

学術機関における研究教育データの 長期運用保存管理の在り方 TF 成果報告書

2022 年 9 月 30 日

サイエンティフィック・システム研究会
学術機関における研究教育データの長期運用保存管理の在り方
TF

目 次

1. はじめに	2
1.1. TF 発足の背景	2
1.2. 活動方針	2
1.3. 活動内容	2
1.4. 活動期間	2
1.5. TF メンバー	3
1.6. 活動実績	3
2. 総括	6
2.1. チェックリストから洗い出された確認すべき内容	6
2.2. チェックリストには無く別途検討が望まれる内容	10
3. おわりに	12
A. チェックリスト	13
A.1. 京都大学 研究データマネジメントルーブリック	13
A.2. ANDS Checklist	14
B. TF メンバー個別報告書	17
B.1. 名古屋大学 小川泰弘 個別報告書	17
B.2. 広島大学 渡邊英伸 個別報告書	21
B.3. 沖縄科学技術大学院大学 Eddy Taillefer 個別報告書	26
B.4. 北見工業大学 馬場智之 個別報告書	29
B.5. 国立情報学研究所 込山悠介 個別報告書	32
B.6. 豊橋技術科学大学 土屋雅稔 個別報告書	38
B.7. 理化学研究所 舩本現 個別報告書	39
B.8. 富士通 Japan 報告書	44

1. はじめに

タスクフォース（TF）の発足や活動方針について以下に示す。

1.1. TF 発足の背景

学術機関の情報環境はインターネットやWeb 技術の普及により大きく変革している。1970 年代の大型計算機の導入に始まり、その後、汎用機の利用が増えることにより科学技術計算を中心とした研究支援、図書情報など学術情報検索、教務・人事・会計などの各種事務情報の処理などに展開された。1990 年代には、インターネットやWeb 技術が普及したこととともない、学術機関においてもキャンパスネットワークが全学整備され、個々に稼働していた計算機システムを遠隔から利用する機会が増加するとともに、教職員や学生のコミュニケーションにおいても電子メールなどのネットワークを中心とするアプリケーションが日常的に利用されるようになった。2000 年代に入るとミッションクリティカルな情報基盤へと成長し、柔軟かつ迅速なサービス提供を可能にしつつ大規模災害等における事業継続に対応すること等を目的に情報基盤はクラウド上で展開されるまでに至っている。ネットワーク基盤も有線環境のみならず無線環境が整備され、スマートフォンやタブレット等の登場によりユーザ利用環境も大きく変わり、SNS などの普及もあいまって、誰もが場所を選ばず気軽に情報を生成し発信することができ、また、他者が発信した情報も加工され容易に共有することも可能になったことで、コミュニケーションは送り手から利用者の一方通行型だったものが、送り手と利用者間の双方向型となり、情報発信が組織主体から利用者主体にシフトするまでに至っている。

インターネットを中心とした社会において、研究活動や教育活動においてもデジタル時代に適合することが重要になってきている。2009 年に米国の計算機科学者である Jim Gray は、これまでに発展してきた実験科学（第 1 の柱）、理論科学（第 2 の柱）、計算科学（第 3 の柱）に対して、第 4 の柱としてデータ指向型科学を提唱した。また、2013 年の G8 の科学技術大臣会合の共同宣言（※1）をきっかけに、ICT を活用してあらゆる人々が学術的研究や調査の成果等を発信・共有を可能にすることで科学を変容させるオープンサイエンスの取り組みが国際的に加速している。2021 年 3 月 26 日に閣議決定された、「第 6 期科学技術・イノベーション基本計画」（※2）では、Society 5.0 の実現に向けた科学技術・イノベーション政策の一つに、知のフロンティアを開拓し価値創造の源泉となる研究力の強化が掲げられており、強化に向けた一つの取り組みとして新たな研究システムの構築（オープンサイエンスとデータ駆動型研究等の推進）が示されている。このように、これからの学術機関は、社会全体のデジタル化や世界的なオープンサイエンスの潮流を捉えた研究そのもののデジタルトランスフォーメーション（DX）を通じて、より付加価値の高い研究成果を創出することが要求されている。2019 年 5 月には、大学 ICT 推進協議会より「学術機関における研究データ管理に関する提言」（※3）が刊行され、2022 年 1 月にはデジタル庁、総務省、文部科学省、経済産業省より教育データ利活用ロードマップ（※4）が公表されている。現在、研究教育分野の DX 化による改革を目標に研究教育分野のデータ利活用に関する議論が活発に行われており、研究や教育データの利活用ガイドラインやポリシー策定のためのひな型も提供されたことにより、多くの学術機関では、研究データ管理ポリシーや教育・学習データ利活用ポリシーの策定が進められている状況にある。

研究 DX や教育 DX を支える情報基盤の構築が徐々に増えつつある一方で、研究データの長期保存や生涯学習など数十年以上の長期的なデータを利活用する際の情報環境の運用管理についてはあまり検討されていない。国が一元的に研究教育データを管理するデータベースを構築することは考えられていないことから、学術機関が連携して研究教育データ利活用の情報基盤をどのようにして長期的に運用管理していくべきかについて議論することは重要と考える。

本 TF は、情報環境の運用管理の観点から長期的な研究教育データ運用保存管理について様々な学術機関の意見をもとに要件を洗い出し報告書としてまとめることを目的に発足するものである。

※1

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/206801/G8_Science_Meeting_Statement_12_June_2013.pdf

※2 <https://www8.cao.go.jp/cstp/kihonkeikaku/index6.html>

※3 <https://axies.jp/report/publications/proposal/>

※4 https://www.digital.go.jp/news/a5F_DVWd/

1.2. 活動方針

研究DXや教育DXにおいては、学術機関の情報基盤を整備し運用管理する情報系センターに加え、データ公開やリポジトリ管理を担う図書館、研究者を教育支援するUniversity Research Administrator (URA) など様々な組織や教職員が同じ目的意識を持ちながら推進するものであり、分野・職種の壁を越えた横断的な連携が必要不可欠である。そのため、トップダウン的な取り組みが重要となることから、研究データの利活用に関する提言（※3）や教育データの利活用に関する提言（※5）、研究データポリシー策定のためのガイドライン（※6）や教育・学習データ利活用ポリシーのひな型（※7）などが既に公表されている状況にある。

一方で、ICTに関するテクノロジーは日々目まぐるしく変わり、セキュリティ脅威が増している中、技術への対応やセキュリティ対策のためにサービスの更新スピードが増しており、情報環境の更新にともなうユーザサポートの負担も増すばかりである。さらに、学生数の減少による収入減や高等教育機関に国から配分される資金の削減が続くという状況の中、学術機関の情報環境整備は人材・予算の両面において危機的な状況に陥っている。2020年12月に大学ICT推進協議会から刊行された「多様な教育研究活動の高度化を支える大学ICT基盤の集約化・共通化・協働化」に関する提言（※8）にも示されているように、学術機関におけるDXが本格的に始まろうとしている中で、情報環境整備の限界を乗り越えるためのパラダイムシフトを可能にするためには、ICT 基盤を学内・地域内・国内の様々なレベルで集約化・共通化・協働化すべきである。

そこで、本TFでは、学術機関の情報基盤を運用管理する情報系センターの教職員を中心に、情報環境やシステム運用管理の観点から、ボトムアップ的に研究教育データの長期運用管理で今後求められる要件を洗い出し、報告書としてまとめることを方針とする。本報告書により、学術機関における研究教育データの長期運用保存管理の情報基盤整備の集約化・共通化・協働化に寄与することを目指す。

※5 <https://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-24-t299-1.pdf>

※6 https://rdm.axies.jp/_media/sites/14/2021/07/urdp-guideline.pdf

※7 <https://axies.jp/report/publications/formulation/>

※8 <https://axies.jp/report/publications/dxtf/>

1.3. 活動内容

TF前半では、研究データ管理に関連する取り組みを調査し、議論する対象範囲や方向性を整理した。その中で、全ての研究データおよび教育データにおいて議論することが難しいことから、議論する対象を研究データに絞ることにした。また、京都大学が和訳して公表した研究者観点で研究データ管理要件の対応状況を確認できるルーブリックとAustralian National Data Service (ANDS※現在はAustralian Research Data Commons (ARDC) に統合されている) が公開したシステム管理者観点で研究データ管理要件の対応状況を確認できるチェックリストを題材に議論する方針とした。

TF後半では、各TFメンバーが自組織の状況をチェックリストで確認することを通じて、チェックリストの不明瞭な点や不足している事項を洗い出すとともに、研究教育データの長期運用管理で今後求められる共通の要件および追加で求められる要件を整理した。2022年9月に活動を終了し、これまで行ってきた調査、検討、議論をもとに本報告書としてまとめた。

本TFの報告書が、分野を超えて研究DXの情報環境整備および長期運用管理に関する検討の一助となることを期待する。

1.4. 活動期間

2020年7月～2022年9月

1.5. TF メンバー

		氏名	機関/所属(2022年9月15日現在)
会員	担当幹事	小川 泰弘	名古屋大学
	推進委員	○渡邊 英伸	広島大学
		Eddy Taillefer	沖縄科学技術大学院大学

		馬場 智之	北見工業大学
		込山 悠介	国立情報学研究所
		土屋 雅稔	豊橋技術科学大学
		舩本 現	理化学研究所
賛助会員 (富士通)	推進委員	井上 俊治	富士通 Japan (株)
		滝沢 亮	富士通 Japan (株)
		乃万 智洋	富士通 Japan (株)
		坂口 吉生	富士通 (株) 官公庁事業本部官庁第六事業部

○印：まとめ役

1.6. 活動実績

- 第1回会合：2020年7月9日(木) オンライン (Zoom)
 - ・活動方針の確認
 - ・TFとしての狙い/スケジュールについて議論
 - ・テーマ設定/次回に向けた宿題確認
- 第2回会合：2020年9月28日(月) オンライン (Zoom)
 - ・京都大学のデータマネジメントポリシーについて情報共有、議論
 - ・NIIのRDMポリシーに基づくシステム構築手法について情報共有、議論
- 第3回会合：2020年12月15日(火) オンライン (Zoom)
 - ・「クラウド時代のデータ共有技術WG」情報提供・意見交換
 - ・第2回会合で整理した検討すべき4つの項目 (ループブリック、キュレーション、セキュリティ、学認RDM) 以外に検討項目がないか議論
- 第4回会合：2021年2月26日(金) オンライン (Zoom)
 - ・各機関のデータポリシーの共有
 - ・メンバーでANDSのループブリック (システム管理者観点) に回答した結果を共有
- 第5回会合：2021年4月28日(水) オンライン (Zoom)
 - ・京都大学の研究データマネジメントループブリックを各メンバーの研究プロジェクトに当てはめて研究者の観点で検証
- 第6回会合：2021年6月30日(水) オンライン (Zoom)
 - ・各メンバーの回答から京都大学の研究データマネジメントループブリックを「再利用のために最適化」というレベルに持っていくために何が必要か、というポイントや技術的な観点から意見を整理
- 第7回会合：2021年8月23日(月) オンライン (Zoom)
 - ・課題の深掘りとセキュリティ、マネタイズ、人材などの追加観点を洗い出し
- 第8回会合：2021年10月19日(火) オンライン (Zoom)
 - ・更なる課題の深掘りと追加観点を洗い出し
- 第9回会合：2021年11月30日(火) オンライン (Zoom)
 - ・ANDSチェックリストの要求項目でどのような技術やサービスが利用可能かを洗い出し
 - ・ID管理、認証、アクセス
 - ・内部・外部ネットワーク接続性
 - ・分析支援用の分野固有のツールへのアクセス
- 第10回会合：2021年12月21日(火) オンライン (Zoom)

- ・NII込山先生による「研究データポリシーの作成に向けて」の情報共有
- ・ANDSチェックリストの要求項目でどのような技術やサービスが利用可能かを洗い出し
 - ・ソフトウェア開発

■第11回会合：2022年3月1日(火) オンライン (Zoom)

- ・ANDSチェックリストの要求項目でどのような技術やサービスが利用可能かを洗い出し
 - ・可視化
 - ・コラボレーション環境

■第12回会合：2022年4月27日(水) オンライン (Zoom)

- ・ANDSチェックリストの要求項目でどのような技術やサービスが利用可能かを洗い出し
 - ・ハイパフォーマンスコンピューティング
- ・報告書目次検討

■第13回会合：2022年6月10日(金) オンライン (Zoom)

- ・報告書のまとめ方の方針決め
- ・個別報告書の内容の整理
- ・個別報告書のテンプレートの共有

■第14回会合：2022年8月5日(金) オンライン (Zoom)

- ・各メンバーの個別報告書の内容の共有
- ・報告書の構成決め

■第15回会合：2022年9月15日(木) オンライン (Zoom)

- ・報告書の内容整理

2. 総括

本報告書は、ANDS のチェックリストで提示されている確認項目をベースに、「京都大学の研究データマネジメントループリックで研究者に求められている要求事項を満たすかどうか」、「国内の学術機関で情報基盤を運用管理する部署にとって活用できるものかどうか」、「別途確認すべき項目があるかどうか」について、議論したものである。なお、「●」の項目はANDS のチェックリストの項目、「✓」の項目は本 TF で追加して確認すべき事項である。

2.1. チェックリストから洗い出された確認すべき内容

1. Storage for data and metadata

本 TF では、ANDS のチェックリストで提示されている以下の確認事項は研究データ管理のインフラという点において情報系センターなど情報基盤を運用管理する部署にとって重要な要件であると考ええる。これらの確認事項を満たすことにより、京都大学の研究データマネジメントループリックの「データの保存とバックアップ」の項目で「再利用のための最適化」で示される要求に対しても応えることができるものと判断する。

- 特定のデータセットに十分なストレージがありますか？もしそうなら、それは永続的なストレージですか、それとも一時的なものですか？
- データセットは複数の場所でバックアップまたはアーカイブされていますか？
- データセットとそのメタデータが分離している場合、それらは一緒に保存されますか、それとも分離される可能性がありますか？データとメタデータはどの程度密接にリンクされていますか？
- 新しいデータストレージテクノロジーが利用可能になったときに、データセットをそれに移行できますか？
- データセットは、基本的なアクセスのために圧縮またはデータ変換ツールと標準に依存していますか？もしそうなら、それらのツールと標準へのアクセスは継続されますか？
- データセットは、基本的なアクセスのために独自のツールまたはフォーマットに依存していますか？もしそうなら、それらのツールやフォーマットへのアクセスは継続されますか？

また、情報系センターなど情報基盤を運用管理する部署にとって上記以外にも必要に応じて追加で確認すべき事項があると考ええる。

- ✓ 一人あたりの容量制限をきちんと定めることができるか
- ✓ 容量の拡張性を有しているか
- ✓ 検索性があるか
- ✓ 障害時などにデータの復旧が確実にできるか
- ✓ 平時の保守・管理コストを数値化できるか
- ✓ 容量の拡張時、データ移行時、他システムとの連携時など一時的な追加コストを数値化できるか
- ✓ ストレージの複雑な構成においても統一的な運用管理が可能な機能を有しているか
- ✓ 保管場所の物理セキュリティや災害対策が確保されているか
- ✓ 保管方法、バックアップ方法、復旧方法は容易かつ安全か
- ✓ 使用状況を確認する機能があるか
- ✓ 容量低下時の通知機能を有しているか
- ✓ 容量不足時に古いデータから削除されることがないか

2. Identity management, authentication & access

本 TF では、ANDS のチェックリストで提示されている以下の確認事項は研究データ管理のインフラという点において情報系センターなど情報基盤を運用管理する部署にとって重要な要件であると考ええる。これらの確認事項を満たすことにより、京都大学の研究データマネジメントループリックの「データの共有と公開」の項目で「再利用のための最適化」で示される要求に対しても応えることができるものと判断する。

- IT インフラストラクチャは、すべてのデータセットに必要なアクセス制限を記録してサポートできますか。パスワード、AAF ID、倫理、アクセス契約および契約？
- IT インフラストラクチャは、研究者がどこにいても、特定のデータセットでのコラボレーションをサポート

できますか？

- IT インフラストラクチャは、認証されたアクセスを追跡および識別するために必要な登録をサポートできますか？
- 将来のある時点で（たとえば、禁輸期間が経過した後）、特定のデータセットのアクセス制限を変更することは可能ですか？

また、情報系センターなど情報基盤を運用管理する部署にとって上記以外にも必要に応じて追加で確認すべき事項があると考えます。

- ✓ 先進認証に対応しているか
- ✓ データに対する暗号化やパスワードロック機能など盗難時のセキュリティ機能を有しているか
- ✓ データの故障や改ざんを検証できる機能を有しているか
- ✓ データに対する操作履歴を確認できる機能を有しているか
- ✓ ストレージの運用管理における必要な種類・期間のログの記録が可能か
- ✓ アカウントの利用停止措置が可能な機能を有しているか
- ✓ 認証連携の可否を確認しているか
- ✓ ORCID など研究者 ID と同一研究者の既存アカウントとの対応付けや制御が可能なことを確認しているか
- ✓ API を介して外部の学術システムとの連携や制御ができることを確認しているか

3. Internal & external network connectivity

本 TF では、ANDS のチェックリストで提示されている以下の確認事項は研究データ管理のインフラという点において情報系センターなど情報基盤を運用管理する部署にとって重要な要件であると考えます。これらの確認事項を満たすことにより、京都大学の研究データマネジメントルーブリックの「データ保存とバックアップ」および「データの共有と公開」の項目で「再利用のための最適化」で示される要求に対しても応えることができるものと判断する。

- データセットが大きすぎたり、コストがかかりすぎて内部ネットワークや外部ネットワークを介して転送できない場合、フォールバック転送メカニズム（宅配便で送信されるハードドライブなど）はありますか？
- あなたの機関は内部データ転送にどのようなデータ転送コストを課していますか？他の機関との間の外部転送はどうですか？
- 特定のデータセットが高い需要を引き付ける場合、IT インフラストラクチャは予想されるアクセスおよびコピー要求に対処できますか？

