

大学・研究機関による データキュレーションの実践に 向けて

国立情報学研究所
南山 泰之

琉球大学
研究データ管理セミナー
2024年2月15日

(於) 琉球大学附属図書館2階ラーニング・コモンズ

目次

1. はじめに
2. 国内の動向
3. 実践に向けて
4. まとめ

1. はじめに

データキュレーションとは

用語の関係性： Curation⊃Archiving⊃Preservation

- **キュレーション**：データが目的に適合し、発見や再利用できるようにするために、データが作成された時点から、データの利用を管理し促進する活動。動的なデータセットの場合、これは目的に合ったデータを維持するための継続的なエンリッチメントまたは更新を意味する。より高度なレベルのキュレーションには、アノテーションや他の出版物とのリンクを維持することも含まれる
- **アーカイビング**：データが適切に選択され、保管され、アクセスできるようにし、セキュリティと真正性を含め、論理的および物理的な整合性が長期間にわたって維持されるようにするキュレーション活動
- **プリザベーション**：データの特定の項目が時間の経過とともに維持され、技術の変化によってもアクセスして理解できるようにするためのアーカイビングの活動



データを解釈可能かつ利用可能にするための一連の処理を
データキュレーションと呼ぶ

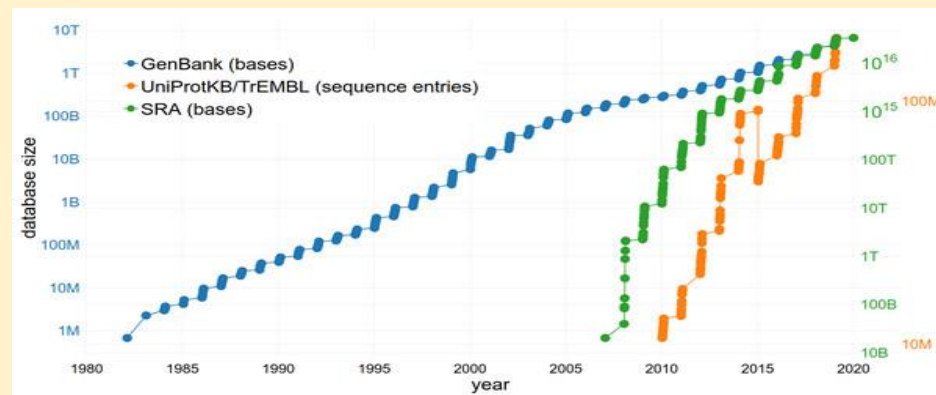
各分野におけるデータキュレーションの実践

生命科学、社会科学などの分野ではデータキュレーションのライフサイクルを定め、効率的なデータキュレーションを実践している

Life sciences

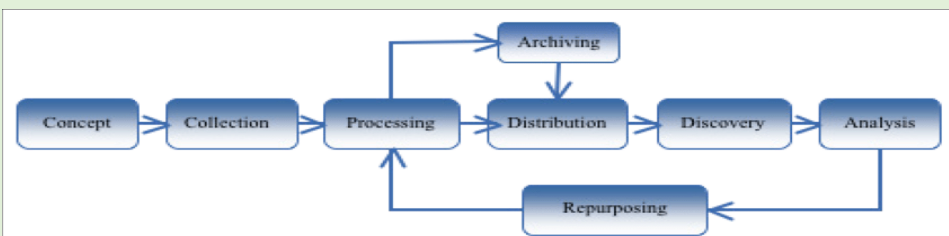


<https://rdmkit.elixir-europe.org/index.html>

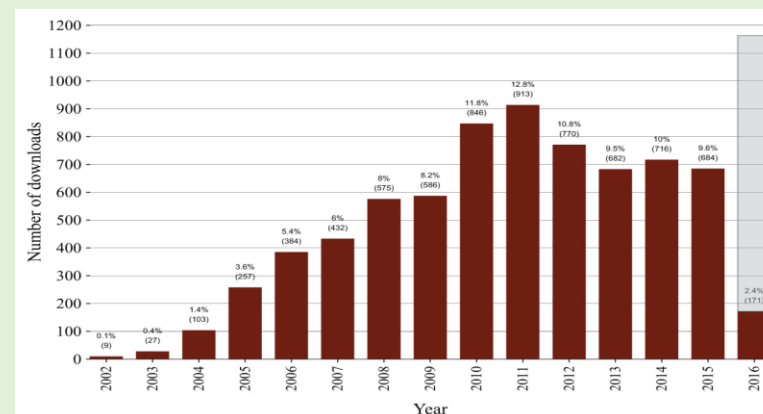


<https://doi.org/10.7717/peerj.9954>

Social sciences



<https://ddialliance.org/Specification/DDI-Lifecycle/>



<https://doi.org/10.1177/2158244016685136>

欧州委員会（EC）によるFAIR原則の費用便益分析

(1) 費用便益分析で対象とした要素と損失値

領域	要素
研究活動へのインパクト	(a) 費やす時間 (b) 保管コスト (c) ライセンスコスト
将来の研究の機会へのインパクト	(d) 研究の撤回率 (e) 助成の重複 (f) 学際性の欠如
イノベーションへのインパクト	(g) 潜在的な経済成長

少なくとも
年間102億ユーロ以上の損失

※【試算方法】(a)～(e)：FAIR原則に従う研究データがないときの非効率性（費やす時間と関連コスト）を評価。研究者が研究データへアクセスするために支払う必要のあるライセンスのコストを見積。FAIR原則が実施されていれば不必要である重複した研究データの保管コストを算出。(f)と(g)：定性的な考察に基づく。



