

資料 2

国際的動向を踏まえたオープン
サイエンスの推進に関する検討会
(第9回)
平成31年2月21日(木)

研究データリポジトリ整備・運用ガイドライン（案）

1 研究データの保存・公開とデータリポジトリの整備・運用

研究環境の ICT 化の急速な進展により研究データの電子化が促進され、研究データは爆発的に増大している。研究データを収集・蓄積して公開することにより、各学問領域の研究活動の迅速化や学際的な研究活動の進展、データ科学による新たな研究領域の創成が期待されている。一方、研究データの信頼性確保や研究不正防止の観点からも、研究データの一定期間の保存と適切な公開や共有・相互利用（以下、公開という）が求められている。これらを背景に、2013年6月のG8ロック・アーン・サミットにおいて「オープンデータ憲章」が合意、公表されるなど、オープンサイエンスの推進に向けた国際的な議論や取組が行われている。こうした国際的なオープンサイエンスの推進の取組の中で、データ共有のための原則が検討され、国際的な認証機関が設立された。

我が国では、2018年6月に閣議決定された「統合イノベーション戦略」において、今日の「知の源泉」の鍵はデータ・情報であるとの認識の下、オープンサイエンスのためのデータ基盤の整備を行う方針を打ち出している。具体的に講ずる施策として、国立研究開発法人（国研）におけるデータポリシー策定の促進等とともに、研究データを格納するリポジトリの整備等のため、文部科学省が「クラウド上で共同利用できる研究データの管理・公開・検索を促進するシステム」を開発すること、また、内閣府が国際認証基準等を参考に「リポジトリの整備・運用のガイドライン」を策定し、大学・国研等に同ガイドラインの適用を推奨することを掲げている。

本ガイドラインでは、国内の研究コミュニティ（大学、国研等）が、国際的にも信頼され、研究データの保存・公開を適切に進めることができるデータリポジトリの整備・運用を行うための指針を示す。なお、本ガイドラインは、リポジトリの国際認証基準の動向や上記の我が国における研究データの管理・公開・検索を促進するシステムの開発の進展等に応じて、必要な改訂を行う。

2 研究データリポジトリの定義及び役割

本ガイドラインでいうデータリポジトリとは、電子的データの保存・共有等を行うための広い意味の情報基盤であり、主に運営体制、計算法基盤（狭義の情報基盤）、人的基盤からなる機関（ある目的を遂行する機能をもって設置された組織）である。本ガイドラインでは、大学・国研等の研究機関の内部の機関として設置されるデータリポジトリを主に想定しており、そのなかでとくに公的資金により得られた研究データを再利用できる形で保存・管理し、機関内または機関外へのデータ利用サービスに供するものを対象とする。（以

下「研究データリポジトリ」、または単に「リポジトリ」という。) なお、リポジトリは、大学・研究機関の成果を保存・管理・公開することを主眼とする「機関リポジトリ」、分野における研究資源を保存して活動促進に資する「分野リポジトリ」、分野や機関を限定しない「汎用リポジトリ」などの分類がある。

本ガイドラインの対象とする研究データリポジトリには、データ駆動型研究開発を促進し、新たな「知の源泉」を創出するため、以下の役割が求められる。

- 当該分野内で標準的または相互利用に適したデータ形式の利用、データ再利用を促進する付帯情報としてのメタデータを整備すること等により、研究データの相互運用性¹を確保し、研究データの共有（公開を含む。）を図ること
- データの品質を管理し、データへのアクセスを継続的に提供すること
- 国際的にも信頼が得られる形で、リポジトリの整備・運用を行うこと

3 FAIR 原則

データを共有するための基準となる国際的な原則として、「FAIR 原則 (FAIR Data Principles)」²がある。「FAIR」は、「Findable (見つけられる)」「Accessible (アクセスできる)」「Interoperable (相互運用できる)」「Re-usable (再利用できる)」の略で、データ公開の適切な実施方法を表現しており、本原則に準拠したデータを作成する機運が国際的に高まっている³。