また、情報系センターなど情報基盤を運用管理する部署にとって上記以外にも必要に応じて追加で確認すべき事項があると考えます。

- ✓ 必要なデータ転送性能を有するネットワークが確保されているか
- ✓ 回線使用量を数値化できるか
- ✓ 回線使用料を確認できるか
- ✓ 通信路の暗号化、アクセス制御などの機能を有しているか
- ✓ ネットワークの運用管理における必要な種類・期間のログの記録が可能か
- ✓ ネットワークの複雑な構成においても統一的な運用管理が可能な機能を有しているか
- ✓ ネットワークのセキュリティ対策や災害対策が確保されているか
- ✓ ネットワークの冗長化が確保されているか
- ✓ バックアップ回線の切り替え方法、復旧方法が確立されているか

4. Access to discipline specific tools to support analysis

本 TF では、ANDS のチェックリストで提示されている以下の確認事項は研究データ管理のプラットフォームという点において図書館などデータ管理を支援する部署にとって重要な要件であると考えます。これらの確認事項を満たすことにより、京都大学の研究データマネジメントルーブリックの「データを解析可能な状態にする」および「データ解析と成果の取り扱い」の項目で「再利用のための最適化」で示される要求に対しても応えることができるものと判断する。ただし、組織によっては分析ツールの利用環境は利用者（研究者自身）の責任のもと準備する方針としている場合がある。その場合は、この限りではない。

なお、GakuNin RDM を利用する場合は、下記の要求事項はGakuNin RDM 側で満たすことが望まれる。参考としてGakuNin RDM のデータ解析機能（※9）の情報も共有しておく。

- あなたの機関は、すべての専門分野のニーズを満たすのに十分な範囲の分析ツールを利用できるようにしていますか？それらを必要とするすべての人がアクセスできますか？
- データセットを分析するために特定のツールへのアクセスが必要な場合、そのツール（統計分析パッケージなど）はアーカイブされていますか？そのツール用に開発された構成またはカスタマイズスクリプトはアーカイブされていますか？
- データセットが商用データ分析ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
- データセットが特定の機器または機器へのアクセスを必要とし、これが特定の構成を必要とする場合、これをデータセットメタデータで指定するための規定はありますか？

また、情報系センターなど情報基盤を運用管理する部署にとって上記以外にも必要に応じて追加で確認すべき事項があると考え。

- ✓ 分析ツールの利用環境に関して責任分解点を明確にしているか
- ✓ 資産管理ともあわせてソフトウェアライセンスの一元管理機能を有しているか
- ✓ ソフトウェアライセンス契約に遵守した運用管理が可能か
- ✓ ソフトウェアライセンスの定期的な棚卸の方法が確立されているか
- ✓ ソフトウェアライセンスの継続性があるか
- ✓ ツールを用いて分析・解析処理するための計算機リソースを必要な時に必要な分だけ柔軟に提供できるか
- ✓ 分析・解析環境は占有して利用でき、他の処理の影響を受けないか
- ✓ 安全なアクセス方法を提供しているか
- ✓ 研究データ管理用のストレージと連携できるか
- ✓ アカウント管理、ユーザ管理、先進認証対応、コスト管理に関する機能を有しているか
- ✓ リソースの管理、監視、アクセス制御、操作履歴、アップデート、バックアップ、セキュリティ対策に関する機能を有しているか
- ✓ 統一的な利用者・リソース管理、不要サービス停止・再開などのシステム管理機能を有しているか
- ✓ データ、メタデータ、分析ツール、処理結果を含む分析環境の提供可否について確認しているか

※9 https://www.nii.ac.jp/openforum/upload/20220601_nii_rdc4_5.pdf

5. Software development

本 TF では、ANDS のチェックリストで提示されている以下の確認事項は、利用者の責任のもので実施するものであると考え、国内の学術機関において組織的にソフトウェア設計や開発を支援する事例が非常に少ないことやソフトウェア開発を外注するケースが多いこと、独自仕様のソフトウェア開発は望まれないことも踏まえ、省略可能なオプションの要求事項と判断した。組織内にソフトウェア開発部隊が存在する場合や共同研究等を通じてソフトウェアの設計や開発を組織的に支援する場合は、以下の確認事項を参考にしたい。

- 機関には、必要に応じてソフトウェアを開発または変更する能力がありますか？
- 機関には、時間のかかるカスタム開発の開発または変更に対する代替ソリューションを推奨する専門知識がありますか？
- データセットがカスタム開発されたソフトウェアによって生成された場合、データセットを使用してデータセットを構築するためのそのソフトウェアとドキュメントはアーカイブされていますか？
- データセットが市販のソフトウェアまたは既製のソフトウェアによって生成された場合、データセットのカスタマイズまたは構成設定はアーカイブされていますか？商用ソフトウェアにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
- データセットが基本的なアクセス（解凍など）のためにカスタムまたは既製のソフトウェアを必要とする場合、メタデータで指定されたデータにアクセスするために必要な設定またはカスタマイズがありますか？

6. Visualisation

本 TF では、ANDS のチェックリストで提示されている以下の確認事項は、「4. Access to discipline specific tools to support analysis」で示した考えと同様な扱いで問題無いと判断した。

- 機関は、すべての要件を満たすために、十分な範囲の視覚化ツールまたはそれらへのアクセスを利用できるようにしていますか。
- データセットでサポートする視覚化が必要な場合、視覚化アセットとツールはデータセットとともにアーカイブされていますか？
- データセットメタデータを使用して視覚化設定または構成値を記録する機能はありますか？
- サポートする視覚化で繰り返しまたは表示するためにハイパフォーマンスコンピューティングリソース (HPC) にアクセスする必要がある場合、データセットメタデータに関連する要件を記録する機能はありますか？

また、情報基盤を運用管理する部署にとって上記以外にも必要に応じて追加で確認すべき事項も同様であり、組織のサービスとして提供する場合、責任分解点、ライセンス管理、ツールを含むデータ解析・可視化用の処理環境などに関連する項目についても確認するべきと考える。具体的な確認事項は「4. Access to discipline specific tools to support analysis」を参照していただきたい。

7. Collaborative environments

本 TF では、ANDS のチェックリストで提示されている以下の確認事項は研究データ管理のプラットフォームにおけるコラボレーション環境の位置づけのものと考え、図書館などデータ管理を支援する部署にとって重要な要件であると判断した。

多くの学術機関では、Microsoft365 や Google Workspace などの SaaS をコラボレーション環境として情報系センターが提供している場合が多いと考える。ただし、このようなコラボレーション環境は学内の構成員向けに、研究だけでなく教育や普段の大学生活で利用することを前提とした汎用的なコラボレーション環境として提供するものが多いことから、所属機関、専門分野、国境を超えた協同による研究が目的のオープンサイエンスの研究データ管理の観点では適切なコラボレーション環境とは言い難く、研究データ管理専用かつ極力共通のコラボレーション環境が提供されることが望ましいと考えるに至った。コラボレーション環境の実現方法は学術機関や研究者によって多種多様となるため、ここでは、研究データ管理専用かつ共通のコラボレーション環境の一例として GakuNin RDM に焦点をあて、確認事項の対応状況ならびに課題について情報を共有するにとどめることにした。

なお、これらの確認事項を満たすことにより、京都大学の研究データマネジメントルーブリックの「データを解析可能な状態にする」、「データ解析と成果の取り扱い」、「データの共有と公開」の項目で「再利用のための最適化」で示される要求に対しても応えることができると判断する。

(本報告書 B.5. 国立情報学研究所 込山悠介 個別報告書より引用)

1. 機関は、すべてのニーズを満たすために、適切なコラボレーション環境またはそれらへのアクセスを提供していますか。
 - GakuNin RDM ではサービスの導入機関が、研究者がコラボレーションを行なうための作業環境と、学認フェデレーションによるアクセスを提供している。
2. データセットがコラボレーション環境内に保存されている場合、またはデータセットにアクセスまたは分析するためにコラボレーション環境へのアクセスが必要な場合、データセットメタデータに環境のセットアップと構成を保存する機能はありますか？コラボレーション環境を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることはできますか？
 - GakuNin RDM のコラボレーション環境に保存されたデータセットへアクセスすることができ、データ解析機能で分析するための設定ファイルをデータセットと合わせて保存することができる。
 - GakuNin RDM では拡張ストレージ機能とデータ解析機能でそれぞれカスタマイズを保存することができ、コラボレーション環境のためにソフトウェア構成を保存することができる。
3. データセットが商用のコラボレーション環境へのアクセスを必要とする場合、商用環境にアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
 - GakuNin RDM ではパブリッククラウド等の商用のコラボレーション環境へアクセスせずに、他の研究者が利用できる代替のストレージ環境を提供している。また、GakuNin RDM 自体は商用クラウドストレージ環境の代替機能をもつオープンソースソフトウェアといえる。

4. 将来の追加または拡張に必要な場合、データセットを新しいまたは別のコラボレーション環境に移動することは可能ですか。
 - GakuNin RDM では、ユーザは拡張ストレージ領域を、情報システム担当者は機関ストレージ領域を接続して利用することで、データセットを新しいまたは別のコラボレーション環境へ移動することができる。

TF で挙げられた GakuNin RDM のコラボレーション環境としての期待する内容等についても以下に示す。

- ✓ 学認の参加機関であれば認証連携により GakuNin RDM のコラボレーション環境を利用できるが、学認に参加していない機関の場合は別途認証が必要になるため、対応が期待される。

なお、GakuNin RDM は令和 4 年度～令和 8 年度にかけて取り組む内容を公開している（※10）。こちらも参考になれば幸いである。

※10 https://www.mext.go.jp/content/20220701-mxt_jyohoka01-000023472_01.pdf

8. High performance computing (HPC)

本 TF では、ANDS のチェックリストで提示されている以下の確認事項は研究データ管理のインフラという点において情報系センターなど情報基盤を運用管理する部署にとって重要な要件であると考ええる。HPC 施設を所有する学術機関は、これらの確認事項を満たすことにより、京都大学の研究データマネジメントルーブリックの「データを解析可能な状態にする」および「データ解析と成果の取り扱い」の項目で「再利用のための最適化」で示される要求に対しても応えることができるものと判断する。HPC 施設を所有しない学術機関においても、共同施設の HPC 施設を利用できる場合がある。その際に研究者に対して下記の事項の対応状況を共有する目的で確認しておくことは重要と考える。

なお、GakuNin RDM と連携した HPC 施設の事例は現時点ではないものの、現在データ活用社会創成プラットフォーム mdx との連携が進められており、今後に期待する。

- データセットにアクセスまたは分析するために HPC 施設へのアクセスが必要な場合、データセットメタデータで施設のセットアップと構成を指定することは可能ですか。
- HPC 機能を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることは可能ですか？
- データセットが商用 HPC ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
- デスクトップシステムの容量が増えるにつれて、データを移動できるメカニズムはありますか？

2.2. チェックリストには無く別途検討が望まれる内容

本TFでは、ANDSのチェックリストで提示されていない以下の項目についても検討する必要があると考える。

- ✓ データ管理計画 (DMP:Data Management Plan)
 - ANDSのチェックリストでは、京都大学の研究データマネジメントルーブリックの「データマネジメント計画」の項目に該当する確認事項が無かった。日本では、JST、AMED、NEDOなどの資金配分機関が主催する大型研究プロジェクトにおいてはDMPの作成が要求されるようになってきており、研究データ管理計画書の作成を支援するツールも提供されてきているため、今後DMP作成の支援サービスが求められる可能性は高い。研究データ管理計画が曖昧なままプロジェクトが進むと、インフラの管理ポリシーとの不一致が起り、システム管理者、研究プロジェクトメンバー双方に負担をかけることになるため、図書館などデータ管理を支援する部署や情報系センターなど情報基盤を運用管理する部署が連携して事前に運用管理の観点で確認しておくべき具体的な事項の検討が望まれる。
- ✓ メタデータの付与方法
 - ANDSのチェックリストでは、京都大学の研究データマネジメントルーブリックの「データの体系的整理」の項目に該当する確認事項が無いと判断した。研究データマネジメントルーブリックによる自己評価と共通認識の醸成の論文で示されている通り、「データの体系的整理」ではメタデータレベルでデータを記述し整理する方法がルールまたはマナーとして定着していることとされているため、本TFでは、メタ

データの付与がポイントになると考えた。メタデータを個々の研究データに付す作業が面倒だと感じる研究者は多いと想像できるため、今後メタデータ付与の支援サービスが求められる可能性は高い。検索やデータの理解に関するメタデータはもちろんのこと、処理、保存、権利など管理するために必要なメタデータについても、どのように付与し管理・共有していくべきかについてあらかじめ整理されることが望まれる。

- ✓ ユーザインターフェースのシンプルさ、使いやすさ
 - 基本的にデータ管理計画、メタデータ付与、データ管理は研究者の責任のもと実施するものであるが、知識や経験が少ない研究者は多く存在する。長期的に維持管理してもらうためには支援ツールや情報サービスのユーザインターフェースが非常に合理的で使い勝手が良くなるよう設計されることが期待される。加えて、ツールやサービスに慣れてくると、解析手法の自動化・文書化、世代管理、コスト管理などより便利な機能の要望が挙がることも想像されるため、図書館などデータ管理を支援する部署や研究者を教育支援するURAなどが連携してユーザサポートの観点からユーザインターフェースのシンプルさならびに使いやすさを事前に確認することが望まれる。
- ✓ 退職・異動や所有者不在時のデータの移管や権限移行
 - 長期間のデータ管理においてデータ所有者の退職・異動や死亡などによってデータ移管や権限移行の要求があることは容易に想像できる。組織のデータポリシー、研究データ計画、法律・規則など、様々な項目を関係者・関係機関と連絡を取りながら確認し手続きを進めることになる。事前にデータ移管や権限移行の条件、データ移管や権限移行の対象・手順、データ移管や権限移行が不可時の対処などを定めて運用している場合が多いものの、人手が介在するオペレーションとなっており、その作業にともなう負荷やコストは非常に大きい。基本的に研究データは削除することが無く膨れ上がる一方であるため、データの移動やコピーの頻度が少ないデータ移管や権限移行の方法を事前に検討することが望まれる。
- ✓ 教育・訓練・コンサルティング
 - 研究データは重要な情報資産ではあるものの、適切なデータ管理がされない場合、計算機リソースをひっ迫し、最悪の場合サービスの継続性が失われる場合もあるため、研究者に対して定期的に研究データ管理の教育・訓練が重要となる。研究データ管理に関する相談窓口としてコンサルティング担当部署を設置して支援する組織も存在している中、人材不足が大きな課題となっている。多くの学術機関では研究DXを推進していることもあり、研究データ管理に係る業務をICTの技術を活用してデジタルトランスフォーメーションしていくための情報基盤はどうあるべきかを事前に検討することが望まれる。

3. おわりに

本報告書は、本 TF メンバーが所属する国内の情報系センターが運用管理している情報基盤や研究プロジェクトを例に ANDS のチェックリストで提示されている確認項目が、「京都大学の研究データマネジメントループリックで研究者に求められている要求事項を満たすかどうか」、「国内の学術機関で情報基盤を運用管理する部署にとって活用できるものかどうか」、「別途確認すべき項目があるかどうか」について議論したものである。

世界的に最良な研究データ管理の在り方が定まっていない現状において、理想的な研究データの長期運用保存管理の在り方を検討したものの、多種多様な観点があり複雑すぎて具体的な要件まで提示することはできなかった。ただし、研究データ管理ではデータ公開やリポジトリ管理を担う図書館や研究者を教育支援する URA など様々な組織や教職員との連携によって実現されるものであるため、その方々も踏まえてさらなる議論を展開することにつながる報告書にはなつたと考える。また、TF メンバーによる個別報告書も今後チェックリストに回答する機関においては参考になるものであり、GakuNin RDM の対応状況やループリック・チェックリスト自体の改善につながる有益な情報となっている。研究データ管理に携わる全ての方々に、本報告書が一助けになれば幸いである。最後に、本報告書は TF メンバー、大学関係者の議論や助言をいただいたことで纏めることができた。改めてご支援・ご協力頂いたすべての皆様に深謝致します。

A. チェックリスト

A.1 京都大学 研究データマネジメントルーブリック (和訳)

	その都度	1度のみ	有効に活用	再利用のために最適化
データマネジメント計画	データを入手した段階で、自分なりに考える。標準化や文書化はしていない。	最初に計画を作成するが、通常は、途中で見直すことはしない。	詳細な計画を作成し、プロジェクト期間を通じ参照と見直しを行う。	自分あるいは他人が将来スムーズにデータを利用できるように計画している。
データの体系的整理	一貫した手法に従っていない。そのため、しばしば探し出すのに時間を要する。	手法はあるが、それを適用するのは、プロジェクト終了時のみ。	先を見越した手法を導入しているが、必ずしも標準化されていない。	他人が自分でデータを検索、内容を理解し、利用できる。
データの保存とバックアップ	作業中にどのデータが重要であるかを決め、通常はこれを特定の1か所に保存する。	どのデータを保存すべきかを理解している。損失のリスクを緩和するため、作業後にバックアップをとる。	作業中でも定期的にデータを保存するシステムを用いている。複数のバックアップを保持している。	再利用する機会を最大限生かせるよう考慮した方法と場所にデータを保存している。
データを解析可能な状態にする	標準化もしくは十分に文書化された手続きはない。	解析に適した形式を検討し整備するが、事例ごとに異なる方法をとっている。	データ処理方法は標準化され、かつ十分に文書化されている。	将来も自分自身及び第三者が利用しやすい方法で、データ処理を行っている。
データ解析と成果の取扱い	解析や実験がどのような手順やパラメータにより実施されたか確認するために、これらをやリ直すことが多々ある。	解析作業終了後、パラメータの詳細や解析手順、プロトコルを文書化している。	データ解析の際、決められた方法で解析のワークフローや判断のプロセスの詳細をいつも記録している。	自身が行った解析のワークフローや判断プロセスの詳細が他人によっても実行できることが確認できている。
データの共有と公開	研究成果を共有しているが、その根拠となるデータについては共有できていない。	リクエストがあった場合にのみ研究データの共有を行っている。	自身の研究成果や根拠となるデータは、第三者が利用できる形式で共有できる。	優れたデータマネジメントを実践し、自身のデータをいつでもだれとも効率的に共有できる。

引用：青木学聡、他、“研究データマネジメントルーブリックによる自己評価と共通認識の醸成” DEIM Forum 2019 C2-1、<https://db-event.jpn.org/deim2019/post/papers/267.pdf>

A.2 ANDS Checklist

Checklist for assessing IT Infrastructure capability for Data Management

About this Checklist

This document is part of the [ANDS Data Management Framework Guide](#).