(2) 上記の費用便益分析で対象としていない要素と損失の推定値

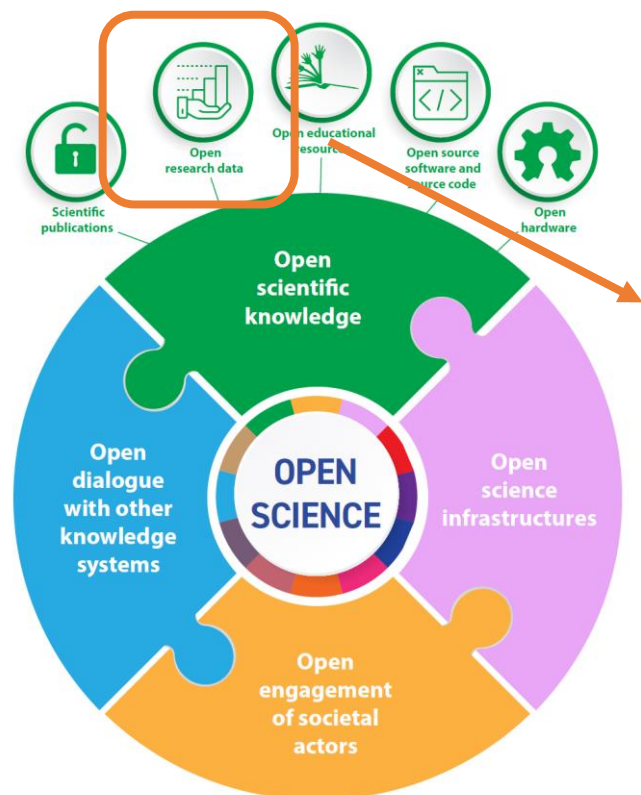
要素
研究の質 経済的取引高 研究データの機械可読性

推定**年間160億ユーロ**の損失

※FAIR原則：研究データの適切な共有・公開を行う
もっともよく知られた基準。学術コミュニティにおける
データキュレーション実践のための事実上の指針。

FAIR原則に沿ったデータキュレーションが行われない際の損失は**年間262億ユーロ**

ユネスコ「オープンサイエンスに関する勧告」



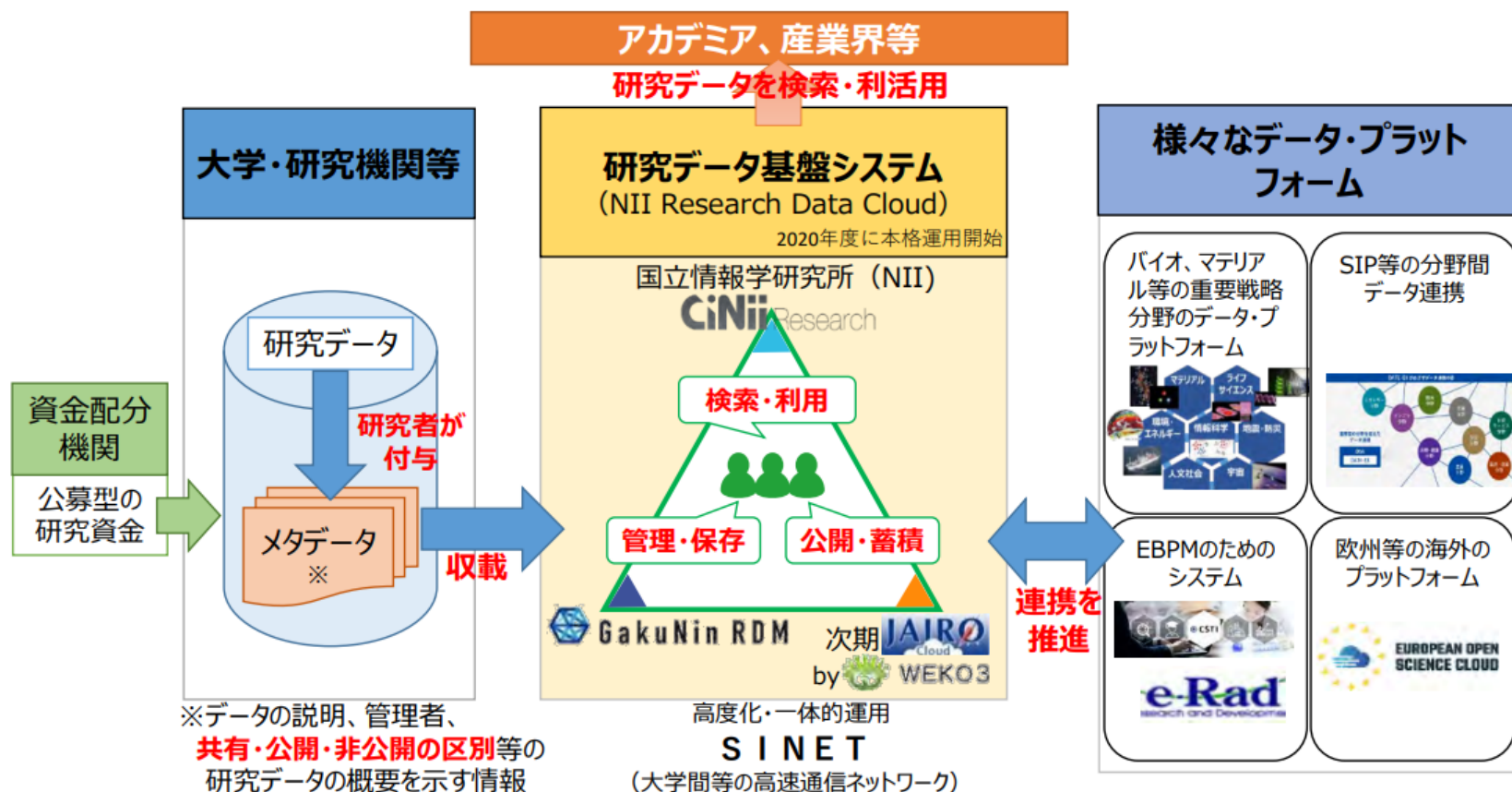
“Open research data are available in a timely and user-friendly, human- and machine-readable and actionable format, in accordance with principles of good data governance and stewardship, **notably the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, supported by regular curation and maintenance”**

FAIR原則に基づくデータキュレーションが社会的にも求められつつある

2. 国内の動向

政策的な要請

研究データ基盤システムを中核としたデータ・プラットフォームの構築
→研究データの公開・共有を推進、産学官のユーザがデータを検索可能



Ref: 研究データの管理・利活用に関する現状と課題について
https://www8.cao.go.jp/cstp/gaiyo/yusikisha/20220428/siryoy3_1_1.pdf

FAIR原則への言及

「研究DXの推進－特にオープンサイエンス、データ利活用推進の視点から－に関する審議について」（2022年12月23日）

内閣府からの審議依頼を受けて、日本学術会議に設置したオープンサイエンスを推進するデータ基盤とその利活用に関する検討委員会、同オープンサイエンス企画分科会及び同オープンサイエンス企画分科会オープンサイエンス・データ利活用推進小委員会が中心となり審議を行った。

【提案1】研究者が容易に利用可能な研究データプラットフォームの構築

【提案2】データプロフェッショナルの育成と多面的な研究評価の実現

【提案3】モニタリング機構に基づくデータ駆動型研究の不断の改善

【提案4】研究自動化(ARW)に向けた情報技術、計算資源の集約

【提案5】分野を越えた連携を実現するFAIR原則の追求

【提案6】法制度面でのデータガバナンスの構築



“多様な分野間の連携においては論文だけではなくデータを介することによりARWの更なる加速が期待されるが、いかなるメタデータ付与がデータの利活用に資するかは分野依存性が高い。**FAIR原則に基づき、各分野における積極的な分野間連携の実践が望まれ、また、その経験値の共有を可能とする取組が推奨される。**日本学術会議や学会等において、FAIR原則を追求する踏み込んだ議論が期待される。”

