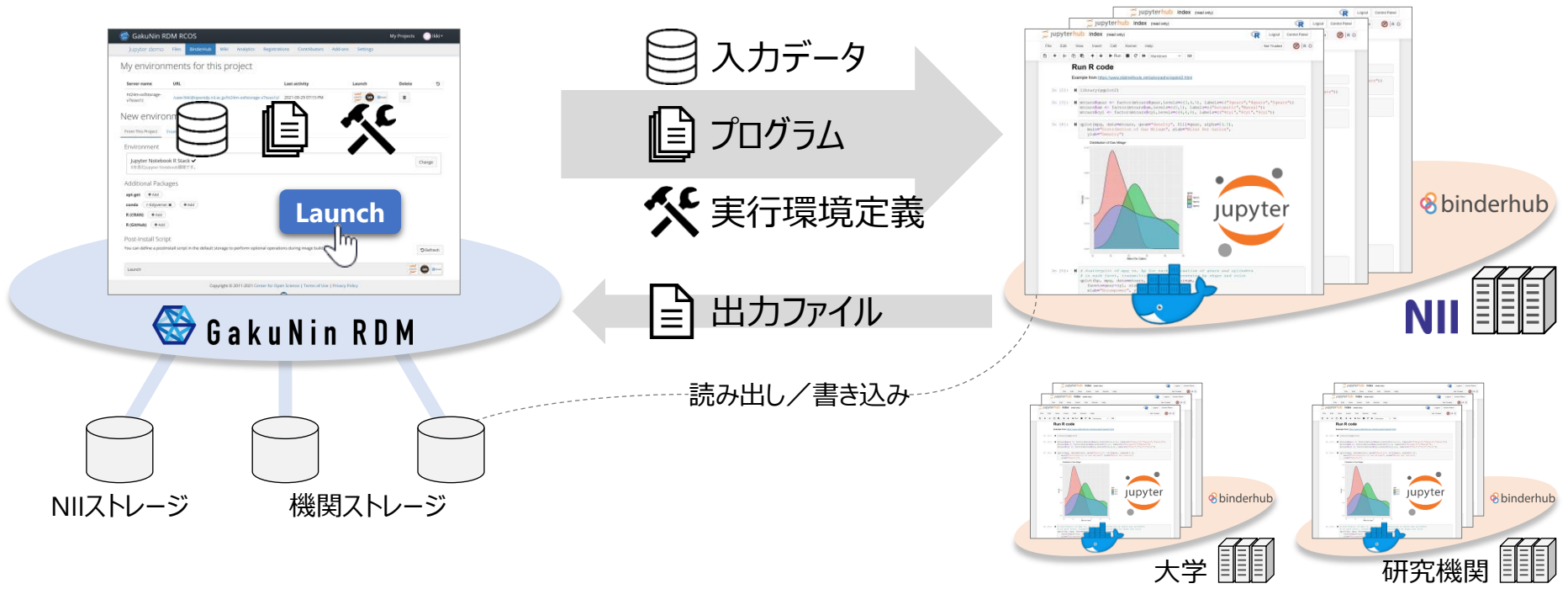


情報処理技術セミナー (クラウド編) 【実習1】 GakuNin RDM データ解析機能の紹介

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
解析基盤チーム
松本正雄

2023年9月1日

GakuNin RDM データ解析機能とは



- 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(後述)

【参照】 データ解析機能の概要: <https://support.rdm.nii.ac.jp/usermanual/DataAnalysis-01/>

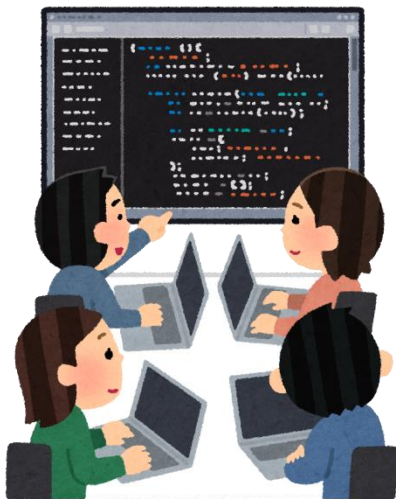
こんなことで困ったことはありませんか？

- 授業のプログラミング演習で
 - 「先生、環境構築ができません」
 - →Macユーザー用のマニュアル作るの忘れてた…。
 - よし、Mac版のマニュアルができたぞ！
 - あれほど言ったのに事前に環境整備してくれなかった。
 - 情報系のスキルの高い学生に待ち時間が発生。
 - スキル、環境の違いによる差をなくしたい。



そんなときは

GakuNin RDMデータ解析機能
ブラウザ操作で環境整備の問題解決！



こんなことで困ったことはありませんか？

- 研究現場で
 - データ解析のプログラムを個々のPCに入れている
 - 信頼できない場所で拾った怪しいインストーラーが蔓延
 - 解析方法や実験データの共有が不十分で結果にズレが…
 - 実験データや解析結果等の情報共有のためのNASのアクセス制御が面倒
 - アカウントの管理も面倒
 - 一人作業になるのでデータの信憑性に疑問が…



そんなときは

GakuNin RDMデータ解析機能！
GakuNin RDMの研究データ管理機能を
そのまま受け継ぎます！



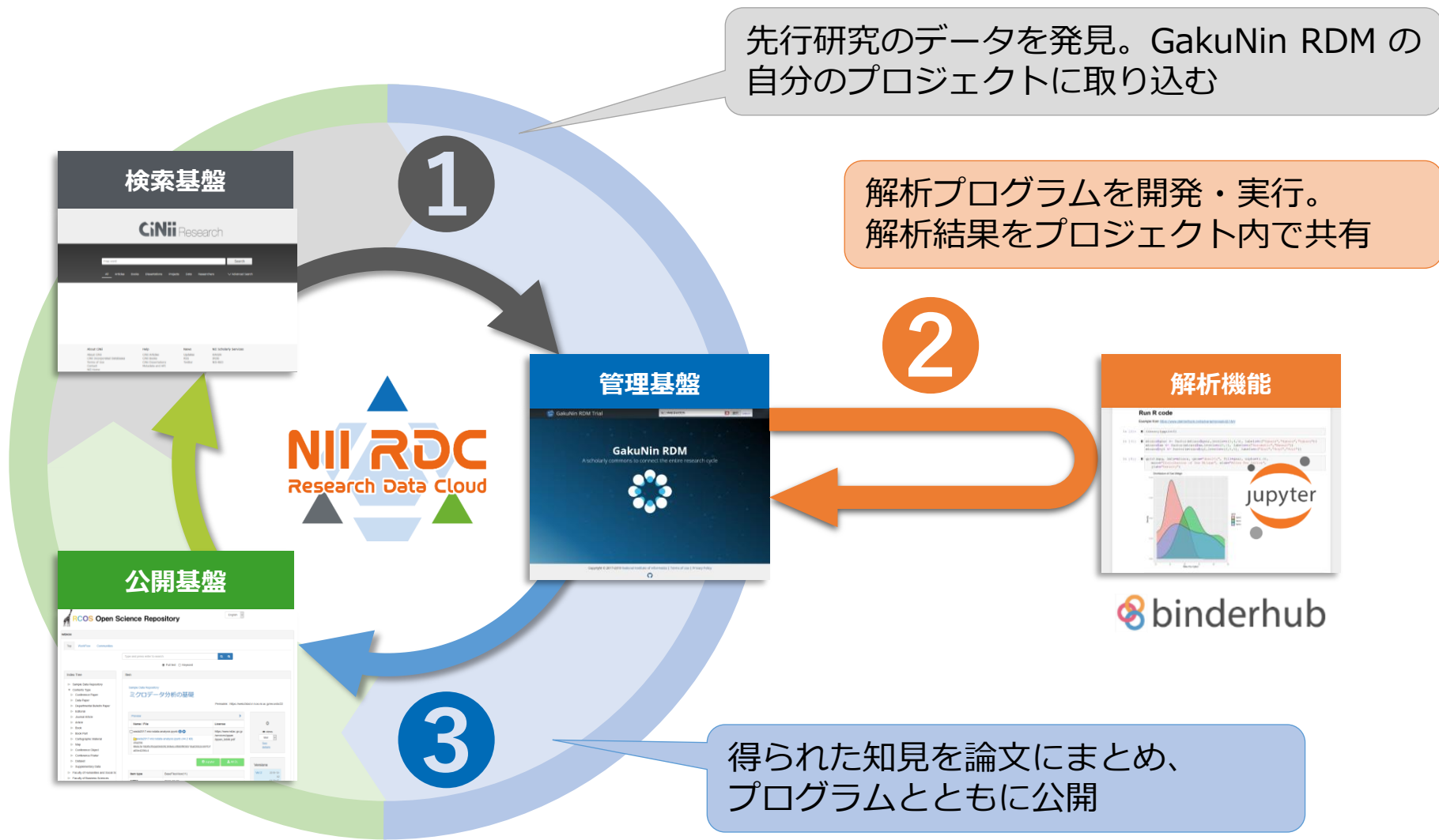
Shibbolethの認証も受け継ぎます。

GakuNin RDMとは

- GakuNin RDM は、国立情報学研究所が提供する研究データ管理システムです。
- 研究チームが持つ多様なデータを一元管理し、効率的で公正なデータ活用をサポートします。
 - 【参照】 GakuNin RDMとは(サポートポータル)
<https://support.rdm.nii.ac.jp/about/>

<p>1</p>  <p>プロジェクト/メンバー管理</p> <p>研究プロジェクトを作成し、メンバーを招待します。デフォルトストレージ、Wiki、ディスカッション機能が使えます。</p> <p>もっと詳しく ></p>	<p>2</p>  <p>ストレージ接続</p> <p>機関が所有するストレージをプロジェクトに接続し、メンバー全員で共有します。外部のクラウドストレージも同様に接続・共有できます。</p> <p>もっと詳しく ></p>	<p>3</p>  <p>証拠保存</p> <p>ある時点でファイルが存在していたこと、改変されていないことを保証します。研究不正の疑いから研究者と組織を守ります。</p> <p>もっと詳しく ></p>
<p>4</p>  <p>データ解析</p> <p>JupyterとRStudioによるデータ解析環境をワンクリックで作成します。他のメンバーが同じ解析環境を複製し、コードを再利用できます。</p> <p>もっと詳しく ></p>	<p>5</p>  <p>Webサービス連携</p> <p>リモート会議やカレンダーなど、使い慣れたさまざまなWebサービスと連携します。(この機能は開発中です)</p> <p>もっと詳しく ></p>	<p>6</p>  <p>リポジトリ連携</p> <p>研究データを機関リポジトリに公開します。(この機能は開発中です)</p> <p>もっと詳しく ></p>

