

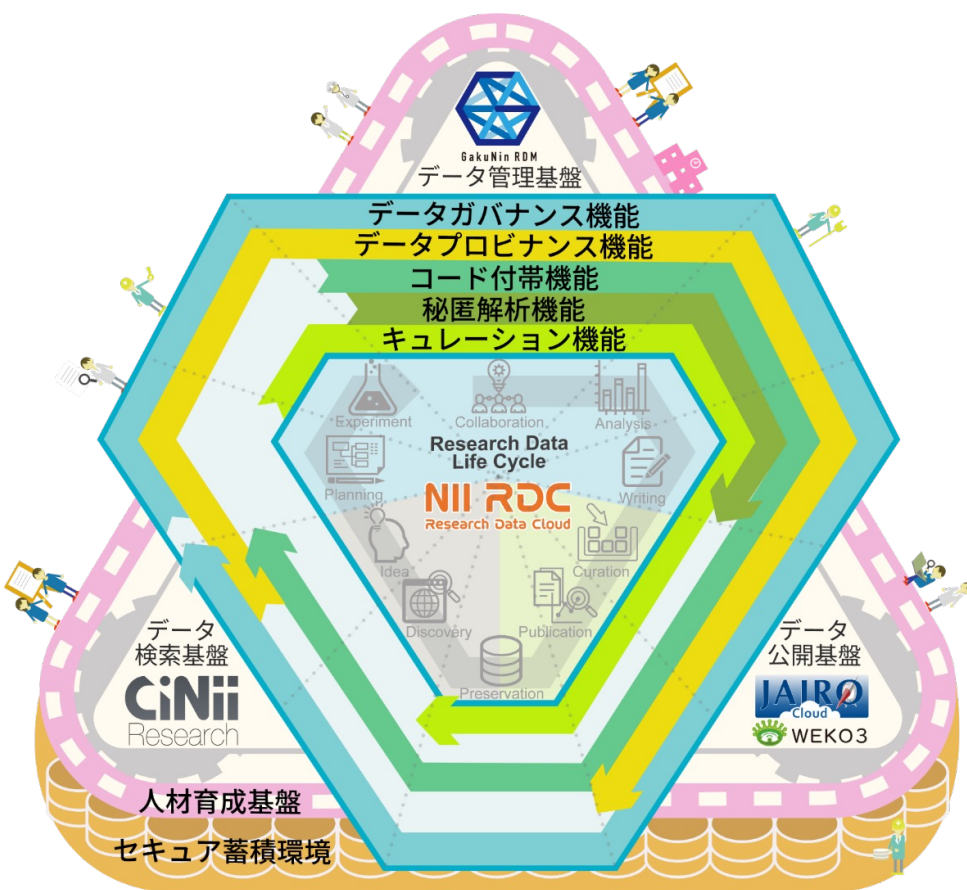
GakuNin RDMと連携する データ解析機能

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

2022/12/14

次世代 NII RDC



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

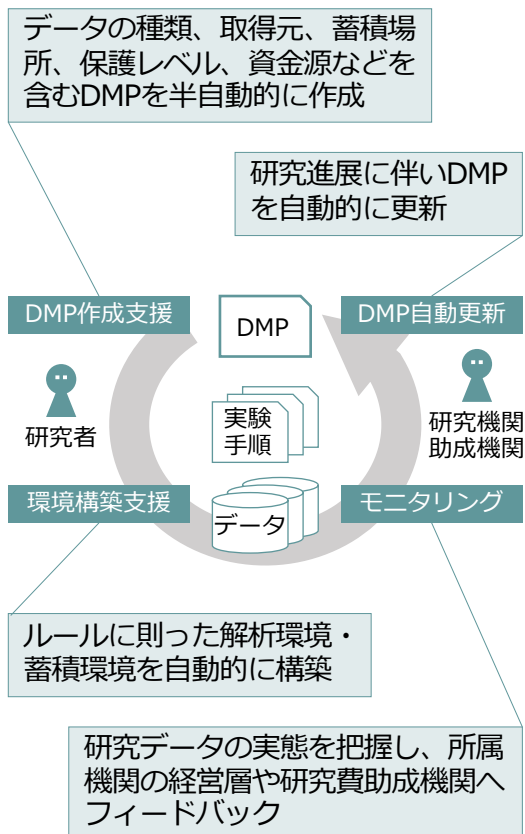
データを利用する研究者がデータの来歴を確認できるとともに、データを提供する研究者が自身のデータの利用状況を確認できる機能。研究不正の疑いから研究者と組織を守り、データ公開への取り組みを後押しします。