Ref: 回答 研究DXの推進－特にオープンサイエンス、データ利活用推進の視点から－に関する審議について

<https://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-25-k335.pdf> 10

研究者からの要望

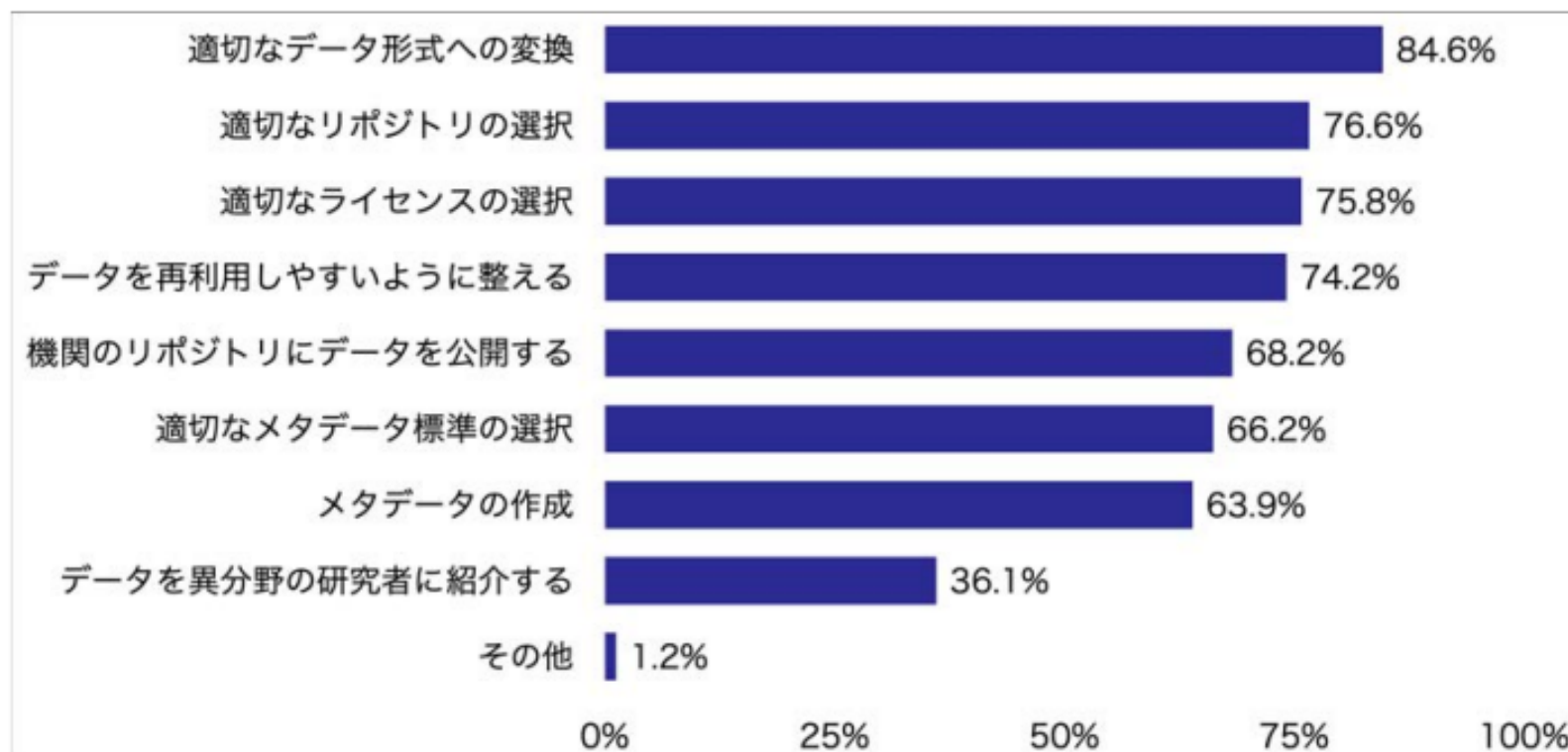
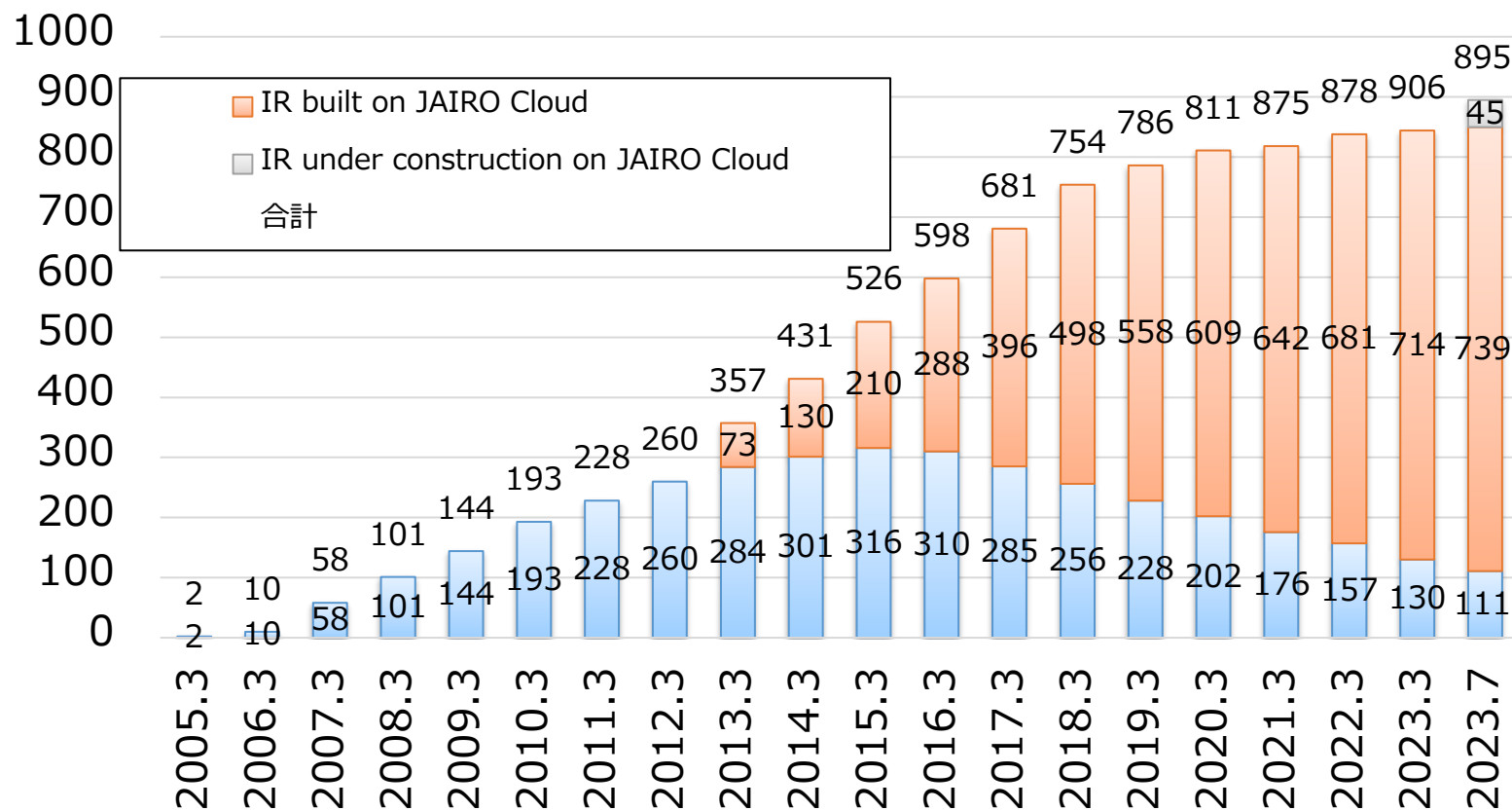


