厚生労働行政推進調査事業費補助金（厚生労働科学特別研究事業）
分担研究報告書

国立研究開発法人におけるデータポリシー策定にあたり検討すべき事項の整理
～「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」を羅針盤として～

木村　映善　国立保健医療科学院保健医療情報管理分野 統括研究官

研究要旨
本文書は国立保健医療科学院のワーキンググループにおいてデータポリシーを策定するにあたって作成した参考資料である。ワーキンググループにおいて「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」を参考にしたときに、当ガイドラインのみでは判断が困難な構成員向けに情報提供をすべく調査した事項をまとめたものを、公開向けに再構成したものである。

A. 本文書の位置付け

本文書は国立保健医療科学院のワーキンググループにおいてデータポリシーを策定するにあたって作成した参考資料である。すなわち、ワーキンググループにおいて「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」を参考にしたときに、当ガイドラインのみでは判断が困難な構成員向けに情報提供をすべく調査した事項をまとめたものである。他の研究組織におけるデータポリシーの検討の際に参考になると判断し、公開するものである。なお、この文書中における提言はワーキンググループへの問題提起として記したものであって、国立保健医療科学院の最終的結論ではないことはご留意頂きたい。

B. ガイドライン検討にあたって背景

統合イノベーション戦略（平成 30 年 6 月 15 日閣議決定）において「オープンサイエンスのためのデータ基盤の整備」、つまり、すべての者がサイバー空間の研究データを利活用し、協働によってイノベーションを創出するというオープンサイエンスを推進するための社会インフラとしてのデータ基盤の整備が明示され、国立研究開発法人は、研究分野の特性、国際的環境、産業育成等に配慮したデータポリシーを 2020 年度中に策定し、これに基づく研究データの管理・公開等を促進することとなった。また、データポリシーの策定にあたっては、内閣府総合科学技術・イノベーション会議（CSTI）が設置した「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」による「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」（以下ガイドライン）（平成 30 年 6 月 29 日）を参考にすることが求められている。

これを受けて厚生労働省は、厚生科学審議会科学技術部会（平成 30 年 7 月開催）での議論を踏まえ、厚生労働省所管の研究機関（医薬基盤・健康・栄養研究所、国立がん研究センター、国立循環器病研究センター、国立精神・神経医療研究センター、国立国際医療研究センター、国立成育医療研究センター、国立長寿医療研究センター）だけでなく、その他の研究機関（国立医薬品食品衛生研究所、国立保健医療科学院、国立社会保障・人口問題研究所、国立感染症研究所、国立障害者リハビリテーションセンター、独立行政法人国立病院機構。独立行政法人労働者健康安全機構、独立行政法人国立重度知的障害者総合施設のぞみの園、独立行政法人労働政策研究・研修機構）においてもデータポリシーを策定し、所管する全ての研究機関でオープンサイエンスを推進することとなった。

上記の「ガイドライン」によれば、データポリシーは研究機関の研究分野の特性やミッション等に基づいて定められるものとされているが、厚生労働行政や他の分野の政策等に資するデータの利活用を促進するためには、研究機関等で一定の共通する事項や内容を設定して、研究データの横断的連携の推進等に向けて相互運用性を高める必要もある。しかしデータポリシーの記載事項や記載内容をどこまで機関共通にすべきか、その具体的な考え方や根拠は明らかにされていない。また、研究データの公開・利活用の基盤となる機関リポジトリに関しても、具体的な整備・運用の方法が確立していない。

そこで本研究では、厚生労働省が所管する研究機関で策定すべきデータポリシーに関して、機関共通で取り組むべき事項、各機関の特性に応じて取り組むべき事項を整理し、厚生労働行政等に資する研究データの利活用を最大限に促進するために必要な、一貫性及び整合性のあるデータポリシーの要件を明らかにするとともに、機関リポジトリの整備及び運用に関する提言を取りまとめることを目的とする。

データポリシー策定にあたって考慮すべきことがらはきわめて広範囲にわたるので、「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」（以下、DPG: Data Policy Guideline）[1]及び当ガイドラインの解説資料　平成 31 年 4 月 2 日版 （以下、AG: Appendix of Guideline)[2]をデータポリシー策定プロセスの羅針盤とし、逐条的なガイドラインの検討をとおして課題を整理し、2020 年度中のデータポリシー策定にむけて準備すべきことを洗い出す。

なお、AG において、各法人の研究分野の特性やミッションの反映を阻害することが懸念されるため、データポリシーのひな型は敢えて提供されておらず、ガイドラインに沿って各法人にて検討すべきことが記載されている。しかしながら、各法人で委細にデータポリシーを検討できる人員は限られているため、本稿において、DPG、AG に加えて独自に調査、分析した事項を提供することによって、各法人での検討の一助になることを期待するものである。

以下、内閣府の「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」(DPG)に記載されている内容を元に、逐条的に当院におけるデータポリシーを検討するにあたっての考察を記述する。DPG の「3. データポリシーで定めるべき項目」以降に提示されている項目（番号）と解説を囲みで引用し、その後我々が検討したことを記述する。

C.　DPG の逐条的検討

C.1　機関におけるポリシー策定の目的について

> ・機関のビジョン、ミッション等を踏まえ、ポリシーを策定した背景と研究　データ利活用の目的について記述する。

・機関が Web サイト等で公開している機関のビジョン、ミッションをベースに解説する。

・公的資金を活用して実施した事業・研究を通して得られた成果を国民に還元し、広く利活用を促進することを通し、機関に求められている役割の強化とアウトリーチを拡大することを説明する。

C.2　管理する研究データの定義、制限事項について

> ・　機関のミッションに従い、ポリシーが対象とする「研究データ」の定義・　範囲を明確にし、利活用が想定されるデータ、将来的に利用の可能性が考えられるデータなど、研究データの種別・内容等について記述する。
>
> ・　研究データの利活用に関する機関の方針や基本的な考え方を踏まえ、また、第５期科学技術基本計画が示すオープンサイエンスの推進に係る方針にも留意し

て、非公開、共有等の対象となる研究データや公開・共有における 制限事項について記述する。

C.2.1研究データ利活用に関する機関の方針・基本的な考え方

第 5 期科学技術基本計画 が示すオープンサイエンスの推進に係る方針[1]に記述されている、オープンサイエンスの推進体制の基本的姿勢は、公的資金による研究成果の利活用の機会を可能な限り拡大することとしている。我々の活動は、国の予算と公的機関からの競争的研究資金に基づいている。それを踏まえて、基本的に保有しているデータは公開が前提であるとし、非公開となるものについて例外的に指定・除外していく手法を採用することとする。