データとコードが循環する世界



GakuNin RDMデータ解析機能と MyBinderとの違い

Mybinder

- 環境は公開
- パッケージを指定したファイルを自分で作り、GitHubに上げてビルド
 - Dockefile、environment.yaml、requirements.txt等のサンプルをGitHubから入手可能
- 10分放置で消滅
- CPU: 1コア
- メモリ: 2GB
- ユーザーストレージ: 不明

GakuNin RDMデータ解析機能

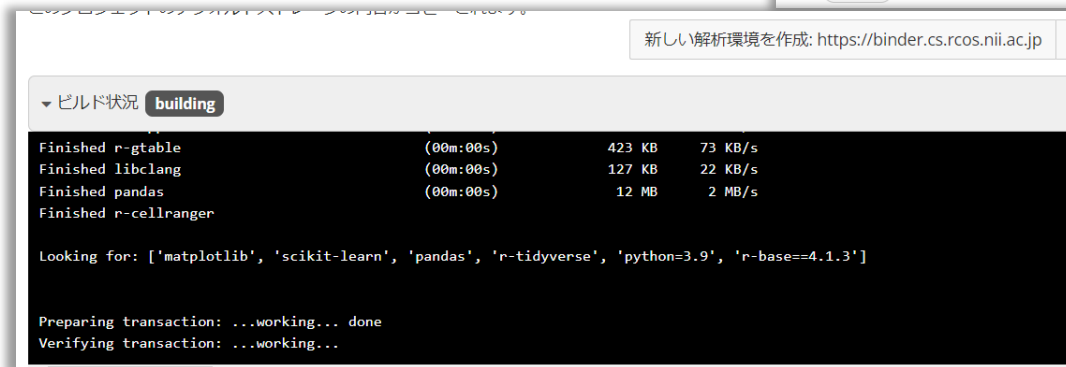
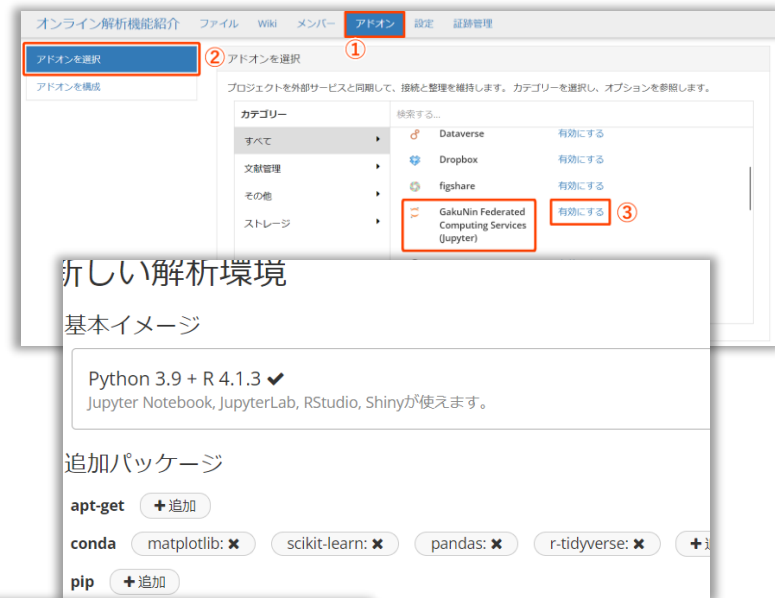
- ビルドした本人しか環境は使えない
- パッケージをブラウザ上で指定し、ボタン一つでビルド
 - Dockefile、environment.yaml、requirements.txt等はGakuNin RDM内に保存、変更履歴も管理
 - 実験環境再構築のパッケージ化を準備中
- 環境は永続的
- CPU: 32コア(共有)
- メモリ: 3GB
- ユーザーストレージ: 10GB

類似サービスとの比較

	GakuNin RDM データ解析機能	GESIS Notebooks	mybinder.org	Google Colab	Microsoft Codalab
対象分野	汎用	社会科学	汎用	主に深層学習	深層学習
提供元	NII (日・学術機関)	GESIS (独・学術機関)	Project Jupyter (任意団体)	Google (米・民間企業)	Microsoft (米・民間企業)
アカウント	学認	GESIS	不要	Google	Codalab
対応言語	Python, R, Shiny	Python, R, Julia	Python, R, Julia	Python, R, Julia, Swift	Python
対応リポジトリ	GakuNin RDM, JDCat (WEKO3), GitHub, Gist, GitLab, Zenodo, Figshare, Hydroshare, Dataverse	GitHub, Gist, GitLab, Zenodo, Figshare, Hydroshare, Dataverse	GitHub, Gist, GitLab, Zenodo, Figshare, Hydroshare, Dataverse	Google Drive, GitHub	?
メモリ CPU ストレージ	3GB 36コア共有 10GB	32GB 2コア 10GB	2GB 1コア ?	13GB 2コア 40GB	?
タイムアウト	半永続 30日不使用で消去	40分	10分	90分 / 12時間	?
インフラ	オンプレ	オンプレ	Google, OVH, Turing Institute	Google	Microsoft

データ解析機能を使い始めるまでの流れ

1. ナビゲーションバーの「アドオン」をクリックします。
2. 「アドオンを選択」をクリックします。
3. GakuNin Federated Computing Services (Jupyter) を有効にします。
4. ナビゲーションバーの「解析」をクリックします。
5. 基本イメージと追加パッケージを指定します。
6. 「新しい環境を作成」ボタンをクリックします。



データ解析機能を使い始めるまでの流れ

1. GakuNin RDMプロジェクトの準備
(新規でも既存でも可)
2. アドオンの追加
 1. GakuNin Federated Computing Services
 1. 【オプション】デフォルト基盤の選択
 2. (必要な場合) 追加のストレージ
3. 「解析」タブに移動
4. 基本イメージの選択
5. パッケージの追加
6. 【オプション】基盤の選択
7. サーバー (仮想環境) のビルド
8. サーバーの実行環境の起動

【参照】データ解析機能の有効化:

<https://support.rdm.nii.ac.jp/usermanual/DataAnalysis-02/>

GakuNin RDM データ解析機能の使い方

データ解析実施の流れ

1. ファイルの準備

1. GakuNin RDMプロジェクトのファイルフォルダに、プロジェクトメンバーに解析環境で使ってほしいデータやプログラムをNII Storageのホーム以下に配置します。
 - ビルドの際、NII Storageのホーム以下がすべてコピーされます。
2. 解析環境のファイル操作でもアップロード可能。
ただし、その場合そのファイルはプロジェクトメンバーに共有されません。

2. 基本イメージの選択

3. パッケージの追加

4. サーバー（仮想環境）のビルド

5. 仮想サーバーの起動

6. 解析プログラムの実行

【参照】 データ解析の始め方:

<https://support.rdm.nii.ac.jp/usermanual/DataAnalysis-03/>

【参照】 フォルダ操作:

<https://support.rdm.nii.ac.jp/usermanual/DataAnalysis-03/>

データ解析実施の流れ (図解)

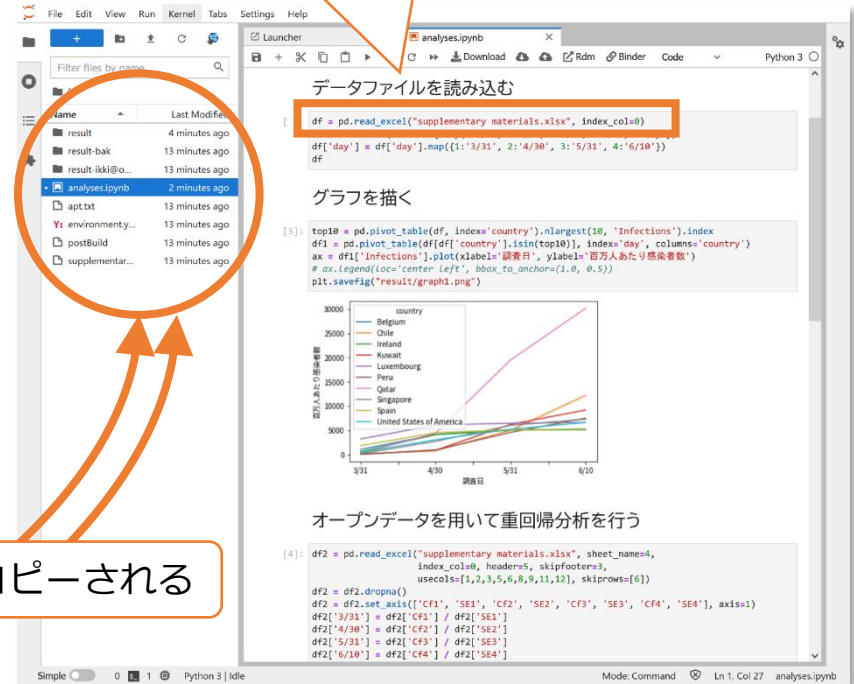
1 分析環境を選んで作成ボタンを押す



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



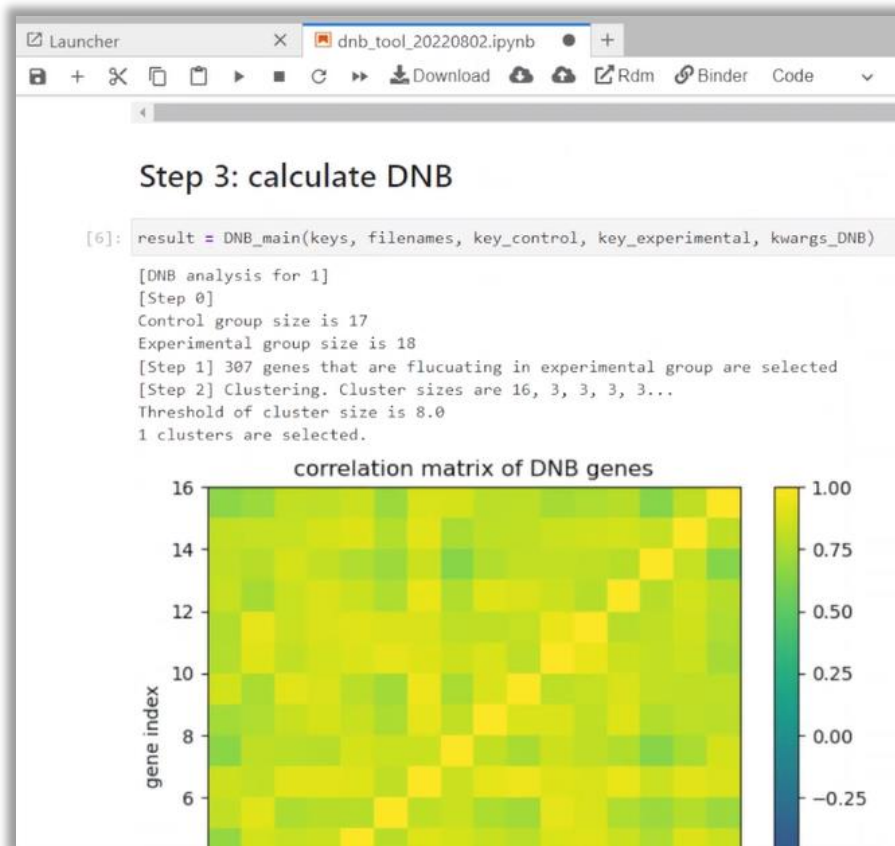
分析結果を管理基盤に書き戻す **4**

データ解析実施の流れ(プログラムの実行)

