NII RCOS
Research Center for Open Science and Data Platform

学術データの オープンアクセス

西岡　千文（国立情報学研究所）

第1回東海地区学術データ基盤セミナー

2022年12月2日

cnishioka@nii.ac.jp

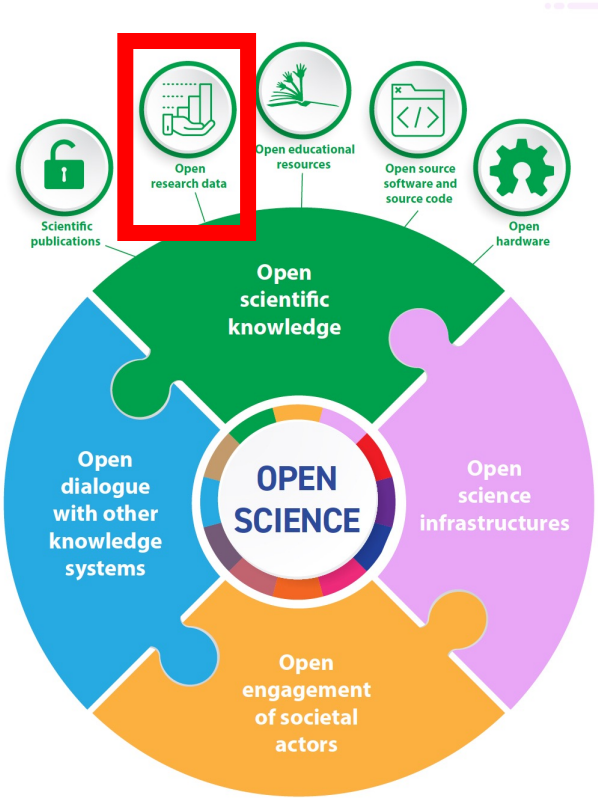
Outline

- <背景> オープンサイエンスと研究データ公開の推進
- 日本における研究データの公開・共有についての考え方
- DataCiteからみる研究データの公開状況

オープンサイエンスと研究データ

オープンサイエンス　多言語の科学知識を誰もが自由に利用・アクセス・再利用できるようにし，科学と社会の利益のために共同研究と情報の共有を増進させ，科学知識の創造，評価，伝達のプロセスを従来の科学界を超えて社会貢献活動に関するすべての人に開放するための様々な運動と実践を統合した包括的な概念

米川和志. E2585 – ユネスコ「オープンサイエンスに関する勧告」. カレントアウェアネスE. No. 433, 2022.



オープンな研究データ　オープンな研究データは、raw/processed、デジタル/アナログのデータ、それに付随するメタデータ、数値スコア、テキストレコード、画像と音声、プロトコル、分析コード、ワークフローが含まれる。それらは引用を条件として、誰でも使用・再利用・保持・再配布できる。オープンな研究データは、人間と機械が読み取り可能で実用的な形式で提供される。それらは、定期的なキュレーションとメンテナンスによってサポートされ、優れたデータ ガバナンスとスチュワードシップの原則、特にFAIR原則に従う。

UNESCO. UNESCO Recommendation on Open Science. 2021.

FAIR原則

- 研究データ公開の適切な実施方法を表現
- 適切な研究データの公開の4原則を満たすために必要となる15項目の要件を定義する
- 多くの国・研究者コミュニティで研究データ公開の基準として認識されている

To be Findable: (見つけられるために)

F1.（メタ）データが、グローバルに一意で永続的な識別子（ID）を有すること。
F2.データがメタデータによって十分に記述されていること。
F3.（メタ）データが検索可能なリソースとして、登録もしくはインデックス化されていること。
F4.メタデータが、データの識別子（ID）を明記していること。

To be Accessible: (アクセスできるために)

A1.標準化された通信プロトコルを使って、（メタ）データを識別子（ID）により入手できること。
A1.1 そのプロトコルは公開されており、無料で、実装に制限が無いこと。
A1.2 そのプロトコルは必要な場合は、認証や権限付与の方法を提供できること。
A2.データが利用不可能となったとしても、メタデータにはアクセスできること。

To be Interoperable: (相互運用できるために)

I1.（メタ）データの知識表現のため、形式が定まっていて、到達可能であり、共有されていて、広く適用可能な記述言語を使うこと。
I2.（メタ）データがFAIR原則に従う語彙を使っていること。
I3.（メタ）データは、他の（メタ）データへの特定可能な参照情報を含んでいること。

To be Re-usable: (再利用できるために)

R1.メタ（データ）が、正確な関連属性を豊富に持つこと。
R1.1（メタ）データが、明確でアクセス可能なデータ利用ライセンスと共に公開されていること。
R1.2（メタ）データが、その来歴と繋がっていること。
R1.3（メタ）データが、分野ごとのコミュニティの標準を満たすこと。

NBDC研究チーム（訳）. FAIR原則（「THE FAIR DATA PRINCIPLES」和訳）. 2019.
https://doi.org/10.18908/a.2019112601

