人文学・社会科学における
データ共有のための手引き

―人文学・社会科学データインフラストラクチャーの構築に向けて―

令和 3（2021）年 11 月

独立行政法人日本学術振興会

人文学・社会科学データインフラストラクチャー構築推進事業

目 次

1 はじめに

1.1 この手引きの目的

この手引きは独立行政法人日本学術振興会（JSPS）が平成 30（2018）年度から実施している「人文学・社会科学データインフラストラクチャー構築推進事業」[1] の成果の一部です。この事業は「人文学・社会科学研究に係るデータを分野や国を超えて共有・利活用する総合的なデータインフラストラクチャー（データ共有基盤）を構築することにより，研究者がデータを共有しあい，国内外の共同研究等を促進する」ことを目指しています。

この手引きは，事業の目指す方向に沿う形で構築されつつあるデータ共有基盤を念頭に置いて，データの共有・利活用そしてそれを可能にするデータの寄託を促進することによって人文学・社会科学分野の研究の振興を図ることを目的としています。

1　事業そのものは５年計画で実施されており，この手引きの策定以外に「人文学・社会科学総合データカタログ」（略称：人社データカタログ），英文では“Japan Data Catalog for the Humanities and Social Sciences”（略称：JDCat）の作成，オンライン分析システムの構築，そしてそれらの広報活動を行っています。詳細については JSPS のホームページ https://www.jsps.go.jp/j-di/index.html を参照してください。

1.2 この手引きが想定している読者

この手引きは，①これまで人文学・社会科学分野において主としてデータの収集・分析を中心に研究を行い論文などによってその成果を公開する活動を行ってこられた研究者，②人文学・社会科学分野の研究者を目指している大学院生など若手の人たちを想定して策定したものです。①の研究者にとって研究成果の独創性とともに先取権はもっとも重要な点であることは当然です。ここで言うデータの共有とは，それらを放棄するということではありません。その意味で研究に用いたデータの共有は，研究成果を公開した後に行うのが一般的でしょう。しかし，成果を公表して一段落した後に次の研究に移るためにはデータを共有するための作業の負担をなるべく軽くした方がよいことは明らかです。また，新たな分野や課題の研究を始める時には過去になされた研究で用いられたデータが利用可能であれば最初からデータの収集方法を考えるよりもずっと効率的に進められるでしょう。そのために,研究者がお互いにデータの共有･利活用できるようにアーカイブなどへのデータの寄託が容易に行うことができるようになっていることが望ましいでしょう。このように考えると，データ作成はその開始時点からなるべくデータの共有・利活用・寄託を意識した形式で作成することが重要です。

②の研究者を目指す大学院生など若手の人たちに対しては，第一に，近年のいわゆる ICT の発展は人文学・社会科学分野の研究においても大きな影響を与えていること，第二に，現時点で大学院生などである立場の人たちが研究者として独立するときにはデータの共有・利活用・寄託がほぼ必須の条件になっているであろうことを指摘しておきたいと思います。

これらのことを一言で言うと，データの収集・作成から適切な管理そして長期的な維持・保存は信頼できる研究成果の検証と新たな研究を可能にするデータ基盤であって，人文学・社会科学分野における研究の発展の基礎となるものです。

1.3 この手引きの策定の経緯と特徴

この手引き策定の根拠となっている事業では，独立行政法人日本学術振興会（JSPS）が中核機関となり，以下の 5 機関が拠点機関として活動しています。

東京大学　　社会科学研究所附属社会調査・データアーカイブ研究センター
一橋大学　　経済研究所
慶應義塾大学　経済学部附属経済研究所パネルデータ設計・解析センター
大阪商業大学　JGSS 研究センター
東京大学　　史料編纂所