- Pythonの実行

- サンプル: DNB解析ツール

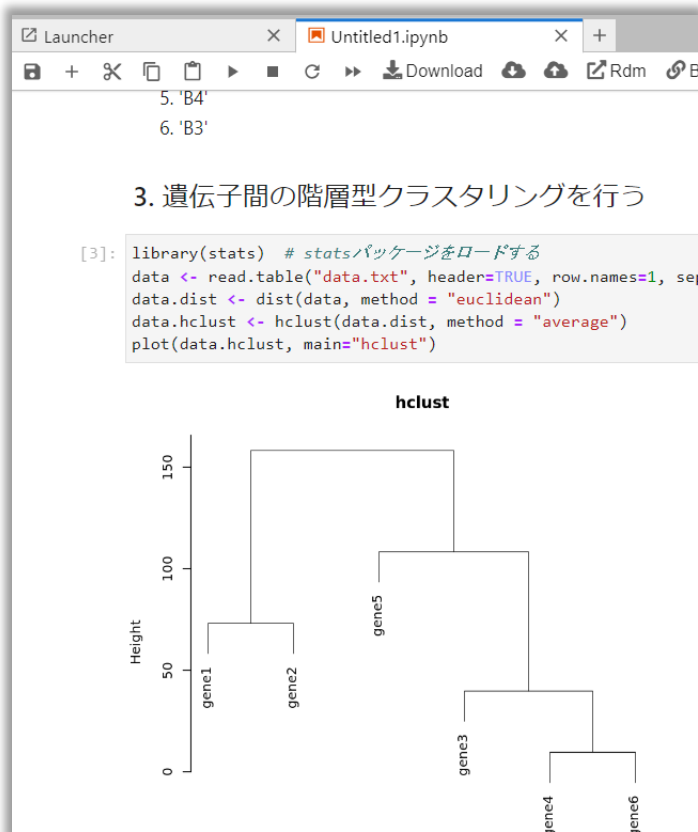
<https://www.sat.t.u-tokyo.ac.jp/moonshot/software/dnb-tool.html>



データ解析実施の流れ(プログラムの実行)




1. Rの実行

- 清水謙多郎先生(東京大学農学部)の遺伝子発現量解析のテキスト
<http://www.bi.a.u-tokyo.ac.jp/~shimizu/genome/web-05/r.html>




データ解析の統合環境(IDE)について

- 解析機能タブから起動できる実行環境は以下の3種類です。

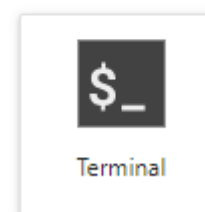
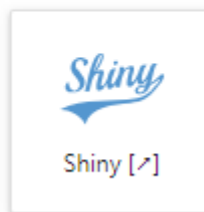
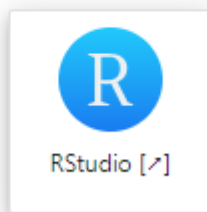
アイコン	名称	特徴
	Jupyter Notebook	シンプルなノートブック形式の開発環境です。
	JupyterLab	Jupyter Notebook の後継となる高機能な統合開発環境です。
	RStudio	R 言語に特化した定番の統合開発環境です。※

- ※基本イメージで「Data Science Notebook」を選択した場合、RStudio は利用できません。

- JupyterLabのLauncherから他のIDEやWebアプリ等、様々なツールを起動することが可能です。

 Notebook

 Other



Jupyter Notebookの実演

jupyterhub dnb_tool_20220802 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

6467 rows x 18 columns

Step 3: calculate DNB

```
In [6]: result = DNB_main(keys, filenames, key_control, key_experimental, kwargs_DNB)
```

[DNB analysis for 1]
 [Step 0]
 Control group size is 17
 Experimental group size is 18
 [Step 1] 307 genes that are fluctuating in experimental group are selected
 [Step 2] Clustering. Cluster sizes are 16, 3, 3, 3, 3...
 Threshold of cluster size is 8.0
 1 clusters are selected.

correlation matrix of DNB genes

jupyterhub Untitled1 (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

```
write.table(tmp, "out2.txt", sep = "\t", append=F, quote=F, row.n
```

3 2 9 9 10 9

'A3' 'A2' 'B3' 'B3' 'B4' 'B3'

3. 遺伝子間の階層型クラスタリングを行う

```
In [3]: library(stats) # statsパッケージをロードする
data <- read.table("data.txt", header=TRUE, row.names=1, sep="\t")
data.dist <- dist(data, method = "euclidean")
data.hclust <- hclust(data.dist, method = "average")
plot(data.hclust, main="hclust")
```

hclust

data.dist
hclust (*, "average")

Rstudioの操作(+Shinyひな形)

The screenshot displays the RStudio interface with a Shiny application in progress. The code editor on the left shows the UI and server logic. The console at the bottom shows the execution of R code to generate a dendrogram. The plot area on the right displays the resulting dendrogram.

```

13 ui <- fluidPage(
14
15   # Application title
16   titlePanel("Old Faithful Geyser Data"),
17
18   # Sidebar with a slider input for number of bins
19   sidebarLayout(
20     sidebarPanel(
21       sliderInput("bins",
22                 "Number of bins:",
23                 min = 1,
24                 max = 50,
25                 value = 30)
26     ),
27
28     # Show a plot of the generated distribution
29     mainPanel(
30       plotOutput("distPlot")
31     )
32   )
33
34
35 # Define server logic required to draw a histogram
36 server <- function(input, output) {
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