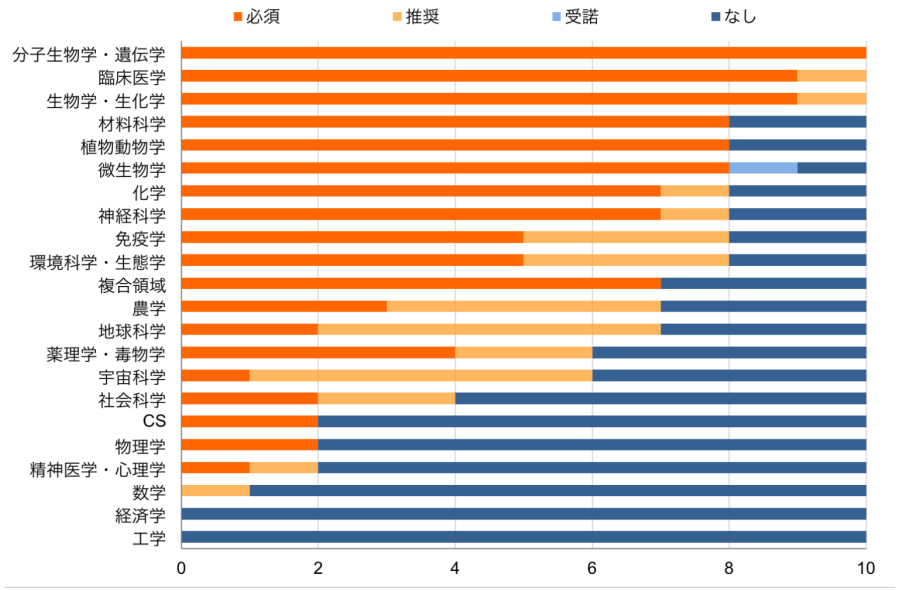
学術コミュニティからの要求

学術雑誌による研究データ公開の要求 → データ公開ポリシー

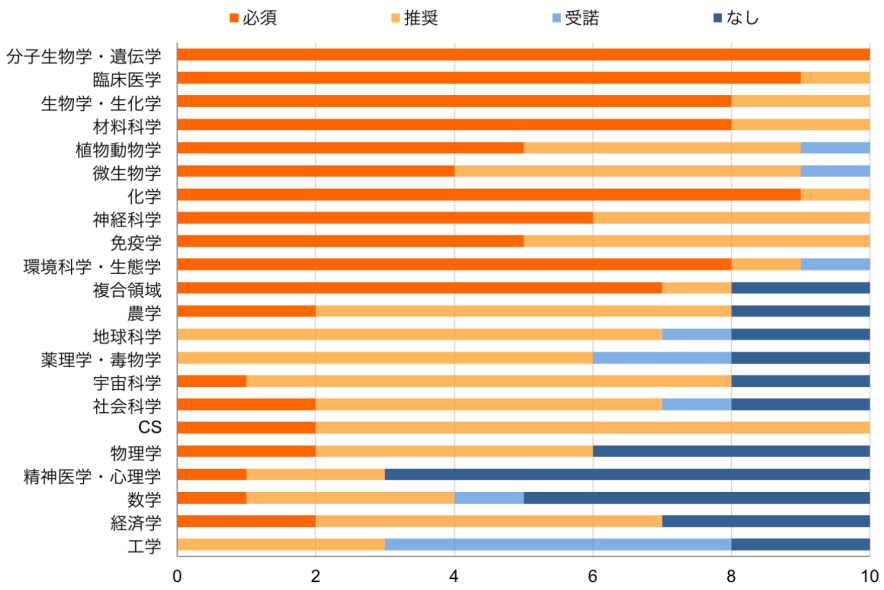
背景として、研究再現性の危機、健全な科学の発展など

データ公開ポリシー（リポジトリ）の強度

2014



2019

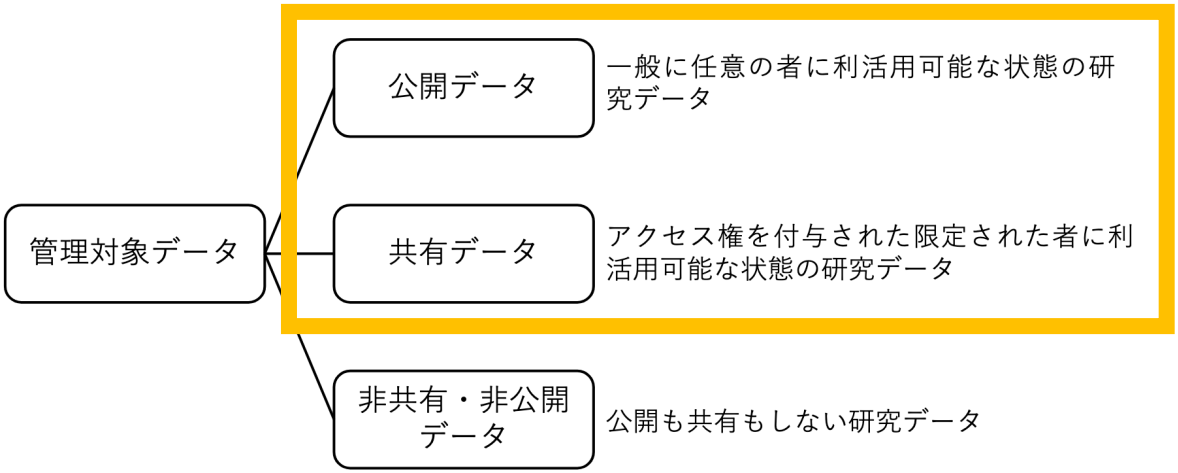


池内有為. 学術雑誌のデータ公開ポリシー：経年変化とデータ引用ポリシーの状況. Japan Open Science Summit, 2019. https://japanlinkcenter.org/rduf/doc/joss2019_rdc_03.pdf

研究データの公開・共有についての考え方 [1/2]

公的資金による研究データについては、
オープン・アンド・クローズ戦略に基づき管理・利活用を行う必要

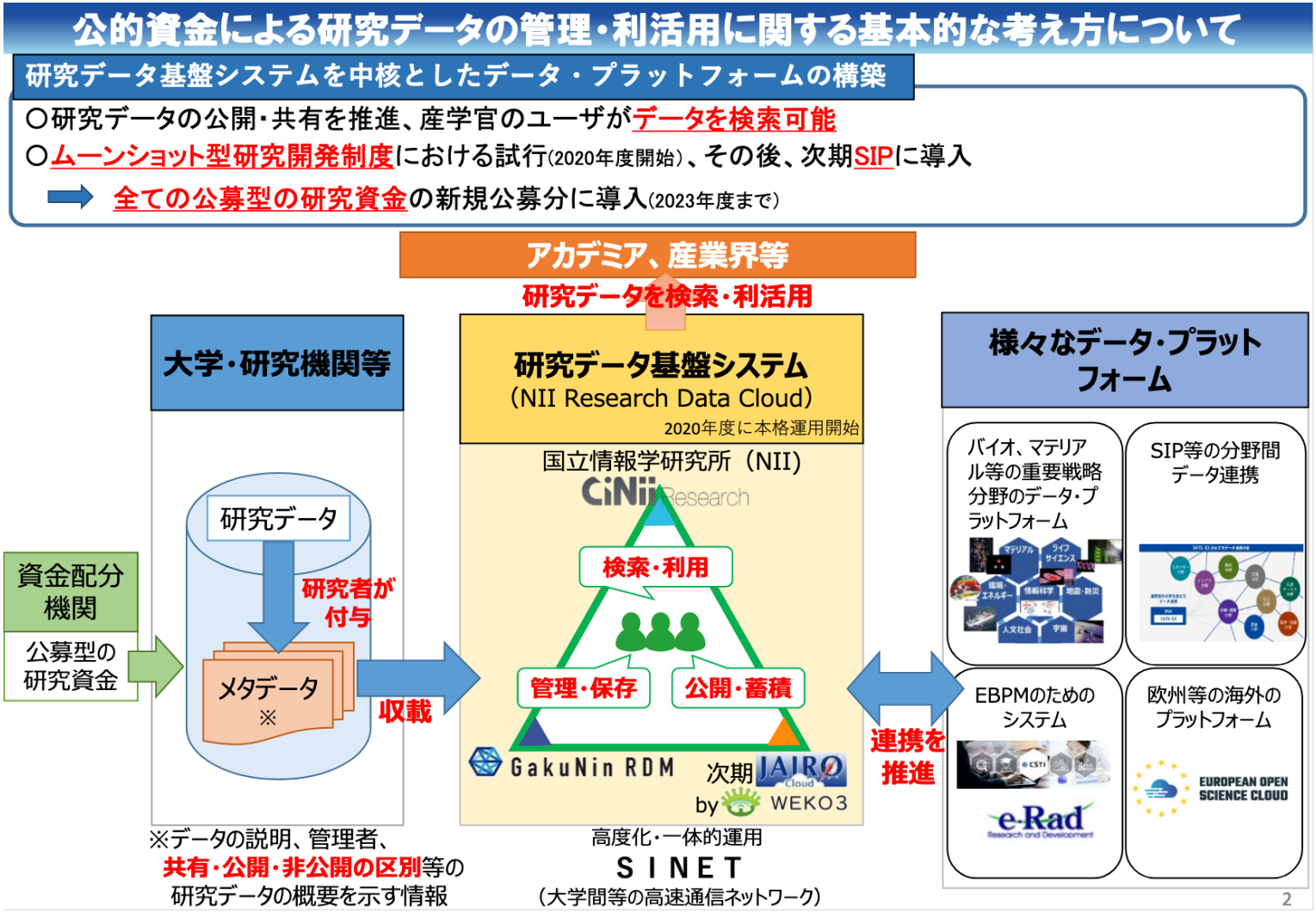
- 公的資金による論文のエビデンスとしての研究データは原則公開
- その他研究開発の成果としての研究データも可能な範囲で公開