これら 5 機関のうち東京大学社会科学研究所附属社会調査・データアーカイブ研究センター，一橋大学，慶應義塾大学，大阪商業大学の 4 機関は平成 30（2018）年度から，東京大学史料編纂所は令和元（2019）年度から活動を始めており，現在これらの機関はそれぞれが保有する独自のデータを基にデータアーカイブを構築しています。

この手引きは中核機関である独立行政法人日本学術振興会（JSPS）が拠点機関の活動と並行する形で策定したものです。したがって，記述にあたっては拠点機関の進捗状況に合わせて，まず社会科学分野のデータに関する記述を行い，人文学分野のデータに関して記載した方がよいものについては，注記・コラムなどで補足する形式を取っています。

ただし,社会科学分野のデータでも量的な社会調査データおよび公的統計のみを取り上げており，事例調査研究やフィールドスタディなどによる質的なデータは含めていません。また，データの性質によってさまざまな制約があり，それにともなう取扱いの違いがあるため，ここではあくまで一般的な事項について記述しています。さらに，各拠点機関のデータ収集に関する個別具体的なルールなどには言及していません。研究者のなかで適当と思われる機関に寄託をお考えの方は事前に寄託希望の機関にルールを確認するようにお願いします。

この手引きはデータ共有基盤構築にともないデータの共有・利活用・寄託が進展することを目指したものであり，あくまで最初のステップです。とはいえ，この手引きによって，研究者のあいだでのデータの共有・利活用さらには寄託が進み，それが大学院生など若手の人たちの糧となり人文学・社会科学分野の研究のより一層の進展に寄与することを願っています。

最後に繰り返しになりますが，この手引きによって人文学・社会科学分野の研究者や大学院生など若手の方々にデータの適切な管理とデータの共有・利活用・寄託の重要性を認識していただき，データの共有・利活用・寄託が進むことにより人文学・社会科学分野の研究が国内的にも国際的にもより一層進展することを期待します。

2 データを共有する意義

自然科学と同様，人文学・社会科学においても，研究結果の根拠となるデータを研究者のあいだで共有できるようにすることは，研究者コミュニティにおける相互批判と研究の発展にとって不可欠のものです。この「手引き」では，「共有」とは，既存のデータを研究者など特定の条件を満たす主体に利用できるようにすることを指し，無制限にだれもがアクセスできる「公開」と区別しています。特に公的資金による学術研究の成果物として作成されたデータセット[1]は，データアーカイブに寄託するなどの方法で，研究者コミュニティのなかで共有されることが望まれます。ここでは，まずデータを共有する意義と，その方法としてデータアーカイブにデータを寄託する意義について述べます。

1　調査において収集された情報の総体のこと。複数のデータファイルを有することがありえる点で，データファイルとは区別されます。

2.1 再現性の保証

データを共有する意義の第一は，分析結果の再現性（replication）を保証することにあります[2]。再現性とは，分析者が使用したのと同じデータに同じ分析手法を適用すれば同じ結果が得られることを意味します。研究成果の根拠となるデータに他の研究者がアクセスできる状態にあることによって，分析結果の再現性が担保されます。海外のジャーナルのなかには，論文に使われているデータが，他の研究者にとってアクセスできるような状態にあること，より具体的には，信頼できるデータアーカイブに寄託されていることを，投稿の条件としている場合もあります[3]。

2　ICPSR , 2020, *Guide to Social Science Data Preparation and Archiving: The Best Practice Throughout the Data Life Cycle·6th edition*, Ann Arbor, MI: University of Michigan, (Retrieved June 22, 2020, https://www.icpsr.umich.edu/files/deposit/dataprep.pdf).

3　Eynden, Veerle Van den, Louise Corti, Matthew Woollard, Libby Bishop and Laurence Horton, 2011, *Managing and Sharing Data: Best Practice for Researchers*, Colchester, Essex: UK Data Archive University of Essex, (Retrieved February 19, 2020, https://ukdataservice.ac.uk/media/622417/managingsharing.pdf).