```

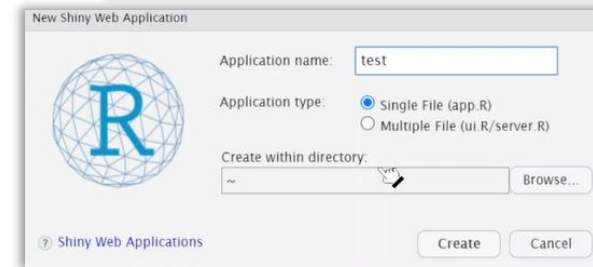
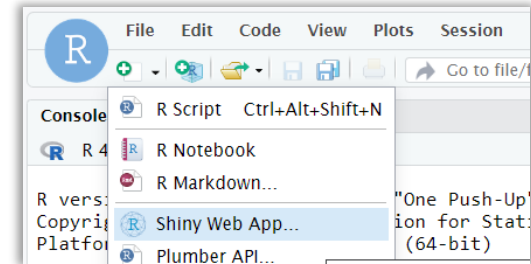
> # 最大発現量を示す組織名を表示
> colnames(data)[max.col(data)]
[1] "A3" "A2" "B3" "B3" "B4" "B3"
> # [行の名前]と「最大発現量を示す組織の番号」と「組織名」を列ベクトル単位で結合し、
> # 結果をtmpに格納
> tmp <- cbind(rownames(data), max.col(data), colnames(data)[max.col(data)])
> # tmpの中身をout2.txtというファイル名で保存。
> write.table(tmp, "out2.txt", sep = "\t", append=F, quote=F, row.names=F)
> # statsパッケージをロードする
> library(stats)
> # ファイルの読み込み
> data <- read.table("data.txt", header=TRUE, row.names=1, sep="\t")
> # 遺伝子間の距離を計算。デフォルトはユークリッド距離
> data.dist <- dist(data, method = "euclidean")
> # 平均距離法を適用
> data.hclust <- hclust(data.dist, method = "average")
> # 結果を表示
> plot(data.hclust, main="hclust")
connected to your session in progress, last started 2023-Aug-31 08:49:00 UTC (1 hour ago)

```

The plot area shows a dendrogram titled "hclust" with a y-axis labeled "Height" ranging from 0 to 150. The x-axis labels are "gene1", "gene2", "gene3", "gene4", "gene5", and "gene6". The dendrogram shows the hierarchical clustering of these genes based on their pairwise distances.

RstudioからShinyのひな形作成

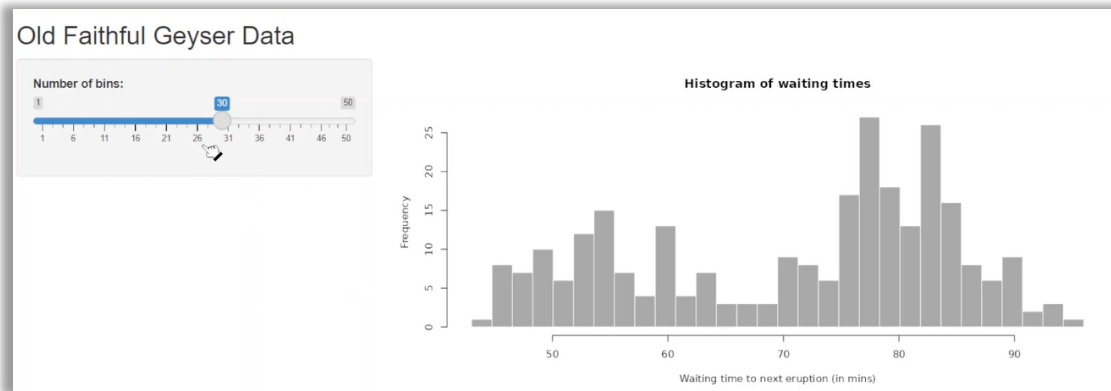
- 「File」 → 「New」 または 「+」 アイコン (New File)から「Shiny Web App...」を選択します。
- 「Application name」 に目的のアプリ名を、ディレクトリを作製する場所 (親ディレクトリ) を「Browse...」をクリックして選択します。
- Shinyを起動します。
- アプリケーション名となっているディレクトリをクリックします。



Index of /

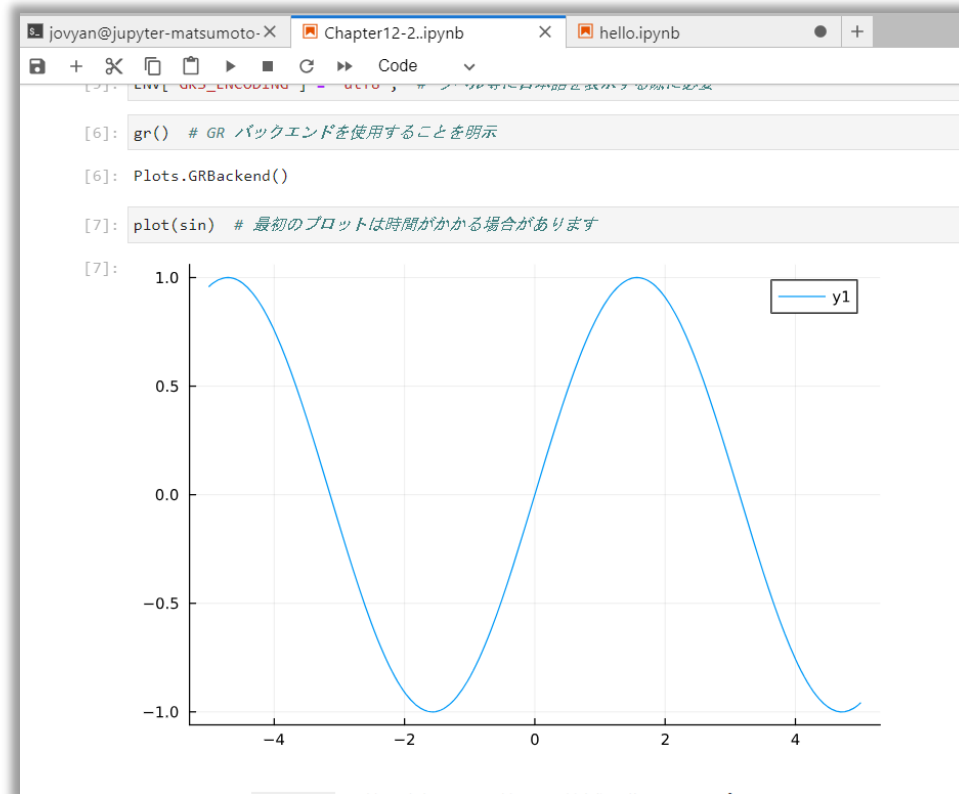
- dnb_tool/
- R/
- test/

- ShinyのWebアプリが起動します。



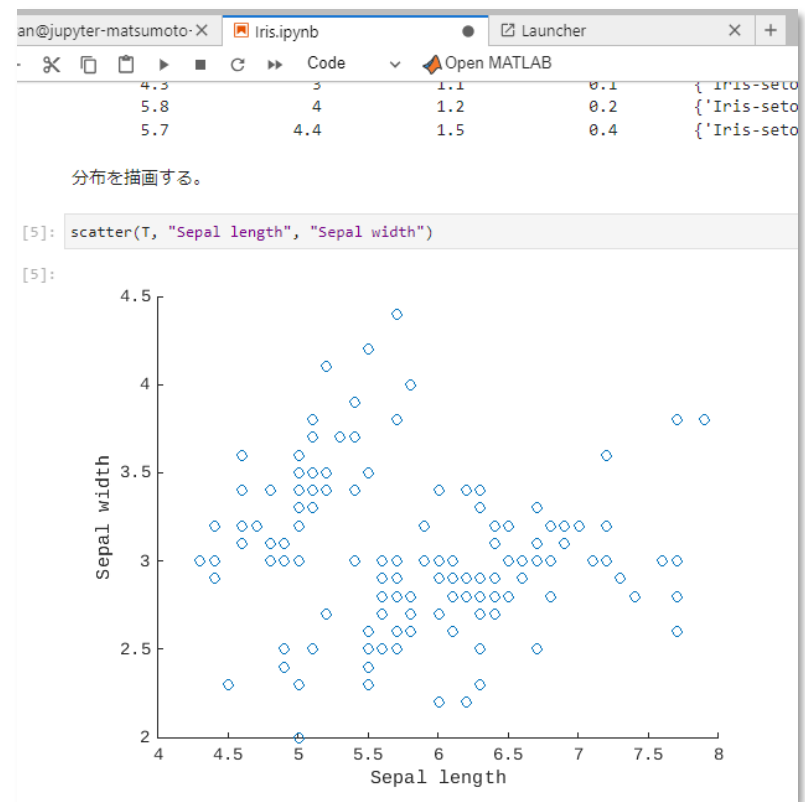
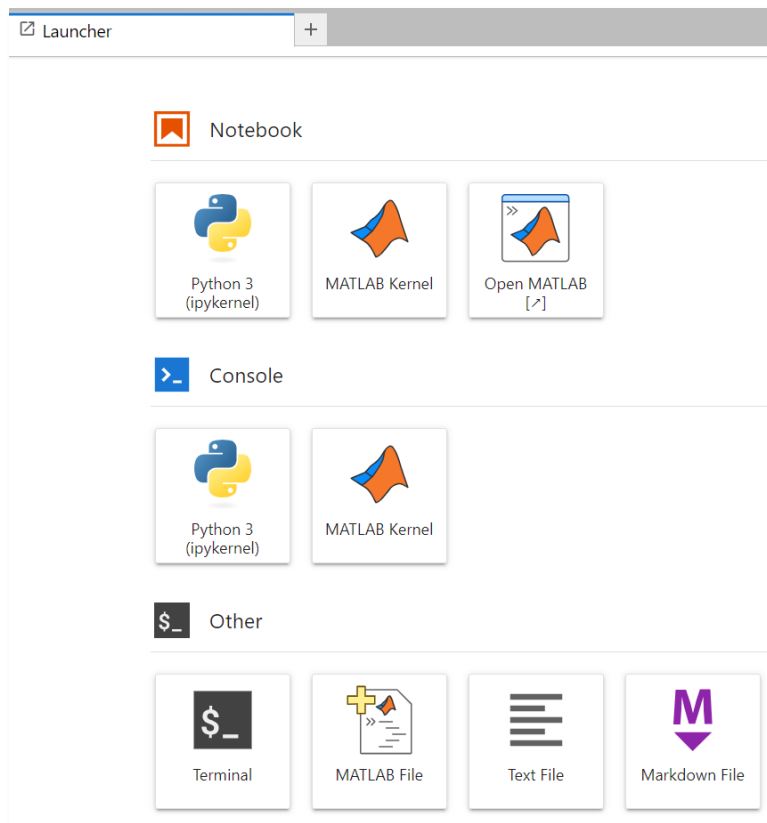
Juliaの実行

- Juliaの実行環境構築は基本イメージに「Data Science Notebook」を選択してください



MATLAB対応予定

- サービス開始に向け準備中



便利機能

- GakuNin RDMへの解析結果の同期
- 拡張ストレージの使い方
- ターミナル
- ファイル操作

GakuNin RDMへの解析結果の同期

- 同期方法
 - Jupyter Notebookの場合
 - JupyterLabの場合
 - Rstudioの場合
- 同期用ディレクトリの解説
 1. 共有範囲
 2. ディレクトリ名の規則
 3. 新規仮想サーバーへの書き出し

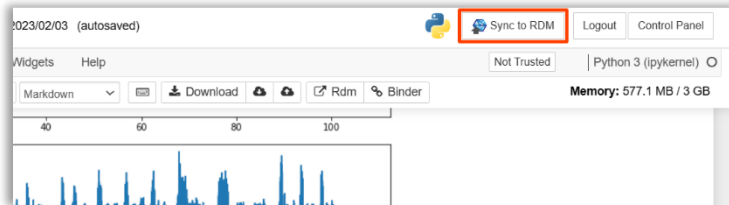
同期方法

1. 準備(共通)

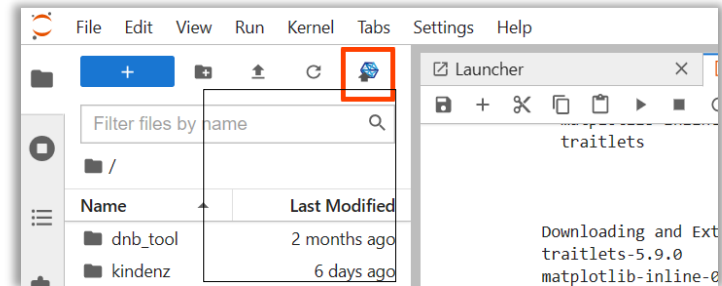
1. ホームディレクトリに result という名前のフォルダを作成します。
2. アップロードしたいファイルを result フォルダに置きます。