The Guide is for research institutions which are intending to assess the capability maturity of their current infrastructure supporting the management of institutional research data assets. A Capability Maturity Model is used as the instrument to support institutions to investigate their research data capability maturity using:

- **4 elements of data management capability:** Policies and procedures; IT infrastructure; support services; managing metadata
- **assessed across 5 levels of maturity:** Initial; Development; Defined; Managed; Optimised

This document is a basic starting point for institutions who wish to assess their **IT infrastructure capability maturity** in terms of research data management.

	Suggested questions to assess capability maturity of IT infrastructure
Storage for data and metadata	• Do you have adequate storage for any given dataset? If so, is it permanent storage or temporary? • Are datasets backed up or archived in multiple locations? • If datasets and their metadata are separate, are they stored together or could they become separated? How closely are the data and metadata linked? • When new data storage technology becomes available, will you be able to migrate datasets to it? • Do datasets rely on compression or data transformation tools and standards for basic access? If so, will there be continued access to those tools and standards? • Do datasets rely on proprietary tools or formats for basic access? If so, will there be continued access to those tools or formats?
Identity management, authentication access &	• Can the IT infrastructure record and support the access restrictions needed for all datasets, e.g. passwords, AAF identity, ethics and access agreements and contracts? • Can the IT infrastructure support collaboration on any given dataset with researchers wherever they are located? • Can the IT infrastructure support the registrations needed to track and identify authenticated access? • Is it possible to change any access restrictions for a given dataset at some time in the future (e.g. after an embargo period has passed)?

Internal & external network connectivity	• If a dataset is too large or expensive to transfer over internal or external networks, is there a fall back transfer mechanism (e.g. hard drives sent via courier)? • What data transfer costs does your institution levy on internal data transfers? What about external transfers to/from other institutions? • If a given dataset attracts high demand, can the IT infrastructure cope with the access and copying requests expected?
Access to discipline specific tools to support analysis	• Does your institution make available a sufficient range of analytical tools to meet all disciplinary needs? Can all those who need them get access? • If a dataset requires access to a particular tool in order to analyse it, has that tool (e.g. statistical analysis package) been archived? Have any configuration or customization scripts developed for that tool been archived? • If a dataset requires access to a commercial data analysis tool, are there alternative open-source tools available for other researchers without access to the commercial tool? • If a dataset requires access to a specific instrument or piece of equipment and this in turn requires a specific configuration, is there provision for this to be specified in the dataset metadata?
Software development	• Does the institution have the capacity to develop or modify software as required? • Does the institution have the expertise to recommend alternate solutions to time-intensive custom-developed development or modification? • If a dataset is generated by custom-developed software, has that software and documentation to build it with the dataset been archived? • If a dataset is generated by commercial or off-the-shelf software, have any customization or configuration settings with the dataset been archived? Are there alternative open-source tools available for other researchers without access to commercial software? • If a dataset requires custom or off the shelf software for basic access (e.g. decompression) have any settings or customizations needed to access the data been specified with the metadata?
Visualisation	• Does the institution make available a sufficient range of visualisation tools, or access to them, to meet all requirements? • If a dataset requires supporting visualizations, are the visualisation assets and tools archived with the dataset? • Are there facilities to record any visualisation settings or configuration values with the dataset metadata? • If the supporting visualisations require access to high performance computing resources (HPC) to repeat or view, are there facilities to note the relevant requirements in the dataset metadata?
Collaborative environments	• Does the institution provide adequate collaborative environments, or access to them, to meet all needs? • If a dataset is stored within a collaborative environment or requires access to a collaborative environment in order to access or analyse it, are there facilities to store the environment setup and configuration in the dataset metadata? Can any customized software components needed to use the collaborative environment be archived with it? • If a dataset requires access to a commercial collaborative environment, are there alternative open-source tools available for other researchers without access to the commercial environment?

	Is it possible to move a dataset to a newer or different collaborative environment if required for future additions or expansions?
High performance computing (HPC)	- If a dataset requires access to HPC facilities in order to access or analyse it, is it possible to specify the facilities setup and configuration in the dataset metadata? - Is it possible to archive any customized software components needed to use the HPC facilities? - If a dataset requires access to a commercial HPC tool, are there alternative open-source tools available for other researchers without access to the commercial tool? - As desktop systems capacities grow, are there mechanisms to allow data to be shifted across?

B. TF メンバー個別報告書

B.1 名古屋大学 小川泰弘 個別報告書

B.1.1. 現状の研究データ管理に関連する取り組み・提供しているサービスの紹介

名古屋大学における研究データ管理の取り組みの経緯

そもそも研究データ管理は、各個人が研究を進める上で個別に実施してきたものである。近年のオープンサイエンス推進の高まりを受けて、名古屋大学でも教職員が関連会議に参加したり、GakuNin RDM のクローズドテストに参加したりしてきた。そして 2018 年度には、学内において情報基盤センターの教職員、図書館および URA の職員からなる RDM プロジェクトを開始した。

2019 年度にはアンケートを実施して、学内における研究データ管理の認知度や要望を調査した。こうした内容を大学 ICT 推進協議会 (AXIES) などを通じて他の大学に情報提供するとともに、AXIES の研究データマネジメント部会における「学術機関における研究データに関する提言」の策定にも加わった。また、学内での RDM 推進のため、各ステークホルダーを集めて研究データ基盤整備部会を発足させた。システム面では、引き続いて GakuNin RDM のオープン実証実験に参加し、RDM 用のストレージも導入した。また、学内利用者向けの手引きを作成・公開した。

2020 年度には、名古屋大学学術データポリシーを作成し、また、研究データ基盤整備部会の下に学術データ基盤整備 WG を発足させた。システム面では、GakuNin RDM で使用するストレージを学内クラウドとして利用していたストレージと統合し、利便性を向上させた。

2021 年度には、GakuNin RDM を利用するための学内の利用規程を確認の上、正式運用を開始した。名古屋大学学術データ基盤整備基本計画を策定し、2021 年度から 2023 年度の期間に実施する予定である。システム面においては、GakuNin RDM の利用において多要素認証を導入した。また、これまではオンプレミスのストレージを利用してきたが、多要素認証への対応や今後のメンテナンスも考えて、One Drive へ移行した。

B.1.2. 現状のシステム運用管理状況を踏まえての ANDS チェックリストの実施結果

現状の GakuNin RDM を対象とした ANDS チェックリストの結果は以下の通り

○Storage for data and metadata

・特定のデータセットに十分なストレージがありますか？もしそうなら、それは永続的なストレージですか、それとも一時的なものですか？

→ OneDrive の最大容量は一人あたり 5TB であり、一般的なデータセットに対しては充分だと思われるが、大規模なデータセットに対しては足りない。そうしたデータには学内の HPC に付属するストレージの利用を提案している。

OneDrive のストレージは一時的なものであるが、HPC に関しては一時的ものと永続的なストレージの両方を用意している。永続的なストレージはソニー社製のオプティカルディスクアーカイブ PetaSite 拡張型ライブラリーシステムを採用している。

・データセットは複数の場所でバックアップまたはアーカイブされていますか？

→データセットのバックアップは実施している。複数の世代はバックアップしているが一箇所である。

・データセットとそのメタデータが分離している場合、それらは一緒に保存されますか、それとも分離される可能性がありますか？データとメタデータはどの程度密接にリンクされていますか？

→GakuNin RDM の管理に必要なメタデータは別途保存される。現在、データセットは OneDrive、メタデータは NII のディスクに保存される。

GakuNin RDM 以外のメタデータは、今後 GakuNin RDM に導入予定の DPM 支援ツールや、JAIR0 での公開システ

ムに依存する。

- ・新しいデータストレージテクノロジーが利用可能になったときに、データセットをそれに移行できますか？

→GakuNin RDM に依存するが、現在までにGakuNin RDM で使用するストレージシステムを変更してきているため、移行は可能だと考えられる。

- ・データセットは、基本的なアクセスのために圧縮またはデータ変換ツールと標準に依存していますか？もしそうなら、それらのツールと標準へのアクセスは継続されますか？

→圧縮またはデータ変換ツールは利用していない。

- ・データセットは、基本的なアクセスのために独自のツールまたはフォーマットに依存していますか？もしそうなら、それらのツールやフォーマットへのアクセスは継続されますか？

→アクセスはGakuNin RDM に依存しているが、OneDrive を利用する場合、同じデータにはOneDrive 経由でアクセス可能である。

○Identity management, authentication & access

- ・IT インフラストラクチャは、すべてのデータセットに必要なアクセス制限を記録してサポートできますか。パスワード、AAF ID、倫理、アクセス契約および契約？

→認証は学認を利用する。名古屋大学側のシステムの変更により、2022 年9 月から多要素認証を導入した。

- ・IT インフラストラクチャは、研究者がどこにいても、特定のデータセットでのコラボレーションをサポートできますか？

→GakuNin RDM ではデータを直接編集できる訳ではないので、コラボレーションという点では、Google Drive などより劣る。今後の発展に期待する。

- ・IT インフラストラクチャは、認証されたアクセスを追跡および識別するために必要な登録をサポートできますか？

→GakuNin RDM を管理するNII である程度可能だと思われるが、各大学でどこまで可能かは不明である。

- ・将来のある時点で（たとえば、禁輸期間が経過した後）、特定のデータセットのアクセス制限を変更することは可能ですか？

→各ユーザがある程度可能であると考えられるが、どこまで細かな指定ができるかは不明。また、GakuNin RDM の管理者である大学が個別のデータのアクセス制限を変更することはできないと思われる。

○Internal & external network connectivity

- ・データセットが大きすぎたり、コストがかかりすぎて内部ネットワークや外部ネットワークを介して転送できない場合、フォールバック転送メカニズム（宅配便で送信されるハードドライブなど）はありますか？

→GakuNin RDM では想定していない。HPC などではコールドデータに光ディスクを採用しているので、それを各ユーザが自分で配送することは可能だが、大学のサービスとしては考慮していない。

- ・あなたの機関は内部データ転送にどのようなデータ転送コストを課していますか？他の機関との間の外部転送はどうですか？

→OneDrive の利用は無料。GakuNin RDM は将来的に課金されるものと思われるが、各ユーザからどのように徴収するかは未定。

- ・特定のデータセットが高い需要を引き付ける場合、IT インフラストラクチャは予想されるアクセスおよびコピー要求に対処できますか？

→これはGakuNin RDM ではなく、公開基盤であるJAIRO Cloud で検討する課題である。

○Access to discipline specific tools to support analysis

- ・あなたの機関は、すべての専門分野のニーズを満たすのに十分な範囲の分析ツールを利用できるようにしていますか？それらを必要とするすべての人がアクセスできますか？

- ・データセットを分析するために特定のツールへのアクセスが必要な場合、そのツール（統計分析パッケージなど）はアーカイブされていますか？そのツール用に開発された構成またはカスタマイズスクリプトはアーカイブされていますか？

- ・データセットが商用データ分析ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？

- ・データセットが特定の機器または機器へのアクセスを必要とし、これが特定の構成を必要とする場合、これ

をデータセットメタデータで指定するための規定はありますか？

→GakuNin RDM 上で、BinderHub との連携アドオンが開発されており、これによりある程度の分析が可能である。それ以外の対応はGakuNin RDM の今後の開発次第である。

○Software development

・機関には、必要に応じてソフトウェアを開発または変更する能力がありますか？

・機関には、時間のかかるカスタム開発の開発または変更に対する代替ソリューションを推奨する専門知識がありますか？

・データセットがカスタム開発されたソフトウェアによって生成された場合、データセットを使用してデータセットを構築するためのそのソフトウェアとドキュメントはアーカイブされていますか？

・データセットが市販のソフトウェアまたは既製のソフトウェアによって生成された場合、データセットのカスタマイズまたは構成設定はアーカイブされていますか？商用ソフトウェアにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？

・データセットが基本的なアクセス（解凍など）のためにカスタムまたは既製のソフトウェアを必要とする場合、メタデータで指定されたデータにアクセスするために必要な設定またはカスタマイズがありますか？

→基本的にソフトウェアの開発は研究者個人が対応しており、機関として対応はしていない。

○Visualisation

・機関は、すべての要件を満たすために、十分な範囲の視覚化ツールまたはそれらへのアクセスを利用できるようにしていますか？

・データセットでサポートする視覚化が必要な場合、視覚化アセットとツールはデータセットとともにアーカイブされていますか？

・データセットメタデータを使用して視覚化設定または構成値を記録する機能はありますか？

・サポートする視覚化で繰り返しまたは表示するためにハイパフォーマンスコンピューティングリソース（HPC）にアクセスする必要がある場合、データセットメタデータに関連する要件を記録する機能はありますか？

→基本的には、Access to discipline specific tools to support analysis の項目で記述したBinderHub との連携アドオン次第である。

○Collaborative environments

・機関は、すべてのニーズを満たすために、適切なコラボレーション環境またはそれらへのアクセスを提供していますか？

→GakuNin RDM の機能により実現しているが、現状GakuNin RDM に参加していない機関に在籍するユーザは利用できない。GakuNin RDM では個人単位での利用が想定されていないため、その点の改善が望まれる。

・データセットがコラボレーション環境内に保存されている場合、またはデータセットにアクセスまたは分析するためにコラボレーション環境へのアクセスが必要な場合、データセットメタデータに環境のセットアップと構成を保存する機能はありますか？コラボレーション環境を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることはできますか？

→BinderHub との連携アドオン次第。

・データセットが商用のコラボレーション環境へのアクセスを必要とする場合、商用環境にアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？

→BinderHub との連携アドオン次第。

・将来の追加または拡張に必要な場合、データセットを新しいまたは別のコラボレーション環境に移動することは可能ですか？

→BinderHub との連携アドオン次第。

○High performance computing (HPC)

・データセットにアクセスまたは分析するために HPC 施設へのアクセスが必要な場合、データセットメタデータで施設のセットアップと構成を指定することは可能ですか？

・HPC 機能を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることは

可能ですか？

- ・データセットが商用 HPC ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？

- ・デスクトップシステムの容量が増えるにつれて、データを移動できるメカニズムはありますか？

→現状、名古屋大学の HPC（不老）と GakuNin RDM は連携していない。将来的に BinderHub との連携で利用可能になるかもしれない。

GakuNin RDM の容量はユーザが管理するため、容量に応じてデータを移動できるメカニズムはない。

B.1.3. システム運用管理者視点から今後求められる研究データの長期運用管理の在り方についての考えや意見

ユーザ視点から見た場合、研究データ管理において煩わしいのは、管理するためにデータをアップロードすることである。その意味では、管理下にあるデータを普段使いすることが望ましい。現状の GakuNin RDM では、その点では単なるデータ保管庫であり、研究データをいちいちアップロードする点で不便である。

複数のユーザがデータを共有するという目的の場合は、普段の研究データがそのまま管理対象になりうるが、現状の GakuNin RDM では、アップロードされたデータを直接変更することができない。この点で、Google Drive などのデータ共有システムに劣っている。BinderHub との連携や、現在開発中のオフィスソフトとの連携機能により、こうした点が改良されることを望む。ユーザにとって使いにくいシステムの利用を強制するのは望ましくない。ユーザが使いたくなるシステム、もしくは普段から使っているシステムで、そのままデータ管理ができるのが望ましい。

システム運用管理者からみた場合、ユーザの多様な要求にどのように応えるかが問題となる。ANDS チェックリストはその観点に立った項目が多い。しかし、ユーザが多様であるがゆえ、すべての要求に応えるのは不可能であるし、またそれぞれの機能を開発するコストも高い。その観点からすれば、システム運用管理者側は、どのサービスを実現すれば良いかの見極めが重要になる。ANDS チェックリストは商用システムをオープンソースツールで代替できるか重視していたが、そこは各ユーザに任せる問題であり、システム運用管理者が対応する問題ではないと考える。現状では、GakuNin RDM の機能拡張に期待することになるが、その際に、どの機能を優先的に要求するかが重要である。その点では、管理下のファイルを直接編集可能になるという点は、ユーザの利便性と GakuNin RDM の普及のために重視すべき点と考える。

B.2 広島大学 渡邊英伸 個別報告書

B.2.1. 現状の研究データ管理に関連する取り組み・提供しているサービスの紹介

広島大学における、研究データ管理に関する取り組みの概要を示す。

2019 年 11 月 8 日～26 日にかけて、大学 ICT 推進協議会 (AXIES) 研究データマネジメント (RDM) 部会からの依頼に基づき、「広島大学における研究データ管理」に関するアンケートを実施した。調査対象は教育研究情報収集システムに登録されている 1,982 名の研究者であり、全国共通の設問で得られた結果をもとに、本学で組織的に研究データを保管・共有・公開する環境を整備するための検討を開始した。

広島大学は、AXIES の RDM 部会との関わりが深く、2017 年の部会の設置、2019 年の「学術機関における研究データ管理に関する提言」刊行や「大学での研究データ管理に関するアンケート」(雛形)作成などに本学も携わっている。なお、現在も本学の教員が RDM 部会の担当理事でもあるため、RDM 部会と継続して情報共有を行っている。

2021 年 1 月に、「広島大学 DX 推進基本計画 令和 2～4 年度版」を策定した。本基本計画は 10 年後の情報環境及びデジタル技術を活用した教育・研究・支援業務等のあり方を見据えて策定したものであり、10 年後を見据えた基本方針に加え、今すぐ(令和 2～4 年度)に優先して実施する全学的事項を示している。優先して実施する事項は、教育・研究・支援業務等、大学の活動全般に渡っており、いずれも全学的重要事項となっている。研究データ管理については、優先して実施する全学的事項の「教育・学習データの活用と教育コンテンツのデジタル化」、「構成員の健康管理のデジタル化」、「大学 IR データ分析」とともに、以下の内容が明記されている。