- 以下との整合性に留意
  - 関係諸法令、データの取り扱いに関する各国の国内法及びEU規則
  - FAIR原則等の国際的な規則や慣行

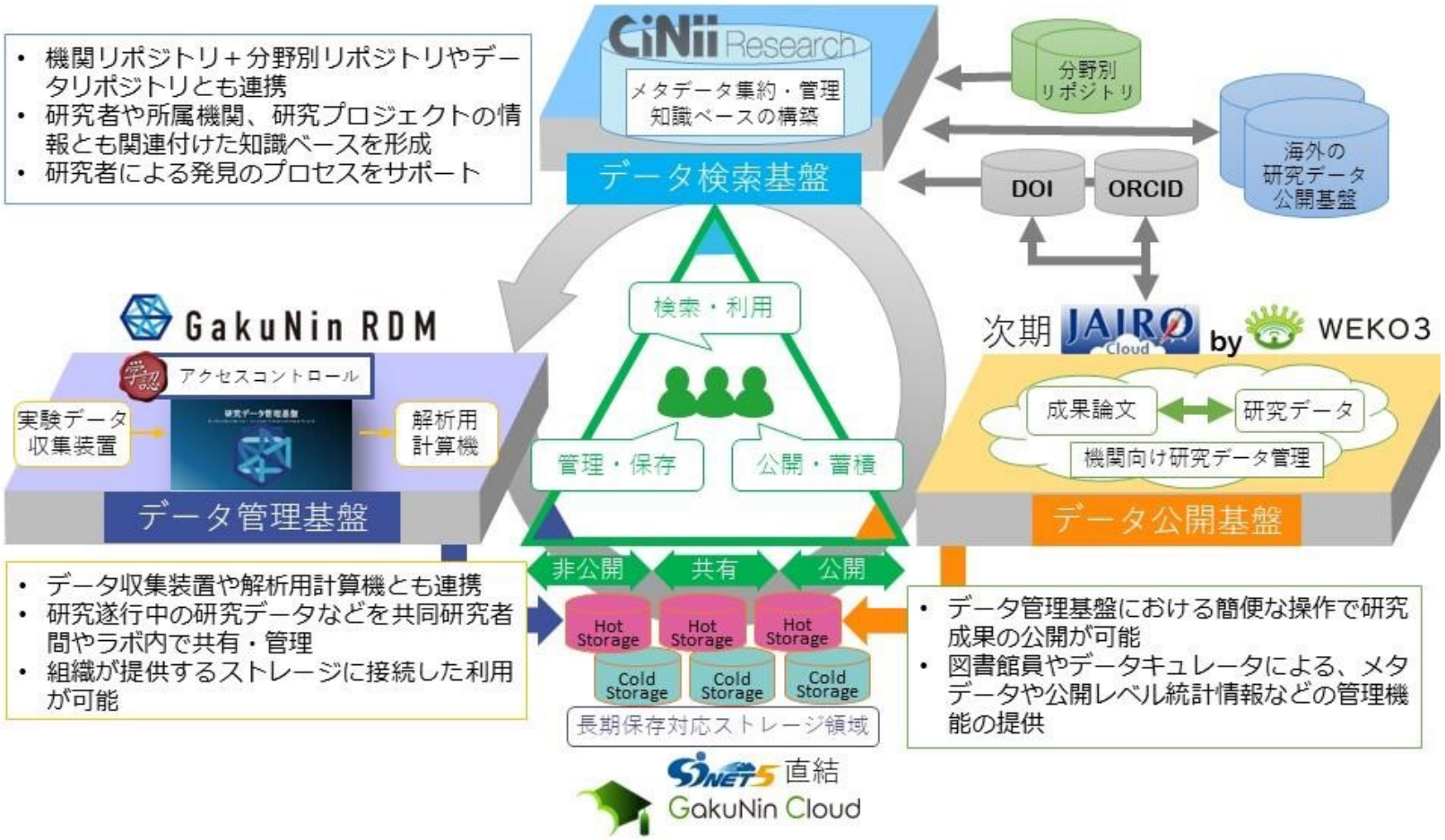
統合戦略イノベーション会議. 公的資金による研究データの管理・利活用に関する基本的な考え方. 2021.
https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/sanko1.pdf
科学技術・イノベーション推進事務局.公的資金による研究データの管理・利活用に関する基本的な考え方について. 2021.
https://www8.cao.go.jp/cstp/datapolicy_outline.pdf

研究データの公開・共有についての考え方 [2/2]



科学技術・イノベーション推進事務局.公的資金による研究データの管理・利活用に関する基本的な考え方について. 2021.
https://www8.cao.go.jp/cstp/datapolicy_outline.pdf

NII研究データ基盤（Research Data Cloud）



研究者と研究機関の役割

研究者

- 管理対象データの決定
- メタデータの付与
- データマネジメントプラン（DMP）の作成
- 研究データの保存
- オープン・アンド・クローズ戦略に基づく研究データの公開・共有
- 公募型の研究資金によるプロジェクト等の終了後の取扱い

研究機関

- データポリシーの策定

> 機関リポジトリを有する全ての大学・大学共同利用機関法人・国立研究開発法人においては、2025年までにデータポリシーを策定する

- 機関リポジトリへの研究データの収載と研究データへのメタデータの付与の推進
- 研究データマネジメント人材・支援体制の整備及び評価
- セキュリティの確保、関係諸法令の遵守等

統合戦略イノベーション会議. 公的資金による研究データの管理・利活用に関する基本的な考え方. 2021.
https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/sanko1.pdf

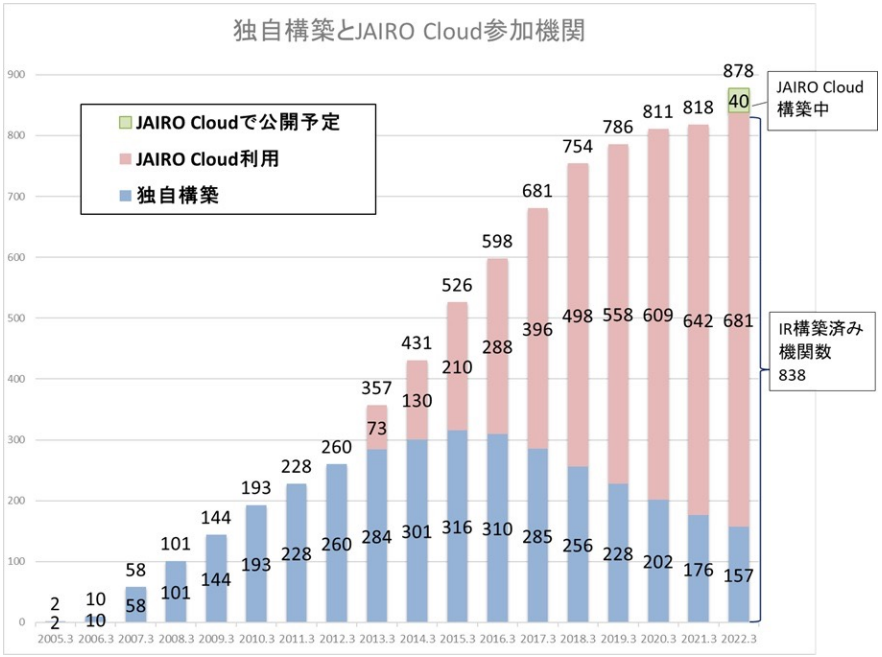
オープンアクセス方針と研究データ方針

	オープンアクセス（OA）方針	研究データ方針（公開に関する箇所）
いつ	出版後/採択後	研究遂行後（？）
どこで	機関リポジトリ/OAジャーナル	機関リポジトリ/分野別リポジトリ/汎用リポジトリ
誰が	研究者	研究者
何を	学術雑誌掲載論文	研究データ
なぜ	大学で創造された知の社会への還元/国・助成機関からの要請	大学で創造された知の社会への還元/国・助成機関からの要請/健全な学術の発展
どのように		FAIR原則に準拠等