2. 同期操作(IDEごとに操作方法が異なります)

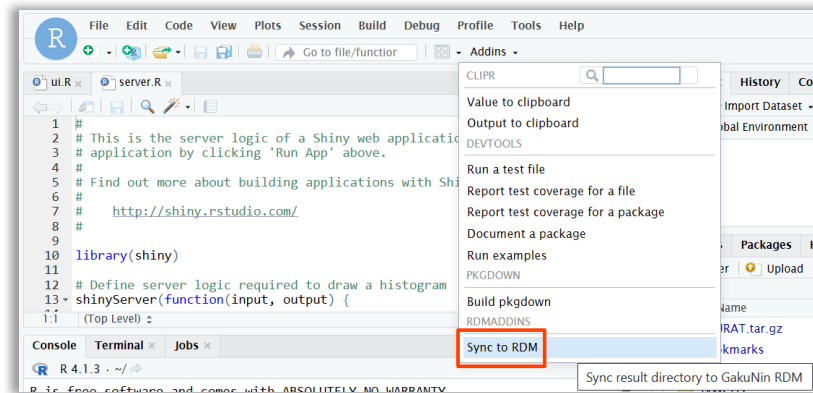
- Jupyter Notebook では、画面上部にある「Sync to RDM」ボタンをクリックします。



- JupyterLab では、左ペインの上部にある GakuNin RDM アイコンをクリックします。



- RStudio では、ツールバーの「Addins」プルダウンメニューの「Sync to RDM」を選択します。



ディレクトリ名の規則

- 解析結果はプロジェクトメンバーに共有されます。
 - 変更の権限などはプロジェクトの設定に従います。
- 同期フォルダ名は解析環境毎に規則性のある固有の名前が付きます。
 - 命名規則は「result-ユーザー名-日付-環境名」です。
 - 例えば以下のような名前になります。
[result-matsumoto@nii.ac.jp-20230901-ab123-osfstorage-abcdefgh](#)
- 新たな仮想サーバーを作る際は、同期したディレクトリとファイルもそこに書き込まれます。

拡張ストレージの使い方

- NFS(標準)
- 外部ストレージ(ユーザーによる設定)
- マウント先 (ツリー図)
- ワーキングディレクトリに表示させる

NFS(標準)

- 1つのアカウントに10GB付与されています。
- マウント先は /mnt/user です。(全ユーザー同じ)
- 解析環境を削除しても消えません。
データを永続化できます。
- 自分が作成したすべての解析環境から
読み書きできます。
- このディレクトリで他のユーザーとの
ファイルの受け渡しはできません。

外部ストレージ (ユーザー設定)

- アドオンにより外部ストレージを追加できます。
- アドオンのマニュアル
<https://support.rdm.nii.ac.jp/usermanual/FileHandling-03/>
 - アドオン全体のマニュアルですが、ストレージについて詳しく記載されています。
- /mnt/rdm/ 以下にマウントされています。
 - Google Drive を追加した場合、/mnt/rdm/googledrive
 - OneDrive を追加した場合、/mnt/rdm/onedrive
- 外部ストレージはGakuNin RDMにおいて、プロジェクトメンバー全員がプロジェクトで与えられた権限でアクセス可能です。
 - データの機密性に留意して親ディレクトリなどを設定してください。

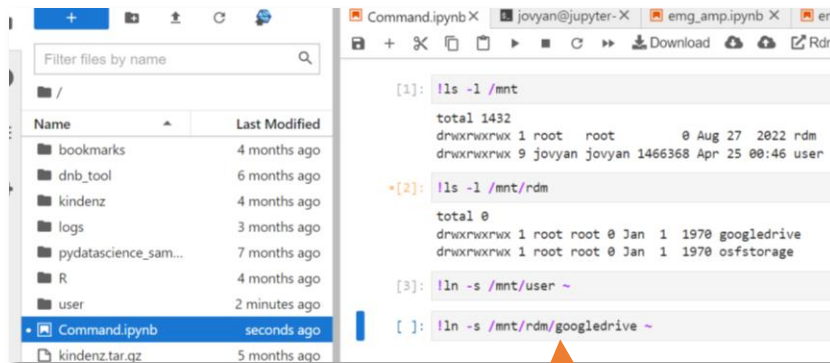
マウント先 (ツリー)

- Google DriveとOneDriveを追加した場合のディレクトリ構成
 - /mnt
 - user : NFS
 - rdm/
 - googledrive : Google Drive
 - onedrive : OneDrive
 - osfstorage : GakuNin RDM FS

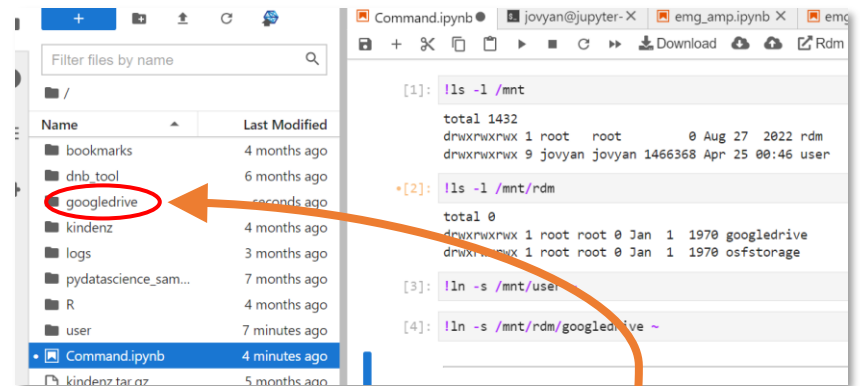
ワーキングディレクトリに表示させる

- NFS、外部ストレージなどの外部ストレージはIDEのワーキングディレクトリに表示されません。
- シンボリックリンクを使ってワーキングディレクトリに拡張ストレージを表示させることができます。

• Before



• After

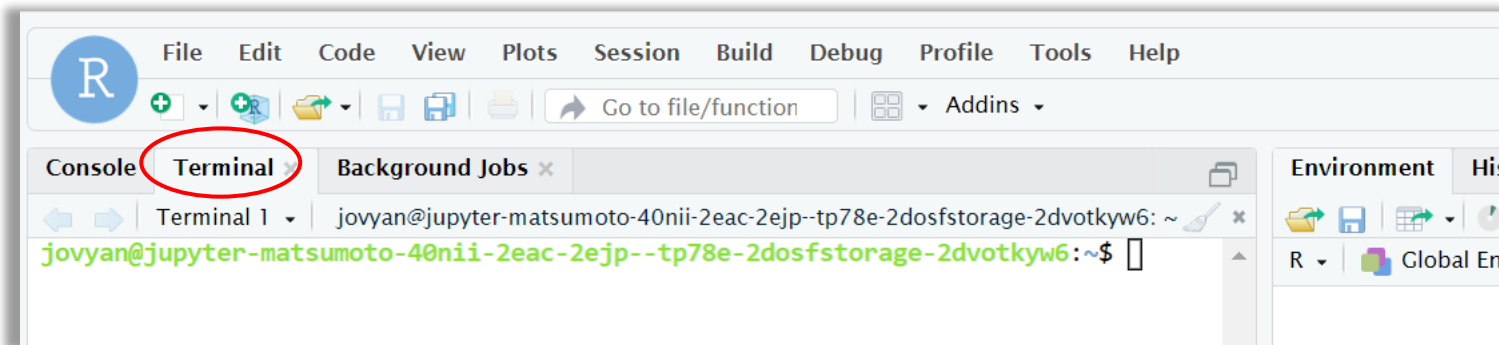
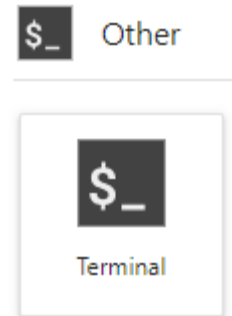
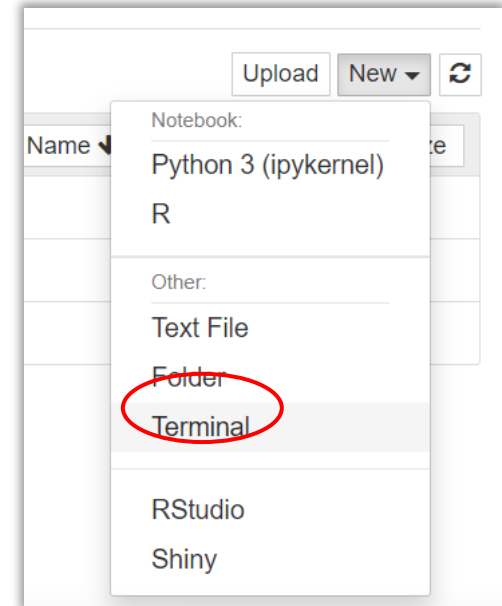


以下のコマンドを実行

`ln -s /mnt/rdm/googledrive ~`

ターミナル (端末)

- Linuxコマンドが実行できる標準の機能です。
- 起動方法
 - Jupyter Notebook では、右上方の「New」をクリックしてプルダウンメニューを表示し、「Terminal」を選択します。
 - JupyterLab では、Launcher の Other のグループにある「Terminal」のアイコンをクリックするとターミナルが開きます。
 - RStudio では、左ペインの「Terminal」タブをクリックします。



ファイル操作の実例

- Google Driveを介して他の解析環境にファイルをコピーする

シナリオ：ホームディレクトリ下のdir1ディレクトリの中にあるfile1を他の環境にコピーしたい

□ コピーしたいファイルがある環境のシェル