- 4.2 研究データ管理 デジタル化の進展により、研究の各段階において研究データを的確に管理する必要性が増している。分野を問わず、研究の確実な積み上げと再現性のために研究データを研究過程の段階ごとに適切に保存・管理し、必要に応じて提示することも求められている。また、オープンサイエンス推進のため、他者が生成した研究データを利用することや、研究者自身が生成した研究データを共有・公開する機会も増えている。そこで、本学の全研究者が利用する研究データ管理のためのデジタルプラットフォームを構築し、活用を推進する。なお、本事項の実施にあたっては「研究データ管理ポリシー(仮称)」等の策定を行う。

「広島大学 研究データ管理・公開・利活用ポリシー」は、2022 年 6 月に策定された。なお、「広島大学教育・学習データ利活用ポリシー」は 2021 年 11 月に策定し、2022 年 4 月 1 日から情報システムにおける個人情報を含む教育・学習データの利活用に関する同意書の取得を開始している。「健康データ利活用ポリシー」は、2022 年 3 月に策定済みの状況である。

また、2021 年 1 月 15 日～2021 年 2 月 1 日の期間に公募された「デジタルを活用した大学・高専教育高度化プラン」に応募し、252 件の申請から広島大学は以下 2 件が採択された。

- 取組①「学修者本位の教育の実現」に「次世代オンライン教育を実現する『バーチャルクラスルームデジタルラーニング(VCDL)』環境の構築」
- 取組②「学びの質の向上」に、「DX で拓く学びのパラダイムシフト～ニューノーマルの新たな教育実現に向けた『バーチャルクラスルームデジタルラーニング(VCDL)』環境の構築～」

参考：https://www.mext.go.jp/a_menu/koutou/sankangaku/1413155_00010.htm

そのため、「教育・学習データの活用と教育コンテンツのデジタル化」に関する情報基盤整備を優先したこともあり、「研究データ管理」に関する情報基盤整備はこれからという状況である。なお、「教育・学習データの活用と教育コンテンツのデジタル化」に関する情報基盤整備は 2022 年 3 月に終えている。

B.2.2. 現状のシステム運用管理状況を踏まえての ANDS チェックリストの実施結果

広島大学のシステム運用状況を踏まえて、ANDS チェックリストの結果を以下に示す。

1. Storage for data and metadata

- 特定のデータセットに十分なストレージがありますか？もしそうなら、それは永続的なストレージですか、それとも一時的なものですか？
 - パブリッククラウドストレージとして、Microsoft 365 が提供する OneDrive と Google Workspace for Education Fundamentals の Google Drive が利用可能である。OneDrive の一人当たりの最大容量は 3TB、Google Drive の一人当たりの最大容量は 50GB である。いずれも本学構成員とのみ共

有が可能である。加えて、プライベートクラウドストレージとしてNextcloudも提供しており、学外者とのファイル共有として利用可能である。一人当たり最大容量は30GBである。なお、30日間経過したファイルは自動で削除される機能を有している。

- 小規模な研究データセットにおいては、十分なストレージを提供していると考えるが、大規模な研究データセットに対しては十分とは言えない。また、本学の構成員でなくなればストレージサービスは利用できないため、永続的なストレージとして利用はできない。
- GakuNin RDM用の永続的なストレージとして、どこまでパブリッククラウドストレージ（Dropboxなど）が利用可能かを検証している。
- データセットは複数の場所でバックアップまたはアーカイブされていますか？
 - パブリッククラウドストレージは、各サービスが複数のデータセンターにデータセットを二重化している。また利用者各自がバックアップまたはアーカイブを実施することも可能である。
 - プライベートクラウドおよびオンプレのストレージは、データセンター内でサーバの冗長化構成を組んでいるが、データセットの冗長化はしていない。データセットの冗長化を含めたデータ管理は利用者の責任で実施する方針となっている。
- データセットとそのメタデータが分離している場合、それらは一緒に保存されますか、それとも分離される可能性がありますか？データとメタデータはどの程度密接にリンクされていますか？
 - 現状、データセットおよびメタデータは、研究者自身が保存管理している。
 - 今後GakuNin RDMを利用した場合は、データセットは本学が用意するストレージに保存され、メタデータはGakuNin RDMに保存されることになる予定。
- 新しいデータストレージテクノロジーが利用可能になったときに、データセットをそれに移行できますか？
 - GakuNin RDMでは、複数のパブリッククラウドストレージを接続できない現状があると聞いているが、GakuNin RDMやストレージサービスに依存することであるため、現時点で移行可否は不明。
- データセットは、基本的なアクセスのために圧縮またはデータ変換ツールと標準に依存していますか？もしそうなら、それらのツールと標準へのアクセスは継続されますか？
 - データ管理は利用者の責任で実施する方針となっているため、組織としてデータセットに対する圧縮機能またはデータ変換ツールを提供していない。
- データセットは、基本的なアクセスのために独自のツールまたはフォーマットに依存していますか？もしそうなら、それらのツールやフォーマットへのアクセスは継続されますか？
 - データ管理は利用者の責任で実施する方針となっているため、組織としてデータセットに対する特別なツールを介するアクセスまたはフォーマット機能を提供していない。

2. Identity management, authentication & access

- ITインフラストラクチャは、すべてのデータセットに必要なアクセス制限を記録してサポートできますか。パスワード、AAF ID、倫理、アクセス契約および契約？
 - 本学の情報サービスはSSOに対応し、学外からのアクセスの場合はMFAが必須となっている。なお、SSOやMFA非対応の情報サービスは学内限定にしている。加えて、学認と認証連携をしている。
 - GakuNin RDM利用時の認証は学術認証フェデレーション（学認）の利用を予定している。
- ITインフラストラクチャは、研究者がどこにいても、特定のデータセットでのコラボレーションをサポートできますか？
 - 本学の情報サービスは利用者がどこにいても利用可能である。
 - GakuNin RDMおよび連携するストレージサービスにおいても同様と考える。
- ITインフラストラクチャは、認証されたアクセスを追跡および識別するために必要な登録をサポートできますか？
 - 本学の情報サービスでは、いつ、だれが、どの端末で、どのサービスにアクセスしたかの追跡および識別は可能である。
 - データに対する追跡および識別は、パブリッククラウドストレージならびにプライベートクラウドストレージのアクティビティ機能で利用者自身が確認することも可能である。
 - GakuNin RDMの管理機能は不明のため調査が必要である。
- 将来のある時点で（たとえば、禁輸期間が経過した後）、特定のデータセットのアクセス制限を変更すること

は可能ですか？

- 本学の情報サービスでは、通信制限やアカウントロックによるアクセス制限は可能である。パブリッククラウドストレージ、プライベートクラウドストレージ、オンプレのストレージは、利用者自身でアクセス制限を行うことも可能である。
- GakuNin RDM と連携して、特定のデータセットまでを対象としたアクセス制御が可能かは不明のため調査が必要である。

3. Internal & external network connectivity

- データセットが大きすぎたり、コストがかかりすぎて内部ネットワークや外部ネットワークを介して転送できない場合、フォールバック転送メカニズム（宅配便で送信されるハードドライブなど）はありますか？
 - 現状、本学の情報サービスにおいてフォールバック転送メカニズムは無い。
 - データ管理は利用者の責任で実施する方針となっているため、基本的に利用者自身で対応することになるが、個別相談で対応するかは検討が必要である。
- あなたの機関は内部データ転送にどのようなデータ転送コストを課していますか？他の機関との間の外部転送はどうですか？
 - 本学の情報サービスに関するデータ転送コストを利用者が負担することは無い。
 - GakuNin RDM の費用負担については確認が必要である。
- 特定のデータセットが高い需要を引き付ける場合、IT インフラストラクチャは予想されるアクセスおよびコピー要求に対処できますか？
 - 本学のクラウド上で稼働している情報サービスにおいては、対処可能と考える。
 - GakuNin RDM においては、検証が必要である。

4. Access to discipline specific tools to support analysis

- あなたの機関は、すべての専門分野のニーズを満たすのに十分な範囲の分析ツールを利用できるようにしていますか？それらを必要とするすべての人がアクセスできますか？
- データセットを分析するために特定のツールへのアクセスが必要な場合、そのツール（統計分析パッケージなど）はアーカイブされていますか？そのツール用に開発された構成またはカスタマイズスクリプトはアーカイブされていますか？
- データセットが商用データ分析ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
- データセットが特定の機器または機器へのアクセスを必要とし、これが特定の構成を必要とする場合、これをデータセットメタデータで指定するための規定はありますか？
 - 本学ではすべての専門分野のニーズを満たすのに十分な範囲の分析ツールを提供していない。
 - 利用者自身が用意したツールについては、大学側で禁止としていないため、利用者の責任のもとツールを利用することは可能である。
 - GakuNin RDM の機能の一つとして分析ツール、あるいは分析環境が提供される場合は、本学でも利用可能になると考える。

5. Software development

- 機関には、必要に応じてソフトウェアを開発または変更する能力がありますか？
- 機関には、時間のかかるカスタム開発の開発または変更に対する代替ソリューションを推奨する専門知識がありますか？
- データセットがカスタム開発されたソフトウェアによって生成された場合、データセットを使用してデータセットを構築するためのそのソフトウェアとドキュメントはアーカイブされていますか？
- データセットが市販のソフトウェアまたは既製のソフトウェアによって生成された場合、データセットのカスタマイズまたは構成設定はアーカイブされていますか？商用ソフトウェアにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
- データセットが基本的なアクセス（解凍など）のためにカスタムまたは既製のソフトウェアを必要とする場合、メタデータで指定されたデータにアクセスするために必要な設定またはカスタマイズがありますか？
 - ソフトウェア開発は、利用者の責任のもと実施する方針となっているため、組織としてソフトウェア開

発に関するサービス提供や支援を行っていない。

- GakuNin RDM の機能の一つとしてソフトウェア開発支援ツール、あるいは開発環境が提供される場合は、本学でも利用可能になると考える。

6. Visualisation

- 機関は、すべての要件を満たすために、十分な範囲の視覚化ツールまたはそれらへのアクセスを利用できるようにしていますか。
- データセットでサポートする視覚化が必要な場合、視覚化アセットとツールはデータセットとともにアーカイブされていますか？
- データセットメタデータを使用して視覚化設定または構成値を記録する機能はありますか？
- サポートする視覚化で繰り返しまたは表示するためにハイパフォーマンスコンピューティングリソース (HPC) にアクセスする必要がある場合、データセットメタデータに関連する要件を記録する機能はありますか？
 - 視覚化に関しても同様に、利用者の責任のもと実施する方針となっているため、組織として視覚化に関するサービス提供や支援を行っていない。
 - GakuNin RDM の機能の一つとして視覚化支援ツール、あるいは処理環境が提供される場合は、本学でも利用可能になると考える。

7. Collaborative environments

- 機関は、すべてのニーズを満たすために、適切なコラボレーション環境またはそれらへのアクセスを提供していますか。
 - NextCloud や Microsoft 365, Google Workspace など学外者とのコラボレーション環境は提供している。
 - ゲスト用のアカウントを発行し、本学の情報サービスを利用することも可能となっている。
 - GakuNin RDM においては、学認に参加機関であれば認証連携により本学のコラボレーション環境を利用可能になると考えられるが、学認に参加していない機関の場合の対応は検討が必要である。
- データセットがコラボレーション環境内に保存されている場合、またはデータセットにアクセスまたは分析するためにコラボレーション環境へのアクセスが必要な場合、データセットメタデータに環境のセットアップと構成を保存する機能はありますか？コラボレーション環境を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることはできますか？
 - メタデータの管理は利用者の責任のもと実施する方針となっているため、組織としてメタデータ管理運用に関するサービス提供や支援を行っていない。
 - GakuNin RDM を利用した場合は、GakuNin RDM やストレージサービスに依存することであるため、現時点では不明。
- データセットが商用のコラボレーション環境へのアクセスを必要とする場合、商用環境にアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
 - 商用環境との連携は利用者の責任のもと実施する方針となっているため、組織として商用環境との連携に関するサービス提供や支援を行っていない。
 - GakuNin RDM を利用した場合は、GakuNin RDM やストレージサービスに依存することであるため、現時点では不明。
- 将来の追加または拡張に必要な場合、データセットを新しいまたは別のコラボレーション環境に移動することは可能ですか。
 - コラボレーション環境の移行は利用者の責任のもと実施する方針となっているため、組織として商用環境との連携に関するサービス提供や支援を行っていないが、利用者が移行作業を行えば可能である。
 - GakuNin RDM を利用する場合は、GakuNin RDM やストレージサービスに依存することであるため、現時点では不明。

8. High performance computing (HPC)

- データセットにアクセスまたは分析するために HPC 施設へのアクセスが必要な場合、データセットメタデータで施設のセットアップと構成を指定することは可能ですか。
- HPC 機能を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることは可能ですか？

- データセットが商用 HPC ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
- デスクトップシステムの容量が増えるにつれて、データを移動できるメカニズムはありますか？
 - 本学は九州大学情報基盤研究開発センター「スーパーコンピュータシステム ITO」と包括契約を結び、広島大学構成員（教職員・学生）に対して利用負担金無しで研究教育用計算機システムを利用できる環境を提供している。
 - GakuNin RDM は「スーパーコンピュータシステム ITO」と連携していない。
 - GakuNin RDM と「スーパーコンピュータシステム ITO」が連携した場合は、GakuNin RDM や ITO に依存することであるため、現時点では不明

B.2.3. システム運用管理者視点から今後求められる研究データの長期運用管理の在り方について

での考えや意見

学術機関の研究データ管理は、本学と同様に GakuNinRDM の活用を想定している学術機関は多いと考えられる。利用者（研究者）の観点で見た場合、Web ブラウザ以外のデータ転送に対応できるかは重要な点と考える。研究環境では、研究者の PC 以外にも、研究機器から直接データを保存する要望や独自に管理するサーバやストレージとの連携も想定されたため、Web ブラウザを介さず GakuNinRDM 間とのデータ保存が可能であることが望ましいと考える。

また、業務の自動化に関する機能を柔軟に活用できる環境であることが望まれると考える。研究データのライフサイクルにおいて、手動で対応したい場合や自動化で対応したい場合など研究者毎に様々な要望が存在するため、より多くの要望に応えられる柔軟な自動化機能を簡易に扱えることができるかが重要と考える。

その他に、データ移行においても Web ブラウザを介さない方法も重要と考える。ブラウザの仕様やストレージの仕様などによってダウンロードやアップロードの限界があり、大量のデータやサイズが大きいデータを扱う RDM では、例えばマルチクラウドストレージ間で同期が可能など、移行も容易な仕組みが組み入れられているべきと考える。

システム管理者の観点では、管理者向けのログがどこまで扱えるかが重要と考える。大学では、認証ログ、ネットワークログ等統合的に管理し、追跡性を確保している場合があるが、RDM の管理者向けのログも同様に扱えることが望ましいと考える。

その他にセキュリティの観点で言えば、DLP などの情報漏洩の対策機能が望まれると考える。大学で定めている研究データポリシーに準拠した管理がされるべきであり、それに違反するデータの扱いを管理者側で制御できることが重要と考える。

ANDS のチェックリストでは、運用管理体制や責任分解点を明確にすることについては触れられていなかった。また、費用や人材についてもチェック項目には無かった、システム運用管理の観点では、必ず考慮すべき点であると考えられる。

B.3 沖縄科学技術大学院大学 Eddy Taillefer 個別報告書

B.3.1 現状の研究データ管理サービスに関連する課題

OIST では、研究の主要なデータを共有領域ストレージに保管するポリシーとしている。認証環境を実装して研究データのストレージにアクセスしやすくしており、また、研究データの保管サービスも提供している。OIST では、課金を行わずに共有領域ストレージを高性能計算機と同じく共通資源として提供し管理しているので、論文に関係するデータの 10 年間保管（アーカイブ）のポリシーは実現できる。しかし、アーカイブされた研究データのメタデータ作成は現在研究員に任されている。そしてアーカイブの研究データが利用しているストレージ容量は増大し続けている。ストレージ媒体を拡張する事と、長い保存期間を維持することが、（管理面、そして経済面での）課題となっている。

○研究データ管理のポリシーについて

OIST が研究データ用のシステムとインフラを提供、管理、運用するという観点でポリシーを作成しても、研究者にとって研究活動と関係のない余計な仕事に労力を割くことになる。そうすると、ポリシーと研究活動の仕組みを合わせる工夫が必要となる。あるいは、研究データ管理のポリシーに生じる必要な作業は研究成果（成績、論文等）のメリットになるように実施する必要がある。

○研究員に任せる研究データ管理の部分について

研究分野ごとに研究データ管理の意識と実現方法が大きく異なるので、原則的に、研究員は事前に教育を受ける必要がある。ストレージの適切な使い方（無駄を避ける等）、データのアーカイブ（保管）とバックアップ（災害復旧）の違い、そしてそれぞれのストレージの種類によって対象になるデータが異なるということ等を含む教育内容になる。また、研究過程におけるデータ管理について、第 1 のステップとして研究計画を作成する時に研究データ計画（DMP）を搭載する習慣を教える。そして、データを生成する時には必要のないデータを削除し、残ったデータは整理して管理を引き継ぐ人の為にメタデータとドキュメンテーションを作成して残すことも習慣化させる。

○研究員による研究データ管理の維持可能性

データの体系的整理は研究室が研究の開始段階で決めても、研究中に守られなくなる傾向にある。大学でそれを把握し定量的に評価することは難しい課題である。研究成果を得る為にはある程度の自由度と柔軟性を認めなくてはならないので、どこまで厳格なデータ管理を行うかは課題である。各項目に対する意識をアンケートをとって理解したり、データ管理を行っている研究と行っていない研究を分類して把握する等の方法が課題として提案できる。

○研究データ共有と公開について

執行する研究データ管理の有効性に直接に影響するので、大学内あるいはラボ間のデータ共有は中央管理型の共有サービスを用いて提供する必要がある。また、外部向けの研究データ公開に関して大学や国レベルでの仕組みを考える必要がある。

