研究データ管理

Carly Strasser 著，機関リポジトリ推進委員会訳

2016 年 3 月発行

Translation of: Carly Strasser. (c2015). ‘Research data Management’. A Primer Publication of the National Information Standards Organization. Baltimore: National Information Standards Organization (NISO). Available online: http://www.niso.org/apps/group_public/download.php/15375/PrimerRDM-2015-0727.pdf

目次

はじめに 4
データ管理計画の策定 5
データ管理計画（DMP） 5
データについての説明 5
標準 6
方針と手順 6
記録と保存 6
必要な資源 7
データ管理計画策定のための最良事例 7
命名法の策定 7
スプレッドシートの作成 7
メタデータ収集計画の策定 8
バックアップ戦略の策定 8
生きた文書としてのデータ管理計画（DMP） 9
データ管理計画策定のための要員 9
研究データの文書化 10
メタデータ 10
インフォーマルなメタデータ 10
フォーマルな（標準的な）メタデータ 10
ソフトウェアの文書化 11
ワークフローの文書化 11
インフォーマルなワークフロー 11
フォーマルなワークフロー 12
ワークフローソフトウェア 12
コンピュータエミュレーション 13
管理 14
データセットのガバナンスと利用契約 14
著作権と所有権 14
独自の権利（*sui generis rights*） 14
データに関する権利に対応する法的仕組み 14
慎重に扱うべきデータ 15
共有のための最良事例 15

データの保管、バックアップ、セキュリティ 16
バックアップ 16
バージョン管理 16
クラウドストレージ 17
保存 18
保存の最良事例 18
リポジトリ 18
学問分野を限定したリポジトリ 18
汎用リポジトリ 19
リポジトリの選定 19
リポジトリソフトウェア 19
利用と再利用 21
データセットの識別とリンク 21
識別子 21
データの引用に関連するもっと複雑な問題 21
引用 22
データの発表（publishing） 22
利用可能 22
引用可能 23
妥当性 23
データの発表モデル 23
功績（credit）とインセンティブ 24
現在の仕組み 24
オルトメトリクス 24
データに対する功績 25
結論 26
附録 A: 資料 27

はじめに

研究の実施方法はこの 20 年で劇的に変化しました。新たな手法とツール（ソフトウェア、ハードウェア、器械、機器）、新しい情報源、インターネットを通じた世界規模の研究のつながりの拡大によって、世界中の研究者が前例のないスピードで進歩しています。しかし、このパラダイムシフトに伴って、重要な課題がいくつも持ち上がっています。そのもっとも顕著なものが、研究の再現性と手法およびワークフローの透明性です。

21 世紀の研究が持つこうした課題に対処するには、健全な研究データ管理が必要です。入念な計画、文書化、データの保存を行うことで、再現性と透明性を有する研究データを得るという目的の達成が、はるかに容易になります。さらに、データを適正に管理することで利用と再利用が容易になり、研究者にとっては研究者間の連携の進展に、研究資金提供者にとっては投資利益の最大化につながります。本手引き書は、研究者と研究者を支援する人々によるデータ管理の向上に役立つことを目的に、研究データ管理の基礎について取り上げます。

データ管理計画の策定

計画策定に関する多くの格言は、もちろん研究データ管理にも当てはまります。計画策定は、データセットの長期保存と有用性を確保するためのもっとも重要なステップです。研究者は最初のデータ点を収集する前に、時間をかけて入念に検討する必要があります。

・データをどのように文書化するのか？　どのようなメタデータを使用するのか？　データ、ファイル、サンプル等の命名はどのように行うのか？
・データを効率的に管理するには、どのようなスタッフやソフトウェア、ハードウェアが必要か？
・プロジェクトの期間中、データ管理の優先的実施を維持する責任を担うのは誰か？
・データの最終的な保管先はどこか？　どのような人々がアクセスするのか？　データの利用と再利用にかかわるのはどのようなポリシーか？　どの資源を利用するのか？
・データが有用性を失うのはいつか？　いつ破棄すればよいか？

データ管理計画（DMP）

多くの資金提供者が、研究データプロジェクト開始前の計画策定を不可欠なものだと認識するようになっており、データ管理計画（DMP）は申請プロセスの一環としてますます必要とされるようになっています。ほとんどの DMP は、プロジェクトデータ管理の主要部に対応するため、次の 5 つの基本要素を有しています。

1.プロジェクトの期間中に収集または生成されるデータの形式についての説明
2.それらのデータと関連メタデータに利用される標準
3.収集または生成されるデータに関するポリシーについての説明
4.生成したデータの記録と保存に関する計画
5.スタッフ、ハードウェア、ソフトウェア、予算要件など、データ管理の実現に必要となる資源についての説明

データについての説明

収集するデータについての説明は、見た目以上に複雑な意味合いを有する場合があります。こうした複雑さが予測される原因は、「データ」の定義そのものに議論の余地があるという点にあります。資金提供者の中には、DMP を念頭に、データとみなすべき見込まれる研究成果をはっきりと列挙する者もいます。成果として、コードやスクリプト、画像やソフトウェア、動画やモデルなどが挙げられる場合があります。これらは従来「データ」として定義されていたものではありませんが、あらゆる研究成果をデータとみなすことにより、歴史的にデータという言葉を用いてこなかった人文科学などの学問分野で、資金提供者向けに DMP を完成させる最良の方法について理解しやすくなるでしょう。また、従来のデータセットに対立するものとして方法論を生み出す理論家などのグループが、プロジェクト完了時にどの成果の共有を検討すればよいのかわかりやすくなるでしょ

う。

一般的に、データの説明には、研究者が何を収集するつもりなのか、どのように収集するのか、ということのほかに、収集したデータの加工・分析計画も含まれなければなりません。データの説明を、申請書の「方法」セクションの付録と考えるべきではありません。むしろ、プロジェクト開始前に、データに関して、適切な収集、十分な文書化、入念な考察を確保するためにどのような活動を行うのかを重点的に取り上げるべきです。詳細については後述の*研究データの文書化*をご覧ください。

標準

プロジェクト開始時に（データおよびメタデータの）標準を特定することは、研究の間、データを適切に管理するのに大いに役立ち、プロジェクト終了後のデータの共有を容易化すると考えられます。利用できるメタデータ標準は多数ありますが、その中から研究者が１つ選ぶのは、データ管理に関する彼らの潜在的な知識が乏しいことを考えると、非常に困難だと思われます。あるデータセットに関して、適切なメタデータ標準を特定する最良の方法は、プロジェクト終了時にデータの保存先として利用するリポジトリを確定することです。こうしたリポジトリには、提出の条件として所定または規定のデータ標準やメタデータ標準がある場合もあります。ほかに、データ標準とメタデータ標準の両方を特定する有益な方法として、同様のデータセットを扱っている同僚に相談してやり方を真似るというものもあります。詳しくは後述の*メタデータ*をご覧ください。