コード付帯機能 活用

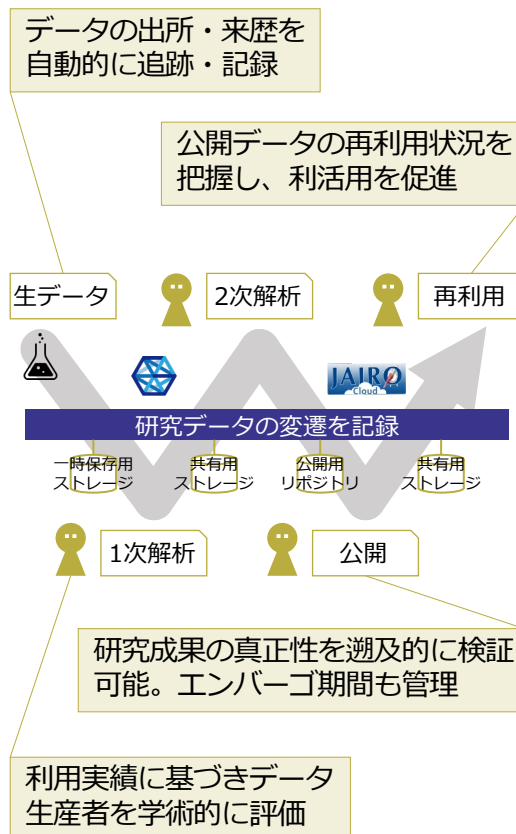
研究者が用いたデータ・プログラム・実行環境定義をまとめて「計算再現パッケージ」として公開・再利用できる機能。先行研究のデータ解析を他の研究者が確実に再現し、発展的な研究を円滑に始められるようにします。