自由度が高い

様々なリポジトリ

機関リポジトリ



学術機関リポジトリ構築連携支援事業. 機関リポジトリ統計.
https://www.nii.ac.jp/irp/archive/statistic/

分野リポジトリ

Subject

UK Data Service
ADS ARCHAEOLOGY DATA SERVICE
EMBL-EBI
The Cambridge Crystallographic Data Centre
Centre for Environmental Data Analysis

汎用リポジトリ

General-purpose

zenodo
figshare
The Dataverse Project
DRYAD

University of Reading. Choosing a data repository.
https://www.reading.ac.uk/research-services/research-data-management/preserving-and-sharing-data/choosing-a-data-repository

研究機関の役割として「機関リポジトリへの研究データの収載」が指摘される一方、分野リポジトリ等の重要性についても認識されている

調査の目的

	オープンアクセス（OA）方針	研究データ方針（公開に関する箇所）
いつ	出版後/採択後	研究遂行後（？）
どこで	機関リポジトリ/OAジャーナル	機関リポジトリ/分野別リポジトリ/汎用リポジトリ
誰が	研究者	研究者
何を	学術雑誌掲載論文	研究データ
なぜ	大学で創造された知の社会への還元/国・助成機関からの要請	大学で創造された知の社会への還元/国・助成機関からの要請/健全な学術の発展
どのように		FAIR原則に準拠等

- 効果的な支援を考えるには、機関の研究者の研究データの公開状況を把握したい
- しかし、文献と異なり研究データの公開状況は把握しにくい
  - 文献と異なり研究データ索引は発展途上
  - 機関リポジトリ以外のリポジトリなど、公開先が多種多様

→ 研究データの流通に大きな役割を果たしているDataCite DOIのメタデータを利用して、研究データ公開状況を把握する

調査方法 [1/3]

① DataCiteの全レコードを取得した（26,446,306件）

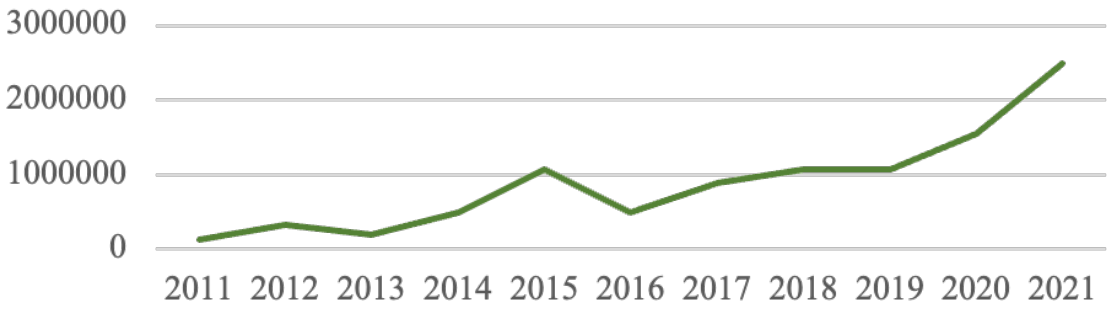
- 2021年10月22日時点のレコード
- 以下のURLより取得：https://archive.org/details/datacite_dump_20211022

【DataCiteとは？】
DataCiteはDOIの登録機関の一つであり、ウェブ上の研究データの発見、アクセス、再利用に係る問題に取り組むことを目的として、2009年12月に設立された。研究データにDOIの割り当てを行う。

◆機関リポジトリ外で公開されている研究データの公開状況を把握

② リソースの種別がDataset、Softwareのいずれかであるレコードを研究データとして対象とする（10,967,592件）

種別	件数
Dataset	10,727,829
Software	239,763



調査方法 [2/3]

③ 各レコードの各著者の各所属を、ROR APIを使用して同定する（表記揺れを解決する）

- ROR（Research Organization Registry）：研究機関の識別子のレジストリ
  - 例：名古屋大学　https://ror.org/04chrp450
- ROR APIは、与えられた文字列に対して合致する機関を返戻
  - 例：”Nagoya Univ.” > https://ror.org/04chrp450
- DataCiteのメタデータでは、著者（creator）のフィールドの下に所属（affiliation）というフィールドがあり、0件以上の文字列が格納

レコード数	10,967,592	著者数（のべ数）	191,927,379
所属の記載があるレコード数	1,167,727	所属の記載がある著者数（のべ数）	96,961,711
RORと結びつく所属があるレコード数	916,262	RORと結びつく所属の記載がある著者数（のべ数）	69,214,833

調査方法 [3/3]

④ レコードのカウントにあたっては、整数カウントを採用した

- 整数カウントでは、1件のレコードを、筆頭著者等の位置づけに限らず、すべての著者・機関・国について重複して1件とカウント
  → 参加度を測る
- 対して、分数カウントでは、1件のレコードを、著者別、機関別、国別等によって按分してカウント
  → 貢献度を測る

- 例：米国2名、日本1名、フランス1名による研究データ

整数カウント

国	件数
米国	1
日本	1
フランス	1

分数カウント（著者別に按分）

国	件数
米国	0.50
日本	0.25
フランス	0.25

cf. 中村優文. (2020). 共著論文の数え方, 整数カウント・分数カウントとその影響の考察. 情報の科学と技術, 70(6), 315-318.

調査の短所

- 著者の所属が入力されていないレコードが多い
  - 著者の所属が入力されているレコードは全レコードのうち10.65%
  - 著者の所属の有無については、リポジトリによって異なる
- DOIを割り当てる粒度がリポジトリによって異なる
  - とあるリポジトリでは、研究データのまとまりに対して、1件のDOIを割り当てる
  - 対して、1ファイルごとに1件のDOIを付与するリポジトリも存在する

どこで研究データは公開されている？（世界）

全レコードを対象

	国	件数
1	ut.ee	2,383,863
2	gbif.org	1,158,774
3	cam.ac.uk	940,895
4	harvard.edu	678,863
5	figshare.com	667,450
6	lfi.ch	614,742
7	dsmz.de	483,762
8	usc.edu	455,181
9	pangaea.de	389,133
10	zenodo.org	382,620
11	ucd.ie	232,285
12	nrct.go.th	170,980
13	osti.gov	154,032
14	datadryad.org	142,043
15	boldsystems.org	126,344
16	plate-archive.org	124,130
17	mendeley.com	118,625
18	pitt.edu	106,469
19	piscoweb.org	104,924
20	hepdata.net	98,673

著者の所属が記載されているレコードを対象

	国	件数
1	harvard.edu	358,538
2	zenodo.org	252,436
3	hepdata.net	91,331
4	dataverse.no	87,667
5	datadryad.org	39,636
6	geus.dk	35,158
7	openforestdata.pl	17,421
8	htejcap.org	14,670
9	syr.edu	13,884
10	gbif.org	13,808
11	inrae.fr	13,294
12	knaw.nl	12,799
13	aip.de	10,854
14	bindingdb.org	8,649
15	piscoweb.org	7,953
16	uni-stuttgart.de	7,344
17	openicpsr.org	7,318
18	chemotion-repository.net	7,045
19	imperial.ac.uk	6,767
20	ieee-dataport.org	6,662