図 61 研究データ管理（RDM）に関して依頼したい項目（n=488：複数回答）

データ利活用に向けた要望が高い

大学・研究機関の準備体制

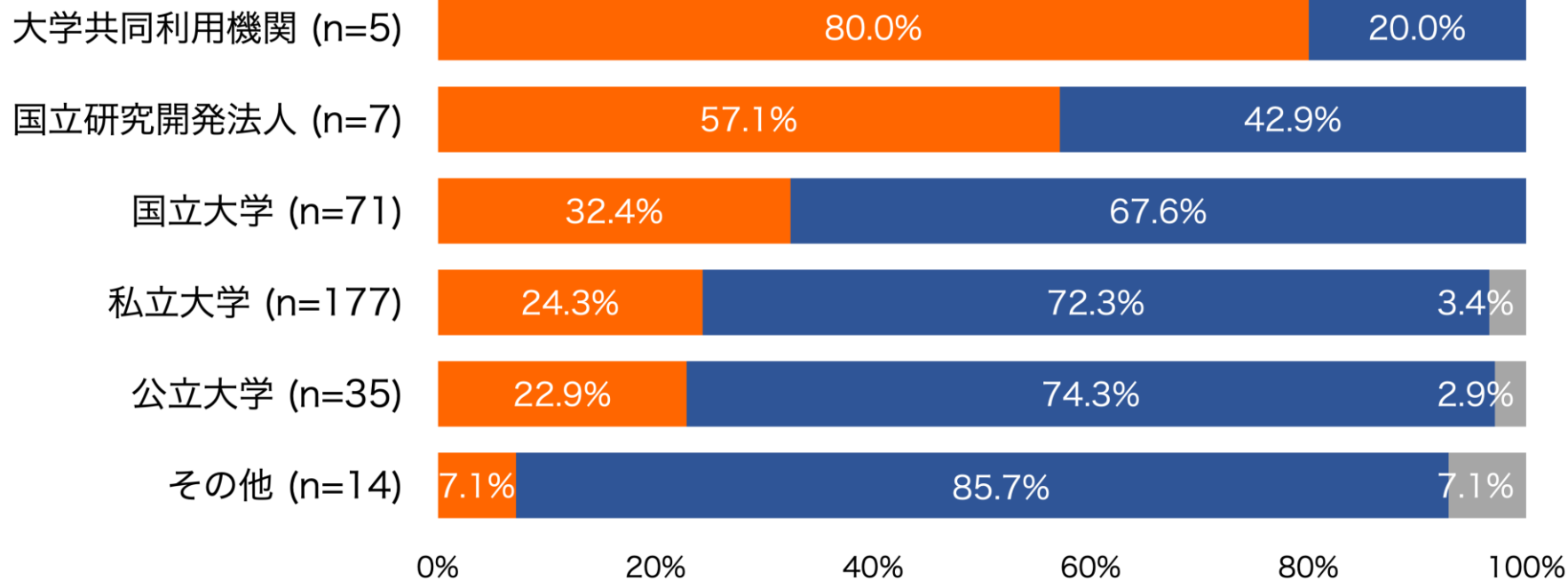


国内の機関リポジトリ数は895を数える
80%以上がJAIRO Cloudサービスを利用

2021～2022年の機関リポジトリによる 研究データ公開事例

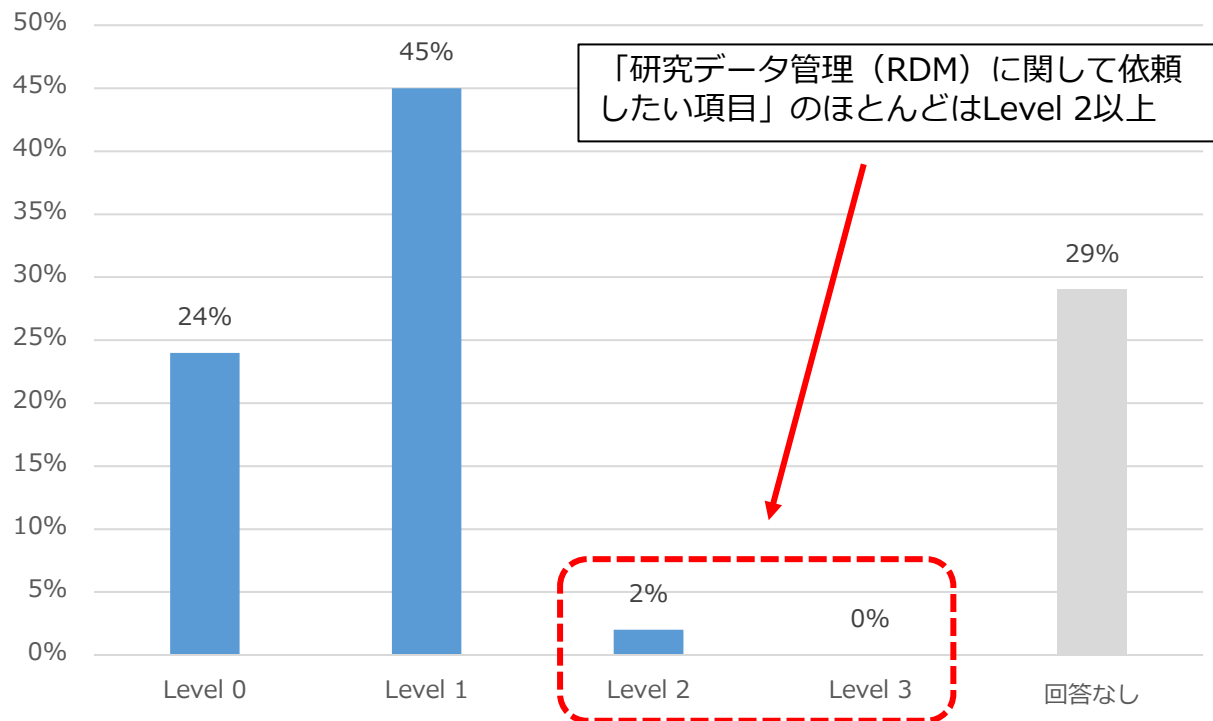
N=309

■ ある ■ ない ■ わからない



全ての機関種別で公開事例あり
大学共同利用機関や研究開発法人等は比率が高い

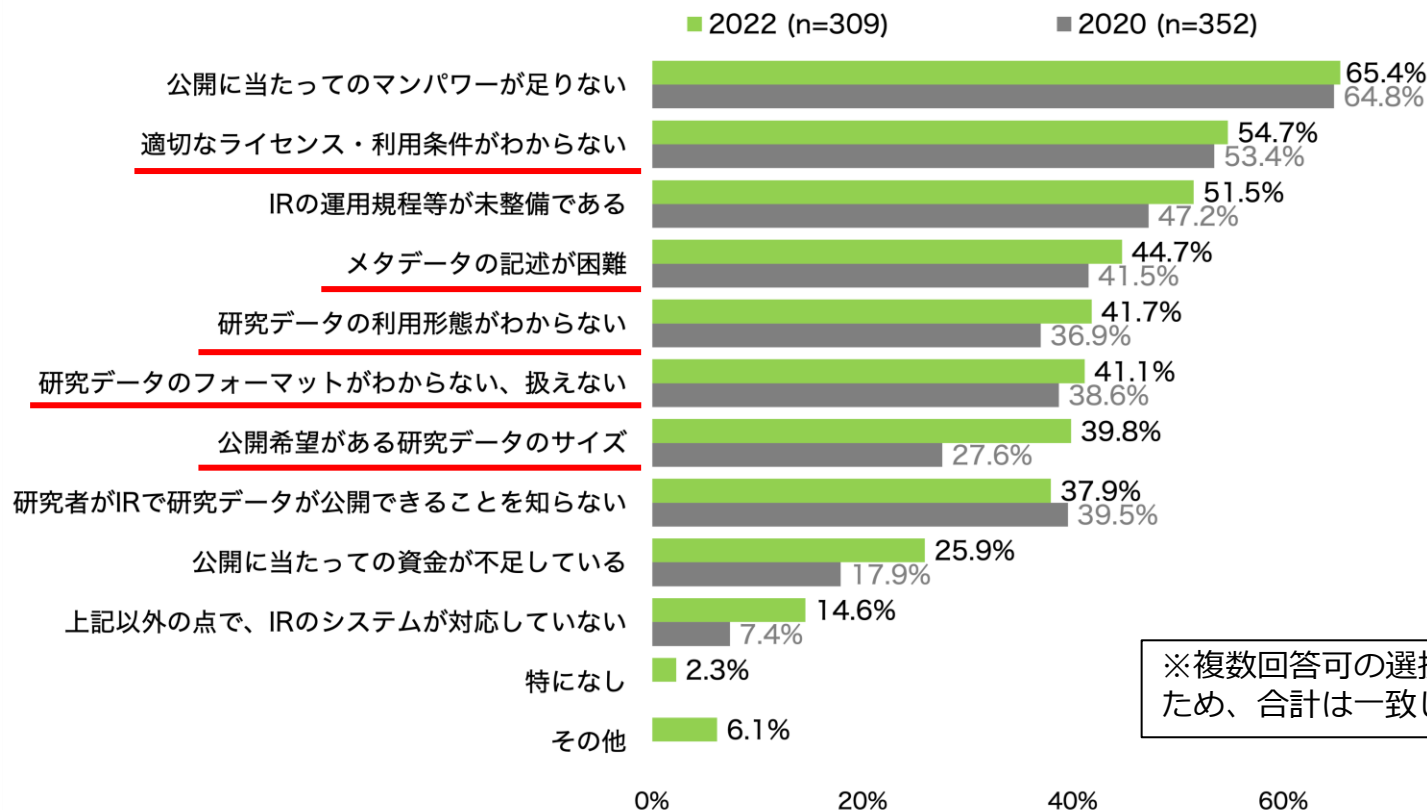
課題：データ公開に向けた作業レベル



レベル	見出し	説明（抜粋）
Level 0	登録データの流通	登録されたままで公開する
Level 1	基礎的な標準に準拠	簡単なチェック、基礎的なメタデータや文書の追加を行う
Level 2	論理・技術面での検証	フォーマット変換、文書やメタデータの強化などを実施する
Level 3	概念的な理解と再利用	Level 2に加えて、データレベルでの編集を行う