C.2.2検討対象となるデータの種類

保健医療科学院に所属する研究者の研究テーマは多種多様である。結果としてデータの種類は多種多様であり、ポリシーレベルで具体的かつ個別に指定することは、煩雑かつ、各部局単位での運用を限定する可能性がある。従ってデータの具体的な種類については言及せず、対象となるデータの範囲のみ提示することとする。データの由来に着目し、下記の通りに公開対象となるデータの分類をした。

・(A) 公開対象となるデータの定義、範囲

・機構の施設・設備を利用して得られたデータ
・外部の公的資金を活用して実施した事業・研究を通して得られたデータ
・外部組織との協業、共同研究等を通して得られ、公開に同意されたデータ
・及び、上記から得られたデータをもとに派生して得られたデータ

・(B) 潜在的に対象となるが一般的に公開せず、審査・契約等締結後に限定公開するもの

・機微な個人情報を含むデータ
・商業利用等に制約を課せられたデータ
・公的機関、国内の研究者等に、開示対象を制限すべきデータ

・(C) 対象から外すデータの定義・範囲

・国家安全保障、個人の安全・プライバシーに係るデータ

但し、個人の安全、プライバシーに関しては、適切な匿名加工処理を加えて公開することを検討し、それでも技術的、制度的に困難なものであると判断したもののみに制限する。技術的困難とは、安全性を担保した匿名加工後にデータとしての有用性を確保できないものである。制度的困難とは、そもそもその存在を秘匿すべきもの、外部への提供が許可されていないものである。

C.2.3対象となるメタデータ

非公開となったものであっても、保有していることの情報開示をするために、メタデータは上記の公開の対象、非対象に関わらず、全てメタデータを作成し公開するものとする。但し、国家安全保障に係るもの

1 第 5 期科学技術基本計画　第４章 科学技術イノベーションの基盤的な力の強化
（２）知の基盤の強化　③ オープンサイエンスの推進（抜粋）
国は、資金配分機関、大学等の研究機関、研究者等の関係者と連携し、オープンサイエンスの推進体制を構築する。公的資金による研究成果については、その利活用を可能な限り拡大することを、我が国のオープンサイエンス推進の基本姿勢とする。その他の研究成果としての研究二次データについても、分野により研究データの保存と共有方法が異なることを念頭に置いた上で可能な範囲で公開する。ただし、研究成果のうち、国家安全保障等に係るデータ、商業目的で収集されたデータなどは公開適用対象外とする。また、データへのアクセスやデータの利用には、個人のプライバシー保護、財産的価値のある成果物の保護の観点から制限事項を設ける

及び院長から具体的に指定のあったものについては、メタデータも非公開の対象となる。非公開対象データについては、行政開示請求の文脈で対応することとする。

メタデータとは、検索システムの対象となるデータに関する情報である。しかし、検索システムによってメタデータの構造が異なるとメタデータの交換や横断的な検索が困難となるため、メタデータの構造を標準化する動きがあり、例えば書誌情報では書誌名、著者、日付、出版社等の標準的な15要素を定めたDublin Core[3]が知られている。図書以外のデータについてのメタデータについても検討がなされており、採用するメタデータについては後述する。

以下、データ、メタデータの対象について整理した表を掲示する。

メタデータは基本的に公開するが、存在自体を知られるべきではないデータについてはメタデータも非公開とする。メタデータはリポジトリや検索システムに登録され、利用者が検索した時に検索結果として表示される。しかし、データの本体がリポジトリに登録されていなければ、データ本体へのアクセスは不可能である。このような状態が発生するのが、「公開範囲」が限定的公開・非公開に設定されている状況である。個人情報が含まれており、匿名加工がなされるべきものについては、匿名加工の処理が終了して公開の許可が下りたときにメタデータとデータ本体が同時に公開される。

匿名加工が困難あるいは加工するとデータの有用性が著しく落ちる場合は、限定的公開（メタデータのみ公開）での運用を検討する。

エンバーゴ期間（後述）が指定されているものについてはメタデータを先に公開し、データ本体はエンバーゴ期間経過後にアップロードして公開するか、リポジトリの機能を利用してエンバーゴ期間経過後に自動公開する設定を適用する。

なお、機微な情報を含むデータの公開に関する判断はANDSのガイドラインで解説されており、メタデータとデータを公開するパターンの組合せを検討する際に参考にした[4]。

エンバーゴ（Embargo）については、複数の意味で使われており、どの文脈で使用されているのかを区別する必要があることに留意されたい。

（1）論文公開前の広報（情報解禁日）としての意味

出版社側から論文公開前に論文に関するプレスリリースを著作者・機関側で出すことを自粛するよう要請しているもの。例えば、Nature research pressのEmbargoに関するポリシーは下記の通りである。

Nature research Press and embargo policies

https://www.nature.com/nature-research/editorial-policies/press-and-embargo-policies

We strongly discourage authors and

表　公開対象とメタデータ、データの関係

公開範囲	一般公開（個人情報のないもの）	一般公開（個人情報があるが匿名加工がなされたもの）	限定的公開	非公開	非公開（機密）
メタデータ	公開	公開	公開	公開	非公開
データ	公開（エンバーゴ指定されているものは期限後に公開）	匿名化後公開	非公開・契約締結後に公開	非公開	非公開

potential authors from direct solicitation of media coverage of material they have submitted to Nature and the Nature Research journals. Accepted contributions can be discussed with the media only once the publication date has been confirmed and no more than a week before the publication date under our embargo conditions. Please refer to the ″Communications between scientists″ section for more information about our embargo policy as it pertains to conference presentations and preprints.