方針と手順

適切なデータ収集計画の策定には、収集するデータを管理する何らかの規定や方針が存在するのかどうかを明らかにする必要があります。それらについては、どのようにして第三者がデータにアクセスするのか、第三者のアクセスを認めるまでに公開禁止期間を設けるのか、資金提供者や組織の方針によりデータを共有する（または共有しない）義務がすでにあるのかについての計画とともに、DMP に記載することが望まれます。慎重に扱うべきデータ（被験者や絶滅危惧種に関するものなど）は、連邦法や州法の適用対象になる場合があり、施設内倫理委員会の承認が必要となることが考えられます。データに関する知的財産権というテーマについては、米国著作権法をどの程度データセットに適用するかについてと同様、今なお議論が行われています。しかし一般的には、データに適用する予定の使用許諾や適用免除の種類について、研究者が説明しなければなりません。

記録と保存

DMP には、データをどこで長期的に保管するつもりなのかについても、情報を盛り込まなければなりません。その情報として、データの長期的な利用性とアクセス性を確保するための保存戦略を実施している 1 つ以上のリポジトリが考えられます。データの最終的な保存先を明確化すること

で、研究者はファイルフォーマット、データを寄託する前に必要となりうる変換、メタデータの標準とコンテンツなど、データ管理の他の側面について計画を立てやすくなります。またDMPでは、特定のリポジトリでのデータの保存に対して責任を負うプロジェクトスタッフを明確にすることも必要です。

必要な資源

優れたデータ管理は高コストな取り組みになる可能性があるため、DMP では計画を適切に実施するために必要となる資源について明確にしなければなりません。そうした資源には、スタッフの作業時間、ハードウェア、ソフトウェア、データの整理・管理・バックアップ・保護のための保管にかかる費用などが考えられます。こうした費用は小さなものではないため、研究者は施設サービス提供者（図書館、IT 部門）などに意見や情報を求め、データ管理活動のために適切な予算を計上するようにしなければなりません。

データ管理計画策定のための最良事例

データ管理計画の策定は、資金提供者が求める最低限の DMP の作成に限定するべきではありません。むしろデータが、プロジェクトの間、研究責任者にとって、また将来のいかなるデータ利用者にとっても、使いやすいものであるために、データライフサイクルの不可欠な一部として扱わなければなりません。

命名法の策定

多くの研究者が、効率と使いやすさを最大限に引き出すための組織的な仕組みを構築せずに、データの収集とファイルの蓄積を始めます。サンプルやデータ（物理もしくはデジタル）の命名法の策定に時間をかけることで、名称が重複している、見分けにくい、といった問題や、将来の名称変更や分類に関する作業で生じる問題を避けることができます。説明的で重複せず、コンテンツやサンプルを表すように名称を付けなければなりません。たとえば、ある土壌サンプルを sample 1 と命名するのは、sample 1 と命名された水のサンプルがある場合、有益ではありません。Soil01_SiteB_2014 や Water01_SiteB_2014 の方が選択肢として適しているでしょう。

デジタルファイルに関する組織的な枠組みを決定する際、将来の依存性（スクリプトに特定のデータファイルが必要になるだろうか？）やファイルフォーマット（.csvファイルはすべてグループ化すべきか？）、分析順序に関して検討する必要があります。フォルダとタグを使用すると、ファイルを論理的に整理するのに役立ちます。

スプレッドシートの作成

学問分野を問わず、多くの研究者がデータの整理や視覚化にスプレッドシートを利用します。スプレッドシートを利用しすぎて、データの文書化やデータセットの来歴に支障が出ることもあります。

スプレッドシートをユーザーフレンドリーでデータ操作にとって好ましいものにしている特徴が、成果につなげるために経た手順の取り消しをほぼ不可能にすることもあるのです。

理想としては、そうした操作と分析は来歴を保証するスクリプトを用いて実施し、ワークフロー全体の文書化を確保することが望まれます。しかし、研究者が近い将来スプレッドシートの利用をやめることは考えにくいでしょう。それを踏まえて、スプレッドシート関連の最良事例で、データの利用を長期的に可能にするのに役立つものを、いくつか以下に挙げます。

1.別のタブに生データを（未加工かつ未分析の状態で）残しておく。あらゆる操作は、別のタブ上で、もしくは別のファイル内で行い、元データが失われないようにすべきである。

2.スプレッドシート内の情報を「細分化」する。つまり、情報を分割して、どのセルにも１種類のデータのみ入力するということである。たとえば、１つのセルに「Austin, TX（テキサス州オースティン）」と場所を入力するのではなく、都市（Austin）と州（TX）を２つのセルに分けて入力すべきである。

3.スプレッドシートを連続していない単独のテーブルに分割することを検討する。収集場所に関するあらゆる情報（用地名、用地の住所番号、緯度、経度、都市、郡、州、国）を直接そのデータテーブルに保存するのではなく、こうした情報を収集場所の説明のみ目的とする別のテーブルに格納する。そうすることで、主データテーブルには収集場所の番号を記すだけで足り、ユーザーはその用地に関するテーブルから詳細情報を参照することができる。こうしたやり方はリレーショナルデータベースに似ている。

メタデータ収集計画の策定

メタデータ標準を決定するだけでは十分ではありません。研究者はデータを収集する前に、データの説明方法を厳密に検討しなければなりません。学問分野によっては、「データディクショナリ」という概念を用いるところもあり、それには意味、他のデータとの関連性、出所、用途、フォーマットなど、データに関する情報が含まれます。他の情報としては、ヌル値をどのように識別するのか、サンプルにどのような命名規則を用いるのか、ある器械の出力の有効桁数は何桁にすべきか、どの単位を用いるべきか、などが考えられます。（デジタルまたはアナログの）データ収集シートを利用する場合、データセットのために選定したメタデータスキーマに準拠するよう、データの収集前にこうしたことを策定しなければなりません。