「FAIR 原則」の要件はデータ自体の相互運用性を意識して定められているが、その性質上、信頼性の高いデータリポジトリに求められる要件とも極めて親和性が高い⁴。

4 信頼できる研究データリポジトリの整備・運用に関する要件

「FAIR 原則」及び国際的な認証基準 (Core Trust Seal⁵による「信頼できるデータリポジトリを認定するための中核的な統一要件」)を踏まえ、また、研究データ利活用協議会の小委員会である「ジャパン・データリポジトリ・ネットワーク」による中間的な検討結果を参考に、以下のとおり、信頼できる研究データリポジトリの整備・運用に関する要件を示す。

¹ ここでいう相互運用性は Interoperability の訳語であるが、通常想定される計算機やソフトウェアを相互に接続して運用・運転できることとは異なり、データが相互に通用する、データ同士をあわせて用いる場合に有効に利用されることを指す。

² 『データ共有の基準としての FAIR 原則』 <https://doi.org/10.18908/a.2018041901>

³ 欧州委員会が、欧州の研究データ公開のための情報基盤である European Open Science Cloud について、「FAIR 原則」をデータ共有と管理の基礎に位置付けているなど。

⁴ <https://ipres2017.jp/wp-content/uploads/keynote1.pdf>

⁵ <https://www.coretrustseal.org/>

4.1 運営体制

4.1.1 研究データリポジトリのミッション

研究データリポジトリには、研究データを適切に保存・管理し、提供する役割がある。研究データリポジトリには、保存・管理や継続的なデータアクセスの提供を可能とするようなミッションを定めて、公表することが求められる。研究データリポジトリの整備・運営に責任をもつ機関（以下「ホスト機関」という。）は、これに協力し、これを支援しなければならない。

4.1.2 運営組織

研究データリポジトリは、中長期的に（基本的に5年以上）安定して運用されることが求められる。このため、研究データリポジトリはその運営に関し、以下について配慮することが求められる。

- 永続的な組織によって運営されること
- 人的リソース、ITリソース、資金を十分に有すること
- スタッフが継続的な研修と専門的な能力開発を受けられること

また、研究データの急速な変化に対応するため、研究データリポジトリの運営組織及びその職員が専門家のガイダンスを受けられることが不可欠である。例えば、機関内部の専門家、外部諮問委員会、対象コミュニティから、研究データリポジトリの運用に対するフィードバックを受けられるようにする、国内外の相互運用性を低下させないため定期的なデータ整備・管理に関する国際会議への参加を促す等のしくみが必要である。

4.1.3 運営計画

研究データリポジトリは、そのミッションを達成するため、リポジトリの運営計画を策定し、その実施状況を把握し、必要に応じて改定を行わなければならない。ホスト機関はこれを支援し、協力する必要がある。また、データへの継続的なアクセスを保証する観点から、予期せぬ資金撤退といった場合のデータの移転やデータ提供者へのデータの返却についても、その際の対応を予め示すことが望ましい。

4.1.4 データポリシー

研究データリポジトリは、その整備・運営にあたり、基本的なデータポリシー⁶について

⁶ データポリシーに含まれる項目の例を以下に挙げる。

- リポジトリが受け入れるデータの範囲（分野・種類・品質・ファイルフォーマット・サイズ、メタデータ（形式・スキーマ）、データ提供者の要件）
- データの公開・非公開および猶予期間（エンバargo）
- データの帰属・知的財産の取り扱い
- 倫理規定
- データの品質管理（評価方法、責任）
- データの利用条件（ライセンス）

検討し、文書化しておくことが大切である。ホスト機関がデータポリシーを策定している場合は、それに準拠しつつ、必要に応じて、研究データリポジトリ単位で、より具体的なデータポリシーを定めることも有用である。権利や免責事項に関しては、法律の専門家の助言を得ることを推奨する。

4.2 情報基盤

4.2.1 研究データリポジトリの ICT インフラ（狭義の情報基盤）

研究データリポジトリには、データ預け入れ先としての信頼性と持続性の確保が必要である。そのためには、研究データリポジトリは自らのサービスレベルを検討し、サービスレベルに応じた計算機・ネットワーク等の情報基盤（以下「ICT インフラ」という。）を確保・運用する必要がある。研究データリポジトリの ICT インフラに関して検討すべき項目を以下に挙げる。