このように期間を指定されたものについては、その期間以前にメタデータやデータをリポジトリに置いても公開、露出状態にしないことが求められる。

(2) 公開猶予期間

無料公開されない購読型論文が、一定期間を経て無料公開される際に設定される期間。ただし、無料公開されたからといって、その論文をリポジトリにても公開してよいとは限らないので、出版社の規約を確認されたい。

(3)著者の権利や契約にもとづくもの

例えば、データを公開する研究者が、研究の先行利益を確保するために、リポジトリに登録してもただちに公開を希望しない場合に、公開予定日を設定する。リポジトリの機能によって公開予定日を過ぎた時に初めてデータが公開されるといった設定が可能である。

C.3　研究データの保存・管理・運用・セキュリティについて

・　研究データの特性に応じたデータの保管、運用方針と国研としての取組について記述する。

（記述上の留意点）

・　機関内で実施される研究活動において順守すべき研究データの保存・　管理・運用・セキュリティに関する対応についての方針、及びこれらを実施するための体制、並びにワークフローについて記述する。その際、　研究データの特性、運用のフォローアップ、その他のポリシーとの整合性に留意する。

・　研究データを登載するリポジトリ等について記述する。なお、特定のリポジトリ等名のほか、リポジトリ等が備えるべき条件について記述することが望ましい。

・　研究プロジェクト終了後における研究データの保存・管理等の継続性にも考慮することが望ましい。

C.3.1　運用管理規程について

機関内で実施される研究活動において遵守すべき研究データの保存・管理・運用・セキュリティについては、既存の機関のセキュリティポリシー、システム運用管理規程に従うこととし、研究者の業務の増加を最小限に抑える。但し、行政文書についての分類やライフサイクルについての具体的な規程はあるが、研究、教育に関するデータについては、研究者の自主的管理に依存するところが多く、具体的な規程に乏しい。状況を調査し、必要に応じて規程を整備する必要がある。別途記述するようにデータ公開にそなえて、データを保管する期間・場所を定めることになるが、このデータ保管・管理を研究者に要求するのは業務圧迫につながる怖れがある。また人事異動、退職、研究プロジェクト終了等に伴い、データ管理の継続性が損なわれる可能性があることから、個人従属的な業務形態にすることは避けるべきである。従って、機関内あるいは「機関横断的」にデータのアーカイブに関わる担当者の設置を検討し、研究者からデータを引き渡されたあとは、管理責任は機関が一義に負うような体制、管理規程を整えることが望ましいと考えられる。「機関横断的」とは、省庁下の国研群にお

いて、各々でデータを管理する担当者を雇用するのではなく、省庁単位で国研群のデータを一括して管理する部署・担当者を設置するアプローチである。現在、我が国においてはデータを管理できるスキルを備えた人材が不足しており、また雇用が不安定な傾向がある 。長期的視野に立てば、専門スキルを持つ方を集約して安定して雇用できるような環境を創出することが望ましいと思われる。

C.3.2リポジトリ等が備えるべき要件等について

オープンアクセスにむけた研究データの管理方式は、Open Archive(OA)ジャーナルに投稿するGold OA方式と、機関リポジトリ等にセルフ・アーカイビングするGreen OA方式がある[5]。投稿した論文誌の出版社のポリシー、研究助成機関による論文の公開に関する方針もあり、出版された論文をセルフ・アーカイブすることが必ずしも認められていないため、どちらかの方式のみ採用するということは不可能である。ただ、Gold OAの掲載にはコストがかかり研究者の負担、引いては税金の有効活用の観点から、可能な限りGreen OAへの掲載も認める出版社への投稿を推奨し、同時に機関におけるセルフアーカイビングを推進することが望ましいと考える[6]。また、今後の研究資金提供者によってOAセルフアーカイビングが義務つけられるようになった時に研究者を支援すべくセルフアーカイブが無償でできる機関リポジトリの存在は必要になると思われる。

Green OAのリポジトリの種類は、大学・研究機関みずからが保有する機関リポジトリ、研究者コミュニティによって運用される分野別のリポジトリ(arXiv等)、助成機関によって運営される、助成を受けた研究の成果を格納するセントラルリポジトリ(PubMed Central (PMC)等)の形態がある[7]。本データポリシーの検討の対象となっているのは国立研究開発法人であり、国の事業として実施している研究、調査にもとづいて開示すべきデータ等もあることが予想されることから、研究者コミュニティや助成機関等の第三者が運用しているリポジトリで開示するのではなく、自らあるいは委託したリポジトリで開示するという運用を採るのが適切なケースがあると思われる。なお、NIHがFunding Agencyとして助成した研究の成果を公開させるためにPMCをセントラルレポジトリとして提供しているように、国立研究開発法人がFunding Agencyとしての役割を果たしているならば、セントラルレポジトリの運用も担うことを検討すべきであると考えられる。

ただし、現状の研究機関の予算と人員状況下では、Green OAを実現するための体制を構築することは困難である。他の国立機関によって提供されているリポジトリ等のサービスを活用し、研究機関内にそのサービスを管理したり研究者によるデータ登録を支援したりする人員を配置、育成していくという方針を採用すれば、持続性のある運用が可能であると考える。もし、研究機関ごとに担当者を配置することが困難であれば、厚生労働省あるいは厚生労働省から指定された国立研究開発法人に人材を集約し、他の研究開発法人のデータ管理をもまとめて引き受けるとことで運用の最適化を図る必要があるかもしれない。

機関リポジトリのサービスについては、自前で機関リポジトリを構築する他、J-StageやCiNii等の外部リポジトリが多数使われている[8]。ただ、自前で機関リポジトリといっても完全に独自に構築している事例は少なく、外部の機関リポジトリサービスを利用しているのが実情である。国立研究開発法人の機関リポジトリとしては、オープンアクセスリポジトリ推進協会(JPCOAR)[9]下に、国立情報学研究所が運営している Japanese Institutional Repositories Online Cloud (JAIRO Cloud)という、機関リポジトリ環境を提供するサービスが検討に値すると考える。理由は下

記の通りである。

(a)低廉なシステム運用コスト - 事業継続性の確保

JAIRO Cloud の利用は、JPCOAR 基本会費と JAIRO Cloud 利用料金の組合せの年間会費を支払えば可能であり、構成員が 200 名規模では年間 10 万円程度、1000 名規模でも数十万円程度と低廉な利用料金の設定がなされている(2019 年度時点)。また、JPCOAR 会員に加入するため、JPCOAR で開催される人材育成や啓蒙活動に参加することで、独自に取り組むよりもオープンデータに関わる人材育成のマネジメントの負担も軽減されることが期待される。