バックアップ戦略の策定

研究者は、データのバックアップ方法について、明確で詳細な計画を確実に備えておく必要があります。検討事項としては、必要なハードウェアとソフトウェア、バックアップの頻度、こうしたバックアップの実施責任者、バックアップをどこにどれだけの期間保管するのか、データのメインインスタンスが失われた場合にどのような措置を講じるべきか、ということなどがあります。バックアップが自動的に実施される場合、こうしたバックアップが実際に行われており、必要な情報を正確かつ着実に保存していることを検証・確認するための計画を整備しなければなりません。

生きた文書としてのデータ管理計画（DMP）

適切な研究データ管理計画の策定は、プロジェクトが開始すれば終了するというわけではありません。むしろ、研究者は計画を頻繁に見直して、その実施と整合性を確保しなければなりません。リマインダーを設定して、毎週 DMP を見直し、必要に応じて更新するのが最良の慣行です。分析、戦略、データ収集はプロジェクトの間に大幅に変化する場合があります。こうした変化について、DMP を通じて記録に残し、検討する必要があります。

データ管理計画策定のための要員

研究データ管理計画の策定責任は、研究者だけが担うのではありません。資金提供者だけでなく、研究者を支援する機関や組織も、データ管理に重要な役割を果たします。補助金コーディネーターとプロジェクトオフィスの出資者は、補助金と並んで、提出された DMP にも細心の注意を払い、研究者の計画について、資源を当該計画に注ぐことになる他の組織と共有すべきです。こうした組織には、図書館、機関リポジトリ、情報技術オフィス、研究支援サービス、データ保護オフィス、施設内倫理委員会、弁護士などが考えられます。

機関は研究の基本的インフラ（実験室、インターネット接続、図書館へのアクセス）を提供しなければならないのと同様に、研究ライフサイクルを通して、適切なデータ管理の準備を整えなければいけません。それには、データ管理に関して研究者のサポートを行うための訓練を受けたスタッフ、プロジェクトの間利用できるデータの保管とバックアップのためのインフラ、長期的に研究データの保存と利用を確保する機関リポジトリの利用などが考えられます。

研究データの文書化

研究データは、長期的に有用であるために、識別のための特定のガイドラインに従わなければなりません。

メタデータ

文書化が不十分な研究成果（データやコードなど）は、ラベルを剥がされた缶詰製品のようなものです。内容は望ましいものなのかもしれませんが、メタデータがなければそれを伝えることはできません。高品質のメタデータは、データそのものと同じくらい、効果的なデータの共有に不可欠です。なぜなら、メタデータの品質が高ければ高いほど、データセットが再利用されやすくなるからです。

インフォーマルなメタデータ

研究データ管理の最良事例に詳しくない研究者は、インフォーマルなメタデータに関して、彼ら自身がそれをそのように呼ぶことはないかもしれませんが、詳しいのが普通です。一般的に、インフォーマルなメタデータとは、実験ノートやフィールドノート、それに「readme」ファイルに含まれる情報のことです。研究を実施するにつれて、主任研究者とその同僚がデータとワークフローを理解し、意思決定を文書化し、プロジェクトの進捗状況について注釈をつけるのに役立つ記録が蓄積されます。こうしたインフォーマルなメタデータは、個人的に構築された細かな差異のあるものであるとはいえ、研究成果を理解するのに不可欠といえるでしょう。こうした情報は、可能な場合は「最初からデジタル形式で作成（born digital）」すべきですが、それが可能でない場合でもデジタル化すべきです。そうすることで、こうした重要な情報を最初からデジタル形式で作成されるデータセットに格納するのに役立ちます。インフォーマルなメタデータは、相互運用性の確保とデータ発見という点では理想的ではありませんが、可能な場合は関連する研究成果と一緒に記録すべきです。

フォーマルな（標準的な）メタデータ

研究成果に関するメタデータのもう 1 つの形態は、標準化されたものです。標準的なメタデータは、標準化フォーマットに準拠した研究成果についての情報を含み、統制語彙を有し、関係するコミュニティに受け入れられ、利用されています。利用可能なメタデータ標準は多数存在します。特定のプロジェクトに関してもっとも適切なメタデータ標準は、主にそのデータに関係する学問分野と、そのデータセットの長期保存を考えるリポジトリによって変わります。学問分野に特定のメタデータスキーマとしては、Ecological Metadata Language（生態学と環境学で利用）や FGDC 19115（地理空間データで利用）などの例があります。標準化されたメタデータの主なメリットは、類似するデータセットとの相互運用性が確保される点と、データの容易な発見を可能にする点です。発見可能性は、メタデータ標準という基礎的な符号化によって、メタデータが機械可読になることで高まります。

標準的なメタデータの生成はプロジェクトの間に行うべきですが、残念ながら通常は、保存先として選んだデータリポジトリが、データの寄託に関して特定のスキーマを要求する場合にのみ考慮されます。望ましいのは、プロジェクト開始時に 1 つの標準を決定し、その標準に準拠するのに必要なメタデータ要素を収集の時点で文書化することです。研究者はプロジェクトの早い段階で、選んだリポジトリに連絡を取り、助言や提案を求めるべきです。リポジトリの中には、標準的なメタデータの生成に利用できるソフトウェアツールやアプリケーションを無料で提供しているところもあり、研究者はプロジェクトの間を通してメタデータの作成が容易になります。

ソフトウェアの文書化

従来の定義のように、一揃いのデータだけが研究成果になるわけではありません。別の共通の成果として、ソフトウェアがあります。これは単に「コード」と呼ばれることもあります。研究がますます計算集約的になるに従って、コードは多くの場合、データセットを整理整頓し、分析し、グラフや表などの最終生産物を生み出すために生成されます。こうした最終的な研究成果は、データの要約や再加工されたデータの形態を取るため、こうした最終形態から元の生データまでたどるのはほぼ不可能です。以前は、研究者も一般の人々も、発表された結果は、元々収集されたデータがありのまま正確に提示されていると信じていました。しかし、研究の撤回が増えて世論が懐疑的になるにつれて、研究業界は発表する研究結果について、これまで以上にしっかりした来歴を示すことが必要になっています。

プロジェクトの完全な再現性を可能にするには、こうしたタスクを完了するために用いたコードも、データと一緒に保存しなければなりません。そうしたコードの文書化は、それを第三者に理解可能なものにするために不可欠です。コードの文書化には、コードに注釈を付けたり、readme ファイルを提供したりすることで、コードスクリプト、データセット、およびアウトプットの間の関連性を説明することなども含まれると考えられます。

ワークフローの文書化

ワークフローは再利用と再現性に役立ちうるもう１つの研究成果です。メタデータ同様、ワークフローにもインフォーマルなものとフォーマルなものがあると考えられますが、どちらも研究プロジェクトとその成果を理解するのに不可欠なものだといえます。