○研究データアーカイブと文献のリポジトリについて

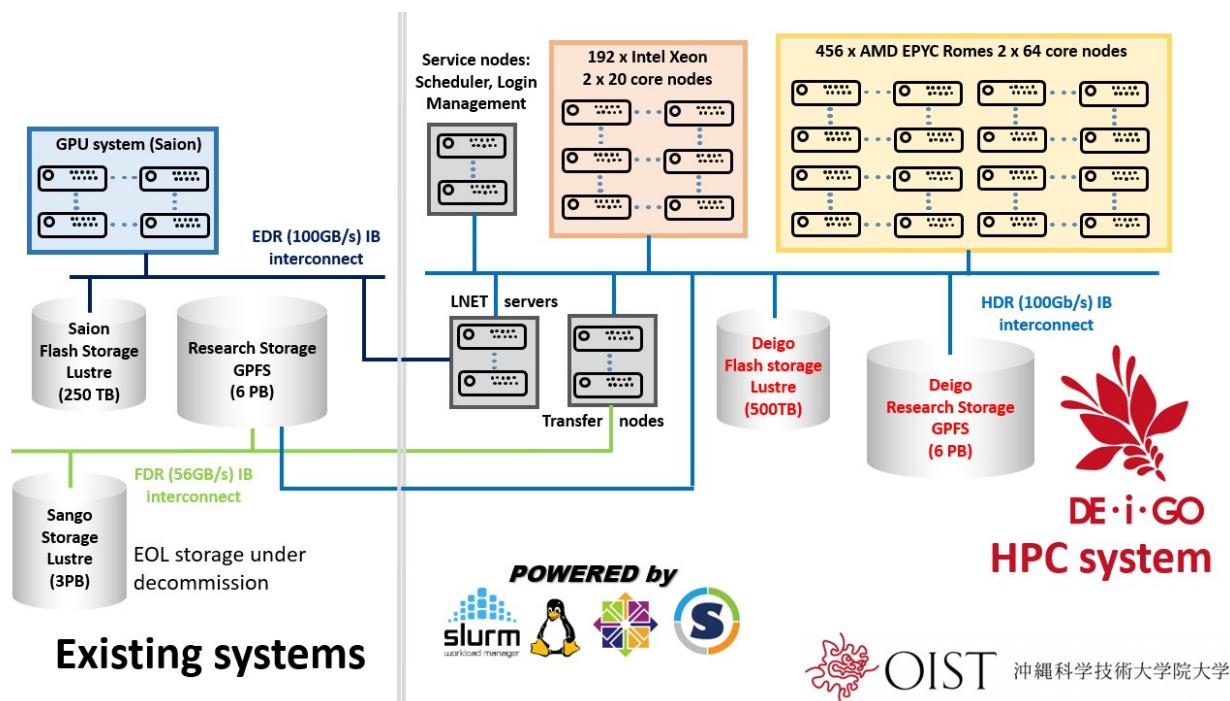
文献のリポジトリとして論文の論拠となるデータのリポジトリが要求されている。研究データの中でどれを研究の論拠とするかは研究機関が決めるものとなっている。オープンデータとして公開する場合、リポジトリとしてのデータストレージはかなりの容量が必要となり、専用の予算を付ける必要がある。また、研究データの有用性が管理サイドで判断はできないため、研究者自身の教育から始めなくてはならない部分が多く見られる。研究者の教育・意識改革が重要である。

データ形式に関しては、圧縮と編集がしやすいテキスト形式データが適切である。その一つの理由には、データ作成や編集の際に用いられたソフトウェア（商用でもコミュニティソフトでも）は提供が終了する可能性があるということが挙げられる。

そして、リポジトリに保管するデータの切り分けは管理者サイドでは困難なため、研究員や大学機関に担わせる。例えば、衛生による望遠鏡の画像データなどはなかなか捨てることが困難である。

B.3.2 現状のシステム運用管理状況

(OIST のシステム関連図を共有) OIST では共有資源として研究用のストレージと計算システム(HPC)を管理している。計算システムとストレージは高速回線で繋がれている。データバックアップとアーカイブのサービスも提供しており、ユーザが間違えて消してしまったデータのリストア、終了した研究データのアーカイブも行っている。データ共有(外部との共同研究)のためクラウドに Dropbox や github 等も使っている。よりよい運用のため、これらはできる限り研究データ管理教育とセットで提供している。



○ストレージの利用

OIST では課金を行わずに各ユニット(研究室の単位)に容量を割り当てている。8年前に1PB(1000TB)を越えていた容量の使用率は現在10PBを越えた。ストレージシステムをあるタイミングで交換や拡張を行って容量増加の要求に応じている。共有ストレージはほとんど100%の研究ユニットが日常的に使用し、SMBやNFSのアクセスを可能とすることで所内のどこからでもアクセスできるようにしている。

○データとメタデータ管理

OIST ではデータは機関の持ち物であるが、分野ごとでしかデータの重要度は判断できないのでデータの管理はユーザにまかせている。データ管理にはメタデータの利用を推奨している。メタデータの管理方法は分野により異なる。システム管理側でメタデータを管理するより、ユーザ自身に管理してもらう運用にする際は、そのためのツールを提供する方針にしようとしている。ストレージをスキャンしてファイルシステムによりメタデータを作成するツールがあり、それを使えばある程度ユーザでメタデータの管理ができる。ツールを使ったとしてもデータ管理においてユーザの労力は必要になる。例えば、別の研究員がデータを理解し再使用するためには、メタデータに分野による追加の情報を付け加える必要がある。

○ソフトウェア開発のサービス

計算システムの上にソフトウェアをインストール、またはコンパイル、そして実行するユーザーサポートを提供している。また、研究論文に関係するソフトウェア設計や開発の事例も数例あるが、それはサポートサービスではなく共同研究の形で行っている。

○可視化のサービス

主にリモートデスクとして可視化を提供している。ここでは、科学研究計算による可視化(シミュレーション中に中間データの観測等)と科学データ解析による可視化(電子顕微鏡で再構成したタンパク質の3次元構造の観測と解釈等)の2種に種別している。そして、研究データ管理の観点で考えると、可視化サービスそのものを中央管理された共有サービスとして提供することは、データ管理が含まれたソフトウェア基盤と円滑なデータ操作ができる統合システムをユーザーに使用させられるメリットがある。

○ハイパフォーマンスコンピューティング(HPC)

OIST は、中間データとして高性能計算機クラスターのストレージ上に生成したデータを共通ストレージに移

動して保存管理できる。これらの計算機やストレージは OIST の構成員であれば誰でも使える。ストレージは拡張可能でテープのアーカイブサポートもある。データ移動、コピーの頻度を少なくさせること、すべての処理が確実に完了すること、がスパコンならではのデータ管理のポイントである。公平な研究計算活動を保護する為にクラスタのスケジューラーを使用し、ユーザーにはベスト・プラクティスのルールに従ってもらっている。

B.4 北見工業大学 馬場智之 個別報告書

B.4.1. 北見工業大学の研究・教育データ保存の体制

北見工業大学ではkitCloud と呼ばれるクラウドサービスを学内仮想基盤システム上に構築し、教職員に 50GB、学生に 2GB のストレージ容量がそれぞれ付与されている。また、学内・学外を問わず、パスワード保護付きでファイル共有も可能である。しかし、あくまで情報処理センターが提供するサービスの一環であり、大学側の方針として研究・教育データの保存を目的として運用されている訳ではない。実際、筆者もデータ保存を目的としては利用しておらず、教員ごとの利用率にもばらつきがあると考えられる。

本学はNII の GakuNin RDM にも登録しており、教職員および研究室に所属する学生は利用可能である。しかし、こちらから大学側から利用に関する公式の告知はなく、人づてにごく一部に広まったのみで、利用率はおよそ 3 割程度である。加えて、この数字は一度でも登録・利用した人数の概算であり、実際に継続利用している数字はさらに少ないと考えられる。一部利用者からは、「使いづらい」「アップロード・ダウンロードに時間がかかる」といった声も挙がっている。

一方、本学の教育データ保存は、富士通の提供する学習管理システム(LMS)、CoursePower が大きな役割を担っている。2019 年以前は、上記のクラウドサービス同様、LMS 利用は各教員の裁量に委ねられており、利用率は高くなかった。しかし、2020 年以降、新型コロナウイルス蔓延の影響でオンライン授業が主流となったのちは、ほぼ全ての教員に授業設計として CoursePower が利用され、利用率はほぼ 100%まで増加した。加えて、コロナ禍以前は、Word や PowerPoint、PDF ファイル等の比較的小容量データの保存が主であったため、コロナ禍初期の頃の 2020 年 5 月時点における CoursePower のストレージ容量使用率は 10%程度であったのに対し、コロナ禍以降は主として授業の様子を撮影した動画ファイルや音声解説付きスライド等、大容量データがアップロードされたため、2021 年 2 月時点で 80%程度(およそ 2TB 相当)と 1 年足らずでストレージ使用率は爆発的に増加し、ストレージ不足の問題が浮き彫りとなった。2022 年 7 月現在においてもこの問題は根本的な解決には至っておらず、昨年以前の全授業データを消去する等の対処療法でしのいでおり、LMS に頼った教育データの”長期”保存は実現できていない。

本学は 2022 年から、小樽商科大学、帯広畜産大学とともに三大学法人統合をした。これに伴い、三大学の有する研究成果を地域社会に還元することを目的に『オープンイノベーションセンター』が発足した。同センターには大学の垣根を超えて研究プロジェクトを促進させる目的があり、当然、共同研究データの保存や効率的に研究プロジェクトを推進させるシステムの構築が課題となる。発足して間もないため活動実績は少ないが、現在、複数人で行う研究プロジェクトを対象に、GakuNin RDM を含むいくつかのサービスを同一画面で表示できるポータルサイトを構築中である。

B.4.2. 京都大学ループリック評価

『京都大学研究データマネジメントループリック』を n.2. 表 1 に示し、筆者の研究データ保存体制を紹介する。

筆者の専門は原子核物理であり、大規模数値計算を用いて、原子核の構造を調べる理論研究を行っている。大規模数値計算は他大学の共同利用計算機を用いて行い、得られた計算結果の解析はローカルコンピュータで行う。分子動力学と呼ばれる、シミュレーション計算の一種ではあるが、画像・動画ファイルの生成はほとんどなく、大容量データ保存は必要としない。ただし、数値データやバイナリデータは大量に生成されることが多い。共同研究も行うが共同研究者とは議論が主であり、研究データを共有し合うケースは少なく、研究データの作成や解析は筆頭著者のみで進めるケースがほとんどである。以上をふまえて、『京都大学研究データマネジメントループリック』の項目を順に検討する。

【データマネジメント計画】においては、『その都度』の「文書化はしていない」が該当する。突発的なアイデアや共同研究者からの提言などによって、新たに研究データを追加することは頻繁にあり、当初の研究計画通りに進むことはほとんどない。したがって、データマネジメント計画についても厳密に定めた経験はない。

【データの体系的整理】においては、どれにも該当しない、あるいはあらゆる項目を横断して該当すると解釈できる。自分なりに統一した体系的整理の手法があり、標準化とまでは言えないが同じ研究手法を採用していれば、他人が自分の研究データを検索し内容を理解することは可能である。より広い定義の一般的な他人が利用で

きるようにはなっていないが、そもそもほとんどの研究内容には独自の専門性があるため、研究データの整理以前に他人が理解・活用できるかといった問題もあり、他人の定義をどこまで含めるかによって『再利用のために最適化』の項目は変わりうると言える。また、上述の通り突発的にデータを追加することも多く、そのような場合には一貫した整理手法を設定せず、あとになって探し出すのに時間を要する場合もある。そのような場合でも、研究終了時(論文執筆完了時)には体系的整理を行うことが多い。以上のように、あらゆる項目の内容を含んでいる。

【データの保存とバックアップ】においては、『一度のみ』が該当する。ただし、『再利用のために最適化』の項目の「再利用する機会を最大限生かせるよう考慮した方法と場所」の内容が不明瞭であるため、一考の余地はある。筆者の行う研究におけるデータの種類には大別して「プログラミングコード」、「数値計算結果バイナリデータ」、「解析データ」の三種がある。「プログラミングコード」は最重要データであり、損失すると研究の続行が著しく困難になるおそれもあることから、複数箇所にバックアップを行っている。「数値計算結果バイナリデータ」の重要性は最も低く、損失したとしても作業のやり直しに多少の時間を要する程度であり、また、データ数も膨大であることからバックアップは基本的に行っていない。本タスクフォースで議論すべき長期保存データにも該当しないと考える。「解析データ」は研究終了時(論文執筆完了時)にまとめて整理してバックアップを行うことが多い。

【データを解析可能な状態にする】においては、『その都度』が最も該当するが、『再利用のために最適化』における「将来も自分自身が利用しやすい方法でデータ処理を行っている」も部分的に該当する。すなわち、標準化もしくは十分に文書化された手続きは存在しないが、自分では一貫した解析手法が確立しており将来にわたり不便はない。そして、第三者が利用しやすいかどうかは上述の通り、専門性による困難さなのか、データ解析や整理の不備によるものなのかの区別が明確でなく線引きが難しいため、一概には言及できない。

【データ解析と成果の取扱い】においては、『その都度』が該当する。数値シミュレーションにおいては解析パラメータを変えて計算し直すことが頻繁にあり、細かく文書化しては煩雑になるため必ずしも記録は残っていない。したがって、以前行った解析のパラメータを確認するために、同じ内容の数値計算をやり直すことも稀にある。

【データの共有と公開】においては、『一度のみ』が該当する。ただし、これまでに共同研究者を除いて、公開のリクエストがあったことはない。

B.4.3. 研究・教育データ保存の今後と課題

研究者が研究データを長期保存しなければならない動機付けは、大別すると3点あり、

- (1) 『将来、自分で再利用するため』
- (2) 『共同研究を含め、広く他者に利用してもらうため』
- (3) 『不正防止のため』

となる。このうち、筆者のように基礎研究のために社会への即時還元が難しく、共同研究者も少ない場合、(2)は動機にはなりにくい。また、多くの研究者は研究不正に関わることはなく、真面目に研究を行っているため、(3)も自発的に研究データを保存する動機にはならない。したがって、大きな共同研究プロジェクトに参加していない研究者にとっては、(1)で不便さを感じていない限り、そもそも研究データを長期保存する必要性に迫られていない。この点において、研究データを長期保存して欲しい国や大学の学術機関と意識のずれが生じているように感じられる。前述のオープンイノベーションセンターで構築中のポータルサイトも、複数人が参加する研究プロジェクトを対象としており、個々の研究者の研究データ保存に関する施策としては同センターも頭を悩ませているが、今後、啓蒙活動を行っていく予定であるとしている。

前節で検討した結果から、表1のループリックも不明瞭な点や線引きの難しい点等があり、今後も改良の余地は多くあると考える。例を挙げると、【データ保存とバックアップ】の項目に「再利用する機会を最大限生かせるよう考慮した方法と場所」とあるが、データ保存に関連した技術に詳しくない限り、どのような方法が「データを最大限生かせているか」の判別は難しいだろう。また別の例として、「標準化」という単語が複数使用されているが、データ保存関連の知識がない限り、何が標準なのかを知らない研究者も多いだろう。また、研究者の専門分野や研究環境によってループリックの内容も大きく変わりうる。現状のループリックでは、より広い範囲の専門をカバーしようと抽象的にするあまり、結局どの分野の研究者から見ても回答しにくい内容になっているように見受けられる。例えば、物理学分野内においても、理論物理学者と実験物理学者では、データを共有すべき共

同研究者の多寡や研究プロジェクト規模の違い等から、遵守すべきループブリックも大きく変わりうるのではないだろうか。もちろん、あらゆる専門分野に対してループブリックを設定するというのは現実的ではないが、ある程度研究者の専門や環境を反映したループブリックが複数用意されても良いと考える。

また、どのようなデータマネジメントの在り方が正しいのか定まっていない点も課題の一つである。すなわち、ループブリックに回答することで、自らのデータ管理の現状を把握することはできるかもしれないが、得られた回答結果を踏まえて、それでは何をどう改善すべきなのかが示されていない。一見すると、n. 2. 表1において右側に進むほどデータ管理が徹底されている印象を受ける。しかし、研究者によっては研究の柔軟性・流動性を大きく損なうことになりかねないため、全ての項目において最右端に到達することが必ずしも最良とは限らない。また、ループブリック内に【データマネジメント計画】とあるが、そもそも多くの研究者はデータ保存に関する計画を真面目に立てた経験は少ないと思われ、理想的な計画とは何を指しているかも一般に認知されていないだろう。これらの点も、研究者の専門分野や研究環境に影響を受けやすい内容であることから、複数のループブリックを用意すると同時に、理想的なデータ保存の在り方をそれぞれ具体的に定義する必要がある。

研究・教育データ保存にかかるプロセスが煩雑か否かも、今後の技術的課題として見過ごすことはできない重要な要素の一つである。近年の研究不正防止強化の観点や、オープンサイエンス・オープンイノベーションの流れを汲み取ると、いくら煩雑だからといって「面倒」の一言で片付けることはできないが、それでも技術的困難や煩雑さが付きまとうと、長期データ保存の重要性は認識されにくい。大学教員の校務や雑務の多さが指摘されて久しいが、研究・教育データ保存がさらにその現状に拍車をかけてしまうと敬遠されてしまうだろう。極端な例を挙げると、メタデータを個々の研究データに付す作業ですら面倒だと感じる研究者もいるだろう。

以上をまとめると、今後の課題として以下の3点に集約される。

- ① 『個々の研究者のデータ保存に対する意識改革』
- ② 『研究者の専門・環境を反映して回答しやすく、また今後のデータ保存の指針が把握しやすいデータマネジメントループブリックの設定』
- ③ 『複雑かつ煩雑な手続きを要しないデータ保存プロセスの確立』