(b)科学技術基本計画のプロットへの追従

JPCOAR は、2019 年〜2021 年度にかけての戦略[10]として、JPCOAR はオープンサイエンスの推進に寄与するため、研究データの公開、流通に関する先導的な取組み、 オープンアクセスを推進する学術情報流通の基盤を整備し、コンテンツの流通、活用を促進する、 オープンアクセス、オープンサイエンスの推進に対応できる人材育成を行う、ことを挙げている。JPCOAR の会員となり、活動に参加することで最新のオープンサイエンスに関する動向の把握と、最新の機関リポジトリ環境の整備、人材育成に寄与することが期待される。

(c)外部への委託による機関内業務の削減

国立情報学研究所によって情報セキュリティ、システムの継続的運用上の配慮が十分になされたクラウド環境に機関リポジトリ業務を委託することにより、機関におけるシステム管理者の雇用、情報セキュリティの確保に係る諸業務を節約できる。また、クラウドにデータを展開することにより、自然災害、停電等によるデータの損壊、損失の機会を最小化し、オープンサイエンス時代に求められているデータ保存・管理の継続性に対する要件に対応することが可能になる。すなわち、DPG で求められている「研究プロジェクト終了後における研究データの保存・管理等の継続性への考慮」にも応えられると考える。

(d)データの真正性、保存性の確保

データの保存期間は事業が継続する限り、半永久的としたい。保存に関するコストへの配慮もそうであるが、長期間保存によるデータの消失、損壊を回避する必要がある。オンプレミスの運用であればサーバやストレージの更新によるマイグレーションが必要であるが、データ移行はデータの欠損、破壊の最も大きな契機の一つであり可能な限りさけたい。また、ストレージに高品位なものを利用することでデータそのものが損傷を受ける機会を削減したい。このような要求を鑑みれば、データの多重複製、データ損傷のモニタリング、自動修正を行う高機能なストレージに保管することが望ましい。近年のクラウドサービスはビッグデータを確実に保存するためにオブジェクトストレージ等の仕組みを取り入れてデータ保存に関する品質を飛躍的に高めている。そういう意味でも最先端のストレージ技術の運用が期待されるクラウドベースでの運用としたい。

なお、DPG には「実施するための体制、並びにワークフローについて記述する。」という言及はあるが、これは院内で運用される対策基準あるいは実施手順レベルにおいて記述すべきものであり、対外的に公開されるポリシーにおいて記載するような内容（粒度）ではないと考える。

C.3.3研究データのマネジメントについて

研究データの管理は研究を開始する前に計画するべきであり、標準的な研究上の実務に組み込まれるのであれば、必ずしも時間や費用を大きく費やすことにはならないという主張がある[11]。しかし、現時点

では我が国の Funding Agency でも Research Data Management (RDM)の提出を求めるようになっているが、必ずしも実際の研究上の詳細な計画まで踏みこんだ内容になっておらず、また研究者が利用できる標準的なデータ管理環境も存在しておらず、依然として研究者自身による管理に委ねられている。

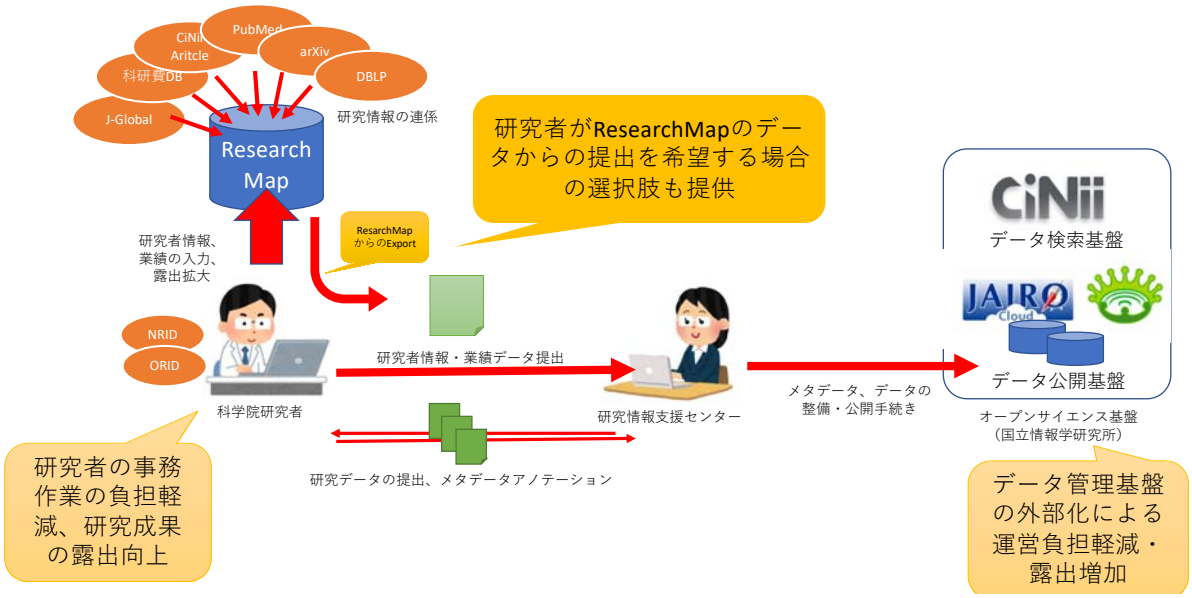

図 1 研究データベースとリポジトリの連携を活用したオープンデータ登録支援

我が国では、国立情報学研究所オープンサイエンス基盤研究センターが米国 NPO の Center for Open Science が開発している研究データ管理用の Open Science Framework[12]をカスタマイズして GakuNin RDM[13]という研究データ管理基盤を提供している。研究データ管理基盤をあらためて導入するメリットは、研究者の本来の業務ではない研究データを安全に管理するためのシステムの運用・保守を外部化できること、データの永続性を保証し研究資料の消失・散逸を抑制するというセキュリティ保全[14]の他に、研究者コミュニティでのデータ共有を促進し、研究プロセスを共用化・再利用することで研究を加速する効果が期待される。

ただ、この GakuNin RDM を用いた先行検証を含めて、一般的に利用出来るかたちで提供されている研究データ管理用基盤を用いた研究活動およびその成果として得られたエビデンスについて確認されていない。研究データ管理用基盤の導入について検討するのはよいとしても、この数年間以内に研究データ管理用基盤を利用したマネジメントが確立されている状態を想定するのは困難である。