インフォーマルなワークフロー

インフォーマルなワークフローは単純なフローチャートの場合もあり、情報（データ）の生成または収集の時点から、最終的に作り出される成果（グラフ、出版物）に至る経路について説明します。そうしたワークフローは手書き、コードを使った文書化、readme ファイルの形式を取ることもあれば、研究者によって、研究から導き出された結論を第三者が理解しやすくなるような方法で文書化

されることもあります。

フォーマルなワークフロー

インフォーマルなワークフローとは対照的に、フォーマルなワークフローはソフトウェアによって作成され、第三者による再利用を目的としています。ワークフローを文書化するソフトウェアシステムには評判のよいものがいくつかあり、研究者にとって再現性の重要度が高まるにつれて、その数は増加し続けています。こうしたシステムは複雑で、特定の学問分野に共通するワークフローに特有のものである場合があるため、一般的には利用されていません。フォーマルなワークフローで学問的に利用されているものに Taverna があり、遺伝学とゲノミクスの世界ではこのワークフローソフトウェアを利用して、共通の計算タスクを自動化しています。フォーマルなワークフローシステムの利用は普及してはいませんが、その利用は、研究が最初からデジタル形式で生成され、計算集約的な方向に向かい続けるため、今後数年間で増加すると考えられます。

ワークフローソフトウェア

学術研究の重要な成果としてのデータの認識が高まり、データの管理と共有の支援に関心を持つ資金提供者だけでなく、データのためのアプリケーションの構築やサービスの提供を行う人々の関心も引いてきました。それは、研究者と研究者を支える団体（図書館、データセンター、出版社）の双方に向けて、データを念頭に置いて作成されてきた多数の新しいツールを見れば明らかです。

ほとんどの学問分野にとって、データの収集・分析・発表というプロセスには、多様なコンピュータプログラムが必要であり、そうしたプログラムはそれぞれが異なるファイルフォーマットを要し、そのいずれもが、収集から結果の発表まで、データがたどる経路（すなわち、データの来歴）を複雑なものにする可能性があります。こうした複雑なワークフローは、多くの研究者が、とりわけ独自のワークフローを適切に文書化しないことから、再現性と再利用の妨げになります。

研究者が研究の際に利用するソフトウェアツールの数が増加傾向にある一方、コンピュータ科学者が多数の入出力の捕捉に役立つソフトウェアを制作しています。ワークフローソフトウェアは、10 年以上の間にあちこちで使われるようになりましたが、制作されたシステムの多くが、平均的な研究者にとって今なお複雑すぎるため、効果的に活用できていません。研究者にとって説明責任と再現性を示す重要性が高まるにつれて、この状況は変化すると考えられます。

ワークフローを捕捉するソフトウェアは非常に多様で、多数のアプリケーションが特定の学問分野を念頭に作成されています。多くの場合、それらは電子実験ノートの形態を取り、研究者がプロジェクトの進行に合わせて、コンピュータ上に記録を書き留めるのに役立つよう意図されています。iPhython はその 1 例です。ほかには実行環境があり、研究者はフォーマルなワークフローを設計

し、実行することができます。例として Taverna と Kepler があります。これらのシステムは有望ですが、LaTeX や Git、R などのシステムへの統合を重視しているものが多いため、コーディングに慣れていない研究者は慎重に利用しなければなりません。

コンピュータエミュレーション

複雑なソフトウェアの連携に対処する 1 つの方法として、仮想マシンをエミュレータとして利用し、元のデータセットが加工された環境を作るというものがあります。

管理

データセットのガバナンスと利用契約

研究データを共有するということは、将来、第三者がそのデータを検証、ダウンロード、利用のいずれか、もしくは複数を行う可能性があるということです。データを利用・再利用可能なものにするためには、そうした行為を可能にする適切な使用許諾もしくは権利放棄が必要です。データについての規定や規則に対する研究業界の規範、すなわちガバナンスは、今なお現在開発途上であり、データ特有の多数の要因から複雑なものになっています。

著作権と所有権

データとは事実であり、それはつまりデータは著作権に関連する規制や保護の対象にならないということです。しかし、著作権法や知的財産法は、データ、コード、およびその他の研究成果の集合体には適用されます。たとえば、実験ノートの個々のデータ点は著作権法の対象になりませんが、実験ノートそのものは著作権法によって保護されます。

大学の学術研究者は、データを含む彼らの研究成果の所有権を大学が主張するのを目にすることがよくあります。一般的に研究者は、大学で研究を始める際にこうした取り決めに合意しているのです。しかし、この点については広く知られても理解されてもいないため、多くの研究者は彼らが収集したデータの集合体、実験ノート、その他の研究成果の所有権は自分たちにあると考えます。こうした誤解は、大学側がデータの集合体に対するその権利をめったに行使しないことから深刻化しますが、こうした状況は新たな権限が実施されるにつれて変化すると思われます。

独自の権利（*sui generis rights*）

データセットに適用されうる他の一連の権利は、*sui generis* として知られています。このラテン語は、「その特性において他に類がない」という意味です。*独自の*権利とは、データベースの創作性や独創性に関係なく、そのデータベースの抽出や再利用を禁じる知的財産権法です。著作権法が重視するのは創作性と独創性ですが、*独自の*権利はそうした区別をしません。欧州連合は、データベースの作成後 15 年間、*独自の*権利を認めていますが、米国はこれまでのところそうした法律を制定していません。こうした権利は、国際的なパートナーが関与するプロジェクトにとって重要になるでしょう。

データに関する権利に対応する法的仕組み

第三者が法的措置を心配せずにデータを利用できるようにするために、利用の許容範囲について理解しなければなりません。データセットの所有者は 3 つの仕組みのうちの 1 つを利用することで、それを実現することができます。

その最初の 1 つは契約であり、データアクセスポリシー、データ使用ポリシー、データ利用契約などと呼ばれる場合もあります。契約は完全にカスタマイズ可能であるため、データがどのように、そして誰に利用される可能性があるのかに関心を持つ研究者や機関にとって、魅力的な選択肢になります。しかし、契約はデータの再利用を困難にするため、その結果としてデータセットの長期的な価値を引き下げる恐れがあります。2つ目の仕組みは使用許諾であり、その条項は普遍的であるため、データの再利用が容易になります。3 つ目の仕組みは権利放棄です。データセットに対する権利が放棄された場合、そのデータは公知とみなされ、利用が自由になります。権利放棄はデータセットの有用性を最大化しますが、著作権の帰属表示や著作者表示が相対的に困難になる可能性があります。