何が課題や障壁となっているのか

機関リポジトリ（IR）による研究データ公開の課題や障壁となり得ること



研究分野によって作法が異なる項目が多く含まれている

3. 実践に向けて

データキュレーション支援機能

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サービスを提供

活用

コード付帯機能

データ・プログラム・解析環境のパッケージ化と流通機能を提供し、研究成果の再現性を飛躍的に向上

信頼

データプロビナンス機能

データの来歴情報の管理から利用状況を把握でき、データ公開へのインセンティブモデルを提供

蓄積

セキュア蓄積環境

安全で強固なデータの保存・保護機能を有する超鉄壁ストレージを提供し、機微な情報も安心して保全

セキュア
蓄積環境

管理

データガバナンス機能

計画に基づきデータ管理等を機械的に支援し、DMPをプロジェクト管理に不可欠な仕組みへと変革

流通

キュレーション機能

専門的なキュレーションを実践できるエコシステムを構築し、データ再利用の促進に寄与

保護

秘匿解析機能

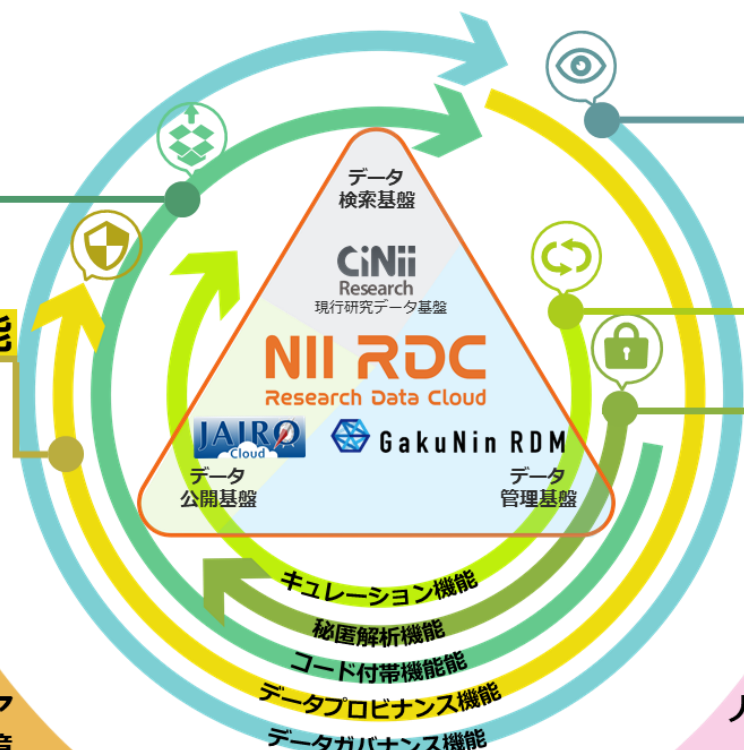
秘密計算技術で機微な情報も安心して解析できる環境の提供で、新しいデータ駆動型研究の世界を開拓

育成

人材育成
基盤

人材育成基盤

RDMに必要なスキルを学ぶ環境を提供し、
全ての研究者を新しい科学の実践者へと育成



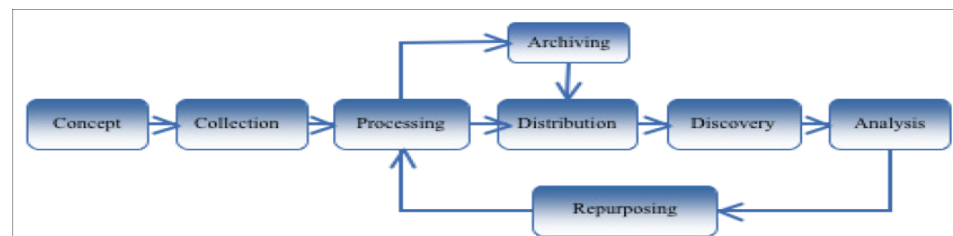
課題：各分野における捉え方の違い

Life sciences



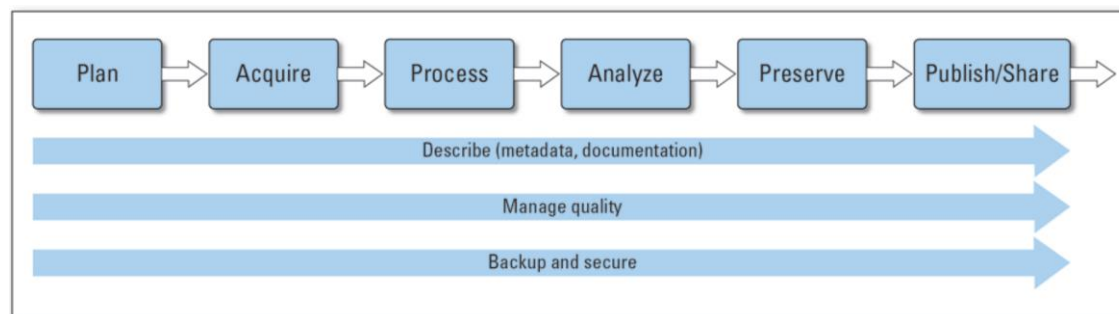
<https://rdmkit.elixir-europe.org/index.html>

Social sciences



<https://ddialliance.org/Specification/DDI-Lifecycle/>

Earth science



<https://doi.org/10.3133/ofr20131265>

各分野で定義されるライフサイクルは、分野内での再利用に最適化されている

データキュレーション活動の形式化

複数分野におけるデータキュレーション活動の実態を調査し、活動に含まれるプロセスを形式化

形式化までのステップ

Data Curation Activities

Data curation activities can be arranged around five steps of data curation life-cycle: Ingest, Appraise (Accept), Curate, Access, and Preserve.

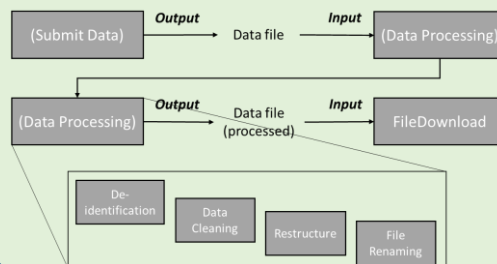
Ingest

- Authentication: The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.
- Chain of custody: Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.
- Deposit agreement: The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.
- Documentation: Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).
- File validation: A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file

分析



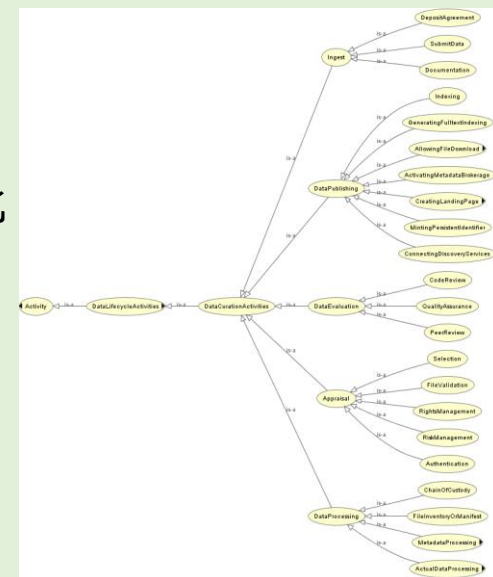
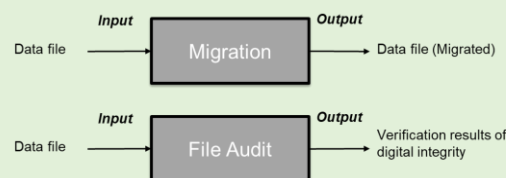
Sequential