従って、当面は RDM システムの活用にむけた取り組みと並行して、従来のプロセスで出来上がったデータをどのように公開するかの検討を進めることとしたい。

対象となる研究論文、データが発生したあと、どのようなプロセスを経て公開に至るかを検討する必要がある。現実的な問題として、対象論文・データの把握は研究者の自己申請に依らざるを得ない。自己申請に関して、研究者やデータ公開を支援する担当者の業務負担を軽減するために下記の様な運用モデルの検討を進めているところである。

別途に記述するように、研究者には国際的な研究者 ID の獲得と研究者データベースへの業績登録を依頼する。その代わり、これまで年度の事業報告や業績評価の為に作成していた資料から研究、論文に関する業績の報告にかかる業務を免除するようなサポートをする。具体的には、研究機関の

データ公開支援者は、研究機関に所属している研究者が登録している研究者データベースをクローリングし、新たな業績が登録されたことを確認次第、データ公開申請書を起票し、当該業績に関する書誌情報等を予め埋めた上で、研究者に公開の可否、成果物の提出について申請させるWebサイトへの案内をする。例えば、研究者データベースについては、国立研究開発法人科学技術振興機構が提供している次世代研究基盤リサーチマップ(researchmap)[15]によって、各種データベースと連携し、各種研究補助金の書類や評価データの提出フォーマットをダウンロード可能なサービスが提供されている。また、researchmap のデータベースからデータを入手するWebサービスも提供されている。このWebサービスを利用して複数の研究者に関する情報をXML形式で一括して入手可能とされている。このXMLデータから研究機関が使用しているリポジトリサービスに登録するためのメタデータの雛形を自動生成する仕組みを開発することで、データ公開に関する業務効率化が期待される。論文データベースは DBLP、PubMed、ORCID、Web of Science等多数あるが、researchmap に注目する理由は、複数の論文データベースや書誌情報と連携して業績を取り込めること、科研費や業績報告など、我が国に関する研究者の各業務を支援する機能が提供されている等、研究者にとっても機関にとっても業績の管理をワンストップで提供するプラットフォームであると考えている。

C.3.4　TODO

・来年度以降、データポリシーにもとづいた運用を開始する以前に、研究データの管理にかかわる情報システムの管理規程等の確認・整備が必要である。

・Jairo Cloud等、外部のリポジトリサービスを利用してリポジトリを運用する場合は、保有データの外部組織への管理委託に関する規程の確認が必要である。

・researchmap の研究データベースから入手できる情報とリポジトリに登録するメタデータについての調査とコンバートプログラムの開発

C.4　研究データに対するメタデータ、識別子の付与、フォーマットについて

・　研究データに対するメタデータ及び識別子付与についての方針を記述する。また、研究データの特性に応じた標準的なフォーマットが存在する場合　は、それも併せて記述する。

C.4.1　取り扱うメタデータの規格

国際的なデータ公開、再利用を推進するために、研究データに対するメタデータ、識別子は国際的に普及している規格を優先的に採用するものとする。

我が国では、機関リポジトリの標準的なメタデータスキーマとして junii2[16]が流通しており、リポジトリ登録に使用されている。しかしながら、研究データの記述や OA 状況をモニタリングするために必要な要素、各種識別子を記述するための要素を十分に備えていないとされる[17]。国際的には研究データのメタデータとしてDataCite のメタデータスキーマ[18]が標準的なものとして定着しており、junii2 もDataCite のメタデータとのハーモナイズが検討されていた[17]。この流れをうけて、junii2 に代わるメタデータスキーマとして2017年10月にJPCOARスキーマ ver 1.0、そして 2018 年 8 月に現時点で最新版となるver1.0.1 がリリースされている[19]。Confederation of Open access Repositories (COAR)やDataCiteで利用されている統制語彙を当面必要とする用途に絞って採用し、研究データ等の新しいコンテンツへの対応、OA の状態 (OA かどうか、エンバーゴ終了日など)、公的研究助成に関する管理情報、多様な論文の種類を網羅するなど、JPCOAR コアスキーマは学術情報の国

際的な流通性を高め、かつオープンサイエンスに対応可能なメタデータの設計がなされている[20]。我々は国際的な研究機関とも協調して事業にあたることが期待されているから、このような国際的な流通性に配慮したメタデータを積極的に採用すべきであると考える。また、共用リポジトリサービスの JAIRO Cloud では 2019 年度に JPCOAR スキーマに対応した JAIRO Cloud を運用開始予定であるとのことである[20]。

TODO:[17]によれば、現在の機関リポジトリは研究データの DOI 付与には対応していないとのこと。現在の機関リポジトリの実装について状況確認。NII 等の共同研究を検討し、機関リポジトリに必要な要件の確認と機能の開発に協力する。

TODO:JPCOAR メタデータ規格について検討し、研究者の保有データ調査の為の基礎資料とする。(分類手法など)

C.4.2研究者の識別子

著者を管理する ID スキームは複数提案されている。研究者の負担軽減のために、シェアの大きい ID スキームの採用を検討する。日本における研究助成申請時に使われる科研費研究者番号をベースとした NRID と、国際的な研究者識別子付与活動をリードしている ORCID [21] (Open Researcher and Contributor ID)を中軸として利用することが望ましいのではないか。また、科学研究助成事業データベース上で NRID から ORCID のリンケージが可能となっている[22]。研究機関全体の取り組みとして、着任時の研究者向け研修（図書館、論文データベースの操作研修等）等の機会を利用して、ORCID の取得、そして NRID と ORCID のリンケージ設定を推奨するのがよいのではないか。

メモ：

・NRID: 科研研究者番号 KAKEN データベース等で使用

・e-Rad: 府省共通研究開発システムで使われる研究者用 ID

・ORCID 世界中で研究者を一意に特定するための研究者用 ID 非営利団体 ORCID,Inc によって運用されている。

C.4.3組織の識別子

科研費電子申請システムで利用されている機関番号[23]と世界の研究機関データベース GRID (Global Research Identifier Database) [24]を採用するものとする。また、2019 年 2 月に CrossRef から GRID をベースとした Research Online Registry(RoR)が発表されており、こちらの動向も注目していきたい。

例） 国立保健医療科学院の GRID は grid.415776.6
RoR は https://ror.org/0024aa414

なお、識別子については採用するメタデータ規格の入力ガイドライン等で推奨されているものが存在する可能性があるので、メタデータ規格の検討時に識別子についてもあらためて検討することとする。