慎重に扱うべきデータ

その他の複雑な問題に、慎重に扱うべきデータをめぐる問題があり、こうしたデータには、被験者や絶滅危惧種、保護地域に関する情報などが入ると考えられます。こうしたデータセットは、場合によっては、情報の広範囲におよぶ共有を防ぐことを目的とする法律や規則の適用対象となり、著作権にかかわらず、研究者がデータセットを共有する能力を制限することもあります。その結果、オープンデータ、オープンサイエンス、オープンリサーチに向かう現在の動きと、プライバシーとをめぐって緊張が生じています。使用許諾と権利放棄ではプライバシーと機密保護に対応できないことに注意することが重要です。こうした類の問題が重大な影響を持ちうる場合、契約を利用しなければなりません。

共有のための最良事例

（慎重に扱う必要のない）データの再利用を保証するためのデータ共有に関する最良事例は、そのデータセットを、再利用を妨げるいかなる規定や制限事項もないパブリックドメインに公開することです。データをパブリックドメインに置くことで、第三者がその利用に関する規定を気にすることなく、データを随時ダウンロード、統合、リファクタリングすることが可能になります。それは一般に、クリエイティブ・コモンズ・ゼロ（CC0）よる権利放棄を利用することで実現します。CC0 は当該データがパブリックドメインにあることを明示します。

権利放棄（および使用許諾）については、著者への連絡を求める条件やデータの利用方法に関する制限を追加せず、機械可読でなければなりません。研究者や機関は独自に作成した利用契約の使用を好む場合もありますが、それらは機械可読ではなく、データの利用に対する制限やデータセット作成者への連絡を条件にしています。こうした独自の規約は、データの再利用を大幅に妨げるため、避けなければなりません。そうした規約の代わりに、研究者はクリエイティブ・コモンズやオープン・データ・コモンズなど、一般的に利用されているライセンススキーマを選ぶべきです。

データをパブリックドメインに置くことで、著作権の帰属表示をめぐる文化規範の問題を避けられる

わけではないことに注意してください。データの引用が今後さらに一般化するにつれて、研究者は従来の学術出版物に対して功績（credit）を認められるのと同じように、データセットについても功績を認められるようになるでしょう。

データの保管、バックアップ、セキュリティ

データの保管とバックアップは、デジタルデータを伴ういかなる研究プロジェクトにとっても必要な問題です。しかし、この一見ごくありふれた課題は、研究者の日常業務において顧みられないことがあまりにも多いのです。

バックアップ

完全なデータセット、関連するコード、ワークフローに関して、最低でもオリジナル（original）、ニア（near）、ファー（far）の 3 つのコピーがなければなりません。最初のコピー、「オリジナル（元）」は作業用のデータセットと関連ファイルです。このオリジナルのコピーは通常、研究者が主に使用するコンピュータに保管されます。2 つ目のコピーはオリジナルの「ニア（近く）」になければなりませんが、物理的な場所が同じでないことが理想です。このコピーは自動バックアップソフトか手動のどちらかで毎日更新されます。多くの場合、このニア・コピーは外部ハードドライブや研究者の属する施設内の共有ファイルサーバに保管されます。3 つ目のデータのコピーは、オリジナル・コピーからもニア・コピーからも物理的な場所が「ファー（遠く）」でなければなりません。同じ建物内に保管すべきではありませんし、間違っても同じ室内で保管してはいけません。ファー・コピーは災害の脅威が異なる場所に置くのが理想的です。このファー・コピーの形態として考えられるものに、自動的にバックアップが行われ、システム内にデータの複数のコピーが保管されるクラウドベースのバックアップシステムの形態があります。

バージョン管理

データのコピーを複数保管することで生じる主な課題の 1 つは、バージョンの管理です。非常に多くの場合、研究者は独自の方法を利用して、データセットを始めとするファイルのさまざまなバージョンを識別しています。彼らが目を向けるべきなのはソフトウェア業界です。長年、バージョン管理の問題に取り組んでおり、プロジェクトの自然な流れを記録するのに役立つ、適切に設計されたバージョン管理システムを導き出しています。こうしたシステムを利用して、ファイルに加えられた変更をバージョン番号、タイムスタンプ、変更箇所の説明とともに記録します。こうして行った変更は容易に比較することができ、必要に応じて復元することも可能です。バージョン管理システムの一般的な例に、Subversion、Git、Mercurial があります。バージョン管理システムは、コンピュータサイエンスの知識のない平均的な研究者にとって、必ずしも使いやすいものではありませんが、GitHub や Bitbucket などの新しいウェブツールは、初心者にも使いやすい強力なツールです。

クラウドストレージ

データ保管の選択肢としてますます人気が高まっているのは、Dropbox や Google Drive などのクラウドベースサービスの利用です。これらは多くの研究者にとってすばらしい解決策です。しかし、慎重に扱うべきデータを取り扱っている場合、こうした大衆向けの選択肢には慎重になるべきです。データのセキュリティについては必ずしも保証されていないため、パスワードのログのハッキングなど、セキュリティ侵害は現実的な脅威です。データセットに慎重な扱いの必要なデータ（被験者の個人情報、追跡中の絶滅危惧種、遺産地域の地理的情報など）が含まれる場合、無料のクラウドベースシステムではなく、高度なセキュリティを有するシステムに保存すべきです。こうした種類のデータを扱う研究者は、所属する機関や組織の情報技術担当グループと連携して、データのコピーがすべて間違いなく安全に保管されるようにしなければなりません。

保存

研究データの適切な管理の最終段階は、データの長期的な保存です。データの保存はデータの保管と同じではありません。これら 2 つを区別する要素は、データの保存が、当該データがある期間に渡って正確に提供されるようにすることを目的としている点です。それには、当該データの継続した有用性を可能にする戦略と方針のほか、複数のコピーを管理し、必要に応じて新しいメディアにコンテンツを移行させるなどの保存方法が必要です。研究データの保存は、長期的キュレーションを目的とする信頼できるリポジトリにデータを格納することで可能になります。

保存の最良事例

プロジェクトの終了時にすぐ保存できる状態にしておくために、研究者に実施できる簡単なステップがいくつかあります。それを以下に挙げます。

・研究成果（データなど）のフォーマットを選択する。標準的で一般に使用されていて、独占所有ではなくオープンソースで、バイナリではなくテキストベースのもの。

・データに独自の識別子が付与されるようにする。それによって当該データが引用可能で功績をもたらすものになる（詳しくは*利用と再利用*のセクションを参照）。

・高品質で機械可読なメタデータを作成する（前述の*研究データの文書化*を参照）。

・データが許可条項により適切に使用許諾されることを確認する（前述の*データセットのガバナンス*のセクションを参照）。

リポジトリ

データ管理計画の立案には、データセットの最終的かつ長期的な格納場所、すなわちリポジトリを選定することも必要です。リポジトリを選定する際、研究者は以下の点について考慮する必要があります。