```
$ cp ~/dir1/file1 /mnt/rdm/googledrive
```

□ コピー先の環境のシェル

```
$ mkdir -p dir1  
$ mv /mnt/rdm/googledrive/dir1/file1 ~/dir1
```

※cpでも良い。

- NIIのNFSを介して他の解析環境にディレクトリをコピーする

シナリオ：ホームディレクトリ下のdir1ディレクトリ(シンボリックリンクあり)を他の環境にコピーしたい

□ コピーしたいファイルがある環境のシェル

```
$ cd ~/  
$ tar czf /mnt/user/dir1.tar.gz dir1
```

□ コピー先の環境のシェル

```
$ cd ~/  
$ tar xpf /mnt/user/dir.tar.gz
```

※ディレクトリ内にシンボリックリンクがない場合は `cp -r` でも良い

※コマンドラインで標準出力と無名パイプ「|」の組み合わせが使えないことがあります。
`tar cf - dir1 | tar xpf - -C /mnt/user` といった方法でコピーはできません。

Jupyterによるファイル操作(推奨)

Terminalは履歴が消えてしまうことがあるため、コマンド実行用のJupyter Notebook (.ipynb)ファイルを作って、セルの中でコマンドを実行することをおすすめします。

コマンドの先頭に「!」を付けるとLinuxコマンドを実行します。

- Google Driveを介して他の解析環境にファイルをコピーする

シナリオ：ホームディレクトリ下のdir1ディレクトリの中にあるfile1を他の環境にコピーしたい

コピーしたいファイルがある環境のNotebookのセル

```
!cp ~/dir1/file1 /mnt/rdm/googledrive
```

コピー先の環境のNotebookのセル

```
!mv /mnt/rdm/googledrive/dir1/file1 ~/
```

- NIIのNFSを介して他の解析環境にディレクトリをコピーする

シナリオ：ホームディレクトリ下のdir1ディレクトリ(シンボリックリンクあり)を他の環境にコピーしたい

コピーしたいファイルがある環境のNotebookのセル