C.4.4データフォーマットについて

標準形式への研究データの準拠については、データを生成した部局において個別に判断するものとする。既存フォーマットを標準規格に準拠した別のフォーマットに変換することは追加のコストや労力がかかる可能性があり、利用者側のニーズ（何のデータをどのような形式で利用したいか等）が把握出来ていない現状では優先度は低く、まずはデータの存在について周知していく取り組みについて優先度を割り振るべきであると思われる。

但し、再利用性が高い状態で公開することを要請されているため、ポリシーでは原則としてスター・スキーム[25]の 3 段階目

以上の標準的形式での公開を目指すことにしたい。

例えば、国立保健医療科学院は公衆衛生分野を主に担当するが、公衆衛生分野では、近年の国境を越えた人や資源の移動にともなう自然環境、感染症などについて対策がボーダレス化しつつある今、データの迅速な公開、共有にむけて標準規格の準拠についての重要性が認識されつつある。しかし、分野によってはデータの構造についてコンセンサスが得られている標準規格は殆ど存在しない。各分野においてのデータの標準化に関するベストプラクティスが蓄積され、相互運用性の確保の努力が行われることを期待し、現時点では「当研究分野において標準化された規格があれば、可及的にそれに準拠したデータ形式で公開すること」と述べるのに留めるのが望ましいと考える。

なお、Resource Definition Framework (RDF) [26]及び、RDF によって記述された Semantic Network からデータを抽出する SPARQL クエリ言語がオープンかつ標準的なデータとして公開する手法として知られているが、結局のところ当該研究ドメインに関するリソースの表現を統一するための語彙定義がなされていることが前提であり、先述したように公衆衛生分野では発展途上である。所謂「標準規格に準拠したデータ」のありかたについては、理想と実利のバランスをとりつつ、下記の取り組みを並行して進めるのが望ましいと思われる。なお、前述した「スター・スキームの 3 段階目以上の標準的形式」に関しては、以下の条件を基準に選定することが望ましいと思われる。

・長期的な互換性を確保することを努力しているソフトウェアで使われている形式であること。

全てのデータはデータアクセスに関わるハードウェアおよびソフトウェア環境が旧式化するというリスクにさらされている[11]。そのため、過去のバージョンで作成されたデータをインポートする下位互換性を確保し、長期的なデータアクセスを保証することをサポートしているソフトウェアがあるならば、そのソフトウェアが利用している形式を優先して採用することを検討する。

例えば PDF 形式文書はスター・スキームで推奨されているフォーマットではないが、やむをえず利用する場合は 長期保存のための国際規格 ISO 19005-1(PDF/A)に準拠した PDF 形式で保存する等の配慮が必要である。これは長期にわたって保存しても表示される内容，色等が変化しないで再生できる PDF 文書を提供することを目標として策定された規格である。

・データ形式に関する仕様が公開されていること。

ソフトウェアがオープンソースのものであればソフトウェアの開発が停止してもソースコード等からデータ形式の仕様を確認し他のデータ形式に移植する可能性が残されている。しかし、商用ソフトウェアではサポート中止後、仕様が開示されずブラックスボックス化し、データへのアクセスが困難になることが考えられる。現時点でデータ形式に関する仕様が容易に入手できるものを選択することが望ましい。但し、これは商用ソフトウェアの仕様を排除するものではない。例えば、統計処理ソフトウェア SPSS では公式の公開仕様は存在しないが、GNU PSPP プロジェクトより SPSS Portable Format(.por)[2]、SPSS System file(.sav)[3]の仕様が公開されており、オープンソースの統計処理ソフトウェア R[27]の foreign パッケージや機械学習等に多用されている開発言語 python の pyreadstat ライブラリでサポートされている。同様に統計処理ソフト SAS、Stata、地理情報空間システムに使われる地

2 https://www.gnu.org/software/pspp/pspp-dev/html_node/Portable-File-Format.html#Portable-File-Format

3 https://www.gnu.org/software/pspp/pspp-dev/html_node/System-File-Format.html

理空間データ ESRI シェイプファイル等、プロプライエタリなフォーマットであるが仕様が公開されており、多数の汎用的なツールでサポートしているものは受け容れても良いと思われる。

Excel や OpenOffice のスプレッドシートソフトウェアも Open Document Format for Office Applications (ODF)フォーマットで汎用的なかたちでデータを記述できるようになっているが、Excel は表現力・自由度が高く表や罫線、セルの加工が容易に出来るため、結果としてデータの二次利用に適さないかたちでの公開につながる可能性が大きく[28]、データの二次利用性向上にむけたガイドラインを定義することも困難なことから極力使用しないことが望ましいと思われる。

(i)当座のデータ公開にむけて

最終的には Semantic Web のフレームワークの文脈に従ってデータを公開することが望ましいとされている。そのため、現時点では Semantic Web に則って公開できなくても、現在どのレベルにあるのかを自覚的に認識し、進歩の方向性を見定めることは有用である。オープンデータとして公開するときに、ティム・バーナーズ・リーはスター・スキーム(star scheme)として 5 段階の構造化レベルを考慮し、最終段階である Link of Data (LOD)を意識していくことを提案した[29]。

★ Available on the web (whatever format) but with an open license, to be Open Data
★★ Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★ as (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★ All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★ All the above, plus: Link your data to other people's data to provide context

図 Five Star Scheme Linked Data - Is your Linked Open Data 5 Star?より引用[4]

それによれば、一つ星はどのような形態であれ、オープンライセンスで公開されていること、二つ星は機械可読な構造化データとして利用可能（Excel 等が例示されている）にすること、三つ星は二つ星に加えて非独自フォーマット、すなわち特定のベンダー製品の使用を要求しない（CSV 等）ものにすること、四つ星はさらに W3C の標準規格である RDF と SPARQL を使用して記述すること、五つ星は、これらのデータが他のデータとリンクされる状態で公開されている状態である。オープンサイエンスに資する観点からは、データの再利用の障壁を下げる一方で研究者に負担がかからないような運用が望ましく、最低三つ星の水準を満たすかたちでの公開を求められれば十分であろう。公開データ形式選択にあたっての基準と事例については前項で考察している。