これらは独立ではなく、互いに密接に関わりあった課題である。

表1 京都大学研究データマネジメントループブリック

	その都度	1度のみ	有効に活用	再利用のために最適化
データマネジメント計画	データを入手した段階で、自分なりに考える。標準化や文書化はしていない。	最初に計画を作成するが、通常は、途中で見直すことはしない。	詳細な計画を作成し、プロジェクト期間を通じ参照と見直しを行う。	自分あるいは他人が将来スムーズにデータを利用できるように計画している。
データの体系的整理	一貫した手法に従っていない。そのため、しばしば探し出すのに時間を要する。	手法はあるが、それを適用するのは、プロジェクト終了時のみ。	先を見越した手法を導入しているが、必ずしも標準化されていない。	他人が自分でデータを検索、内容を理解し、利用できる。
データの保存とバックアップ	作業中にどのデータが重要であるかを決め、通常はこれを特定の1か所に保存する。	どのデータを保存すべきかを理解している。損失のリスクを緩和するため、作業後にバックアップをとる。	作業中でも定期的にデータを保存するシステムを用いている。複数のバックアップを保持している。	再利用する機会を最大限生かせるよう考慮した方法と場所にデータを保存している。
データを解析可能な状態にする	標準化もしくは十分に文書化された手続きはない。	解析に適した形式を検討し整備するが、事例ごとに異なる方法をとっている。	データ処理方法は標準化され、かつ十分に文書化されている。	将来も自分自身及び第三者が利用しやすい方法で、データ処理を行っている。
データ解析と成果の取扱い	解析や実験がどのような手順やパラメータにより実施されたか確認するために、これらをやり直すことが多々ある。	解析作業終了後、パラメータの詳細や解析手順、プロトコルを文書化している。	データ解析の際、決められた方法で解析のワークフローや判断プロセスの詳細をいつも記録している。	自身が行った解析のワークフローや判断プロセスの詳細が他人によっても実行できることが確認できている。
データの共有と公開	研究成果を共有しているが、その根拠となるデータについては共有できていない。	リクエストがあった場合にのみ研究データの共有を行っている。	自身の研究成果や根拠となるデータは、第三者が利用できる形式で共有できる。	優れたデータマネジメントを実践し、自身のデータをいつでもだれでも効率的に共有できる。

B.5. 国立情報学研究所 込山悠介 個別報告書

B.5.1. 研究データ管理基盤 GakuNin RDM の紹介

近年では2021年3月に閣議決定された第6期科学技術・イノベーション基本計画の中で数値目標として、「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、2025年までに、データポリシーの策定率が100%になる。公募型の研究資金の新規公募分において、2023年度までに、データマネジメントプラン（DMP）及びこれと連動したメタデータの付与を行う仕組みの導入率が100%になる。」などと示されており、各機関ではこれに向けた研究データポリシーやインフラの整備が課題となっている。これまで、学術機関における研究データ管理は特定の機関、分野、プロジェクトあるいは研究者個人の自助努力で取り組まれている状況にあった。今後、全国的に一律に実施していくためには、担当部署の整理や研究支援者の教育などソフト面の整備の課題もある。一般に大学では研究データ管理のためのストレージは情報基盤センターが学内サービスとして運用しており、研究成果公開のための機関リポジトリは図書館のシステム部門が管轄、研究者への普及・啓蒙などはURA（University Research Administrator）などの研究戦略部門が担当しているが、大学の規模や状況に応じ横連携組織の有無や連携の意識の強弱は異なる。以上の様な理由から研究データ管理業務は研究者、研究支援者と所属機関が共同で進めていく業務ワークフローとなると考えられる。

国立情報学研究所では機関での研究データ管理の業務支援を目的として、2021年から全国の学術機関に向け研究データ基盤 NII Research Data Cloud（NII RDC）を提供している。NII RDC は研究データ管理・公開・検索するための3つの基礎的な基盤から構成されている。そのうちの 하나가研究データ管理基盤GakuNin RDMである。GakuNin RDM は研究プロジェクト実施中に、個人の研究者あるいは研究グループが研究データや関連の資料を管理するための基盤である。GakuNin(学認)とは日本の学術機関の認証フェデレーションであり、RDMは「研究データ管理（research data management）」を意味する。

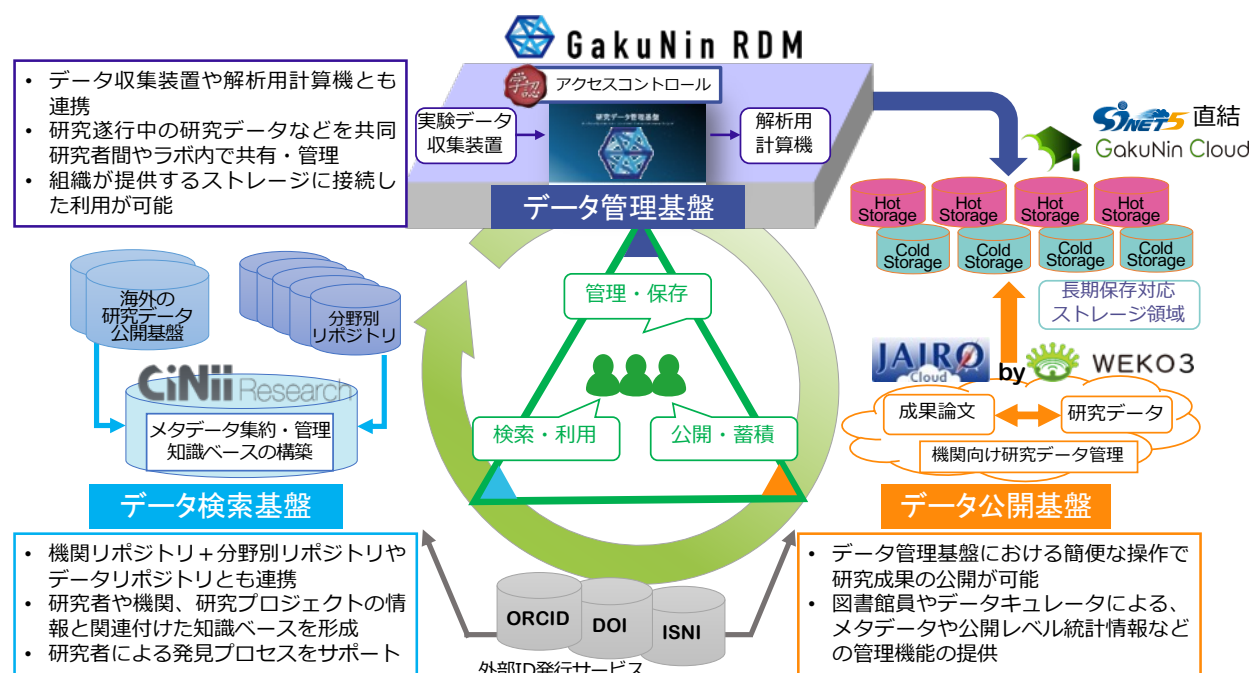


図 X.X.X.n 国立情報学研究所の研究データ基盤 NII Research Data Cloud

B.5.1.1. 管理基盤 GakuNin RDM のシステム概要

GakuNin RDM の基本コンセプトは、学術機関における研究データ管理業務の支援のために、研究不正防止やデータガバナンスのための研究公正の側面、そして、共同研究者間でのデータ共有や研究の協業を促進する研究推進の二つの観点に注力しシステム開発がされている。これらの目的のために GakuNin RDM は、研究推進への対応として、システム外部のクラウドストレージや研究用ソフトウェアと連携し、クローズド空間において共同研究者間でファイル共有やアクセスコントロールする機能を持つ。研究プロジェクトのデータ保存領域では、ファイ

ルのバージョン管理を行いながら、研究公正への対応として研究証跡を記録する機能を有する。

GakuNin RDM はブラウザベースで利用する Web アプリケーションであり、国内の学術機関での導入や運用コスト負担を軽減するためにクラウドサービスとしてユーザ機関へ提供されている。GakuNin RDM のアプリケーションで、サービス利用者は研究プロジェクト毎に研究データ管理用のエントリーページを立ち上げ、デスクトップのファイルをドラッグ・アンド・ドロップし、Web ブラウザ上のファイルマネージャーの様に操作することができる。GakuNin RDM のユーザは標準で NII Storage という最低限の容量分が提供される。また、各機関が個別に導入しているエンタープライズ契約のオブジェクトストレージをアドオン経由で接続することもできる。代表的プロバイダのパブリッククラウドの他、オンプレミスのサーバ上のプライベートクラウドも利用できる。ユーザが GakuNin RDM を通じて保存したデータに対しては時刻認証所業者のタイムスタンプサーバーを用いた証跡管理機能が働き、プロジェクト管理者が意図しないデータの改ざんや削除を検知することができる。

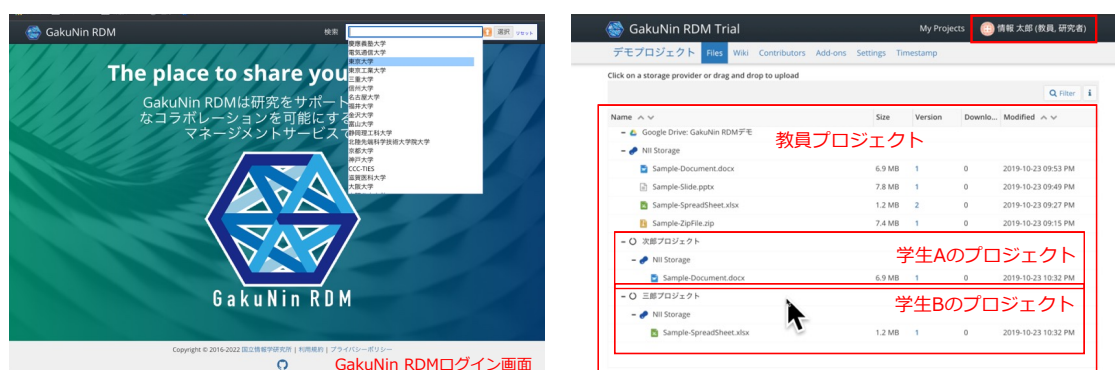


図 X.X.X.n 研究データ管理基盤 GakuNin RDM のポータル画面とデータ管理画面

GakuNin RDM を用い、全学的に研究データ管理業務を普及させるための普及・啓蒙活動は、これまでの大学附属図書館での学生・研究者向けの講習会や説明会のノウハウの活用が期待される。また、学生に対しては研究データ管理の教育プログラムが正規の講義として取り入れられていくように発展していく事も考えられる。実際には、情報基盤センターと図書館だけでは組織を越えた連携が困難なケースも多いため、研究データ管理業務の実施に向けて研究推進・研究支援・研究倫理などの部門が橋渡しとなるケースが増加している。

B.5.2. ANDS の IT インフラチェックリストと現状の研究データ管理基盤の機能の比較

ANDS の機関がデータ管理を実施するための IT インフラチェックリストに GakuNin RDM が現在持つ機能を照らし合わせて、機能の過不足や課題を確認した。

【1】. Storage for data and metadata

1. 特定のデータセットに十分なストレージがありますか？もしそうなら、それは永続的なストレージですか、それとも一時的なものですか？
 - 2022 年 9 月現在、GakuNin RDM は標準ストレージ領域として 1 ユーザあたり 100GB を提供している。オプションで利用機関がクラウドストレージを準備して接続する機関ストレージや、ユーザ自身がクラウドストレージを設定する拡張ストレージを用いることで、データセットを保存するのに十分なストレージを利用できる。
2. データセットは複数の場所でバックアップまたはアーカイブされていますか？
 - GakuNin RDM のデータは複数のデータセンター間でバックアップされている。
3. 可能性がありますか？データとメタデータはどの程度密接にリンクされていますか？
 - GakuNin RDM ではデータセットのリソースはストレージに、メタデータはデータベースにそれぞれ保存されている。Web アプリケーションとしてリソースとメタデータを連携させている。
4. 新しいデータストレージテクノロジーが利用可能になったときに、データセットをそれに移行できますか？

- GakuNin RDM では、ストレージ間でのデータ移行とメタデータデータベース設定の更新により、データセットを新規ストレージへ移行することが可能である。2022 年 9 月現在は、S3 互換ストレージ間での移行実績がある。
- 5. データセットは、基本的なアクセスのために圧縮またはデータ変換ツールと標準に依存していますか？もしそうなら、それらのツールと標準へのアクセスは継続されますか？
 - GakuNin RDM 中ではリソースデータは非圧縮で管理されており、出力時にはデータセットを Zip で圧縮してダウンロードすることができる。圧縮ツールは標準に依存しており、ユーザのデータセットへのアクセスは継続されている。
- 6. データセットは、基本的なアクセスのために独自のツールまたはフォーマットに依存していますか？もしそうなら、それらのツールやフォーマットへのアクセスは継続されますか？
 - GakuNin RDM では基本的なアクセスは Web アプリケーションの UI から行ない、特殊なフォーマットには依存していない。

【2】 Identity management authentication & access

1. IT インフラストラクチャは、すべてのデータセットに必要なアクセス制限を記録してサポートできますか。パスワード、AAF ID、倫理、アクセス契約および契約？
 - GakuNin RDM は認証を要する研究データ管理システムであり、全てのデータセットに対して必要なアクセス制限を設定することができる。学認の運用フェデレーションの SP として振る舞うため、ユーザは利用機関の IdP の認証方式 (AAF ID) に応じてデータセットへアクセスすることができる。機関機関が GakuNin RDM の利用規約を承諾し、国立情報学研究所へ利用申請を行なうことでアクセス可能になる。
 - GakuNin RDM は SP ローカルデータベースでのパスワード認証、倫理等のポリシーレベルの制御の機構、個別契約のオプションは持たない。
2. IT インフラストラクチャは、研究者がどこにいても、特定のデータセットでのコラボレーションをサポートできますか？
 - 研究者はどこにいてもインターネットから GakuNin RDM へアクセスして、特定のデータセットで共同研究者とコラボレーションできる。
3. IT インフラストラクチャは、認証されたアクセスを追跡および識別するために必要な登録をサポートできますか？
 - GakuNin RDM は認証により、ユーザのアクセスの追跡とアカウントを識別することができる。連携している IdP にアカウント登録されることで認証できるようになる。
4. 将来のある時点で（たとえば、禁輸期間が経過した後）、特定のデータセットのアクセス制限を変更することは可能ですか？
 - GakuNin RDM では SP 中での設定または IdP のアカウント設定により、アクセス制限の変更をデータセットの登録後に行なうことができる。
 - 2022 年 9 月現在、GakuNin RDM ではエンバーゴ期間等の自動設定によるアクセス制御の変更は備えていない。

【3】 Internal & external network connectivity

1. データセットが大きすぎたり、コストがかかりすぎて内部ネットワークや外部ネットワークを介して転送できない場合、フォールバック転送メカニズム（宅配便で送信されるハードドライブなど）はありますか？
 - GakuNin RDM では SINET を用いた学術データの流通をサービスの中心としており、巨大なデータセットを物理的に輸送するためのサービスは行っていない。
2. あなたの機関は内部データ転送にどのようなデータ転送コストを課していますか？他の機関との間の外部転送はどうですか？
 - 所内および所外とのデータ転送には SINET を利用しており、機関で SINET 利用料金を支払っている。一般利用者にはデータ転送料は請求していない。
3. 特定のデータセットが高い需要を引き付ける場合、IT インフラストラクチャは予想されるアクセスおよびコピー要求に対処できますか？
 - 特定データセットに対して高頻度アクセスが考えられる場合は、仮想ファイアウォールや仮想ルータの

ネットワーク帯域を広げることができる。より大きな帯域が必要な場合は、特定のデータセットについてストレージを外部パブリッククラウドに切り替えることもできる。

【4】 Access to discipline specific tools to support analysis

1. あなたの機関は、すべての専門分野のニーズを満たすのに十分な範囲の分析ツールを利用できるようにしていますか？それらを必要とするすべての人がアクセスできますか？
 - GakuNin RDM では学術機関向けに、全ての専門分野の需要をみたす汎用的なデータ解析環境として Jupyter を提供している。GakuNin RDM の利用機関の導入担当者がデータ解析機能のオプションの利用申請を行なうことで、学内で誰でもデータ解析機能を利用できるようになる。
2. データセットを分析するために特定のツールへのアクセスが必要な場合、そのツール（統計分析パッケージなど）はアーカイブされていますか？そのツール用に開発された構成またはカスタマイズスクリプトはアーカイブされていますか？
 - GakuNin RDM のデータ解析機能では、データセットとそれを解析する分析環境をあわせて保存・共有することができる。分析環境（ツール）の構成は設定ファイルにより管理されており、カスタマイズされたスクリプトや設定ファイルもストレージ中で保存される。
3. データセットが商用データ分析ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
 - データセットが必要とする商用の分析ツール相当のオープンソースソフトウェアが Python や R 言語などで存在していれば、GakuNin RDM のデータ解析機能の Jupyter 等でカスタマイズすることで、研究者は商用ツールを介さずアクセスすることができる。
4. データセットが特定の機器または機器へのアクセスを必要とし、これが特定の構成を必要とする場合、これをデータセットメタデータで指定するための規定はありますか？
 - 2022 年 9 月現在の GakuNin RDM は、データセットが特定の機器または機器へのアクセスを必要として、特定の構成を持つデータセットを取り込むためのデータセットのメタデータや規定を持たない。

【5】 Software development

1. 機関には、必要に応じてソフトウェアを開発または変更する能力がありますか？
 - GakuNin RDM はオープンソースソフトウェアである Open Science Framework をベースに日本国内の学術機関での研究データ管理に必要な機能を拡張追加されている。自機関にソフトウェア開発能力があると考ええる。
2. 機関には、時間のかかるカスタム開発の開発または変更に対する代替ソリューションを推奨する専門知識がありますか？
 - GakuNin RDM は汎用システムのため、時間のかかるカスタム開発は専門分野の外部機関との共同研究・開発を行なう事で専門知識を補っている。
3. データセットがカスタム開発されたソフトウェアによって生成された場合、データセットを使用してデータセットを構築するためのそのソフトウェアとドキュメントはアーカイブされていますか？
 - GakuNin RDM のデータ解析機能では、データセットとそれを生成したソフトウェアの双方を保存・共有することができる。Jupyter 等の分析環境を利用することでソースコードの行間にドキュメントを記載することもできる。
4. データセットが市販のソフトウェアまたは既製のソフトウェアによって生成された場合、データセットのカスタマイズまたは構成設定はアーカイブされていますか？商用ソフトウェアにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
 - 2022 年 9 月現在、GakuNin RDM は市販ソフトウェアを直接的に起動することはできないが、データセットのカスタマイズや構成の保存は可能。商用ソフトウェアにアクセスせず機能が同等のオープンソースツールがあれば、データ解析機能で環境を構築することで読み込みできると考える。
5. データセットが基本的なアクセス（解凍など）のためにカスタムまたは既製のソフトウェアを必要とする場合、メタデータで指定されたデータにアクセスするために必要な設定またはカスタマイズはありますか？
 - GakuNin RDM は標準でデータセットを Zip アーカイブとしてダウンロードすることが可能。GakuNin RDM のユーザアカウントがあればデータセットへアクセスするためにカスタム、既製のソフトウェアは要しない。