どこで研究データは公開されている？（日本）

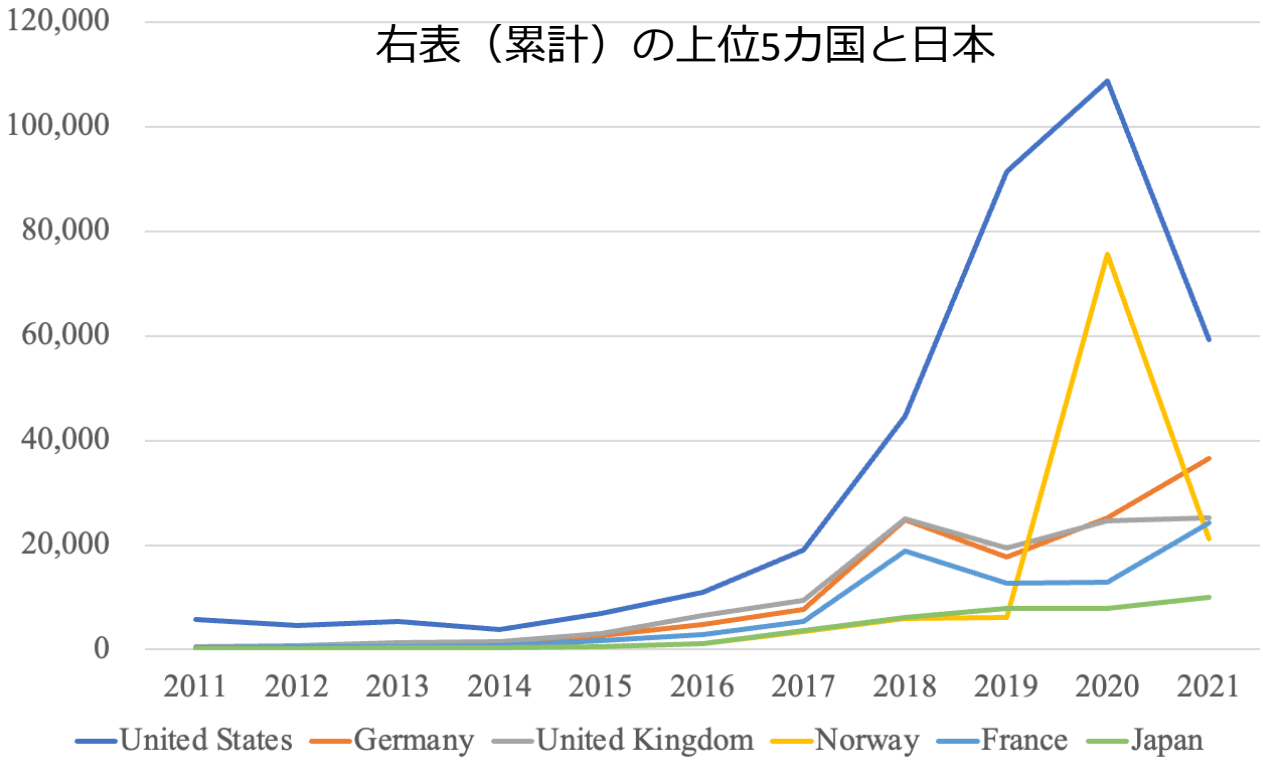
日本のレコードを対象

			件数
1	hepdata.net	論文に紐づく高エネルギー物理学分野のデータリポジトリ	33,429
2	harvard.edu	汎用の研究データリポジトリ（Harvard Dataverse）	4,826
3	zenodo.org	汎用のリポジトリ	3,110
4	datadryad.org	汎用の研究データリポジトリ	1,133
5	gbif.org	地球規模生物多様性情報機構	321
6	dkrz.de	ドイツ気候計算センター	302
7	diasjp.net	データ統合・解析システム	118
8	ieee-dataport.org	IEEEのデータポータル	106

- 機関リポジトリより制約が緩くクイックにアップロードできる汎用リポジトリ
- 分野リポジトリ

誰が研究データを公開している？（世界）[1/2]

国別



	国	件数
1	米国	395,516
2	ドイツ	137,053
3	英国	130,810
4	ノルウェー	115,579
5	フランス	85,791
6	カナダ	75,835
7	イタリア	69,461
8	オランダ	67,803
9	スイス	67,189
10	中国	64,786
11	ロシア	55,657
12	ブラジル	53,816
13	スペイン	50,552
14	ポーランド	50,504
15	スウェーデン	49,928
16	オーストラリア	45,442
17	ポルトガル	44,745
18	**日本**	**44,134**
19	デンマーク	41,776
20	オーストリア	40,250

【参考】日本の論文数(2018-2020年(PY)の平均)は、分数カウント法[…]によると、中、米、独、印に次ぐ**第5位**、Top10%補正論文数では、中、米、英[…]に次ぐ**第12位**、Top1%補正論文数では中、米、英[…]に次ぐ**第10位**である。
NISTEP. 科学技術指標2022. https://www.nistep.go.jp/sti_indicator/2022/RM318_41.html

誰が研究データを公開している？（世界）[2/2]

	機関	件数
1	ベルゲン大学（ノルウェー）	101,802
2	ハーバード大学（米）	60,000
3	ドゥブナ合同原子核研究所（露）	51,386
4	CERN	48,754
5	オハイオ州立大学（米）	46,730
6	国立核物理研究所（伊）	43,266
7	イリノイ大学アーバナ・シャンペーン校（米）	41,651
8	ケンブリッジ大学（英）	40,477
9	国立核物理研究所・ボローニャ（伊）	38,621
10	ウィスコンシン大学マディソン校 （米）	38,555
11	ドイツ電子シンクロトロン（独）	38,100
12	ミシガン大学アナーバー校（米）	36,047
13	アテネ大学（ギリシャ）	35,643
14	イェール大学（米）	35,092
15	プラハ・カレル大学（チェコ）	34,085
16	レベデフ物理学研究所（露）	33,932
17	Genoa（ブラジル）	33,816
18	ボストン大学（米）	33,737
19	ローレンス・バークレー国立研究所（米）	33,622
20	Geological Survey of Denmark and Greenland（デンマーク）	33,509

機関別

米国の大学と物理学系の研究機関が多い

誰が研究データを公開している？（日本）

全レコード（HEPData含）

	機関	件数
1	筑波大学	27,496
2	高エネルギー加速器機構	25,129
3	東京工業大学	23,173
4	京都大学	22,945
5	東京都立大学	22,611
6	名古屋大学	21,494
7	早稲田大学	21,030
8	神戸大学	20,787
9	大阪大学	20,736
10	九州大学	20,352
11	信州大学	20,247
12	京都教育大学	20,100
13	広島工業大学	19,626
14	岡山大学	15,758
15	お茶の水女子大学	10,520

HEPData除

	機関	件数
1	国立精神・神経医療研究センター	2,932
2	京都大学	1,428
3	東京大学	1,141
4	国立環境研究所	983
5	海洋研究開発機構	327
6	産業技術総合研究所	307
7	北海道大学	297
8	名古屋大学	289
9	筑波大学	283
10	東北大学	250
11	理化学研究所	208
12	神戸大学	203
13	北斗病院	198
14	九州大学	186
15	早稲田大学	165