形式化



Occasional



活動を表現する語彙

制御構造別に分類

オントロジー開発

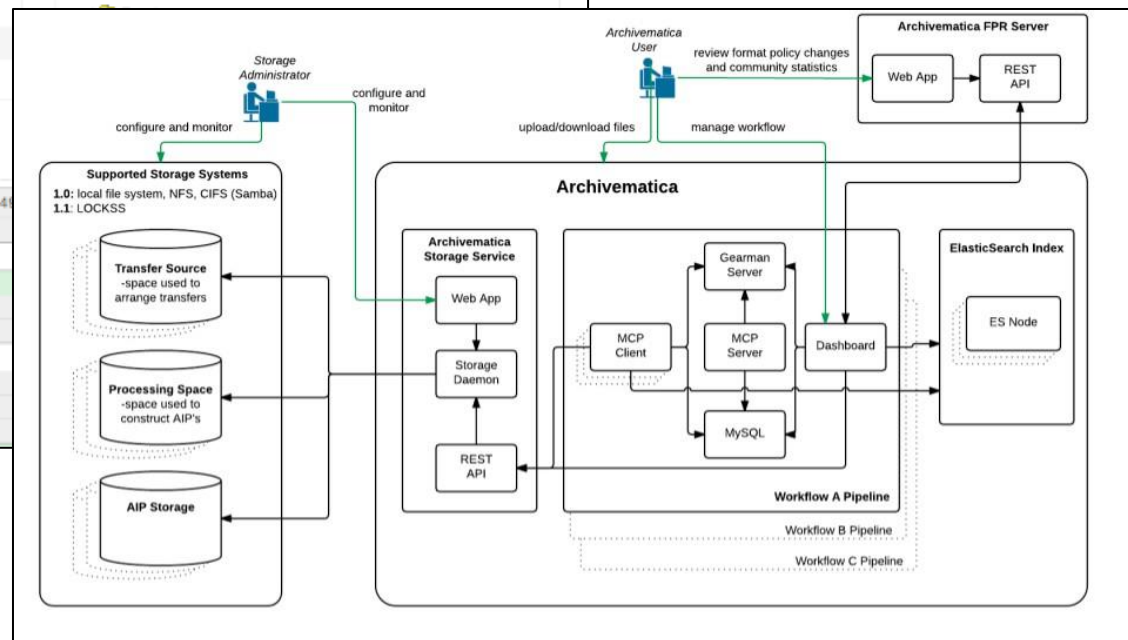
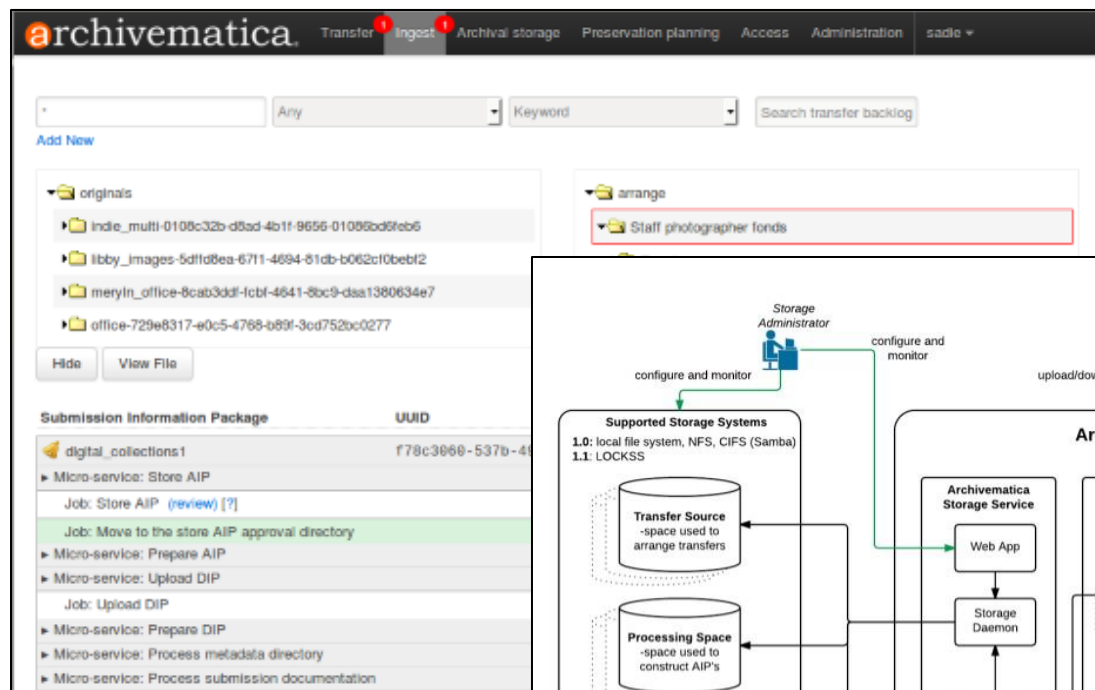
分野横断的なデータキュレーションプロセスを抽出

NII RDCとの対応付け

カテゴリ	活動名称	対応する機能要件(短縮版)	GakuNin RDM	JAIRO Cloud
DataProcessing	配置・記述	あらかじめ定められたパッケージ規格に合わせてフォルダ構造を変換できること	×	×
	コンバージョン	アナログ形式のデータファイルをデジタル形式に転換できること	×	×
	データクリーニング	提出されたデータファイルに含まれる欠陥やエラーを検出できること	×	×
	データの再構造化	提出されたデータファイルの構造を整理できること	×	×
	データ可視化	提出されたデータファイルをもとにしたグラフ・図・表を作成できること	×	○
	匿名化	データファイルに含まれるセンシティブな情報（個人情報、要保護情報など）を書き換えたり削除したりできること	×	×
	ファイルフォーマット変換	データをオープンで非独占的なファイル形式に変換できること	×	×
	リネーム	データのファイル名を変更できること	○	×
	相互運用性	分野別の標準を用いてデータをフォーマットできること	×	×
	証拠保全	一連の処理を通じて、データ一式の履歴情報を記録できること	○	×
	ファイルインスペクション	データ一式を定期的に点検し、データ数、ファイルタイプ（拡張子）、ファイルサイズ等を把握できること	×	×
	コンテキストの付与	データ一式を、関連する出版物、論文、プロジェクト等とリンクさせられること	×	×
	メタデータ生成	所定のスキーマに沿ったメタデータを生成できること	○	○

データ加工に関する活動のカバー率が低い

ソフトウェア候補の選定



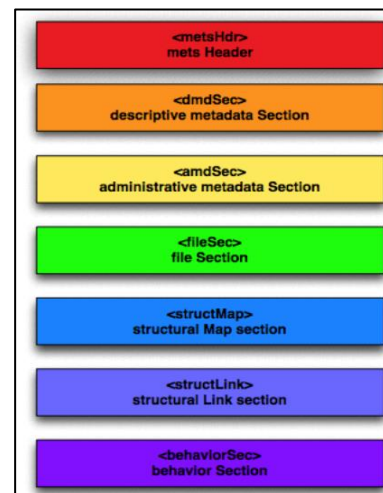
Artefactual Systems社（カナダ）が開発するオープンソース
OAIS参照モデルに準拠し、情報パッケージを作成可能な統合ソフトウェアツール群
分野別のユースケースに合わせた複数パイプラインの制御が可能

Archivematicaができること

1000種類以上のデータフォーマットを特定・変換

入力ファイルに対してMETSメタデータを自動生成

Purpose	Format	Command	Success	Enabled	Actions
Access	Generic JPEG	Transcoding to jpg with convert	0 out of 0	Yes	View Replace Disable
Access	1987a	Transcoding to jpg with convert	0 out of 0	Yes	View Replace Disable
Access	1989a	Transcoding to jpg with convert	0 out of 0	Yes	View Replace Disable
Access	Adobe Photoshop	Transcoding to jpg with convert	0 out of 0	Yes	View Replace Disable
Access	Bitmap 3.0 NT	Transcoding to jpg with convert	0 out of 0	Yes	View Replace Disable
Access	Canon CR2 Raw	Transcoding to jpg with convert	0 out of 0	Yes	View Replace Disable
Access	Digital Negative Format (DNG)	Transcoding to jpg with convert	0 out of 0	Yes	View Replace Disable
Access	Digital Negative Format (DNG)	Transcoding to jpg with convert	0 out of 0	Yes	View Replace Disable
Access	Digital Negative Format (DNG)	Transcoding to jpg with convert	0 out of 0	Yes	View Replace Disable