ただ、スタースキームでは機械可読可能なフォーマットに加えて非独自フォーマット、すなわちベンダー製品の使用を要求しないものを、より上質なデータとして位置づけている。このような論理では商用統計処理ソフトよりベンダー非依存の CSV 形式の方がよいという判断になる。しかしながら、3 つの理由から必ずしもそのような選択が最善とは限らないと思われる。一つ目は先述したとおりプロプライエタリーのものであっても仕様が公開され多くのツールから利用可能な状態であり、長期に渡って情報資産として活用できるものであれば容認しうると考える。もう一つの理由はより積極的なものであり、データそのもの以外にデータ構造、欠損値定義、ラベル、コメント等のメタデータがある場合、CSV に変換す

4 https://www.w3.org/DesignIssues/LinkedData.html

ることによってそれらの一部が失われる可能性があることである。そして最後の理由は、CSV はデータ構造が固定的であり、構造の変化に弱く長期に渡るデータの互換性を確保することが困難であることから、長期間にわたるプロジェクトやデータ構造の要件変更が発生するようなものには向いていない可能性がある。最終的なセマンティックスベースでのデータ処理に対応できるよう、大元のデータおよびメタデータの形態が最大限に保存され、なおかつ長期的にアクセスが保証されている形式を選択することが望ましいと思われる。

(ii)データ構造の標準化にむけて

公衆衛生分野でのデータの標準化は２つの方向性が考えられる。

一つは規制当局や国の事業のための情報収集を用途とし、二次利用性を高めるために内容の標準化をはかることである。CDISC がそのような用途の為の標準化をすすめており、この CDISC の SDTM において公衆衛生分野の語彙を開発、導入することに寄与していくことが考えられる[30]。

もう一つは、LOD/RDF の文脈にあわせて統計データとして公開するためのしくみとして、SDMX(Statistical Data and Metadata eXchange)[31]をベースとした RDF データキューブ[32]の形態での公開を目指すことである[33]。

いずれにしても、公衆衛生分野は LOD の事例は立ち後れている分野なので、研究データを保有している研究者と協業し、公衆衛生分野の LOD を構築する研究プロジェクトと予算獲得にむけた画策は必要であろう。

C.5　研究データの帰属、知的財産の取り扱いについて

・　研究データの帰属及び知的財産の取り扱いについて、国研の関係規程を踏まえた上で、研究データの利活用の方針に応じて記述する。この記述は、 保管に際して遵守すべきルールとして規定するとともに、同ルールと研究データ利活用のルールと整合を取る。 ・　研究データに係る作成者、管理者等の免責事項について記述する。

公的な研究資金にもとづいた研究活動、あるいは機構の施設・設備等を利用した研究、業務の過程で取得されたデータは、当該研究者との特別な取り決めがない限り、機構に帰属するものとされている。

C.5.1　データを公開する際のライセンスについて

データを公開するにあたり、基本的には FAIR の原則にもっとも適うライセンス形態として推奨されているクリエイティブコモンズの CC BY 4.0 に準拠して公開することが望ましいと考えられる。これは、オリジナルから改変された派生物が共有されることと、商用利用を許容するものである。

日本政府は各省庁が保有しているデータをオープンデータとして公開すべく、DATA GO.JP というデータカタログサイト[5]を提供しており、その利用規約では CC BY 4.0 国際でのライセンスが指定されている。

いわゆる「オープン」なライセンスにはオープンソースのライセンス群が知られているが、こちらは主にプログラム、ソースコードに適用されるものであり（もちろん、研究の過程で作成されたアルゴリズム、プログラムを公表するのであれば、オープンソースでの適用となろう)、データや論文にはそぐわないものである。

5 https://www.data.go.jp/

データに関して「オープン」なライセンスには同じくクリエイティブコモンズのCC0[6]、オープンデータ・コモンズの ODC-BY[7]が知られている。但し、CC0 は所謂著作権を放棄する Public Domain であり、我が国では著作権の放棄に関する明文の規定は存在せず、国立研究開発法人が適用するライセンスとしては不適切であると思われる。

ただ、メタデータに関する検討でも議論したように、機微な情報を含むデータを扱うこともあり、ただちにオープンできるデータのみではなく、個別に検討する必要がある。そのため、データポリシーにおいては一律のライセンスを指定するのではなく、FAIR の原則に則したライセンスを優先的に採用することを記述するにとどめることが望ましいと考える。

C.5.2商業利用への提供について

独立行政法人海洋研究開発機構のデータポリシー[34]では、産業利用については原則として有償として、適切な対価を徴収することが記載されている。本稿での検討の対象となっている国立研究開発法人においては、民間より何らかの対価を徴収するような運用についての検討に乏しいところがある。文部科学省の「研究開発成果としての有体物の取扱いに関するガイドライン[8]」において、成果有体物については商業利用を目的とした者への提供の場合は、「提供を要請する者と各機関との間において、成果有体物の取扱に関する必要な条件を明記した譲渡又は貸付契約を締結し、有償で提供する。」とあるとおり、有償での提供が一般的な扱いとなっている。一方、厚生労働省においてはそのようなガイドラインは現存せず、なおかつ法人化された国公立大学とも異なり、国の試験研究機関は現在でも「国」扱いのため、個別の検討が必要である。データポリシー策定と並行して、職務発明規程や就業規則の修正・追補を検討したい。

ただ、商用利用に関する制約を入れることは、CC における「オープン」の定義[35]に反する可能性がある[9]ため、先に推奨ライセンスとして提案した CC BY 4.0 と衝突しうることも念頭におかなければならない。

C.6　研究データの公開、非公開及び猶予期間並びに引用について

・研究データの公開について、機関の研究データの利活用の方針に応じてデータ公開までの猶予期間を適切に設定し、それに基づく公開時期について記述する。

・　公開データの利用に際しては、利用者に対して適切な引用を求める。その際、識別子を用いた引用情報の記載ルールを設けるなど、他のユーザーが引用元のデータを参照できるよう配慮する。

研究データは出版社や研究者自身によって設定されたエンバーゴ期間を過ぎれば、一般公開に支障がない限り遅滞なく公開するものとする。一部の出版社や学協会では出版後一定期間が経過するまで、Green OA で同一論文を公開することを認めない（embargo:公開猶予期間）が存在することがある。また研究者によって合理的かつ公益を損ねない範囲での研究の競争上の優位性を保つための公開までの猶予期間の指定がある場合は、その猶予期間の指定後にデータを公開するものとする。研究者のモチベーション維持のために、公益を損ねない範囲での研究者の利益を擁護することは必要である。