研究再現性を支える3つの機能

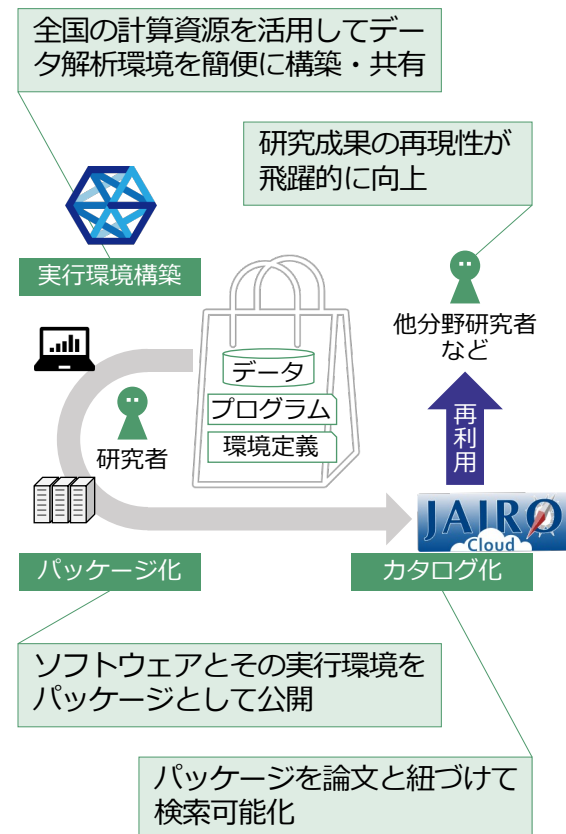
① データガバナンス機能



② データプロビナンス機能



③ コード付帯機能



コード付帯機能: データとコードが循環する世界

1 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

2 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

3 得られた知見を論文にまとめ、
プログラムとともに公開

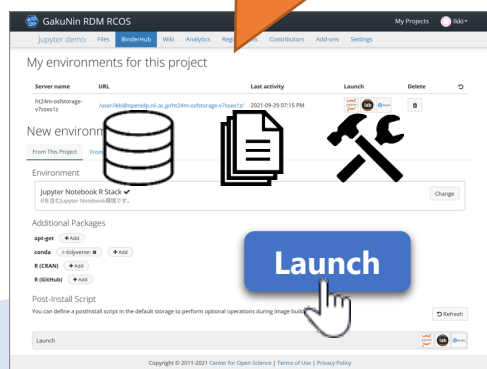
NII RDC
Research Data Cloud



binderhub

コード付帯機能の一部: GakuNin RDM データ解析機能

①環境定義・共有



GakuNin RDM

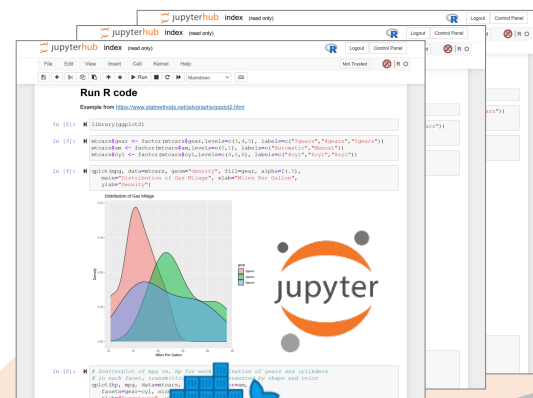
標準ストレージ

機関ストレージ

②取り込み

③書き戻し

④読み書き

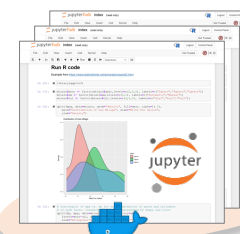


binder

NII



大学等



研究室等

デモ (1/4)

GakuNin RDM RCOS | Reanalysis: X

← → 🔒 🌐 🔍 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 ^ v	サイズ	バージョン	ダウンロ...	最終更新日時 ^ v
 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

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

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。

① 解析環境を構成

② 計算機を選択して起動！

新しい解析環境を作成: <https://binder.cs.rcos.nii.ac.jp>

デモ (3/4)

⑥書き戻しボタン

④ファイルを読み込んで解析

③ファイルがGakuNin RDMからコピーされている

⑤解析結果を ~/result/ に保存

データファイルを読み込む

```
[ ]: 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
```

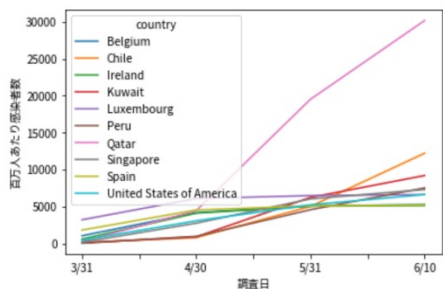
グラフを描く

```
[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")
```

オープンデータを用いて重回帰分析を行う

```
[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])

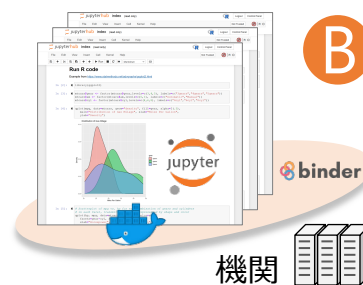
df2 = df2.dropna()
df2 = df2.set_axis(['Cf1', 'SE1', 'Cf2', 'SE2', 'Cf3', 'SE3', 'Cf4', 'SE4'], axis=1)
df2['3/31'] = df2['Cf1'] / df2['SE1']
df2['4/30'] = df2['Cf2'] / df2['SE2']
df2['5/31'] = df2['Cf3'] / df2['SE3']
df2['6/10'] = df2['Cf4'] / df2['SE4']
```