ヘッダ情報

記述メタデータ情報

管理メタデータ情報

ファイルメタデータ情報

ファイル構造情報

ファイル間リンク情報

リソースの動作情報

管理に適した情報パッケージ生成・管理

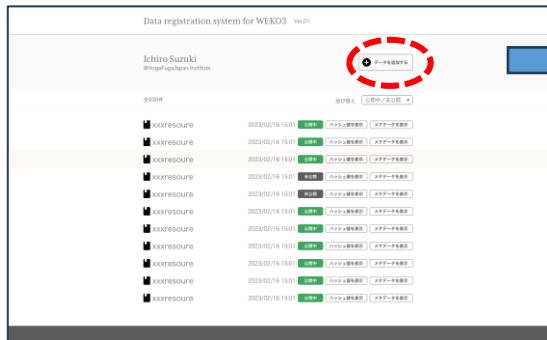
File	File UUID	AP	Actions
arxiv/1008.1792v1.pdf	1d5094d-4800-48d0-aad0-83374b3c3b4	transfer1	Download
arxiv/1008.1792v1.pdf	08323ee-3538-4498-b3ad-924445a2b7f6	transfer1	Download
arxiv/1008.1792v1.pdf	4200a21-2533-47ab-b9c1-833a7783958	transfer1	Download
arxiv/1008.1792v1.pdf	4d8c1de-9890-4ad0-9a47-42215a5c1380	transfer1	Download
arxiv/1008.1792v1.pdf	7300a13-4d34-4342-43a2-712a7843118	transfer1	Download
arxiv/1008.1792v1.pdf	2c3d878-4d48-43ab-4d71-4d4d713d47f6	transfer1	Download
arxiv/1008.1792v1.pdf	11300a7-8407-4d75-adce-25a129a489f	transfer1	Download
arxiv/1008.1792v1.pdf	1131610-21c1-432a-8ad0-9a5052a6a5c	transfer1	Download
arxiv/1008.1792v1.pdf	1c7188b6-2448-4716-b10a-8a7176a736a7	transfer1	Download
arxiv/1008.1792v1.pdf	2a8d443-416d-416d-416d-416d416d416d	transfer1	Download

<https://www.loc.gov/standards/mets/METSPrimer.pdf>
<https://www.archivematica.org/en/docs/archivematica-1.15/>

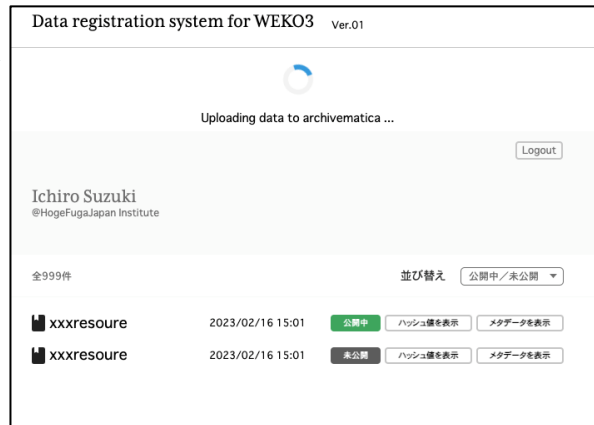
パイプライン制御UIの開発

制御UI

①制御UIへログイン 対象データ選択

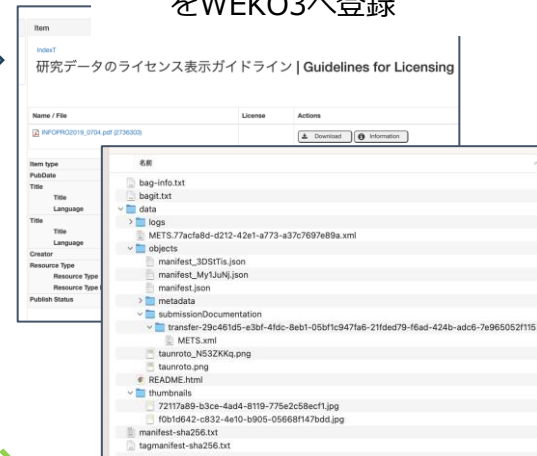


②データ転送



WEKO3

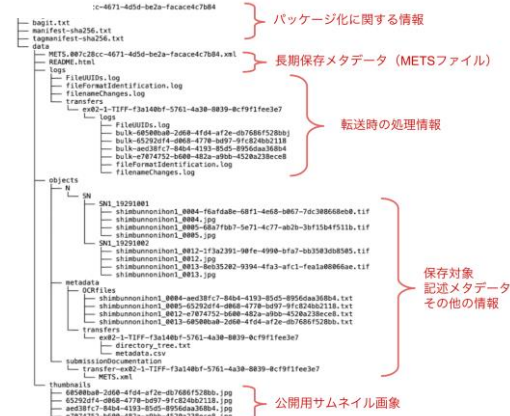
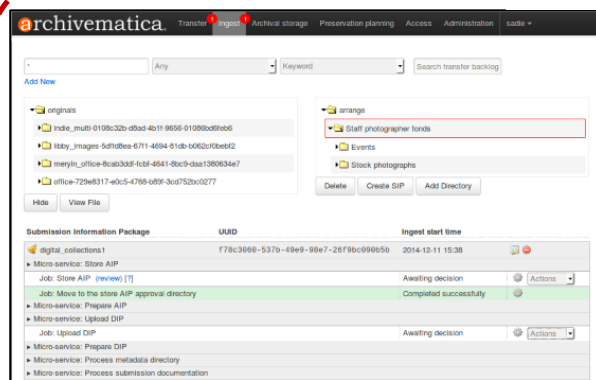
⑥アクセス用パッケージ をWEKO3へ登録



③分野別パイ プライン実行

⑤アクセス用 パッケージ転送

Archivematica



④保存用パッケージ生成 アクセス用パッケージ生成

- ユーザは制御UIを用いて各種リポジトリ (GitHub等) から対象データを取得できる
- 分野別パイプラインはユーザーごとに設定可能
- ユーザ管理はWEKO3と連動して実施可能

人文学データのケーススタディ (データエコ事業ユースケース)

- 渋沢栄一記念財団の文書データ
 - 役職員が在職中に執筆した原稿
 - 渋沢栄一に関する論考や、財団の活動内容を含む記録として長期保存したい
- 文書に多く用いられるフォーマット
 - Microsoft Office (Excel, PowerPoint, Word)
 - プレーンテキスト (txt)
- 階層構造を持つ、合計10ファイルを入力した

```

masahide_DA_1
├── metadata
│   └── metadata.csv
└── ○○執筆原稿
    ├── ○○センター保存分
    │   ├── 200604_慶応大学講座
    │   │   ├── 2006course.doc
    │   │   ├── 慶応大学記念講座開講のお知らせ.txt
    │   │   ├── 慶応講座、栄一、4.11.06.doc
    │   │   └── 添付書類の訂正.txt
    │   └── 201107_新渡戸国際塾
    │       ├── 20110709○○理事長新渡戸国際塾講義 01.pptx
    │       ├── readme02.txt
    │       ├── 新渡戸P P.docx
    │       └── 新渡戸 1 1 .docx
    └── 管理用リスト
        ├── 理事長講演と執筆リスト.xlsx
        └── 理事長講演と執筆リストrk(更新用) .xlsx
    
```

テスト結果

＜作成されたAIPの中身＞

AIP作成後のMETS (PREMIS, DCなどのメタデータを格納)

```
data
├── METS.5fc9f5c4-6f46-4cdf-b00f-7c7b35dff029.xml
├── README.html
├── logs
│   ├── fileFormatIdentification.log
│   ├── filenameChanges.log
│   └── FileUUIDs.log
└── transfers
    └── 2023-10-26_test4-45a464e7-521d-4643-8ee4-613445f4237d
        └── logs
            ├── bulk-2281bbc8-a07e-49e4-8d15-21e7fbee00f
            ├── bulk-266f11eb-752e-4212-a963-40c9d741cb42
            ├── bulk-27380689-38f8-46b4-9cfd-18cff03b87b3
            ├── bulk-2ee70f30-126a-48ce-bc99-9a73c6c10f0f
            ├── bulk-3b459868-46e1-4f38-8291-8dbb84bbe5a9
            ├── bulk-53a03c61-81f0-41f4-88de-317f0572c051
            ├── bulk-8718732f-1c39-4027-a555-3aac307de1ad
            ├── bulk-a1d321eb-601b-45de-aa79-2e8dd3e4c486
            ├── bulk-bcb133ca-345b-457b-ab84-48106c7d1aba
            └── bulk-be67276f-3497-4852-b516-b905a4d58b0d
```