```
!cd ~/  
!tar czf /mnt/user/dir1.tar.gz dir1
```

コピー先の環境のNotebookのセル

```
!cd ~/  
!tar xpf /mnt/user/dir.tar.gz
```

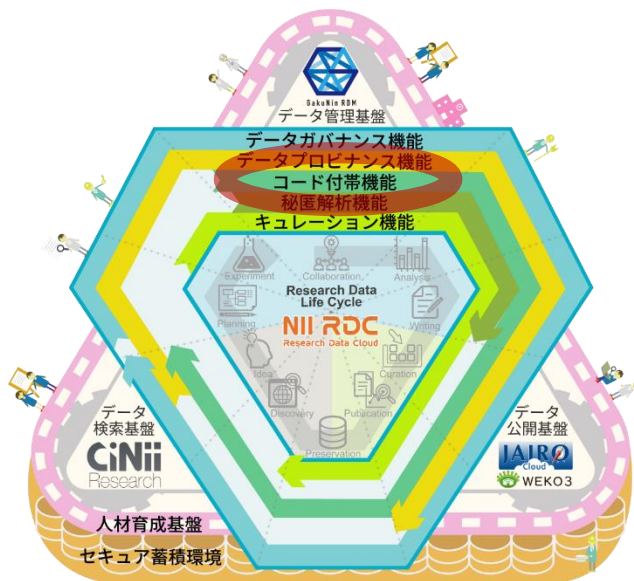
ファイル操作時の注意点

- 通常のLinuxの端末エミュレーターとほぼ同じですが、たまに違う挙動をすることがあります。その際は知恵を絞ってください。
 - 先述の名前なしパイプとtarの組み合わせ等
- Windowsのようにゴミ箱があるわけではないので、ファイルの消去時は注意してください。
- GakuNin RDMプロジェクトのフォルダに解析機能から書き込みをすると、不具合が発生することがあります。
同期機能をお使いください。

GakuNin RDM データ解析機能の構成

施設情報

- 設置場所 = 柏分館 (RDC内)
 - 大学の都合で止まることあり(長期停止時は別クラウドに退避)
- 運営
 - 基本はリモート、物理的作業時以外も定期的に担当者が訪問
- 通信経路
 - 解析環境とSINETは一つ橋の本部経由でつながっている



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。

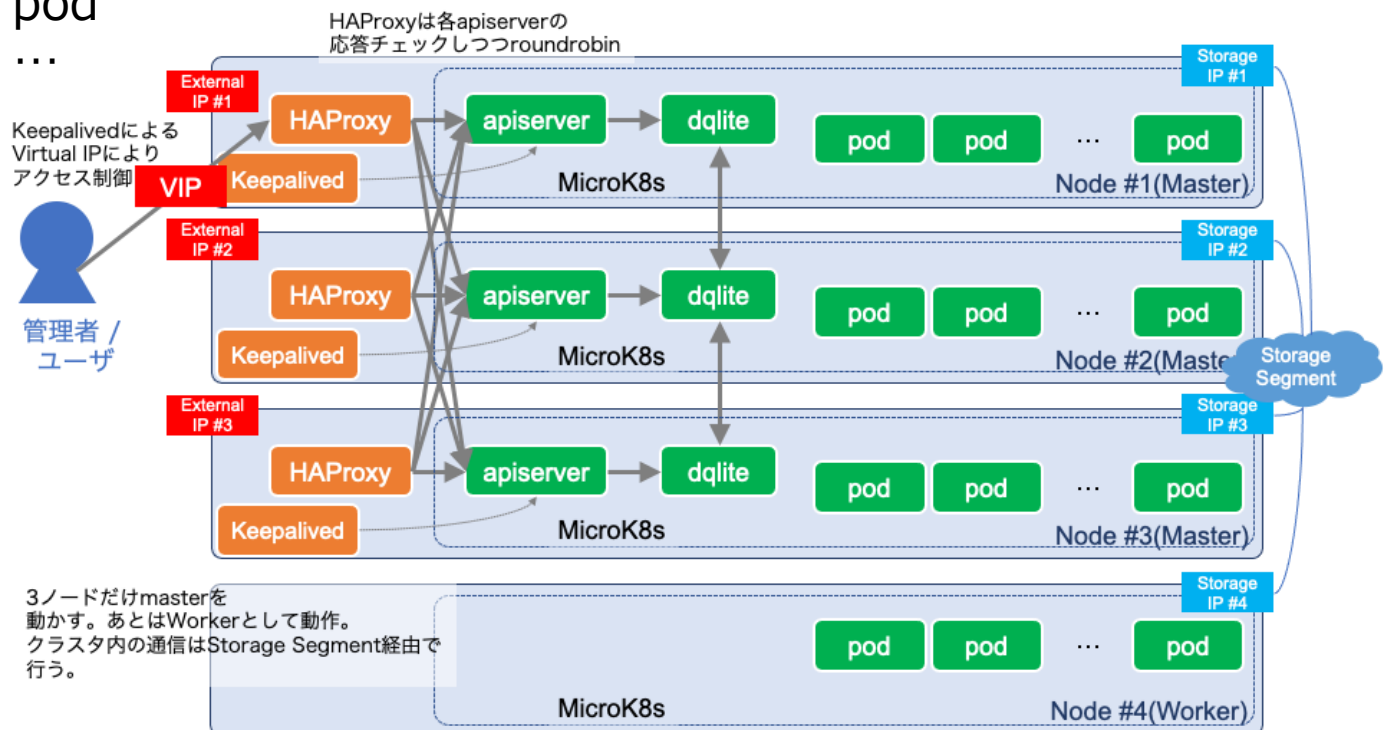
解析機能のクラスタ構成の例

- マスター(兼ワーカー)

- Keepalived
- HAProxy
 - MicroK8s
 - pod
 - pod
 - pod
 - ...

- ワーカー

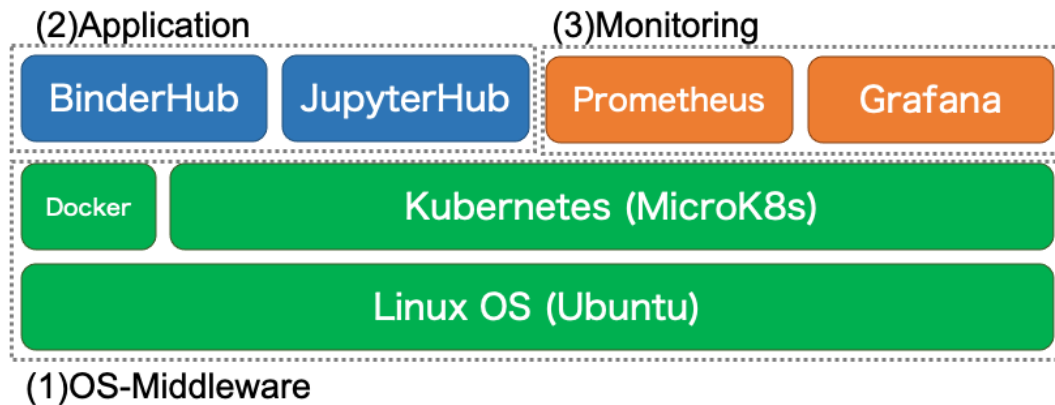
- MicroK8s
 - pod
 - pod
 - pod
 - ...



BinderHub VMの構成

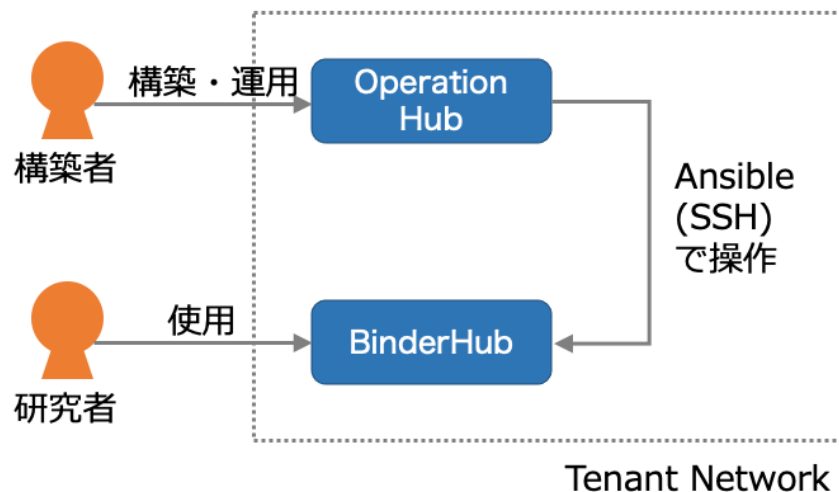
BinderHub VMの構成

- OS (Linux)
 - Docker
 - Kubernetes(MicroK8s)
 - アプリケーション
 - JupyterHub
 - BinderHub
 - 監視系
 - Prometheus
 - Grafana



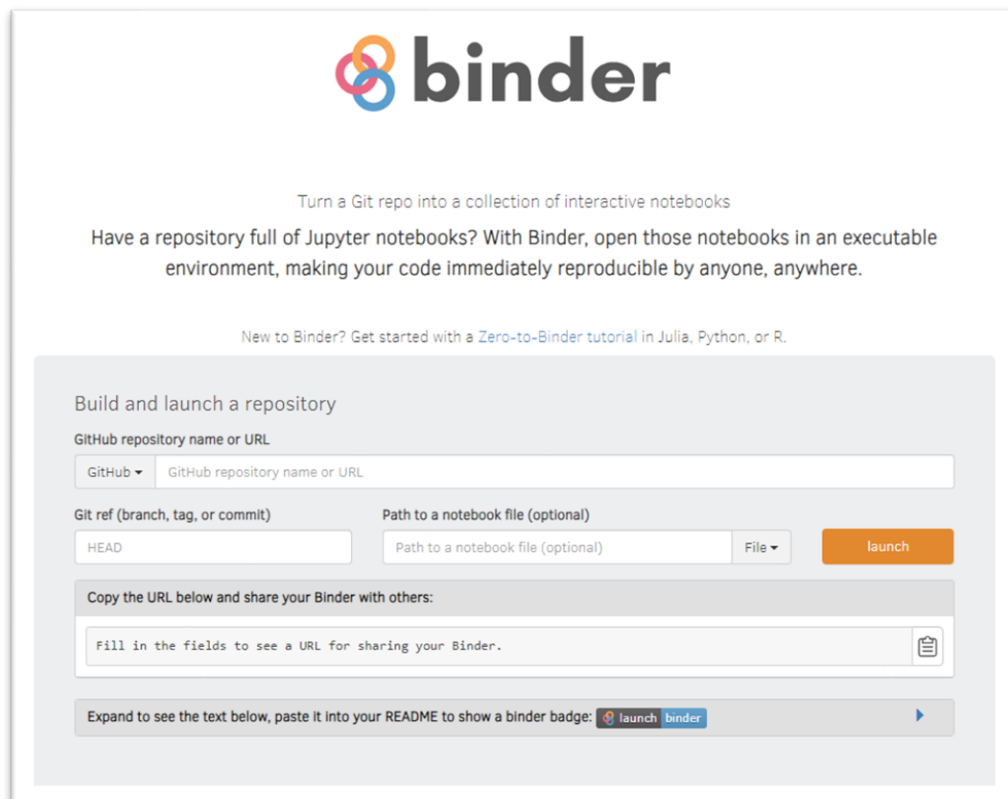
GakuNin RDMデータ解析機能の オペレーション(実務形態)

- OperationHub
 - BinderHub構築オペレーション用VM : [OperationHub](#)
 - 実務はJupyter Notebookで遂行する。
 - SSH経由のデプロイメントツール : [Ansible](#)
- BinderHub
 - クラスタ基盤 : [MicroK8s](#) (helmで管理)
 - オンラインデータ解析機能 : [BinderHub\(RCOS版\)](#)
 - 監視系 : [Prometheus](#)



GakuNin RDM データ解析機能で使っているBinderHubについて

- システムのベースはCS-binderHub
 - <https://github.com/RCOSDP/CS-binderhub>



environment.yaml の例