- 安定的な財源および安定的な運用形態の確保
- 研究データリポジトリのサービスレベルの定義
- 研究データリポジトリのサービスレベルに合わせた機密性、可用性、完全性の確保
- データ保存領域の拡張性および移行可能性の確保

4.2.2 データ管理システム

研究データリポジトリは、提供されたデータの受け入れ・保管、メタデータの作成・更新・保管及びデータの変更を含むデータキュレーションを実施するとともに、その管理等を行うためのデータ管理システムを確保し、運営する必要がある。

データ管理システムは、データの完全性を確保するための機能を備え、リポジトリのデータポリシーに合致したものでなければならない。

さらにデータ管理システムは、「FAIR 原則」を実現するために以下の機能を備えていることが望ましい。

- データ及びメタデータに関連付いた永続識別子の管理（「見つけられる(Findable)」及び「アクセスできる(Accessibility)」）

【永続識別子の例】

- ・ Digital Object Identifier (DOI)
- ・ ORCID
- ・ Global Research Identifier
- ・ Persistent Uniform Resource Locator (PURL)
- ・ Identifiers.org

● データの保存（保存期間、バックアップ方針、不変性検証）

● データの削除（データ削除の条件、判断基準）

その他、「国立研究開発法人におけるデータポリシー策定のためのガイドライン」も参照

- 可能な限りオープン⁷なフォーマットによるデータの管理(「相互運用できる(Interoperability)」及び「再利用できる(Re-usability)」)
- データ及びメタデータのバージョンの管理(「再利用できる(Re-usability)」)
- 分野標準のリッチなメタデータ形式、あるいは分野内でできるだけ広く用いられる形式によるメタデータの管理(「相互運用できる(Interoperability)」及び「再利用できる(Re-usability)」)
- 汎用的なメタデータ形式によるメタデータの管理(「見つけられる(Findability)」)
- データのライセンス情報の管理(「再利用できる(Re-usability)」)

4.2.3 データバックアップシステム

研究データリポジトリは、保管するデータの破損や紛失に備えて、研究データのバックアップの取得・保管、及びデータのリストアを実施するためのデータバックアップシステムの確保を行う必要がある。ここでいうシステムは、計算機システムによるストレージのバックアップのほか、制度・組織間スキームその他幅広くデータ保存に有益なしくみ全体を含む。

バックアップの方法や頻度は、リポジトリのデータポリシーに従う必要がある。また、災害等による障害を考慮し、遠隔地への複製機能を有することが望ましい。データバックアップシステムの構築に当たり検討すべき項目を以下に挙げる。

- バックアップ実行頻度
- 保持方式(フル/差分/増分)
- 保持リビジョン数
- バックアップおよびリカバリ手順
- 遠隔バックアップの有無

4.2.4 データ検索システム

研究データリポジトリは、保管するデータまたはそのメタデータに対してアクセス可能なデータ検索システムを確保し、運用しなければならない。

データ検索システムは、「FAIR 原則」を実現するために以下の機能を備えていることが望ましい。

- データ及びメタデータに関連付いた永続識別子の表示
- 機械判読可能なメタデータおよび永続識別子の提供
- データアクセス手法の表示
- データのライセンスおよび引用方法の表示
- データおよびメタデータの版または変更履歴の表示

⁷ <https://opendefinition.org/od/2.1/ja/>

4.3 人的基盤

4.3.1 研究データリポジトリにおける業務

研究データリポジトリには安定した運営組織とともにそれを担う人材が必要である。運営組織及び運営を担う職員は、以下の3つを円滑に行うことが求められる。

- リポジトリのミッションに則した運営・企画やポリシーの作成等を行う「管理的業務」
- 研究データの受け入れ、データの品質管理、メタデータ管理、データ提供やキュレーションを実際に行う「コンテンツ運用業務」
- 研究データリポジトリの計算機システムやストレージなど、計算機基盤の構築や運用を行う「計算機システム運用業務」