【6】Visualization

1. 機関は、すべての要件を満たすために、十分な範囲の視覚化ツールまたはそれらへのアクセスを利用できるようにしていますか。
 - GakuNin RDM のデータ解析機能では、Jupyter などの汎用の分析環境を使うことができ、様々な分野で利用可能な可視化環境を提供している。
2. データセットでサポートする視覚化が必要な場合、視覚化アセットとツールはデータセットとともにアーカイブされていますか？
 - GakuNin RDM のデータ解析機能では視覚化に必要なアセット・ツール（分析環境とライブラリ）をデータセットと合わせて保存・管理することができる。
3. データセットメタデータを使用して視覚化設定または構成値を記録する機能はありますか？
 - GakuNin RDM のデータ解析機能では分析環境を通じて可視化の構成値を編集し、記録しておくことができる。
4. サポートする視覚化で繰り返しまたは表示するためにハイパフォーマンスコンピューティングリソース（HPC）にアクセスする必要がある場合、データセットメタデータに関連する要件を記録する機能はありますか？
 - GakuNin RDM のデータ解析機能では分析環境を通じて HPC にアクセスするメタデータの要件を編集し、記録しておくことができる。

【7】Collaborative environments

1. 機関は、すべてのニーズを満たすために、適切なコラボレーション環境またはそれらへのアクセスを提供していますか。
 - GakuNin RDM ではサービスの導入機関が、研究者がコラボレーションを行なうための作業環境と、学認フェデレーションによるアクセスを提供している。
2. データセットがコラボレーション環境内に保存されている場合、またはデータセットにアクセスまたは分析するためにコラボレーション環境へのアクセスが必要な場合、データセットメタデータに環境のセットアップと構成を保存する機能はありますか？コラボレーション環境を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることはできますか？
 - GakuNin RDM のコラボレーション環境に保存されたデータセットへアクセスすることができ、データ解析機能で分析するための設定ファイルをデータセットと合わせて保存することができる。
 - GakuNin RDM では拡張ストレージ機能とデータ解析機能でそれぞれカスタマイズを保存することができ、コラボレーション環境のためにソフトウェア構成を保存することができる。
3. データセットが商用のコラボレーション環境へのアクセスを必要とする場合、商用環境にアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
 - GakuNin RDM ではパブリッククラウド等の商用のコラボレーション環境へアクセスせずに、他の研究者が利用できる代替のストレージ環境を提供している。また、GakuNin RDM 自体は商用クラウドストレージ環境の代替機能をもつオープンソースソフトウェアといえる。
4. 将来の追加または拡張に必要な場合、データセットを新しいまたは別のコラボレーション環境に移動することは可能ですか。
 - GakuNin RDM では、ユーザは拡張ストレージ領域を、情報システム担当者は機関ストレージ領域を接続して利用することで、データセットを新しいまたは別のコラボレーション環境へ移動することができる。

【8】High performance computing (HPC)

1. データセットにアクセスまたは分析するために HPC 施設へのアクセスが必要な場合、データセットメタデータで施設のセットアップと構成を指定することは可能ですか。
 - GakuNin RDM ではデータ解析機能にある「計算機持ち込み機能」を用いることで、ユーザ機関や共同利用施設の HPC 施設の計算機リソースをセットアップすることが可能。アクセスのために GakuNin RDM と HPC のそれぞれでネットワーク経路設計、ファイアウォールの設定を行ったのち、解析機能の設定ファ

イルの定義でリソースをプロビジョニングすることができる。

2. HPC 機能を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることは可能ですか？
 - GakuNin RDM のデータ解析機能では Jupyter などの環境をユーザ毎にコンテナを立ち上げて利用することができる。カスタマイズされたソフトウェアを設定ファイルでアーカイブすることができる。
3. データセットが商用 HPC ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
 - GakuNin RDM のデータ解析機能で提供している環境構築ツールは、オープンソースソフトウェアが使われている。
4. デスクトップシステムの容量が増えるにつれて、データを移動できるメカニズムはありますか？
 - GakuNin RDM では拡張ストレージ領域を用いることで、ユーザが利用可能なクラウドストレージ等を接続してデータ移動することができる。また、機関ストレージ領域では組織の情報システム管理者が、ストレージのソフトリミットであるクォータ設定を引き上げる事で、機関単位またはユーザ単位で容量を増設することや、ストレージのリプレース作業等が可能である。

B. 6. 豊橋技術科学大学 土屋雅稔 個別報告書

B. 6. 1. 現状の研究データ管理サービスに関連する課題

豊橋技術科学大学では、研究データの保存については、各研究者が自力で保存用メディアを確保し、10年間保管すること、という規定になっている。このような規定は、研究者の異動に対して頑健ではなく、実現性と継続性に問題がある。そのため、研究データの組織的保存体制を確立することが喫緊の課題である。組織的保存体制の確立には、少なくとも、保存対象となる研究データの範囲の決定、メタデータ付与方針の設定、保存用ストレージの確保、およびこれら全てを継続的に維持する組織体制の整備、が必要である。

○保存対象となる研究データの範囲

保存対象となる研究データの範囲を決定するには、まず、研究データ保存の目的を明らかにしなければならない。研究データ保存の目的は、大きく分けて2通りの考え方がある。第1は、公刊された研究論文について内容に疑義が発生した場合に、その内容を検証できるだけのデータを保存しておかねばならない、という研究公正を目的とする考え方である。第2は、研究データの作成には多額の研究費を要している以上、一度作成された研究データはできるだけ再利用されるべきというオープンサイエンスの推進を目的とする考え方である。

研究公正を目的とする場合は、保存対象となる研究データは、公刊論文の検証に必要な範囲に限定される。また、研究データに付与されるべきメタデータは、当該研究データを利用した論文の書誌情報として自然に定義される。加えて、保存対象となる研究データは、実際に日々作成される研究データに比べれば、おそらくは圧倒的に少量であり、保存用ストレージを確保する技術的・予算的な制約は小さい。

それに対して、オープンサイエンスの推進を目的とする場合は、保存対象となる研究データは、日々作成される研究データ全てとなる。この場合、研究データに付与されるべきメタデータは、当該研究データの作成日時や実験機器、実験条件など、半自動的に付与可能なメタデータに限られる。第1に、このメタデータを安定的に付与することは可能か、という点が問題になる。一部の実験機器は、日時や条件などを自動的に記録する機能を備えるが、そのようなデータのみではおそらく十分ではなく、なんらかの手作業は避けられないと考えられる。第2に、このような限られたメタデータのみを手がかりとして、他の研究者が必要な研究データを検索・再発見し、再利用することが可能か、という点が問題になる。加えて、保存対象となる研究データは圧倒的に大量になり、保存用ストレージを確保する技術的・予算的な制約は大きい。

本学のような小規模大学においては、後者のオープンサイエンス推進を目的とする立場で体制を整備することは、予算や人員などの制限から、非現実的であると考えざるを得ない。前者の研究公正を目的とする立場で体制を整備し、必要に応じてオープンサイエンス推進の目的を取り入れていくことが現実的と考える。

○メタデータ付与方針

研究公正を目的とする場合、保存対象となる研究データは、公刊された論文の内容を検証するために必要な範囲で足りる。したがって、機関リポジトリと連携し、論文と研究データの登録を同時に行うシステムを整備すれば、研究データのメタデータは、当該論文の書誌情報という形で自然に収集・付与される。

○保存用ストレージの確保

研究公正を目的とする場合、必要となるストレージ容量は、比較的少量である。各種コールドストレージサービスを利用すれば、比較的安価に実現可能と考えられる。

公刊済み論文のデータとは言っても、研究上の機密情報とプライバシー情報に関する配慮は必要である。しかし、研究公正を目的とする場合は、広く公開する必要はなく、研究公正に関する調査委員会だけが閲覧できれば良い。したがって、研究者自身と大学本部から許可された人物だけに読み取りを許可するよう権限を設定すれば解決できる。

○問題点

この方針は、技術的・制度的・人員的には極めて現実的な案である。最大の問題点は、研究データの保存を求める各種の要請との適合性である。研究データの保存を求める各種の要請においては、オープンサイエンス推進という立場が強調されていることが多く、上記の方針では対応しきれていない。適合性を高める対応が必要である。

B.7. 理化学研究所 舩本現 個別報告書

B.7.1. 現状の研究データ管理に関連する取り組み・提供しているサービスの紹介

理化学研究所では、情報化戦略の司令塔である ICT 基盤整備・サービス連携委員会の下で専門部会として 2018 年度に研究データ部会を立ち上げてデータポリシーの検討を始め、所内での 2 度のアンケートも踏まえて 2020 年 3 月に「研究データの管理、公開、共有及び利用に関する基本方針（理研データポリシー）」を策定した。

また 2019 年度には情報統合本部の下で基盤研究開発部門に理研の研究データ管理のインフラ構築に取り組むユニットを設けてデータ管理システム導入の検討を進め、理研研究情報管理システム（RIKEN Research Data & copyrighted work Management System (R2DMS)）を開発した。R2DMS は 2021 年度から試験運用として一部のユーザに提供を始め、2022 年度から本格運用を開始した。

R2DMS は、グループ内の研究データを効率よく共有・管理し、理研における研究を促進することを目的としており、GakuNin RDM に独自の拡張を加えたものと、NextCloud から構成されている。GakuNin RDM は参加メンバー、研究概要、研究データを一元的に管理するプロジェクト管理サービスであり、NextCloud は研究データを実際に保管し、プロジェクト管理サービスからも参照可能なオンラインストレージである。2 つのサービスは目的に応じて、単独あるいは連携して利用することが可能である。R2DMS はストレージとして理研の共同利用計算機のデータ科学基盤 HOKUSAI Sailing Ship (SS) を主として想定しているが、その他にも各研究室内の実験機器の一次ストレージや外部の商用クラウドストレージを利用することも可能である。

この他にもオンラインストレージとして Box が導入されている。こちらはストレージ容量としては無制限であるが、アップロード可能な最大ファイルサイズに制限があり通信速度の問題もあるので主として事務部門の利用を想定している。ただし、研究データに関しても小中規模データについては利用されているようである。

B.7.2. 現状のシステム運用管理状況を踏まえての ANDS チェックリストの実施結果

【1】. Storage for data and metadata

1. 特定のデータセットに十分なストレージがありますか？もしそうなら、それは永続的なストレージですか、それとも一時的なものですか？
 - R2DMS が利用している HOKUSAI SS のストレージの総容量は 30PB であり、プロジェクトあたりの容量は原則として制限されていない。クラウドストレージの Box については一人あたりの容量に制限はないが、1 ファイルあたり 150GB という制限がある。永続的なストレージは用意されていない。
2. データセットは複数の場所でバックアップまたはアーカイブされていますか？
 - データセットの地理的なバックアップ、アーカイブは実施されていない。
3. データセットとそのメタデータが分離している場合、それらは一緒に保存されますか、それとも分離される可能性がありますか？データとメタデータはどの程度密接にリンクされていますか？
 - メタデータは理研内 GakuNin RDM に保存され、データセットの実体は NextCloud 内に保存される。両者はシステム間連携によって管理されている。
4. 新しいデータストレージテクノロジーが利用可能になったときに、データセットをそれに移行できますか？
 - NextCloud と GakuNin RDM に依存するが、基本的に可能であると想定している。
5. データセットは、基本的なアクセスのために圧縮またはデータ変換ツールと標準に依存していますか？もしそうなら、それらのツールと標準へのアクセスは継続されますか？
 - 圧縮またはデータ変換ツールは利用していない。
6. データセットは、基本的なアクセスのために独自のツールまたはフォーマットに依存していますか？もしそうなら、それらのツールやフォーマットへのアクセスは継続されますか？
 - アクセスは NextCloud と GakuNin RDM に依存している。GakuNin RDM を利用せずに NextCloud だけでストレージにアクセスすることも可能である。

【2】 Identity management authentication & access

1. IT インフラストラクチャは、すべてのデータセットに必要なアクセス制限を記録してサポートできますか。パスワード、AAF ID、倫理、アクセス契約および契約？
 - 認証は Shibboleth を用いた理研統合認証基盤およびゲストアカウントを利用しており、アクセス制限は GakuNin RDM と NextCloud の機能でおこなう。
2. IT インフラストラクチャは、研究者がどこにいても、特定のデータセットでのコラボレーションをサポートできますか？
 - R2DMS へのアクセスは所外からでも可能である。研究者はどこにいても GakuNin RDM と NextCloud が提供するコラボレーションを利用できる。
3. IT インフラストラクチャは、認証されたアクセスを追跡および識別するために必要な登録をサポートできますか？
 - GakuNin RDM と NextCloud が管理者に対してそれらの機能を提供していれば可能である。
4. 将来のある時点で（たとえば、禁輸期間が経過した後）、特定のデータセットのアクセス制限を変更することは可能ですか？
 - GakuNin RDM の機能で任意の時点でアクセス制限の変更をすることは可能である。

【3】 Internal & external network connectivity

1. データセットが大きすぎたり、コストがかかりすぎて内部ネットワークや外部ネットワークを介して転送できない場合、フォールバック転送メカニズム（宅配便で送信されるハードドライブなど）はありますか？
 - 想定していない。個別相談で対応する可能性はある。
2. あなたの機関は内部データ転送にどのようなデータ転送コストを課していますか？他の機関との間の外部転送はどうですか？
 - 論文の論拠データと公開用の利活用データについては無料。その他の研究データの保存については、ストレージプロバイダーに応じてデータ量に応じて利用負担金がかかる。

3. 特定のデータセットが高い需要を引き付ける場合、IT インフラストラクチャは予想されるアクセスおよびコピー要求に対処できますか？
 - 現状では対応していないが、将来的には検討していく。

【4】 Access to discipline specific tools to support analysis

1. あなたの機関は、すべての専門分野のニーズを満たすのに十分な範囲の分析ツールを利用できるようにしていますか？それらを必要とするすべての人がアクセスできますか？
2. データセットを分析するために特定のツールへのアクセスが必要な場合、そのツール（統計分析パッケージなど）はアーカイブされていますか？そのツール用に開発された構成またはカスタマイズスクリプトはアーカイブされていますか？
3. データセットが商用データ分析ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
4. データセットが特定の機器または機器へのアクセスを必要とし、これが特定の構成を必要とする場合、これをデータセットメタデータで指定するための規定はありますか？
 - 運用側では分析ツールをアーカイブしていない。データを保存したユーザ自身が必要な分析ツールを保存することはできる。GakuNin RDM の機能として分析ツールの利用が可能となった場合は R2DMS でも同様に利用可能となる。

【5】 Software development

1. 機関には、必要に応じてソフトウェアを開発または変更する能力がありますか？
2. 機関には、時間のかかるカスタム開発の開発または変更に対する代替ソリューションを推奨する専門知識がありますか？
3. データセットがカスタム開発されたソフトウェアによって生成された場合、データセットを使用してデータセットを構築するためのそのソフトウェアとドキュメントはアーカイブされていますか？
4. データセットが市販のソフトウェアまたは既製のソフトウェアによって生成された場合、データセットのカスタマイズまたは構成設定はアーカイブされていますか？商用ソフトウェアにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
5. データセットが基本的なアクセス（解凍など）のためにカスタムまたは既製のソフトウェアを必要とする場合、メタデータで指定されたデータにアクセスするために必要な設定またはカスタマイズがありますか？
 - 個別のソフトウェア開発・公開は原則としてユーザ自身がおこなう。個別のソフトウェアの開発・変更の相談は受け付けていない。

【6】 Visualization

1. 機関は、すべての要件を満たすために、十分な範囲の視覚化ツールまたはそれらへのアクセスを利用できるようにしていますか？
2. データセットでサポートする視覚化が必要な場合、視覚化アセットとツールはデータセットとともにアーカイブされていますか？
3. データセットメタデータを使用して視覚化設定または構成値を記録する機能はありますか？
4. サポートする視覚化で繰り返しまたは表示するためにハイパフォーマンスコンピューティングリソース（HPC）にアクセスする必要がある場合、データセットメタデータに関連する要件を記録する機能はありますか？
 - 運用側で用意しているものはない。GakuNin RDM の機能として対応可能なものは利用できる。

【7】 Collaborative environments

1. 機関は、すべてのニーズを満たすために、適切なコラボレーション環境またはそれらへのアクセスを提供していますか？
 - 理研の研究者と共同研究をおこなっている所外のユーザとのコラボレーションにはゲストアカウントを発行することによってアクセスを提供している。またデータセットに対して完全公開を選択することも

可能である。

2. データセットがコラボレーション環境内に保存されている場合、またはデータセットにアクセスまたは分析するためにコラボレーション環境へのアクセスが必要な場合、データセットメタデータに環境のセットアップと構成を保存する機能はありますか？コラボレーション環境を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることはできますか？
 - GakuNin RDM および NextCloud に依存する。
3. データセットが商用のコラボレーション環境へのアクセスを必要とする場合、商用環境にアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
 - GakuNin RDM および NextCloud に依存する。
4. 将来の追加または拡張に必要な場合、データセットを新しいまたは別のコラボレーション環境に移動することは可能ですか。
 - GakuNin RDM および NextCloud に依存する。

【8】High performance computing (HPC)

1. データセットにアクセスまたは分析するために HPC 施設へのアクセスが必要な場合、データセットメタデータで施設のセットアップと構成を指定することは可能ですか。
2. HPC 機能を使用するために必要なカスタマイズされたソフトウェアコンポーネントをアーカイブすることは可能ですか？
3. データセットが商用 HPC ツールへのアクセスを必要とする場合、商用ツールにアクセスせずに他の研究者が利用できる代替のオープンソースツールはありますか？
4. デスクトップシステムの容量が増えるにつれて、データを移動できるメカニズムはありますか？
 - 理研の HPC 施設である HOKUSAI BW とは連携していない。分析ツールは GakuNin RDM が用意する連携アドオンに依存する。

B.7.3. システム運用管理者視点から今後求められる研究データの長期運用管理の在り方についての考えや意見

現状は、これまでの様々な取り組みによって研究データを保存するストレージインフラ（ハードウェア）という点では充実してきており、次のステップとして、保存されたデータの共有・管理を促進するための研究データ基盤の整備を進めている状況だと認識している。そのために GakuNin RDM をはじめとするソフトウェアやツールも活発に開発されている。