NII RCOS
Research Center for Open Science and Data Platform



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



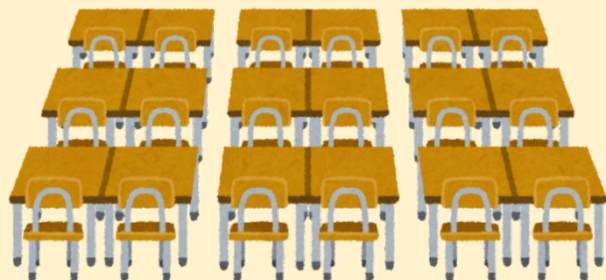
公開・共有

- 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。

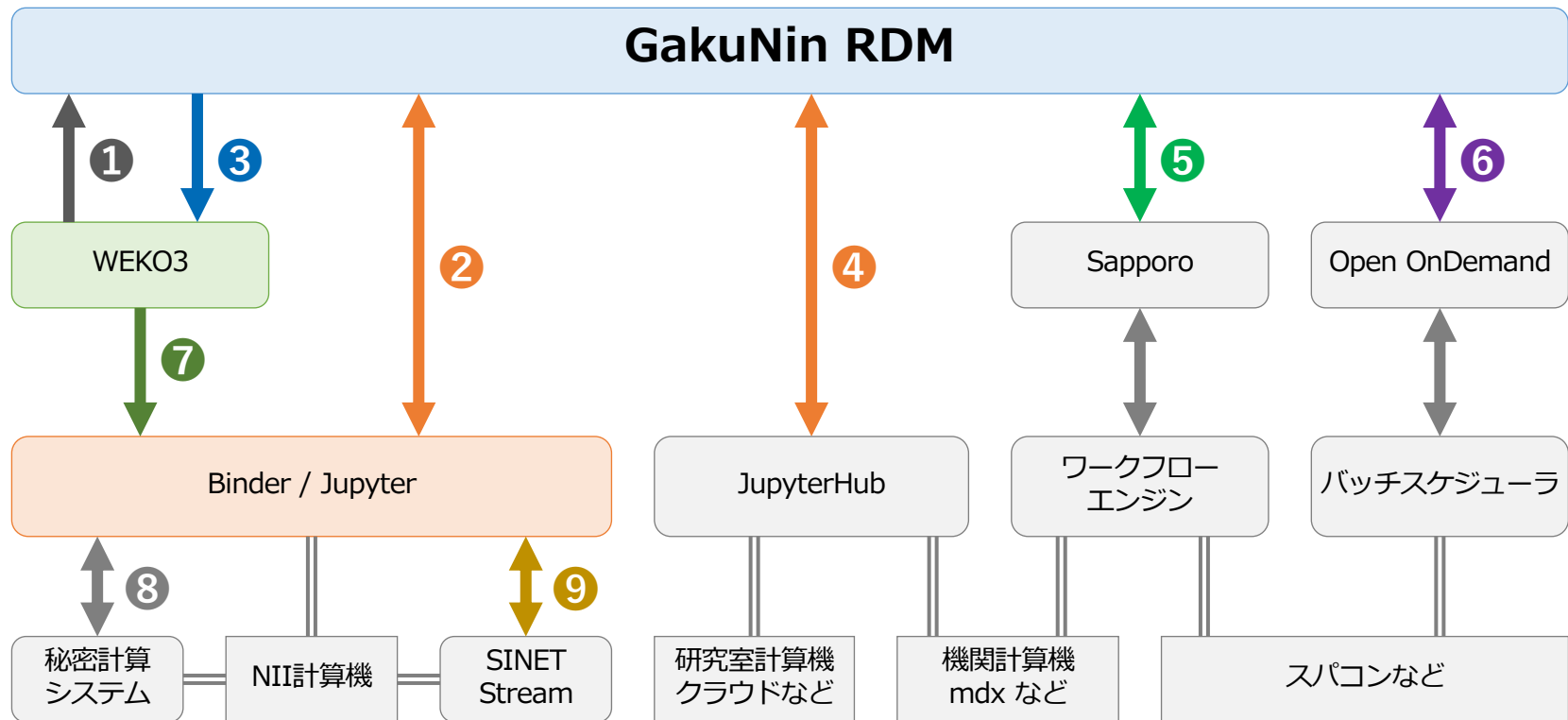