上記3業務の主なものとして、以下が挙げられる。

- 管理的業務： リポジトリの企画・運営、ポリシーの作成、広報・コミュニケーション活動（データ提供者と利用者との綿密なコミュニケーション）など
- コンテンツ運用業務： データ・メタデータの受け入れ・管理、データのキュレーション、データの品質基準の作成など
- 計算機システム運用業務： サーバーの保守や障害復旧、計算機システムのセキュリティ確保、設備の管理など

現状のリポジトリの運営においては、組織の規模や形態により、人員配置のあり方が異なっている。複数の業務を一人の職員が兼務している場合、関連業務を複数の機関・部門・部署で分担している場合、業務の一部を外注やクラウド化している場合がある。リポジトリを設置している法人において、リポジトリ運営業務の位置付けが明確でなく、その重要性がその法人内や関連機関内で十分認識されてないことが、リポジトリ運営の不安定要因の一つとなっている。

リポジトリを設置している機関は、リポジトリの運用業務に多様な専門性を求められることを認識し、必要な人員配置、人員の専門的育成、それらにかかる財源等を措置することが求められる。

4.3.2 職員の配置・育成⁸

⁸ データ専門職（データライブラリアン、データキュレーター）の現状

平成28年に公表された科学技術・学術審議会 学術分科会 学術情報委員会による審議のまとめ (http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu4/036/houkoku/1368803.htm)では、大学等による人材の育成及び確保のための取り組みとして、

- 技術職員、URA及び図書館職員を中心としたデータ管理体制を構築する
- データキュレーター等を育成するシステムを検討し推進する
- 戦略的な人的資源配分を検討し、新たな専門人材の確保に努める

の3点が挙げられている。データの専門職としてデータライブラリアン、データキュレーター等の名称が現れてきており、各大学、研究機関でデータ管理、データ公開を担ってきた人材（技術職員、リサーチアシスタント、大学図書館員等）を認知し、同名称のもと再定義する動きもある。

研究データリポジトリの運用にかかる職員の配置にあたり、一人または少数の職員が兼務して運用業務を担う場合もあるが、他部門の職員の活用も含め複数の職員に分担して運用業務を担わせることを考慮すべきである。このため、運営組織の職員に加え、研究データリポジトリの業務に関連する他部門の職員（技術職員、図書館職員、リサーチアドミニストレータなど）に対し、必要な情報と研修の機会を提供する必要がある。

業務に携わる職員が幅広い関連知識を習得できるよう、研究データ管理（Research Data Management）の基本知識を学べるオンライン教材の活用や受講、またこうした研究データ管理に関する国際会議参加を推奨するなど、職員に対して適切な研修の機会を提供する必要がある。なお、研究データ管理に関する国内教材⁹として、2017年度にオープンアクセスリポジトリ推進協会（JPCOAR）及び国立情報学研究所（NII）によって開発・公開された、以下のオンライン教材（MOOC）がある。

オープンサイエンス時代の研究データ管理

<https://www.nii.ac.jp/service/jmooc/rdm/>

4.3.3 研究データの保存・管理・利活用を支援する人材の配置

研究データが有効に利活用されるためには、研究データリポジトリの適切な運営に加え、研究者が保存したい研究データについて、その保存についての基準に照らした審査、メタデータおよび永続識別子の付与作業の実施または支援、データ論文の執筆時支援、当該データを利用した論文執筆時のデータ保存・公開要請への対応支援、データ利用のライセンス策定支援、データやメタデータの品質評価、データ保存の決定と公開データカタログへの登録、などを行う専門的知見をもち訓練された人材を配置するとともに、保存されたデータを利用する研究者に対してデータ分析支援を行うデータサイエンティストを配置する必要がある。

なお、これらのデータ利活用支援を担当する人材が必要とする知識や情報を共有できるコミュニティを構築することで、リポジトリの相互運用性の確保を促進し、データの利活用を促進することができる。

（参考）研究データリポジトリに関する国内の状況及び国際動向

参考情報として、研究データリポジトリの整備・運用に関する国内の状況及び国際動向について記載予定。

⁹ https://www8.cao.go.jp/cstp/tougosenryaku/tougo_honbun.pdf