処理ログ

受け入れたデータの初期状態と入力したメタデータ

```
objects
├── metadata
│   └── transfers
│       └── 2023-10-26_test4-45a464e7-521d-4643-8ee4-613445f4237d
│           ├── directory_tree.txt
│           └── metadata.csv
├── Shibui_Ze_Ya_Ying_Zhi_Bi_Yuan_Gao_
│   ├── Guan_Li_Yong_risuto
│   │   ├── Li_Shi_Chang_Jiang_Yan_toZhi_Bi_risuto.xlsx
│   │   └── Li_Shi_Chang_Jiang_Yan_toZhi_Bi_risutork(Geng_Xin_Yong_).xlsx
│   └── Qing_Bao_Zi_Yuan_sentaBao_Cun_Fen_
│       ├── 200604_Qing_Ying_Da_Xue_Jiang_Zuo_
│       │   ├── 2006course.doc
│       │   ├── Qing_Ying_Da_Xue_Ji_Nian_Jiang_Zuo_Kai_Jiang_nooZhi_rase.txt
│       │   ├── Qing_Ying_Jiang_Zuo_Rong_Yi_4.11.06.doc
│       │   └── Tian_Fu_Shu_Lei_noDing_Zheng_.txt
│       └── 201107_Xin_Du_Hu_Guo_Ji_Shu_
│           ├── 20110709Shibui_Ze_Li_Shi_Chang_Xin_Du_Hu_Guo_Ji_Shu_Jiang_Yi_01.pptx
│           ├── readme02.txt
│           ├── Xin_Du_Hu_11.docx
│           └── Xin_Du_Hu_P_P.docx
└── submissionDocumentation
    └── transfer-2023-10-26_test4-45a464e7-521d-4643-8ee4-613445f4237d
        └── METS.xml
```

長期保存用
データ

AIP作成前の
METS

全てのログを残した形で階層構造を持つデータのAIP作成に成功
今後、大学図書館におけるユースケース開発に着手予定

4. まとめ

本日の振り返り

1. はじめに

- ✓ データキュレーションとは
- ✓ 国際的な期待の高まり

2. 国内の動向

- ✓ 政策的な要請／研究者からの期待
- ✓ 大学側の準備体制
- ✓ 課題や障壁

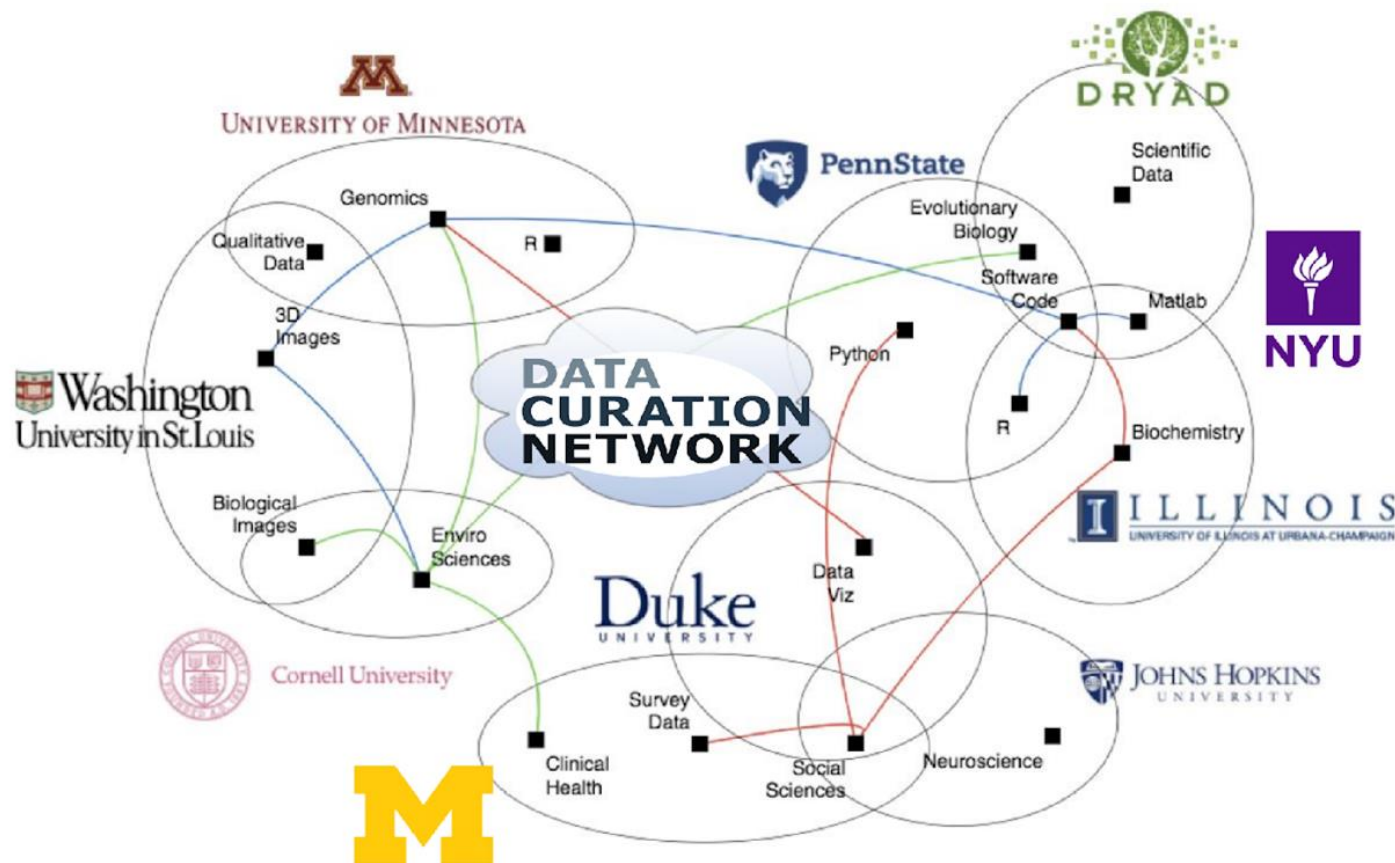
3. 実践に向けて

- ✓ データキュレーション支援機能の開発状況紹介
- ✓ 人文学分野でのケーススタディ

4. まとめ

- 大学図書館との連携を模索中です

データ専門職によるネットワーク構築



データキュレーションの実践には多様なスキルや専門知識が必要
効率的にデータ整備を行うため、専門職ネットワークの構築が進んでいる

将来構想

どの機関でも専門的なデータキュレーションサービスを受けられる体制を構築し、データ再利用の促進に寄与



Japanese
Institutional
Repositories
Online
Cloud

適用



+

データキュレーション
機能

開発

共同運営

J P C O A R
オープンアクセスリポジトリ推進協会

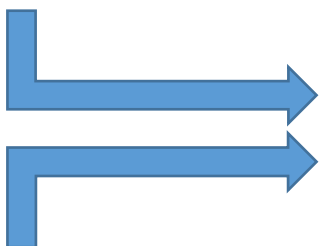
NII

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

開発協力

- 大学図書館員、研究支援職員、研究者（データキュレーター）を含む人的ネットワークを構成
- 単一の機関でカバーしきれない多様な専門分野のデータキュレーションを、大学共同利用機関を含む複数機関間でオンデマンドに支援

参加機関による互助的な
ネットワーク構築



国際連携
(情報交換・技術協力)



Landelijk Coördinatiepunt
Research Data Management

DATA
CURATION
NETWORK



Digital Research
Alliance of Canada

Alliance de recherche
numérique du Canada