・類似するデータセットはどこに保存されているのか？

・検討しているリポジトリのアクセスと利用に関するポリシーはどのようなものか？

・当該データをどの程度の期間、保存するのか？　どの程度の期間、保存しなければならないのか？

・リポジトリの管理者は？　機関か？　営利目的の提供者か？

・リポジトリの利用にかかる費用は？　その支払い方法は？

・複製、永続性、監視、災害復旧、事業の継続性に関するポリシーはあるか？

研究データに関して利用可能なリポジトリには主に 2 種類あります。学問分野を限定したリポジトリ（例、主題リポジトリ）と汎用リポジトリです。

学問分野を限定したリポジトリ

これらのリポジトリは、一定の研究分野を対象に、特定の形式のデータを保存することを目的とし

ています。あらゆる形式のデータセットに対し、多数の主題リポジトリが存在します。研究者は自身の研究分野でもっとも一般的に利用されているリポジトリについて知っているのが普通です。こうしたリポジトリは一般に、ファイルフォーマットやメタデータ標準に特有の要件を課しており、データセットの寄託やダウンロードに対して利用制限を設けていることもあります。データの型、フォーマット、メタデータの提出に関する要件からわかるのは、主題リポジトリが往々にして、類似するデータ型の収集に相対的に優れており、データセットの集積に備えているということです。それによって、検索とメタ分析が可能になります。

たとえば、もっとも広く知られている主題リポジトリの 1 つに GenBank があります。GenBank は遺伝子データを保存しており、データの整形方法のほか、データセットに関係するメタデータの型やフォーマットにも、厳格な基準を設けています。こうした基準のおかげで、GenBank は著名な情報源として、遺伝子データを総合的に扱い、新たな発見をもたらしています。

汎用リポジトリ

汎用リポジトリの方がデータ登録に関して緩やかな要件を課し、幅広いフォーマットを受け入れており、メタデータ要件もあまり厳しくない場合があります。汎用リポジトリの所有者は非営利団体、出版社などの営利団体・企業、機関などの場合があります。機関が所有するリポジトリ（機関リポジトリ）は当該機関の図書館に設置されていることが多く、プロジェクトに伴う研究成果を一括して保管できるため重要であり、こうした保管方法は再現性に不可欠です。

リポジトリの選定

あらゆるデータはリポジトリに保管されるべきですが、データセットに適したリポジトリを選ぶのは難しいこともあります。データが学術分野を限定したあるリポジトリに適しているのが明らかであれば、そこにデータを寄託することで、関係する学問分野の研究者による発見と再利用が保証されるでしょう。しかし、多くの研究プロジェクトが 2 種類以上のデータを有しています（たとえば、ある種のゾウの遺伝子データ、形態学的データ、行動データと、そのデータを分析するための R コードなど）。汎用リポジトリであれば、こうした種類のデータをすべて受け入れ可能で、それによって再現性が促され、実施された研究について、より完全な報告が提供されます。研究者は、これらの（主題か汎用かという）2 つの選択肢の 1 つだけを選ぶのではなく、（1）適切なリポジトリを明確に特定できる学問分野独自のデータを、すべてその主題リポジトリに寄託し、（2）保存先のない研究成果すべてを 1 つのデータパッケージとして、1 つの汎用リポジトリに寄託し、（3）そのデータパッケージのメタデータファイルと readme ファイル内にリンクを張って、複数の主題リポジトリに寄託したデータの所在場所を示すことを検討すべきです。

リポジトリソフトウェア

データ保管の必要性が高まるにつれて、リポジトリソフトウェアの均質化と普及がさらに進んでい

ます。多数のリポジトリがそれぞれのカスタムソフトウェアシステムから、Fedora、DSpace、Dataverse といったオープンソースソフトウェアプロジェクトへと移行しています。こうしたプロジェクトは、コードや付加拡張機能を提供するコミュニティを構築し始めており、データの保管と共有のためのプラットフォームとしての価値を高めています。

利用と再利用

データセットの識別とリンク

第三者がデータを利用できるようにするために、データセットおよびデータセットの一部を識別、引用、リンクするための標準化された方法が必要です。

識別子

識別子とはオブジェクトを一意的に識別する一連の特徴のことです。こうしたオブジェクトとしては、データセットやソフトウェアなどの研究成果があります。ほとんどの研究者に馴染みのある特定のタイプの識別子は、デジタルオブジェクト識別子（DOI®）です。こうした識別子は、デジタル版の学術論文を一意的かつ恒久的に識別する目的で、学術出版団体によって 15 年以上利用されてきました。最近では識別子の利用は、ポスターやデータセット、コードなどの他のタイプのデジタルオブジェクトにも拡大しています。DOI は識別子の種類の中でももっとも広く知られているものですが、識別子にはほかにも種類があります。とはいえ、研究者は必ずしも識別子の細かな差異を理解しておく必要はありません。なぜなら、利用すべき識別子をデータリポジトリが決めることが多いからです。

一般に、識別子はメタデータレコードに含まれ、メタデータレコードはデータセットの場所を示します。つまり、識別子が有用なのは、メタデータが更新され続ける場合のみです。データセットの属性、とくにデータセットの場所が変更された場合、当該メタデータはそれを反映するよう更新されなければなりません。リポジトリはそれぞれが発行する識別子に対応した正確なメタデータを担保する責任がありますが、研究者は識別子によって第三者によるデータセットの発見が保証されるわけではないことを理解しておかなければいけません。

データの引用に関連するもっと複雑な問題

データの引用は学術論文を対象に開発されたモデルに基づいているため、データの引用に関して細かな差異がいくつか存在し、それらに対する最良事例は今なお開発の途上にあります。一般的に問題として認識されているのは、データセットそっくりそのままではなく、データセットの一部のみを引用する必要性（deep citation）です。データベースにはきわめて大きなものもあり、データベースそのものを引用しても、論文で論じられている特定のデータを第三者が再現、再利用するのに役立たないでしょう。動的なデータセットも問題になります。あるデータセットが絶えず更新されている場合（ストリーミング衛星データなど）、ストリーミングデータにタイムスタンプか、そうでなければサブセット化を行う方法が必要です。研究データコミュニティは、データの引用に関連するこれらの問題やその他の問題に対して、解決策を考案しようと取り組んでいるところです。