研究データ基盤の一番下を支えるストレージインフラについては、GakuNin RDM 等のデータ共有・管理ソフトウェアはオンプレと外部のクラウドのどちらのストレージも利用できることが多いが、実際の運用ではこれらをどのように使い分けていくのかが大きな問題だと考えている。本研究所では一部の研究分野ではペタバイトクラスの非常に大規模な研究データを算出・利用するような研究もあり、ストレージの適切な使い分けが必要となってくる。

ANDS のチェックリストは、ストレージインフラに関する項目と研究データの共有・管理に関する項目だけでなく、研究データを利用・分析するためのツールやソフトウェアの整備に関する項目が非常に多く、研究データの保存とそのデータを用いた分析を密接なものとして扱っていることが非常に印象的であった。

単に研究データを貯めて公開するだけではなく、そのデータを使った研究活動をドライブしていくためには、データ分析環境の整備もデータ保存・管理環境に劣らず重要であることは当然である。そのような観点からすると、現状の理研の研究データ管理システムではデータの共有（NextCloud）、データの管理（GakuNin RDM の改修版）までしかカバーできておらず、データの利用をサーバ側でおこなう環境については、整備が不十分と言わざるを得ない。GakuNin RDM 側でも BinderHub との連携アドオンなどで対応していくだろうが、理研に所属している研究者の研究対象や研究手法の多様性を考えると理研も独自にデータ分析環境を整備していく必要があるかもしれない。

いずれにしても、これからは単にデータを保存してアクセスしやすくするだけでなく、データをどれだけ利用しやすくするかを意識して基盤環境を整備していくことが情報システム部門としては重要な課題であると認識を新たにした。

B. 8. 富士通 Japan 報告書

B. 8. 1. はじめに

DX（デジタルトランスフォーメーション）という言葉が一般的となりデータが生み出す価値が注目されている昨今であるが、データの取得から廃棄までデータを管理することには様々な課題が存在する。学術分野の研究データにおいても、適切なデータ管理を行う為のプロトコル整備が行われているが、コスト、ヒューマンリソース、時間等の制約により、プロトコル通りのデータ管理を行うことの難しさが論文(DEIM Forum 2019 C2-1)で報告されている。研究データを適切に管理し研究を行っていくことは、日本の学術及び産業の発展に必要不可欠なことであると考えている。

B. 8. 2. 3つの視点による考察

以下では3つの観点で『京都大学研究データマネジメントループリック』の考察を論じる。

- (1) 学術機関にシステムを納め運用支援を行っているシステム構築技術者の視点
- (2) システム監査技術者の視点
- (3) アプリケーションエンジニアの視点

それぞれの視点から『京都大学研究データマネジメントループリック』に対するコメントを通して、考えを述べていく。

B. 8. 2. 1. システム構築技術者の視点

研究データマネジメントループリックは以下の通り 6 つの項目から構成されている。本タスクフォースに参加している研究機関での研究データ管理にあてはめた結果を参照しても、「有効に活用」や「再利用のために最適化」と評価される項目は少なかった。

学術機関向けにコンピュータシステムを構築し運用支援する技術者として、学術機関が抱える課題も勘案し、データ管理の考察を行う。

表1 京都大学研究データマネジメントループリック (5月)

	その都度	1度のみ	有効に活用	再利用のために最適化
データマネジメント計画	データを入手した段階で、自分なりに考える。標準化や文書化はしていない。	最初に計画を作成するが、通常は、途中で見直すことはしない。	詳細な計画を作成し、プロジェクト期間を通じ参照と見直しを行う。	自分あるいは他人が将来スムーズにデータを利用できるように計画している。
データの体系的整理	一貫した手法に従っていない。そのため、しばしば探し出すのに時間を要する。	手法はあるが、それを適用するのは、プロジェクト終了時のみ。	先を見越した手法を導入しているが、必ずしも標準化されていない。	他人が自分でデータを検索、内容を理解し、利用できる。
データの保存とバックアップ	作業中にどのデータが重要であるかを決め、通常はこれを特定の1か所に保存する。	どのデータを保存すべきかを理解している。損失のリスクを緩和するため、作業後にバックアップをとる。	作業中でも定期的にデータを保存するシステムを用いている。複数のバックアップを保持している。	再利用する機会を最大限生かせるよう考慮した方法と場所にデータを保存している。
データを解析可能な状態にする	標準化もしくは十分に文書化された手続きはない。	解析に適した形式を検討し整備するが、事例ごとに異なる方法をとっている。	データ処理方法は標準化され、かつ十分に文書化されている。	将来も自分自身及び第3者が利用しやすい方法で、データ処理を行っている。
データ解析と成果の取扱い	解析や実験がどのような手順やパラメータにより実施されたか確認するために、これらをやり直すことが多々ある。	解析作業終了後、パラメータの詳細や解析手順、プロトコルを文書化している。	データ解析の際、決められた方法で解析のワークフローや判断のプロセスの詳細をいつも記録している。	自身が行った解析のワークフローや判断プロセスの詳細が他人によっても実行できることが確認できている。
データの共有と公開	研究成果を共有しているが、その根拠となるデータについては共有できていない。	リクエストがあった場合にのみ研究データの共有を行っている。	自身の研究成果や根拠となるデータは、第三者が利用できる形式で共有できる。	優れたデータマネジメントを実践し、自身のデータをいつでもだれとでも効率的に共有できる。

(1) データマネジメント計画

データマネジメント計画の詳細を詰めないままプロジェクトをスタートさせ、プロジェクトが所属する研究教育機関が提供するファシリティ(ストレージ環境含む)を利用していることがある。この場合、プロジェクトが進むごとにデータ保存期間やデータ量が研究教育機関から提供されるファシリティのポリシーと一致していないことが徐々に明確になり、解決の為、ファシリティの管理者、プロジェクトメンバー双方に負担をかけることになる。ファシリティ運用の運用支援を行っている身としては、ポリシーの例外処理にかかる過大な運用コストがかかると感じており、適切なデータマネジメント計画と実行が行われれば、研究教育機関の TCO 削減につながると考える。

(2) データの体系的整備

システム構築及び運用支援のプロジェクトにおいて、プロジェクトデータの管理を行っている。プロジェクトマネージメントでは、プロジェクト立上げ期に発行するプロジェクト計画書に命名規約、階層構造を含めたデータ保存場所を記載し、プロジェクトメンバーに共有することが求められる。しかし、こちらの実践度合いや記載粒度は、プロジェクト規模やプロジェクト内容に依存している。学術分野においても同様に、プロジェクト規模や研究内容によって求められるレベルが異なると考えられる。

(3) データの保存とバックアップ

対応可能なレベルを向上させるためにはストレージシステムにもコストがかかる。そのため、各プロジェクトでは予めプロジェクトオーナーとどのレベルの対応とするのかを決めプロジェクトをスタートすることが重要である。またバックアップ取得有無だけでなく、1 次ストレージ、2 次ストレージといった各ストレージの信頼性を定量的に示し、ステークホルダー間で合意しておくことで、保存データ消失時の納得性を高めることやストレージにかかるコストを抑えることが可能と考える。

(4) データを解析可能な状態にする

特にコメントはない。

(5) データ解析と成果の取扱い

実験環境の説明、実験の内容とその記録を log ノートとして手書きしているような研究分野に関しては、ストレージシステムに保存されている解析データだけでなく、解析のプロセスや判断を記載した解析ノート(log ノート)を電子デバイスに置き換えることで、解析メソッドを含めたデータを保存することが可能となり、解析手法、成果共有の効率化が図られると考える。学術機関への提案においては、電子化する範囲を広げ、より研究支援を行えるシステムを意識する必要があると考える。

(6) データの共有と公開

研究成果を伝える為、データを共有することは重要と考えるが、同時に意図しないデータを公開しないようにするなど、リスク管理をしっかり行うことが肝要と考える。近年は、MS Office などはファイルが流出したとしても適切なアクセス権を持つユーザ以外は中身を確認できないような仕組みを有している。同様の仕組みが MS 製品以外にも拡張され、安全にデータ公開できる環境が整うことを期待している。

学術機関は自由な研究という気風が優先され、組織ガバナンスを効かせにくい部分もあると考えている。一方限られた予算で研究成果の公開、適切なデータ管理を求められており、そちらに対する取り組みが必要である。学術機関にサービスを提供する業者としては、新技術での課題解決、ベストプラクティスの共有を通して、研究教育データの適切な管理に貢献する必要がある。

B. 8. 2. 2. システム監査技術者の視点

次に情報システムの開発フェーズや、維持管理フェーズで行われる『システム監査』の評価指標を示し、『研究データマネジメントルーブリック(和訳)』と比較し考察してみる。

表 2. システム監査指標

	評価指標	評価内容
1	戦略性	組織体の競争優位性の確保・維持への貢献度合い
2	有効性	事業活動に対する目的適合性および関連性
3	効率性	組織体の事業活動における活用度合い、費用対効果
4	信頼性	品質度合いやその品質に影響を与える障害や事故に対する耐性
5	準拠性	システムのライフサイクルにおいて、関連法令、契約、内部規定など守るべきルールを遵守している度合い IT ガバナンスの一環としてコンプライアンス（法令・規範の遵守）に規定されることもある
6	安全性	災害や不正アクセス、破壊行為などのあらゆる脅威から保護するためのセキュリティ対策の度合い

(1) 外観（評価項目）

下記にまとめた通り、『研究データマネジメントルーブリック』と『システム監査指標』の各項目はほぼ合致している。例外としては、IT ガバナンスの中のコンプライアンス規定の整備有無に関わる『準拠性』については『研究データマネジメントルーブリック』では評価項目となっていなかった。また、企業での指標には必須項目である『効率性（費用対効果）』が、『研究データマネジメントルーブリック』評価指標にはないことが分かる。

データマネジメント計画 および データの体系的整備 ⇔ 戦略性

データの保存とバックアップ ⇔ 信頼性、安全性

データを解析可能な状態にする ⇔ 有効性

データ解析と成果の取扱い ⇔ 効率性(活用度)

データの共有と公開 ⇔ 有効性

(2) データマネジメント計画

そもそも学術研究は個人または小集団で実施するものであり、その小組織で管理運営されてきた経緯がある。一般企業をイメージした『システム監査指標』では企業は利益追求集団であり情報システムは他社との競争力醸成の一助であるとの考えであるのに対し、研究業務では相互扶助の考え方もあるため、明確に規定できないのだと考える。

(3) データの体系的整備

前述しているが、研究業務が相互扶助のうでで成り立つものであるが、各研究機関は運営組織が異なることから、標準化を図ることが困難であると考え。例えば、コンピュータを利用した研究業務にも解析結果が重要である研究があれば、数年前に一次計算した収集データや解析データを再利用して進める研究があるなど、数えられない様々な型があることから標準化が困難であると考え。

(4) データの保存とバックアップ

過去の研究データを再利用することが多い研究機関では、長期間に渡るデータの安全性、一意性、信頼性に着目する傾向にあると見受ける。企業でも長期保存の必要性はあるが、製造する製品にはライフサイクルがあることから、研究機関とは異なり、数十年の長期に渡る保存が求められることは少ない。一方、企業では事業活動の源泉でもあるデータの安全性確保に重きを置いており、例えば、秘密情報の外部流出、漏洩の対策に多くのコストをかける傾向にある。

(5) データを解析可能な状態にする (6) データ解析と成果の取扱い (7) データの共有と公開

企業でのデータ共有は古くは自社内でのデータ共有を推奨してきた。活動の中で生成、収集したデータは企業全体の資産であるという考え方から、一つのデータからより多くの成果物が製造されることを推奨しているからである。近年はサプライチェーンマネジメントを構築することで、複数の企業間での統合システム構築がなされているが、複数といっても利害関係が一致する企業間であり、その接続範囲は限られている。一方、研究機関におけるデータ公開はマスコミを通じてのスーパーコンピュータ解析結果の公表などが定着しつつあるが、まだ、黎明期であると考え。データを共有することのインセンティブ、データを公開することでのインセンティブを明確にしないと研究機関におけるデータ再利用向上や共有、公開が進まないと考え。

以上のことから、研究機関が企業の情報システム評価（システム監査指標）に準じる必要はないと考えるが、次の観点のシステム評価を検討することが研究データ管理の今後の一助になると考える。

- ・ データ管理単位毎のデータ再利用ターゲットの明確化
- ・ データ保管に掛かる経費の算出と継続的管理
- ・ データの安全性確保手段の検討と継続的管理
- ・ データ公開、共有推進による研究者へのインセンティブ付与

B.8.2.3. アプリケーションエンジニアの視点

最後にアプリケーションを開発するエンジニアの視点で、データマネジメントにおいて重視するポイントを示し考察してみる。対象とするアプリケーションは、大学経営に資する大学 IR (Institutional Research) ソリューションとしての統合データベースとする。

大学 IR で扱うデータは本論のテーマとなっている研究教育データとは領域が異なっているが、両者のデータマネジメントのポイントに目を向けることで、より深い考察ができると考える。

(1) アクセスコントロールと監査ログ

大学 IR で扱うデータベースには、学内外のあらゆる場面で記録されるデータが収集・蓄積される。データが集約されれば、例えば学生に紐づく活動データを多角的に分析することで、教育改善や志願者増加のための広報活動に活かすことができる。多様な場面でデータを扱えるその反面、センシティブなデータをどのように安全に扱うか、ということも問題となる。「誰がどのデータにアクセスできるか」ということが適切に管理できるよう、データベースに接続するユーザアカウントごとにテーブル単位、レコード単位、カラム単位でのアクセス制限をかけて厳格に管理できることがのぞましい。また、このアクセスコントロールが適切に運用されていることを定期的チェックできる、あるいは問題発生時に過去に遡って「いつ、誰が、どのデータにアクセスしたか」を調査できるよう、監査ログを記録する必要がある。

(2) 拡張性と安定性

大学 IR は 10 年分など長期間にわたって学生の傾向を分析するため、膨大なデータの長期保存が必要となる。近年は IoT の進展により、学生の行動ログを収集しやすくなっているため、技術の進歩により蓄積されるデータのサイズは幾何級数的に増大する。そのような背景からデータベースサイズの拡張の容易さと、データが膨大になっても分析のためのグルーピングやフィルタリングの性能が安定していることは重視すべきポイントである。

(3) ユーザインターフェースのシンプルさ

大学 IR においてデータベースの管理が容易であることは重視すべきポイントである。日々拡充するデータベース設計を理解し、長期間維持していくためには、外部のベンダーに頼らず学内の職員でメンテナンスできなければならない。データを格納するテーブルや活用に適したビューの作成、分析し易い形にデータを加工する際の処理ロジックの記述は、高度な機能を搭載しているよりもデータベースやプログラムの知識を持っていない人員でも理解が可能なほどシンプルであることが優先的に重視される。

(4) 世代管理

大学では「学校基本調査」として年度ごとの在学者数などの統計情報を文部科学省に提出している。学生には在学中に退学や休学、転籍などの異動が発生するため、当該年度の学部別在学者数といった情報は「5 月 1 日現在」として毎年日付を特定した統計情報としなければならない。このようなことを実現するためには、統合データベースのテーブルには時系列項目が付与され、過去のデータ履歴を確認できる必要がある。このようなデータベースを「DWH (データウェアハウス)」と呼ぶ。なお、データの特性に応じてこのような世代管理を必要とする場合と必要としない場合があるため、データベース設計時には注意が必要となる。

(5) コメント管理

大学 IR で扱うデータを集約する際、データの変換処理は少なからず発生する。例えば、データ発生源となるシステムごとにコード体系がバラバラでも、統合データベース同じ意味として解釈される必要があるデー

タ項目は対応表による変換処理を最新データの収集の都度実行する必要がある。統合するデータが多岐にわたると、それらの変換処理ロジックの数も膨大になるため、「どのような目的・意図でそのような変換をしたか」といった情報や、変換後のデータ項目は利用者側でどのように活用されることを期待しているか、が正確に引き継がれる必要がある。そのため、インプットデータ・変換と統合・アウトプットデータなどあらゆる場所でコメントを残し共有できることが必要となる。

(6) 匿名化・秘匿化

大学 IR で扱うデータは、ローデータを利用する価値と情報漏洩リスクを適切に評価し、リスクが価値を上回る場合は匿名化・秘匿化を検討する必要がある。なお、この加工は不可逆的なものとなることから、加工の仕方によってはテーブル同士の結合ができなくなり、データを集約して活用する意義がなくなってしまう可能性があることにも注意が必要である。

以上、大学 IR におけるデータベースが備えるべき機能・非機能的な要件をポイントとして整理した。各項目は研究データマネジメントルーブリックと必ずしも一対一対応するものではないが、これらの要件が求められる背景は通底するものが感じられる。大学 IR は未だ発展途上中であり、これらの要件は組織的にどのレベルまでルール化や標準化されているべきか、という共通認識はない。大学 IR データマネジメントルーブリックの開発にも視野を拡げ、相互に検討を進めることで、より深い考察から評価・検討が可能となるのではないだろうか。

B.8.3. まとめ

3つの視点による考察を述べたが、総括すると以下となる。

- ① 長期データ管理運用保存を一つのプロジェクトとしてとらえ、コスト管理を重視しつつ、
 - ② システム管理者としてシステム評価を意識した運営が必要
 - ③ また、アプリケーション開発と同様に多数の利用者を意識した機能設計も必要
- とまとめられる。研究者の要件を再整理し、システム構築、運用に活用できるものであれば幸いである。

【発行者】サイエンティフィック・システム研究会

【編集】サイエンティフィック・システム研究会

学術機関における研究教育データの長期運用保存管理の在り方 TF

【発行日】2022 年 9 月 30 日



<連絡先> サイエンティフィック・システム研究会 事務局

〒212-0014 神奈川県川崎市幸区大宮町1 番地 5 JR 川崎タワー

富士通株式会社 官公庁事業本部官庁第六事業部第一ビジネス部内

E-mail : ssken-office@ml.css.fujitsu.com

<https://www.ssken.gr.jp/MAINSITE/index.html>

※著作権は各原稿の著者または所属機関に帰属します。無断転載を禁じます。