RU11の大学と国立研究開発法人が多い

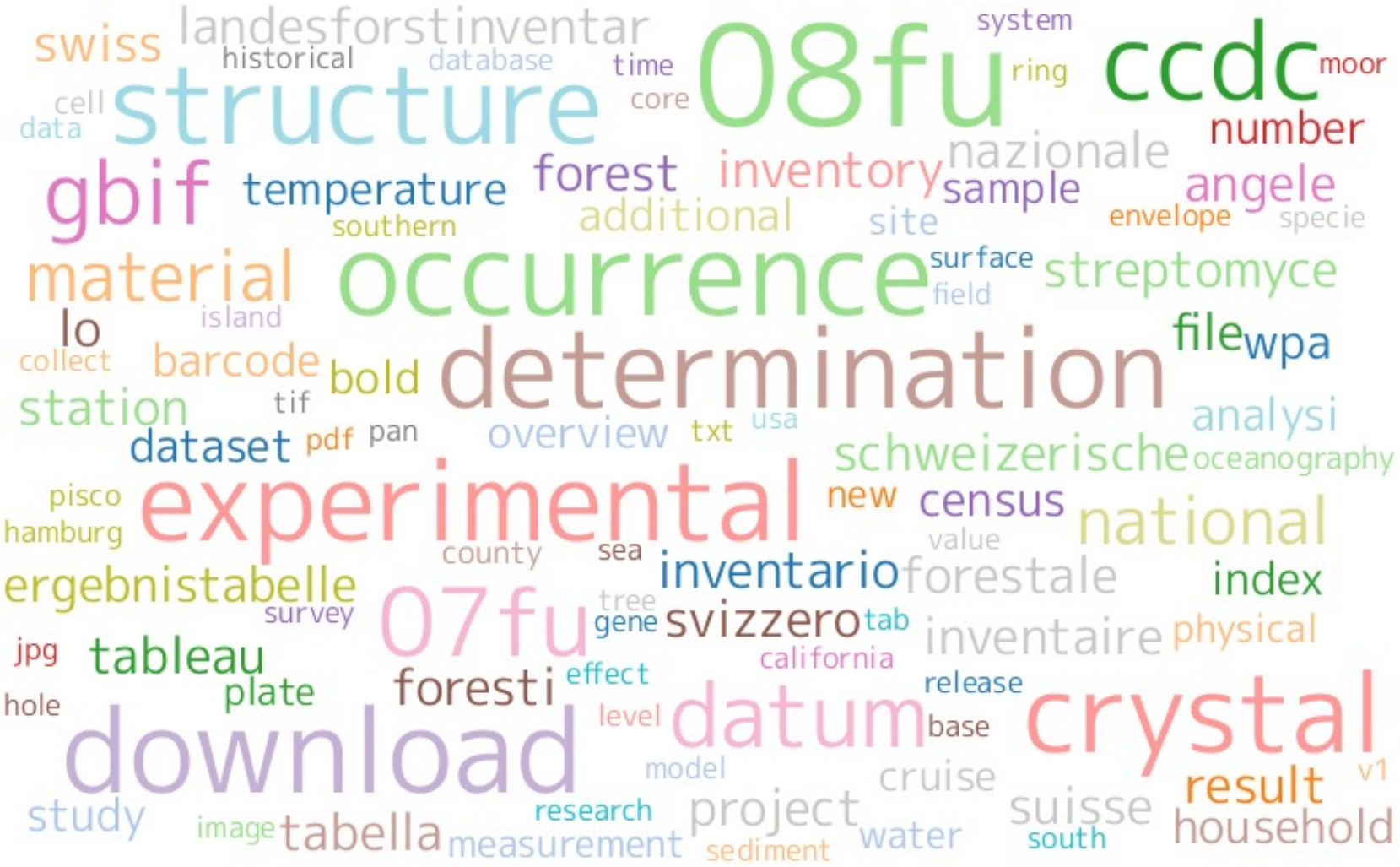
cf. 機関リポジトリ

機関

千葉大学 (48946)
早稲田大学 (20357)
一橋大学 (13758)
農業・食品産業技術総合研究機構 (1295)
岡山大学 (1193)
一般社団法人学術資源リポジトリ協議会 (1022)
京都大学 (251)
奈良文化財研究所 (246)
北海道大学 (210)
名古屋大学 (194)
人文学オープンデータ共同利用センター (140)
国立情報学研究所 データセット共同利用研究開発センター (97)
大阪大学 (86)
人間文化研究機構国立国語研究所 (79)
大阪商業大学 (71)
慶應義塾大学 (62)
地球環境戦略研究機関 (57)
東京大学 (48)
埼玉大学 (43)
旭川医科大学 (41)