引用

データセットがリポジトリを通じていったん共有されれば、その作成者らは、当該データセットを提示できなければなりません。学術論文の場合、これは引用によって行われます。このデータの引用という慣行は、今ではデータセットなどのデジタルオブジェクトを、帰属を表示して提示する方法として奨励されています。著者（データセットの「作成者」と呼ばれることが多い）、日付、データセットのタイトル、公開者であるパブリッシャー（通常は当該データが登録されているリポジトリもしくはデータセットの作成者が属する機関）、識別子などの中核的な要素が共通しているため、データの引用は学術論文の引用に似ています（上述の*識別子*参照）。

データの引用の扱い方に関しては、学術コミュニケーション界で今なお議論が続いています。データの引用は通常、従来の参考文献の一部としては受け入れられませんが、パブリッシャーの中にはこの慣行についてのポリシーを変更し始めているところもあります。論文のランディングページにある要旨より先に、関連するデータセットの引用を明示する学術誌もあります。もっと一般的なこととして、学術コミュニケーション界には、データの引用を通じてデータセットへのアクセスを提供することは、当該学術誌の成果の再現性を保証する重要な方法だという認識があります。

データの引用は研究者や学術出版社の間で論じられるテーマとしては比較的新しく、引用をめぐる問題への取り組みを支えるための団体がいくつか形成されています。研究データ同盟（Research Data Alliance、RDA）や DataCite などの国際団体が、引用データに関連する基準を策定して、データセットへのアクセスの増加を推進しようと取り組んでいます。

データの発表

学術コミュニケーションの従来モデルには、学術論文での研究成果の発表があります。研究成果としてのデータの重要性が増すにつれて、このモデルはデータを伝達する 1 つの方法としても採用されるようになっています。「データの発表」が研究データを利用可能にする最良の方法かという点に関して、研究データコミュニティで多くの議論がなされています。しかし、研究データコミュニティがおおむね受け入れ可能な共通のテーマがいくつか登場してきました。

利用可能

データを発表するということは、そのデータが利用可能なものであるということになります。「発表」の厳密な定義は、何かを公にする、ということです。それゆえ、データを発表するということは、そのデータが公的にアクセス可能であることが求められます。おおまかな解釈では、データが学術誌出版社のウェブサイトの補足資料の中にあるか、作成者のウェブサイトで利用可能であるかということになりますが、データセットへの長期的なアクセスの提供に関しては、これらは理想的なシナリオではありません。データの発表は、信頼できるリポジトリにデータを寄託することによって実現することがほとんどなのです。

引用可能

データを発表するということは、それが引用可能なものであるということになります。一般に公開されているデータは、容易に参照でき、容易に所在場所を見つけられなければなりません。その役に立つのがデータの引用です。データの引用にはパブリッシャーも記載され、データセットの場所を特定するのに利用できます。識別子によって、発表されたデータセットの場所がわかることもあります（データの引用についての詳細は、前述の*利用と再利用*を参照）。

妥当性

データの発表に関しておそらくもっとも論争を呼ぶ側面は、妥当性の検証でしょう。査読のある学術誌に研究成果を発表するというのが、学術コミュニティの規範です。何らかの査読プロセスを経ていない出版物は、相対的に信頼されにくく、参照されることも少なくなります。この発表モデルをデータに拡大することによって、査読（すなわち妥当性の検証）を発表プロセスの一部とすべきことを提案します。

データは査読に関する興味深い問題を提起します。どのようにしてデータセットの重要性を判断するのか？　どのようにすればそのデータセットの将来的な影響を予測することができるのか？　審査や利用、分析に時間とエネルギーを過度に費やすことなく、データセットの妥当性を検証するには、どうすればよいのか？

データについての査読を分析するための 1 つの提案は、技術的な評価と科学的な評価を区別することです。技術的な評価では、容易に検証できる属性、主にデータセットとメタデータの完全性に重点を置きます。この種の評価は、関連する学問分野についての詳しい知識は必要ではないため、比較的容易に実施できます。しかし、科学的な評価はもっと徹底的に行われます。科学的な評価には、多くの場合、方法を評価し、データセットの妥当性を検証し、そして通常はデータセットの再利用の可能性を判断することが必要であり、そのいずれにも専門領域の知識が欠かせません。

データの発表モデル

データの発表という考え方について、学術コミュニティ団体間で依然として議論が行われているものの、データへのアクセスを与え、データへの貢献を示す必要性に合わせて、新たなモデルが登場しています。出版社によっては、データセットと関連する説明の発表のみを目的としたデータジャーナルを発表しているところもあれば、新たな論文の種類として、既存の出版物へのデータ論文の投稿を認めているところもあります。しかし、すべての出版社がデータの発表という役割を担っているわけではありません。既存のデータリポジトリに従って、従来型の論文を投稿しようとする著者に、データを 1 つのリポジトリで発表し、当該データセットへの引用を提示するよう求めている

ところもあります。

データの重要性が増し、独立した学術成果とみなされるようになるにつれて、データの発表モデルが迅速に発展すると考えられます。FORCE11 や研究データ同盟といった新たな団体が、興味深い可能性を開いています。こうした団体は、発表を通じてデータへのアクセスを提供するという既成概念の枠を超えようとしています。しかし、この発表形態は、研究者たちがすでに取り組んでいる使いやすい枠組みであることから、完全に衰退する可能性は低いでしょう。

功績（credit）とインセンティブ

現在の仕組み

最近まで、従来モデルでは学界での功績は出版物が中心でした。この仕組みは何百年も実施されており、研究について広く伝達する唯一の利用可能な手段でした。成果の発表という仕組みは、研究者の影響を評価する必要性に応じるために発展してきました。そして「発表せよ、さもなくば消えよ」という格言が繰り返し言われるようになりました。好ましくないことに、学界は共同研究の増大や研究の斬新さ、その他のもっと特別な意味を持つ要素などではなく、とくに影響力の強い学術誌での出版物の数（および被引用数）を重視してきました。

この仕組みの中心にあるのはインパクトファクター（IF）です。IF は、研究図書館が雑誌購読予算を配分する際に、雑誌の相対的なメリットを判断するためのツールとして、1970 年代に考案されました。現在、IF は研究者を評価するための仕組みに組み込まれており、研究者は研究にもっとも適した学術誌ではなく、IF がもっとも高い学術誌を選ぶ必要に迫られています。研究の功績と奨励の仕組みが理想的なものだと述べる研究者はほとんどいないと思われますが、では、代替案にはどのようなものがあるでしょう？