6 https://creativecommons.org/publicdomain/zero/1.0/

7 https://opendatacommons.org/licenses/by/summary/

8 http://www.mext.go.jp/a_menu/shinkou/sangaku/sangakuc/020901.htm

9 2.1.9 料金領収の禁止 ライセンスは、その条件の一部としていかなる料金支払いの取り決めや、ロイヤリティ、あるいはその他の補償行為あるいは金銭的代償を 規定してはならない。

公開データの利用に際しては、利用者に対して適切な引用を求めることとする。但し、リポジトリによるデータ識別子の永続的な管理体制とデータ引用の標準様式が普及していないため、当面は学術論文の投稿規程等で定められた様式で引用することを要求するところから始めるものとする。リポジトリを構築し、データに対して安定した識別子が付与されるようになれれば、識別子を用いた引用情報の記載ルールを設けることとする。

D. 参考文献

1. 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. *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*. 2018; Available from: https://www8.cao.go.jp/cstp/stsonota/datapolicy/datapolicy.pdf.
2. 内閣府 政策統括官（科学技術・イノベーション担当）. *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～解説資料～*. 2019; Available from: https://www8.cao.go.jp/cstp/stsonota/datapolicy/dpguideline.pdf.
3. Weibel, S., et al., *Dublin core metadata for resource discovery.* Internet Engineering Task Force RFC, 1998. 2413(222): p. 132.
4. Guides, A. *Publishing and sharing sensitive data*. 2018; Available from: http://www.ands.org.au/__data/assets/pdf_file/0010/489187/Sensitive-Data-Guide-2018.pdf.
5. Initiative, B.O.A., *Read the Budapest open access initiative.* Budapest Open Access Initiative, 2002.
6. Harnad, S., *Fast-forward on the green road to open access: the case against mixing up green and gold.* arXiv preprint cs/0503021, 2005.
7. 東京大学附属図書館, *オープンアクセスハンドブック 第2版.* 2017.
8. 課題領域：オープンサイエス(SCPJ)班, 機. *国内学協会のオープンサイエンス対応状況調査（報告）*. 2016; Available from: http://id.nii.ac.jp/1280/00000199/.
9. オープンアクセスリポジトリ推進協会. 2019; Available from: https://jpcoar.repo.nii.ac.jp/.
10. JPCOAR, *JPCOAR オープンアクセスリポジトリ戦略 2019～2021 年度.* 2019.
11. Van den Eynden, V., *Managing and sharing data: Best practice for researchers*. 2011: UK Data Archive.
12. Foster, E.D. and A. Deardorff, *Open science framework (OSF).* Journal of the Medical Library Association: JMLA, 2017. 105(2): p. 203.
13. 込山悠介, *研究データ管理サービス: GakuNin RDM.* 2019.
14. 国立情報学研究所オープンサイエンス基盤研究センター. *GakuNin RDM（研究データ管理基盤）*. 2017; Available from: https://rcos.nii.ac.jp/service/rdm/.
15. 坪井, 彩. and 治. 大須賀, *JST サービス紹介　国内最大級の研究者総覧 researchmap.* 情報管理, 2018. 60(12): p. 906-909.
16. 学術機関リポジトリ構築連携支援事業. *メタデータ・フォーマット junii2（バージョン 3.1）*. 2014; Available from: https://www.nii.ac.jp/irp/archive/system/junii2.html.
17. 大園, 隼., *サンメディアソリューションセミナー オープンサイエンスの最新情報 ： メタデータの相互運用性を中心に.* 薬学図書館 = Pharmaceutical library bulletin, 2017. 62(1): p. 40-47.

18. Group, D.M.W. *DataCite Metadata Schema 4.2*. 2019; Available from: https://schema.datacite.org/meta/kernel-4.2/doc/DataCite-MetadataKernel_v4.2.pdf.
19. オープンアクセスリポジトリ推進協会. *JPCOAR スキーマガイドライン Ver 1.0.1*. 2018; Available from: https://schema.irdb.nii.ac.jp/ja.
20. 大園 隼彦, et al., *JPCOAR スキーマの策定：日本の学術成果の円滑な国際的流通を目指して*. 情報管理, 2018. 60(10): p. 719-729.
21. 蔵川, 圭. and 英. 武田, *研究者識別子 ORCID の取り組み*. 情報管理, 2012. 54(10): p. 622-631.
22. Kurakawa, K., et al., *Researcher Name Resolver: identifier management system for Japanese researchers*. International Journal on Digital Libraries, 2014. 14(1-2): p. 39-58.
23. 独立行政法人日本学術振興会. *機関番号一覧*. 2019; Available from: https://www-shinsei.jsps.go.jp/kaken/topkakenhi/codelist-ka.html.
24. Science, D. *GRID Global Research Identifier Database*. 2019; Available from: https://www.grid.ac/.
25. Berners-Lee, T., *Five star open data*. 2009.
26. Lassila, O. and R.R. Swick, *Resource description framework (RDF) model and syntax specification*. 1998.
27. R Core Team. *R: A Language and Environment for Statistical Computing*. 2019; Available from: https://www.R-project.org/.
28. 奥村 晴彦, *「ネ申 Excel」問題*. 情報教育シンポジウム 2013 論文集, 2013. 2013(2): p. 93-98.
29. Tim Berners-Lee. *Linked Data*. 2010; Available from: https://www.w3.org/DesignIssues/LinkedData.html.
30. 木村 映善 and 上野 悟, *公衆衛生分野でのデータ利活用に貢献する標準医療情報規格と CDISC 標準*. 保健医療科学, 2019. 68(3): p. 212-218.
31. SDMX, S., *Statistical Data and Metadata Exchange*. URL: http://sdmx.org, 2011.
32. W3C. *The RDF Data Cube Vocabulary*. 2014; Available from: https://www.w3.org/TR/vocab-data-cube/.
33. 西村, 正., *Linked Open Data (LOD) による統計データの提供：政府統計データ（e-Stat）の新しい形*. 情報管理, 2017. 59(12): p. 812-821.
34. 独立行政法人海洋研究開発機構, *データ・サンプルの取り扱いに関する基本方針（データポリシー）*. 2007.
35. Foundation, O.K. *Open Definition 2.1*. Available from: http://opendefinition.org/od/2.1/ja/.