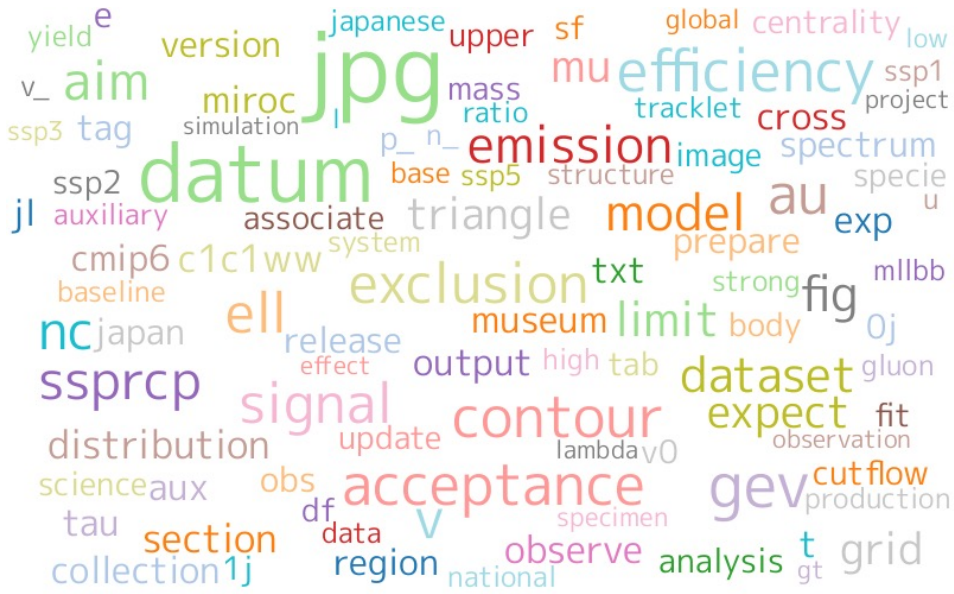
どのような研究データが公開されている？（世界）

全てのレコードを対象

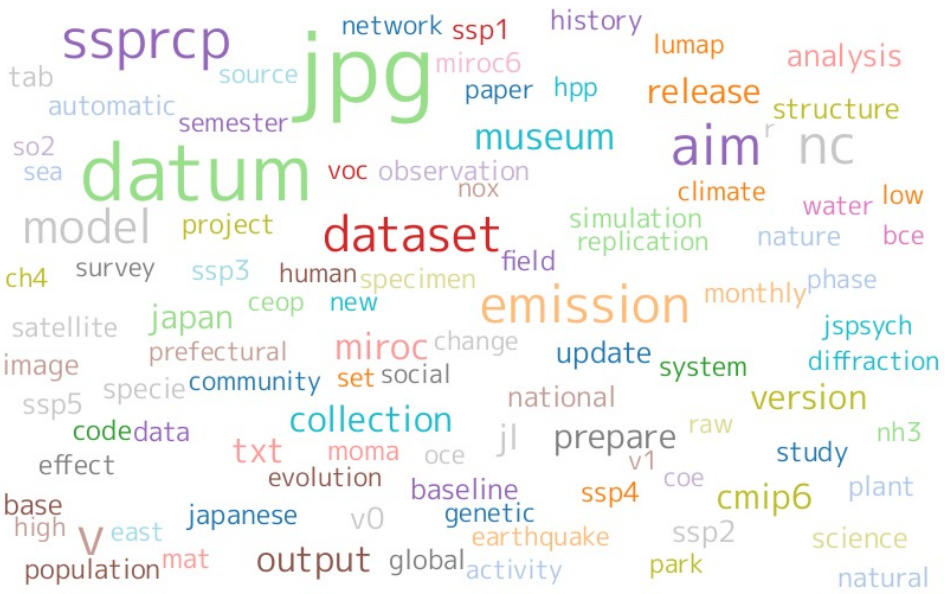


どのような研究データが公開されている？（日本）

全ての日本のレコードを対象



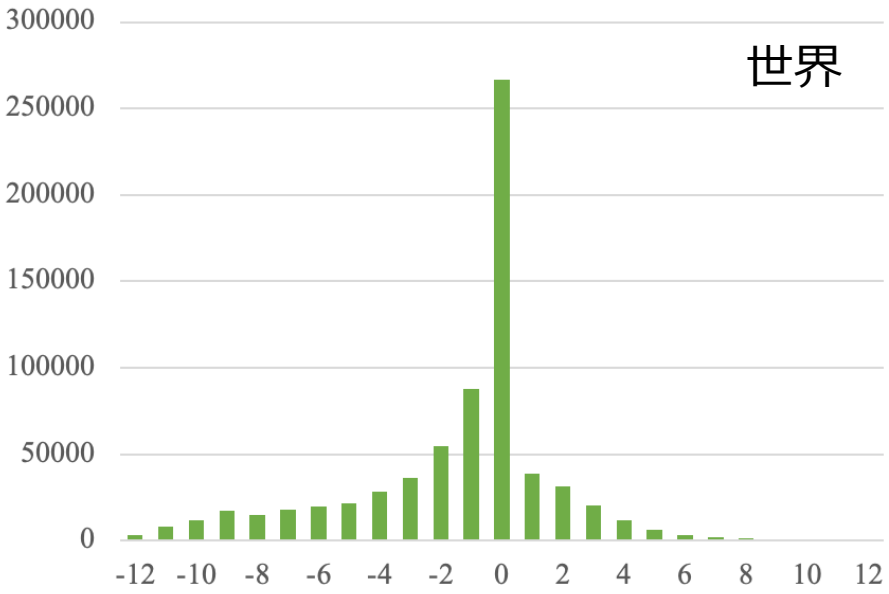
HEPDataを除いた日本のレコードを対象

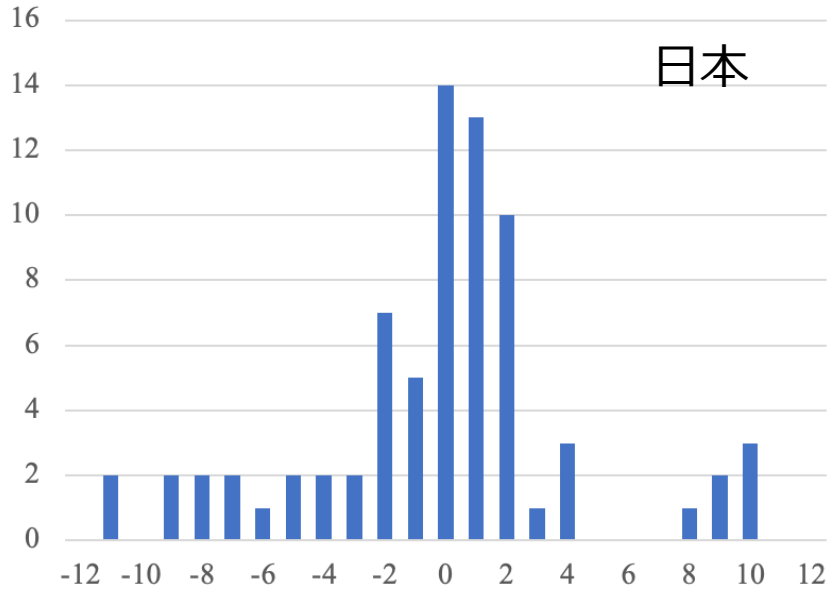


- 科学、技術、医学（STM分野）を中心として、多様なデータが多様なフォーマットで公開されている

いつ研究者はデータを公開している？

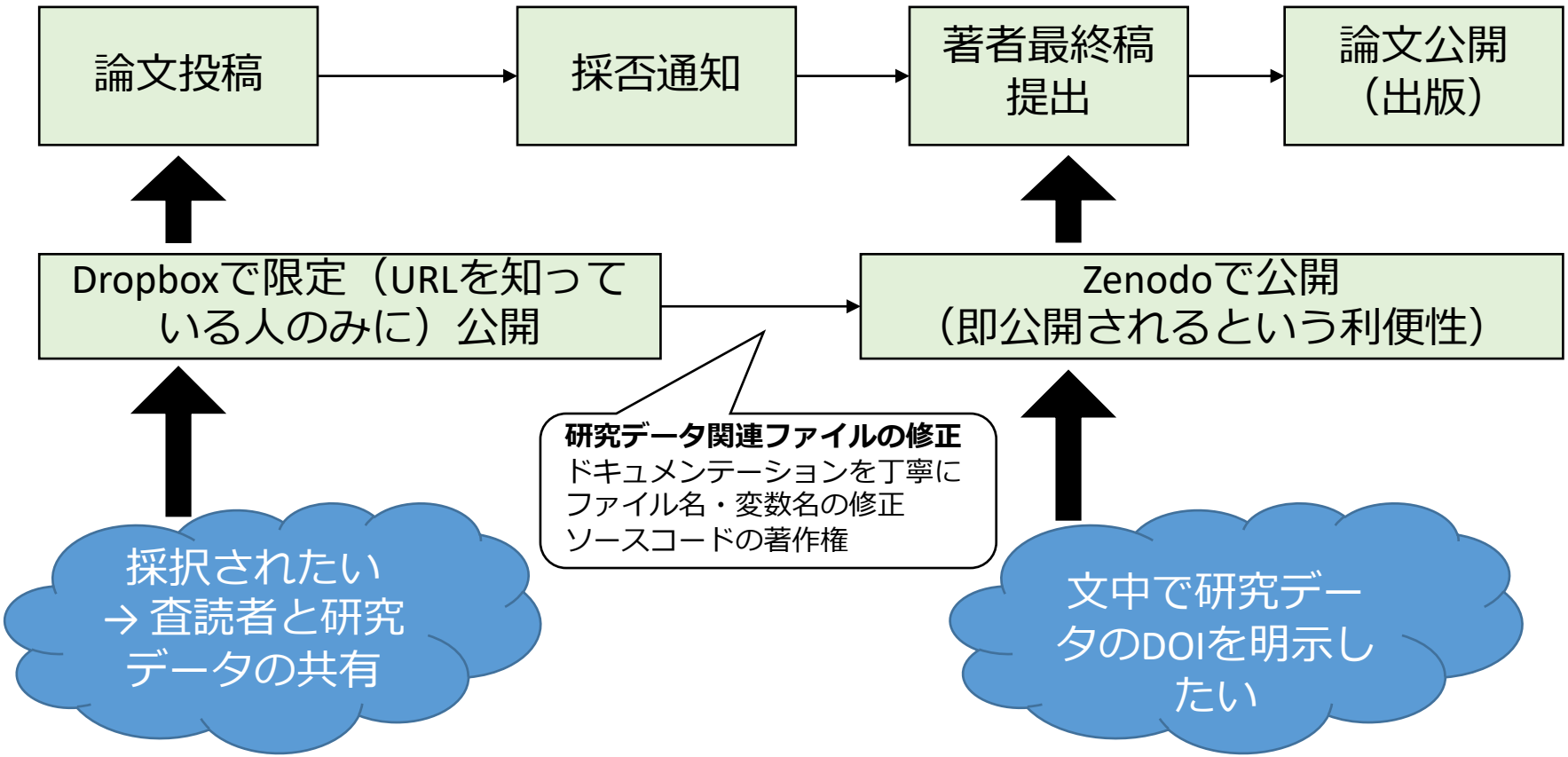
- 文献の出版時期と文献に付随するデータの公開時期の差
  - 関係”IsSupplementTo”で結ばれている研究データと論文を対象
- 横軸は文献の出版時期からの経過月数、縦軸はデータ件数





