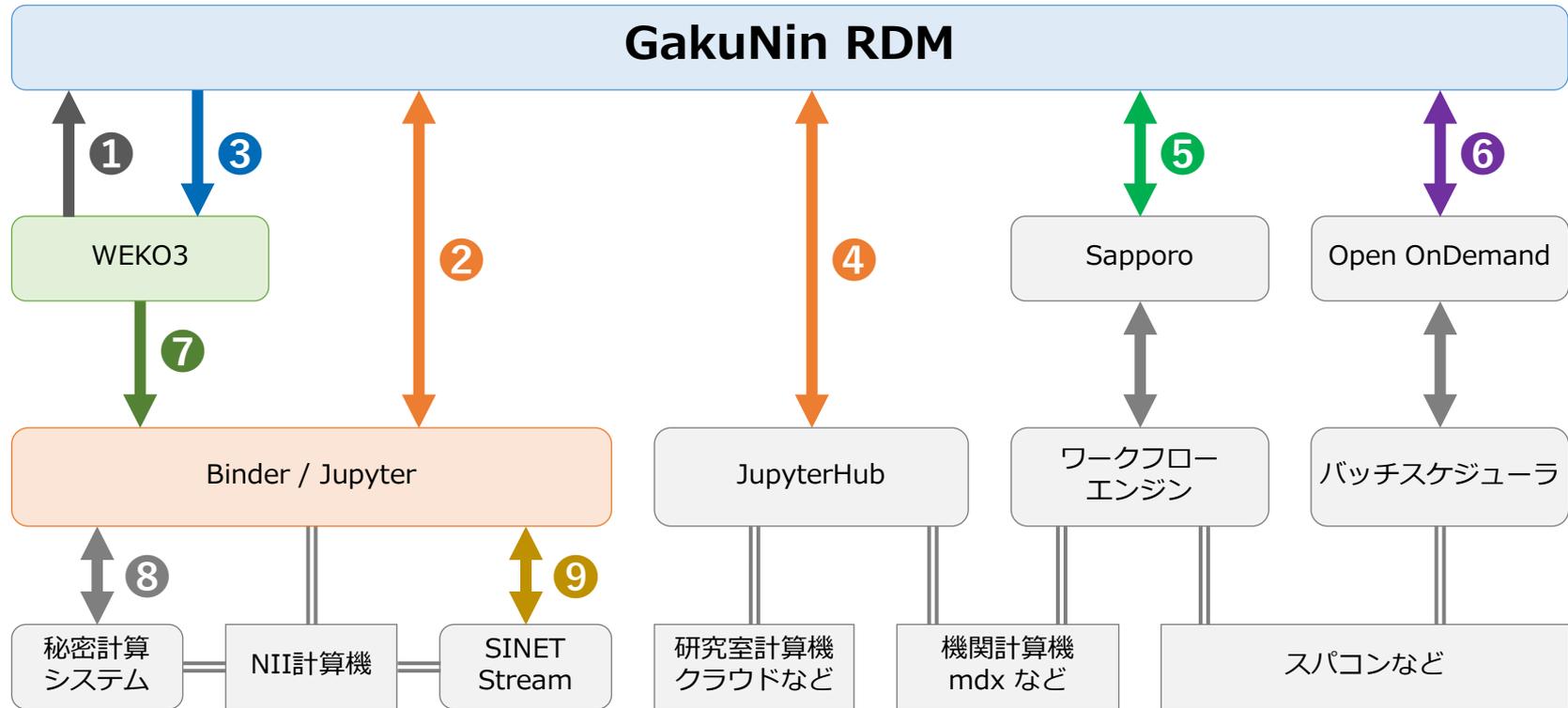
資料

- 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

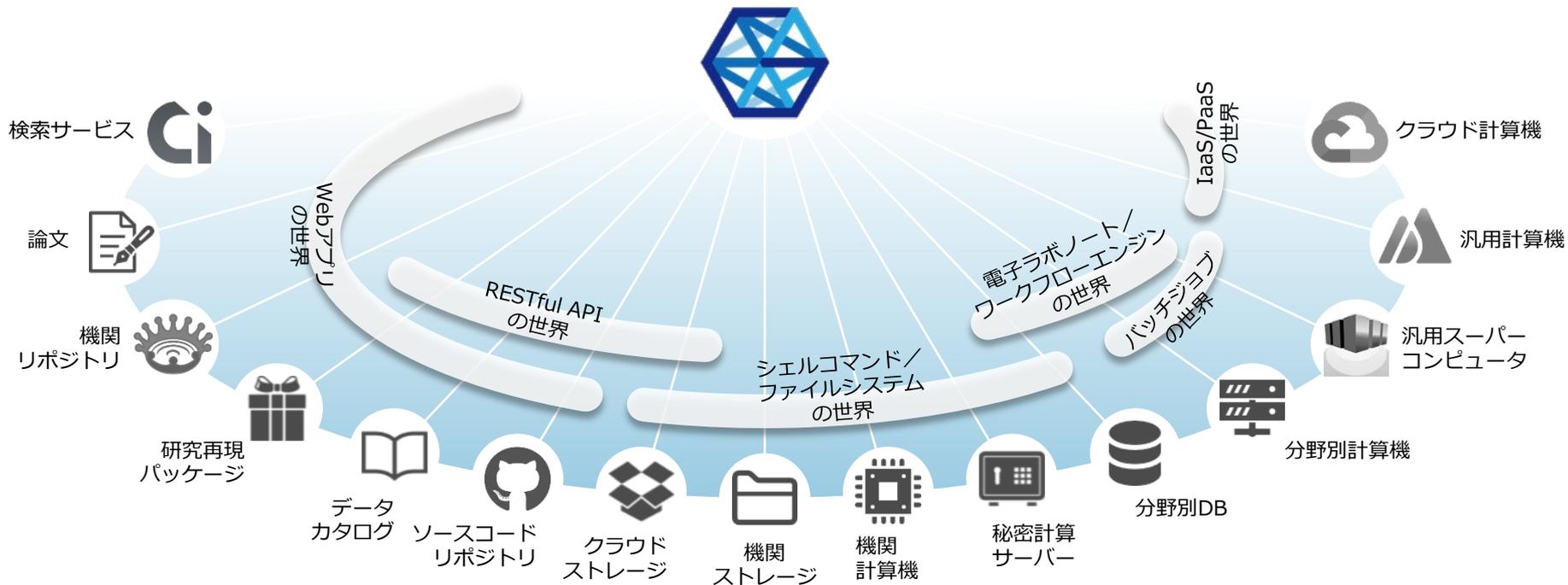
- 開発・活用に関すること → cs-support@nii.ac.jp
- 導入手続きに関すること → rdm_support@nii.ac.jp

コード付帯機能群



①③ 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
②④ 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
⑤ 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
⑥ 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
⑦ 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
⑧ 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
⑨ 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