資料

- マニュアル <https://support.rdm.nii.ac.jp/>
- 解説動画 https://youtu.be/_FzOpDTQrBQ

お問い合わせ

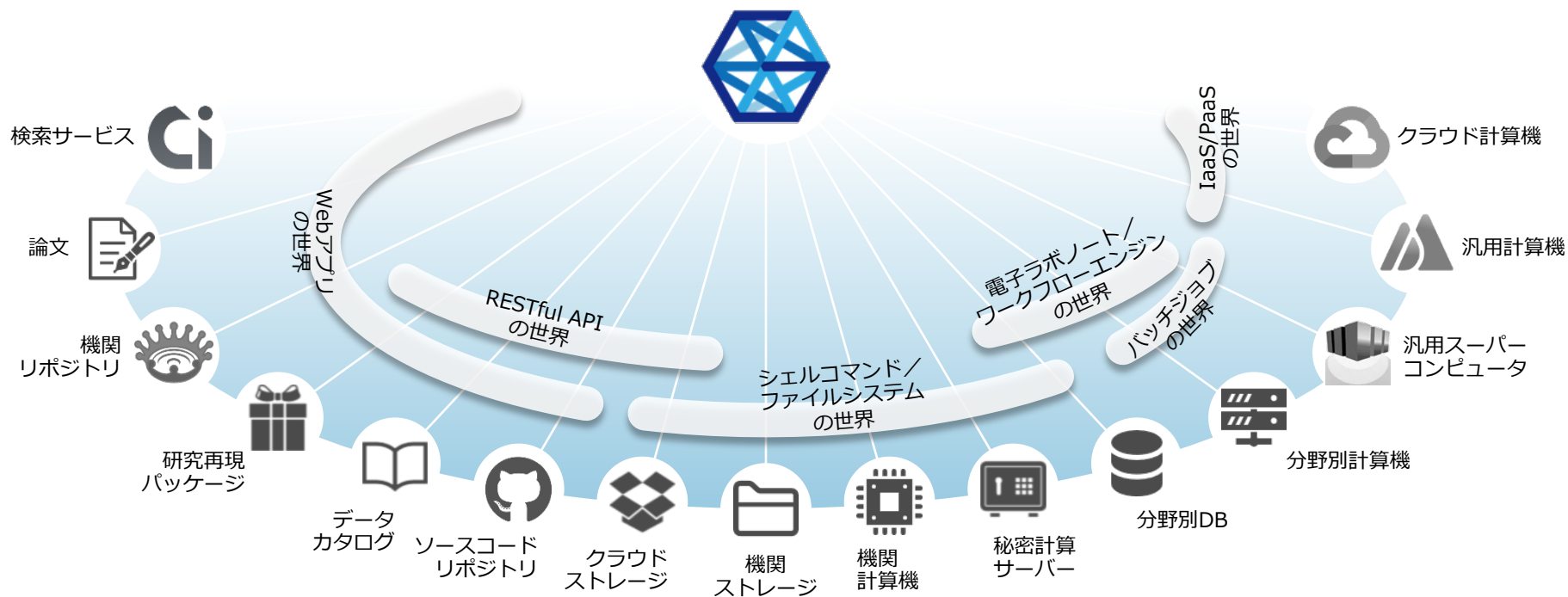
- 開発・活用に関すること→ cs-support@nii.ac.jp
- 導入手続きに関すること→ rdm_support@nii.ac.jp

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**