- 文献出版と同時期、または文献出版より前に公開されることが多い

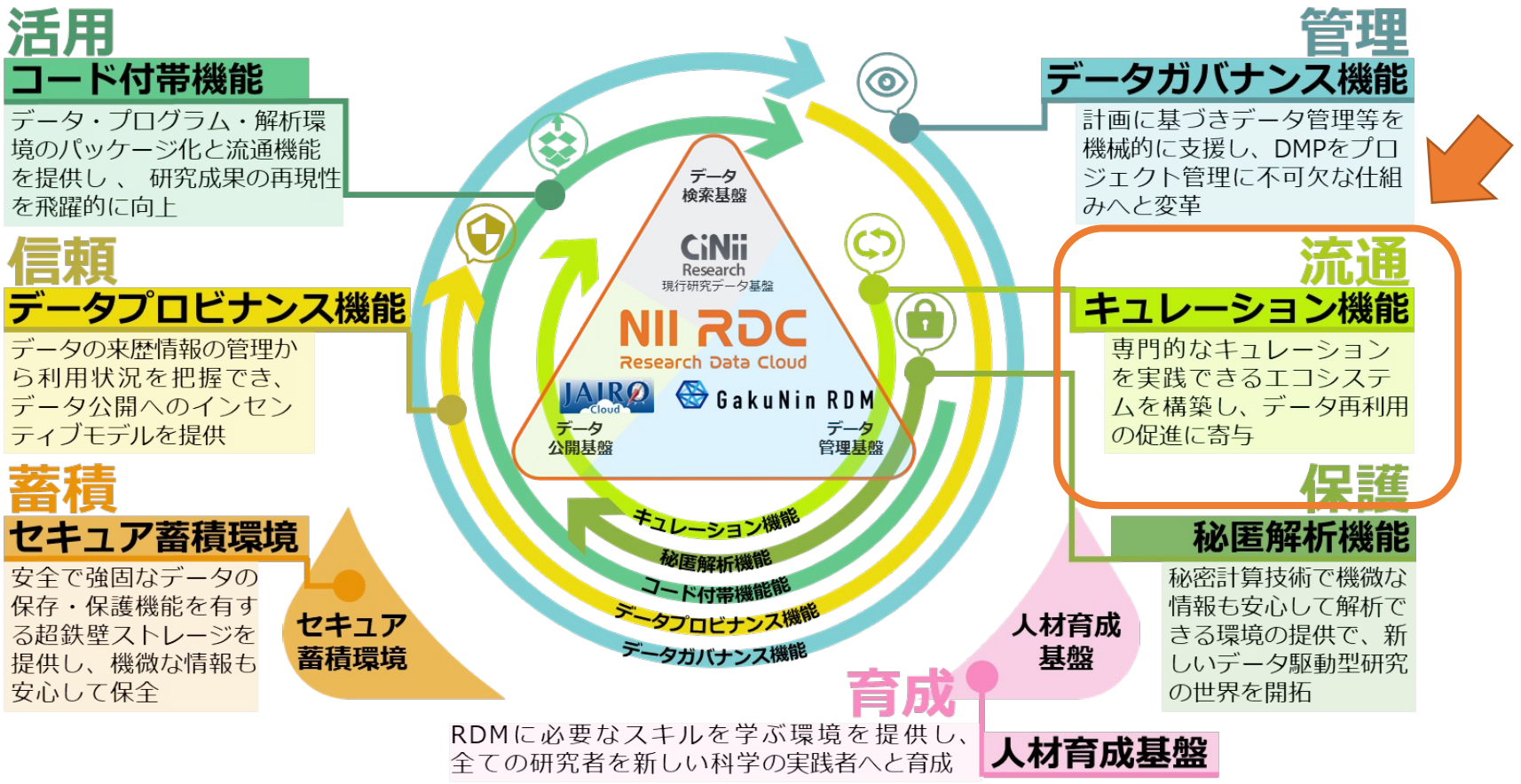
個人談



- 研究遂行中からの支援（ファイル名等）
- 即公開される（あるいは前もってDOIを付与してくれる）プラットフォーム

NII RDCでの研究データ流通の取り組み

汎用的なプロセスをアプリケーションとして実装し、
大学・研究機関へデータキュレーション支援サービスを提供



南山泰之. 機関リポジトリでの研究データ公開. 図書館総合展「機関リポジトリで実現する書誌多様性」, 2022.
https://www.nii.ac.jp/event/other/libraryfair/2022/

研究者はどのように研究データを公開する？

FAIR原則に準拠した形式で公開することが望ましい

To be Findable: (見つけられるために)

F1.（メタ）データが、グローバルに一意で永続的な識別子（ID）を有すること。
F2.データがメタデータによって十分に記述されていること。
F3.（メタ）データが検索可能なリソースとして、登録もしくはインデックス化されていること。
F4.メタデータが、データの識別子（ID）を明記していること。

To be Accessible: (アクセスできるために)

A1.標準化された通信プロトコルを使って、（メタ）データを識別子（ID）により入手できること。
　A1.1 そのプロトコルは公開されており、無料で、実装に制限が無いこと。
　A1.2 そのプロトコルは必要な場合は、認証や権限付与の方法を提供できること。
A2.データが利用不可能となったとしても、メタデータにはアクセスできること。

To be Interoperable: (相互運用できるために)

I1.（メタ）データの知識表現のため、形式が定まっていて、到達可能であり、共有されていて、広く適用可能な記述言語を使うこと。
I2.（メタ）データがFAIR原則に従う語彙を使っていること。
I3.（メタ）データは、他の（メタ）データへの特定可能な参照情報を含んでいること。

To be Re-usable: (再利用できるために)

R1.メタ（データ）が、正確な関連属性を豊富に持つこと。
　R1.1（メタ）データが、明確でアクセス可能なデータ利用ライセンスと共に公開されていること。
　R1.2（メタ）データが、その来歴と繋がっていること。
　R1.3（メタ）データが、分野ごとのコミュニティの標準を満たすこと。

ライセンスが明記されているレコードとその割合

	ライセンス情報有
世界のレコード	5,956,404 (54.31%)
日本のレコード	10,075 (22.83%)
日本のレコード（HEPData除）	5,350 (49.98%)

NBDC研究チーム（訳）. FAIR原則（「THE FAIR DATA PRINCIPLES」和訳）. 2019. https://doi.org/10.18908/a.2019112601

まとめ

オープンサイエンスの潮流の中、研究者はオープン・アンド・クローズ戦略に基づいて研究データの管理・利活用を行う

- 研究データポリシーの策定
- NII RDCを中核とした情報基盤の整備

研究者の研究データの公開の実態を探ることを目的として、DataCiteの調査を実施

- 公開件数：論文と比較すると国際的なポジションが低い
- 公開されているデータ：STM分野を中心として多種多様
- 公開先：分野別リポジトリも多い一方、汎用リポジトリも多い
  → インタフェースを通してクイックに投稿できる環境が必要（？）
- 公開時期：紐づく論文と同時期または前
  → スムーズな公開のために、研究遂行中から支援が必要（？）

ありがとうございました