```
channels:
  - conda-forge
dependencies:
  - r-base=4.2.2
  - r-tidyverse
```

environment.yamlか
Dockerfileを準備し、Git
リポジトリに置いてビルド

My Binder等、無料のお
試しサイトが多数存在
<https://mybinder.org/>

標準的なBinderHubとの違い

- 必要なパッケージを指定し、ボタン一つでビルド
 - environment.yml、 requirements.txt、 Dockerfile等、ビルドに必要なファイルをGitHubに置く必要なし
- GakuNin RDMに解析結果を保存
 - 「GakuNin RDMへの解析結果の同期」のページ参照
- アカウント毎に各環境と共有可能な10GBのNFSディレクトリが利用可能
 - シェルコマンドによる操作が必要
 - クラウドやS3(互換含む)のストレージを追加することも可能

GakuNin RDM

データ解析機能の利用申請

利用申請の流れ

導入手続き

1. サポートポータル確認
<https://support.rdm.nii.ac.jp/>
2. コミュニティサポート登録
<https://community.nii.ac.jp/>
3. GakuNin RDM利用申請
<https://community.nii.ac.jp/s/article/guide-grdm>

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。



資料

- GakuNin RDM導入手続き
<https://support.rdm.nii.ac.jp/participate/>

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

参考リンク

- **オープンフォーラム2023 GakuNin RDMトラック**

- **管理基盤、解析機能の現状と今後について**

https://www.nii.ac.jp/openforum/2023/day1_rcos-kanrikiban.html

- 研究データ管理基盤GakuNin RDMの現状とユースケース
- GakuNinRDMサービス利用申請について

- **GakuNin RDMサポートポータル**

- <https://support.rdm.nii.ac.jp/>

- ユーザーマニュアル

- <https://support.rdm.nii.ac.jp/usermanual/>

- GakuNin RDMページからもリンクされています。



BinderHubユースケース 創出ユーザ支援

BinderHubユースケース創出ユーザ支援

- こんな要望はありませんか？
 - もっと広大なメモリ空間が必要
 - 3GBよりもっと多く！
 - もっと高速な大容量ストレージが必要
 - 計算結果がいつぱいたまるんです。
 - GPUを使いたい
 - AWS？それともmdx？いっそのことオンプレミス？
 - TPUも使いたい
 - Google Cloud？
- 自分でNIIと同様、いや、それ以上の解析基盤を作ってGakuNin RDMデータ解析機能で使いたい！
 - でも構築大変そう…

BinderHubユースケース創出ユーザ支援

- そんなあなたをサポートする準備ができています。
 - 構築前サポート
 - セミナーの開催
 - マニュアルの作成
 - tljh-repo2docker による JupyterHub 環境の構築のマニュアル
 - Binder 構築のマニュアル 一式
 - 質疑応答と提案
 - 構築時サポート
 - 個別支援
 - 構築後サポート
 - 状況の変化に伴うアップデートの提案
 - 機能変更の希望の聞き取り
 - 運用中の不具合の聞き取り

BinderHubユースケース創出ユーザ支援

どうぞお気軽にお問い合わせください。

解析基盤チーム

cs-support@nii.ac.jp

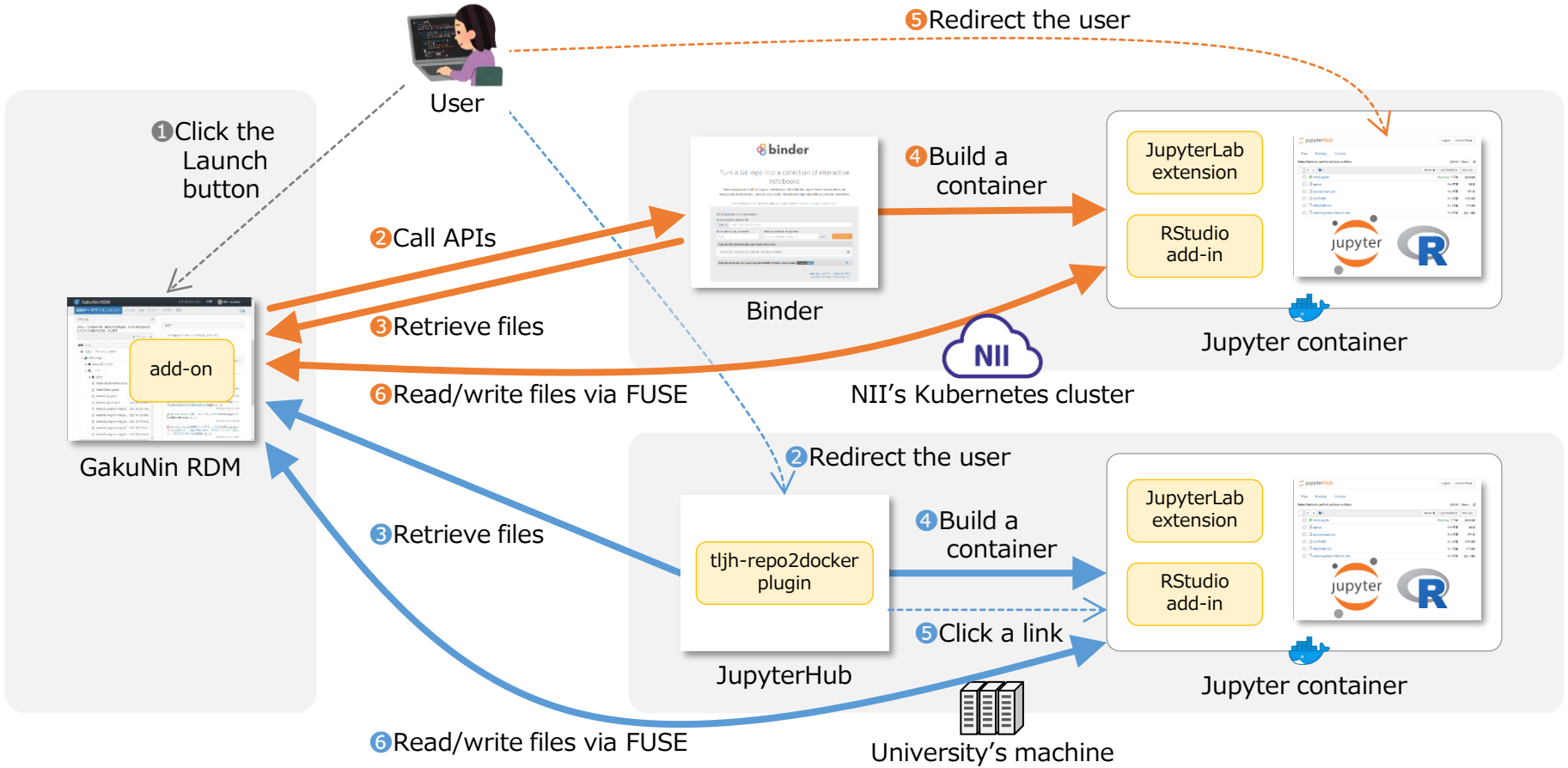
RCOS

rcos@nii.ac.jp

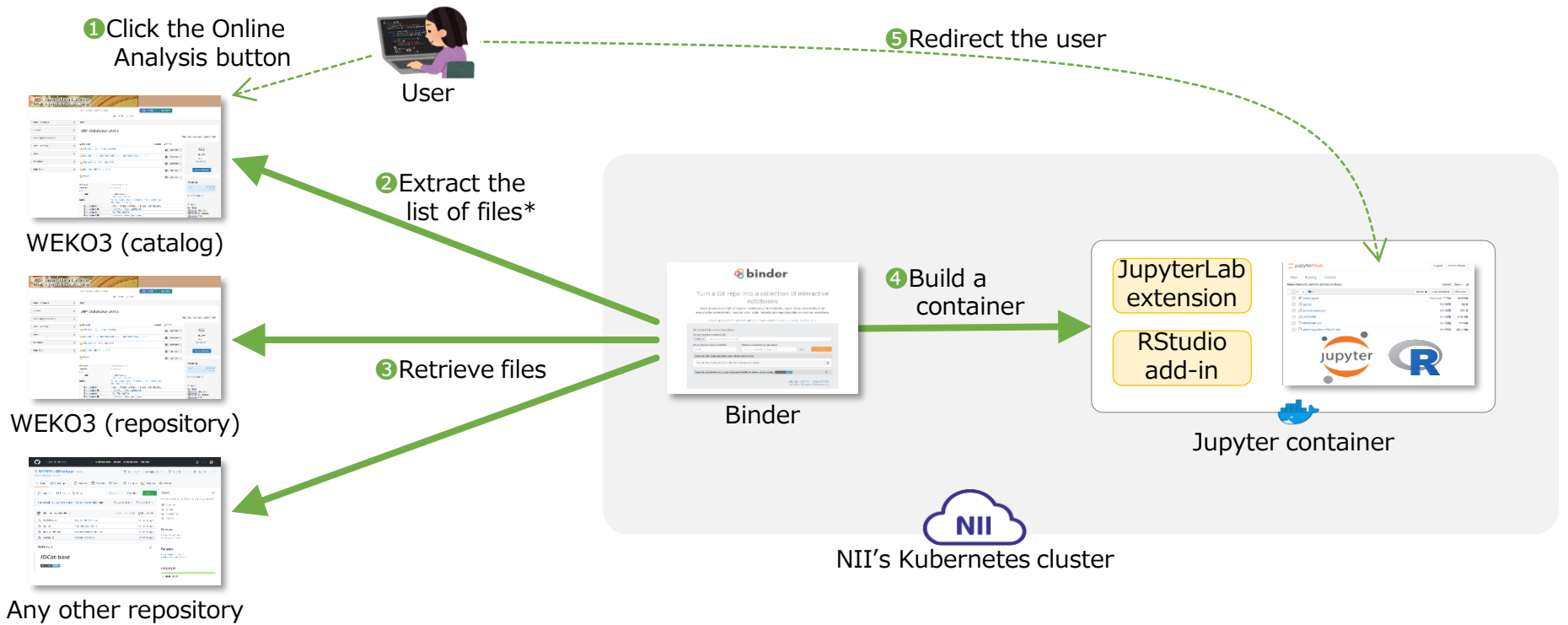
cs-support@nii.ac.jp

Appendix

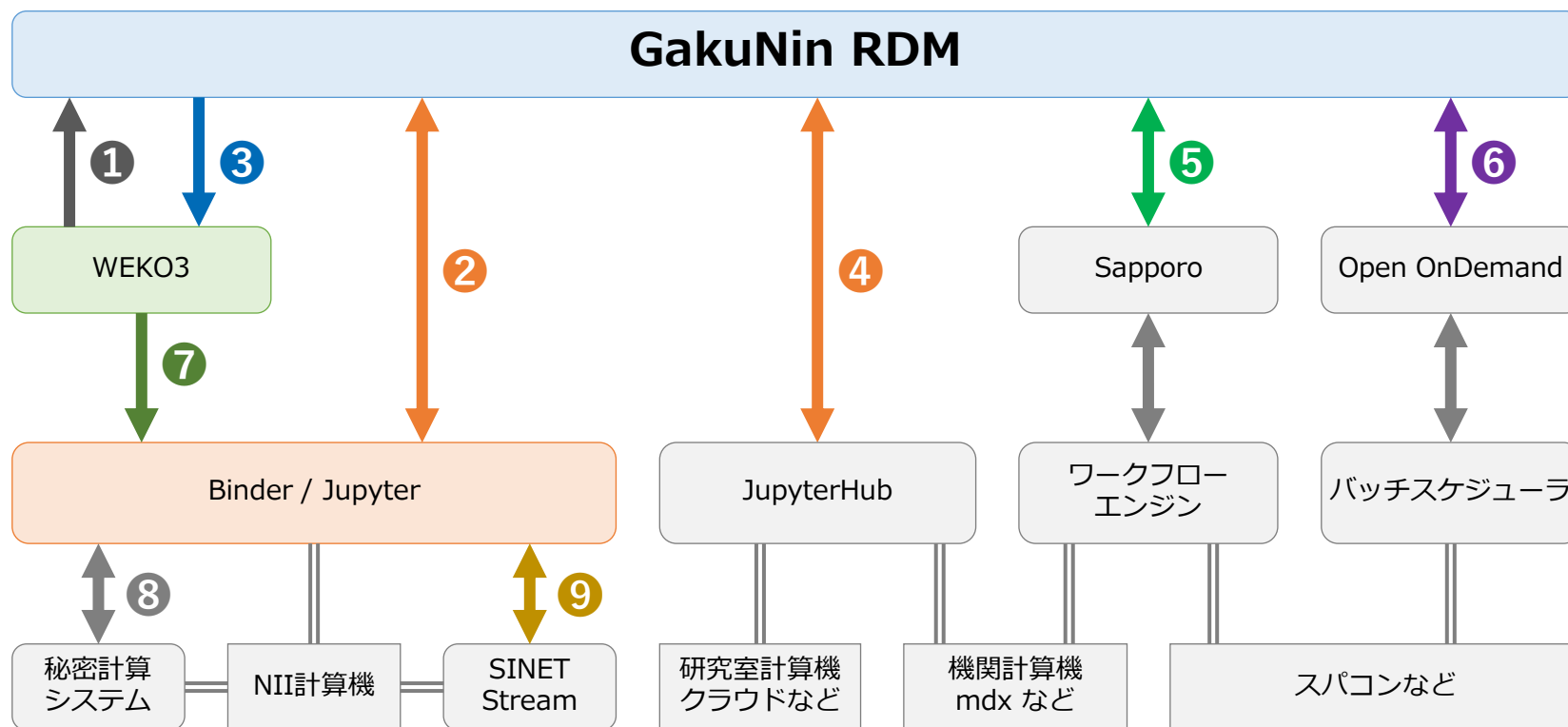
システム構成 (GRDMデータ解析機能)



システム構成 (JDCat分析ツール)



さまざまな計算機システムと連携する解析基盤



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	開発中
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	計画中
⑥ 外部バッチスケジューラ連携機能	Open OnDemand で GRDM のデータをスパコンに転送	開発完了
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	設計中
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	開発中

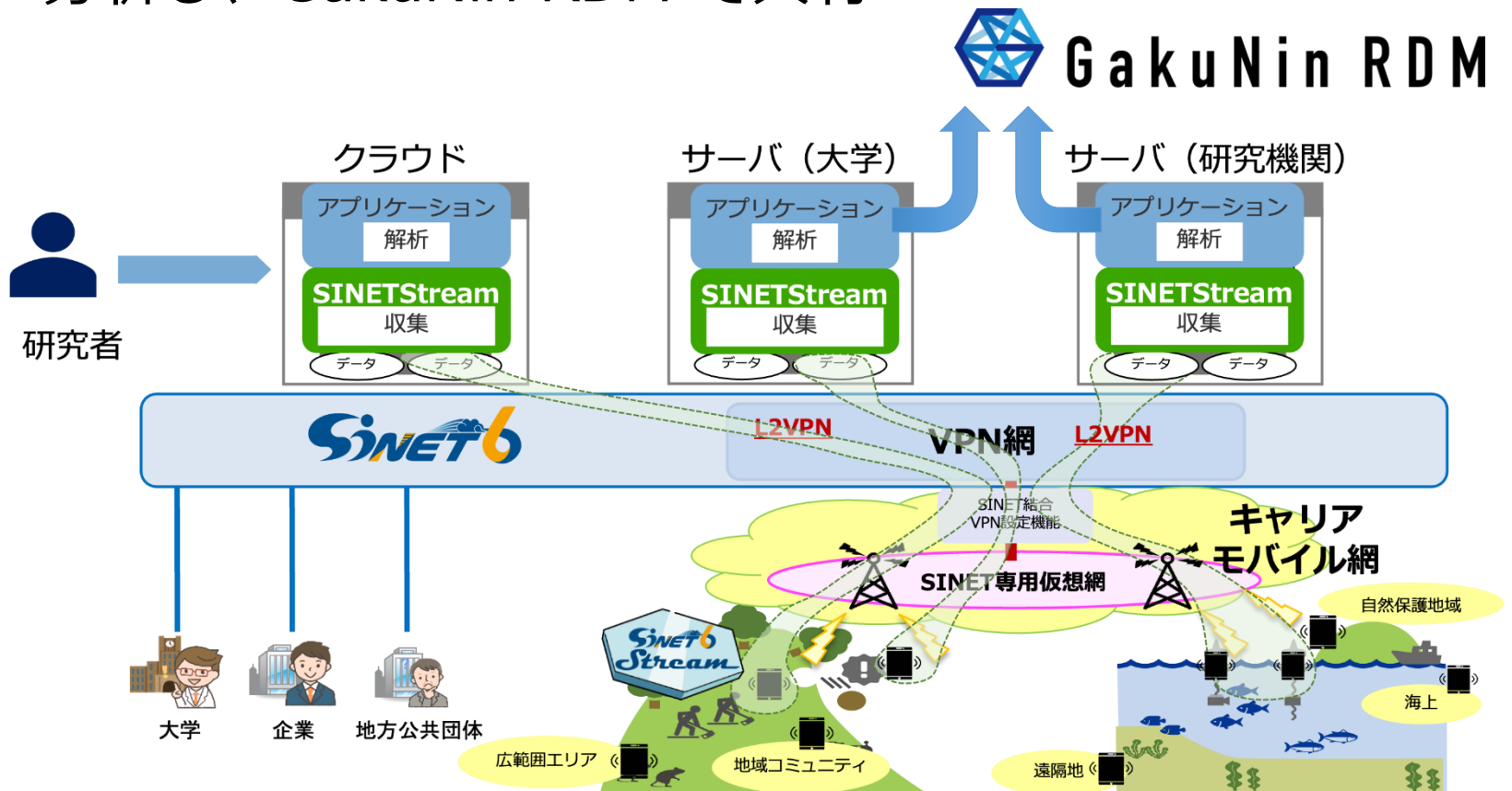
現行サービス／外部連携の展開

- JDCat 分析ツール
 - JSPS 人文学・社会科学データインフラ事業で開発。2023年度からは NII のサービスとして継続
 - 中身は GRDM データ解析機能で使ってるのと同じ BinderHub
- 東大 mdx
 - 外部計算機持ち込み機能の連携先として利用可

導線	想定ユーザー	2021年度	2022年度	2023年度
単体 で試用 https://binder.cs.rcos.nii.ac.jp	人社データインフラ事業の テスト協力教員と学生			
JDCat の一機能として利用 https://jdcats.jp	社会科学分野の研究者・学生等			
GRDM の一機能として利用 https://rdm.nii.ac.jp	GRDM利用機関の研究者等			

応用例: 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



解析基盤に係る用語の整理

管理者向け 名称	コード付帯機能				秘匿解析機能		
目的	汎用	汎用	HPC連携		IoT連携	汎用	
利用者向け 名称	計算再現 パッケージ 機能	GakuNin RDM データ解析機能				秘密計算ト ライアル	秘密計算LA
開発案件名	計算再現 パッケージ	BinderHub / JupyterHub	バッチスケ ジューラ連 携	ワークフ ローエンジ ン連携	SINETStrea m連携	NTT共同研 究	秘密計算 サービス
実装技術	GRDM + WEKO	GRDM + Binder + Jupyter	Open OnDemand	Jupyter + Sapporo	Jupyter + SINETStrea m	算師	析秘
開発元	RCOS	Project Jupyter	Ohio Supercomp uter Center	遺伝研	クラウド基 盤研究開発 センター	NTT社会情 報研究所	NTTコミュ ニケーショ ンズ

用語解説（詳細）

- GakuNin RDMデータ解析機能、データ解析機能
 - GakuNin RDM内でデータ解析のための仮想環境をボタン一つで構築できる機能
- GakuNin Federated Computing Services
 - データ解析機能をGakuNin RDMに追加するプラグインの名前
- オンライン分析機能
 - [JDCat](#)(人文学・社会科学総合データカタログ)の文書DBを分析する機能
- コード付帯機能
 - データ・プログラム・実行環境定義をまとめて公開・再利用できる機能
 - データ解析機能、オンライン分析機能や、解析基盤が提供するその他の機能の総称
- 解析基盤
 - コード付帯機能を提供するコンピューターシステム、サービス群
- 解析環境、サーバー
 - データ解析機能等で構築したJupyter等が動作する仮想サーバー
- NII Research Data Cloud (RDC)
 - NIIの研究用クラウド。解析基盤はこのRDC内にある