研究者がそれぞれの研究分野に影響を与える方法は、新たな手法の確立、ブログへの投稿、科学振興のための一般市民との交流、ツイッターなどのソーシャルメディアを利用した他の学問分野とのやりとりなど、数多くあります。将来の研究に影響を与えるという点で、もっとも重要な研究成果の 1 つはデータです。十分に文書化されて公的に利用可能なデータを提供することで、研究者の研究は今後、計り知れないほど貴重なものになる可能性があります。しかし、この発表を奨励する制度は、データの共有を促してはいません。学界での功績に対する新たな枠組みを構築して、出版数や被引用数では捉えられない種類の影響に対して報いようとしている関係者もいます。

オルトメトリクス

この新しい指標、すなわち「オルトメトリクス（altmetrics）」に向かう動きは、データやコード、ソフト

ウェア、ブログ投稿、ソーシャルメディアとのかかわりなど、「新しい（alternative）」研究成果について情報（metrics）を提供するための枠組みと考えることができます。ウェブサイト分析、ブログ、ツイッターフィードのほか、Mendeley、Zotero、CiteULike などの研究者の情報交換サイトなどの急増により、こうした研究成果の伝達と理解に関する情報を得ることが容易になっています。オルトメトリクスの例としては次のものがあげられます。

・リポジトリからのデータセットのダウンロード回数
・ブログ投稿の閲覧数
・論文へのリンクのリツイート数
・Mendeley での論文の保存数
・ウェブ上の発表の閲覧数

研究者コミュニティにオルトメトリクスに関する不安があるのは当然のことです。昔ながらの研究者は、ソーシャルメディアでの存在感が質の高い研究を行うことよりも重要になることを懸念しています。しかし、オルトメトリクスの目的は、発表に対する既存の功績モデルを補完することであって、それに取って代わることではありません。従来の仕組みでは測れない影響を測定する手段を提供することは、研究評価の選択肢を広げます。論文がツイートされた回数が、終身在職権や昇進を審査する昇進審査委員会に無関係である場合でも、それらの考慮を義務づけるべきではありません。しかし、データセットのダウンロード数を考慮する選択肢を持つことは、データの共有を奨励する重要なステップになるかもしれません。

データに対する功績

多くの研究者が、論文の執筆よりもデータの収集・整理・分析・再分析の方にはるかに多くの時間を費やしています。しかし、研究ライフサイクルの最終段階のみが、現在のところ重要な研究成果とみなされています。データを共有するための準備には時間と労力が必要です。そうした作業を奨励する仕組みがなければ、研究者はデータセットを非公開のままにしておくことになり、再利用と再現性の妨げになるでしょう。

こうした仕組みを変えるには、いくつかの面での歩み寄りが必要です。データの引用に関する基準を確立し、研究者が適切なデータ管理のメリットとデータの取り扱いに関する最良事例について認識を深め、昇進審査委員会が影響力の強い学術誌への出版以外の研究の影響力を考慮しなければなりません。データの管理と共有を求める資金提供者の要求は、研究の重要な要素としてデータを最前線に押し出す土台を作っています。しかし、そうした要求に添うための明確なインセンティブがなければ、高品質で、十分に文書化されたデータが共有される可能性は低いでしょう。

結論

本手引き書で示した指針は、データセットの利用と共有を容易にし、他の研究者との連携の機会を持ちやすくするために、データセットを確実に計画、文書化、保存する最良の方法について洞察を提供しています。研究データコミュニティが取り組んでいる問題はまだあるものの、進展がなされており、データを収集・利用・保管する人々は、それぞれのデータセットを将来、再利用可能なものにするためのツールと資源を、かつてないほど手にしています。

附録 A: 資料

一般

- データ管理計画策定のためのツール
  - Data Management Planning Tool (DMPTool): http://dmptool.org
  - DMP Online: https://dmponline.dcc.ac.uk/
- データ管理手引き書
  - ICPSR Guide to Social Science Data Preparation and Archiving. ICPSR is one of the premier social science data repositories. Their handbook on preparing data for archiving is extremely thorough, and will ensure high quality data.
  - UK data archive. The UK Data Archive has an excellent knowledge base for the creation and management of data. These guides are an excellent place to start for data project management issues.
  - DataONE primer on data management. This PDF covers the basics of data management, arranged in the context of the data lifecycle and geared towards researchers.
  - Data Management for Libraries: A LITA Guide. A short guide on data management topics, focused on helping libraries understand data management to better provide services for researchers that they support.
- トレーニング資料
  - Digital Curation Training for All: a collection of slides, resources, and materials that cover the basics of research data management from the Digital Curation Centre.
  - MANTRA Course. MANTRA is focused on providing educational materials for researchers on important data issues. These lessons are a great start for librarians looking to provide data management workshops.
  - DataONE data management education slide decks
  - UK Data Archive training resources on managing and sharing data

本書で参照したプロジェクトおよびソフトウェア

- EML: https://knb.ecoinformatics.org/#external//emlparser/docs/index.html
- FGDC19115: https://www.fgdc.gov/metadata/geospatial-metadata-standards
- バージョン管理
  - Git: http://git-scm.com/doc

- GitHub: http://github.com
- Bitbucket: http://bitbucket.org
- Mercurial: http://mercurial.selenic.com/
- Subversion: http://subversion.apache.org/

・ソフトウェア

- R: http://www.r-project.org/
- Taverna: http://www.taverna.org.uk/
- iPython: http://ipython.org/
- Kepler: https://kepler-project.org/
- LaTeX: http://www.latex-project.org/

・使用許諾

- Creative Commons: http://creativecommons.org/
- Open Data Commons: http://opendatacommons.org/

・ストレージ

- Dropbox: http://www.dropbox.com
- Google Drive: https://www.google.com/drive/

・リポジトリ

- GenBank: http://www.ncbi.nlm.nih.gov/genbank/
- Fedora: https://getfedora.org/
- DSpace: http://www.dspace.org/
- Dataverse: http://dataverse.org/

・イニシアティブ

- Research Data Alliance: https://rd-alliance.org/
- FORCE11: https://www.force11.org/

・識別子

- DOI: http://www.doi.org
- DataCite: http://datacite.org

・文献管理

- Mendeley: https://www.mendeley.com/
- Zotero: https://www.zotero.org/
- CiteULike: http://www.citeulike.org