2.2 二次分析・二次利用

データを共有する意義の第二は，データの二次分析・二次利用によって，学術研究と教育に貢献できる点です。

二次分析（secondary analysis）とは，すでに分析がなされている既存のデータを再利用して，新たな視点や分析方法にもとづいて分析することをいいます（これに対して，二次利用（secondary use）とは，より広義に既存のデータを当初の調査・研究目的以外に利用することをいいます）。データが共有されることで二次分析が可能となり，既存のデータを活用した新たな研究が促進され

ます。同一のデータをもとに，新たな仮説や競合する仮説を検証したり，新たな分析手法を試したりすることができるようになるのです。

さらに，データを共有することにより，研究者が新たな調査を企画する際の参考としても役立てることができます。たとえば，特定のワーディングや回答選択肢を使った場合に，回答がどのように分布したかをあらかじめ知ることは，新たな調査の企画にとって参考になります。

また，データの共有により，異なる研究者または研究組織が重複するデータの収集やデータセットの作成をしなくてすむのであれば，研究者と研究参加者（被験者・調査協力者）の負担を軽減でき，全体として研究費の節約にもなります。

最後に，データの共有により，教育用に質の高いデータを提供することができ，統計教育・調査教育の発展にも貢献することができます。

これらの点から，データを二次分析・二次利用に供することによって，データの収集やデータセットの作成自体も，研究・教育への貢献として評価されることになるのです。

2.3 データアーカイブに寄託する意義

研究者のあいだでデータを共有できるようにするために，個々の研究者や研究組織が個別にデータを保存し管理することも可能です。しかし，データを保有している側のデータ保存・管理・提供にかかる業務負担とその持続可能性，そしてデータを利用する側のデータへのアクセス可能性などの観点から考えると，データアーカイブに寄託して，データを共有できるようにすることが望まれます。「寄託」（deposit）とは，収集したデータなどをデータアーカイブに預けて，データセットの保存・管理と共有にかかる業務を依頼することをいいます。

データアーカイブに寄託することには次のような利点があります。

- 標準化されたデータカタログの提供により，利用者にとって検索が容易となります（→第 4 章　メタデータ）。
- 標準化されたデータフォーマットでデータを保存することができます（→第 5 章　データのフォーマット）。
- 本来のデータ保有者に代わって，データを長期にわたって組織的に維持・管理することが可能となり，多くの利用者の利用申請に対応することができます（→第 10 章　データアーカイブの役割）。
- データが寄託されていることで，データの質が保証され，データを寄託した研究者・研究組織の研究の信用度・認知度を高めることができます。

社会調査などによって収集したデータを，データアーカイブに寄託するには，データセットや関連情報（メタデータなど：「第 4 章　メタデータ」を参照）を標準的なフォーマットで準備しておくなど，一定の条件を満たす必要があります。これらの条件の詳細はデータアーカイブによって異なりますが，標準的な部分については以下で順次，説明していきます。

【人文学向けコラム 1（人文学におけるデータ共有の意義）】

データ作成の有用性は人文学においても同様です。研究資料を整理・分析するにあたり研究している時点での効率や気づきの可能性を高めてくれるだけでなく，再利用しやすい形で作成しておけば，将来の自分の研究を発展させるにあたって有効に働いてくれます。さらに，そのように

して作ったデータが適切な形で共有されたなら，それは社会科学のデータと同様に研究コミュニティを支える確かな基礎となるだけでなく，人間文化全般に関する応用的・領域横断的な研究全般の発展を促すことにもつながります。一方で，研究倫理がより重視されつつある状況において，データの共有により研究成果の再現性を保証することは，研究者コミュニティが自律的な評価体制を持つことを社会に示していく上で重要であり，人文学においてもこのことと無関係であり続けることは困難でしょう。

本コラムで扱うのは，「第１章　はじめに」に述べられているように，①これまで人文学・社会科学分野において主としてデータの収集・分析を中心に研究を行い論文などによってその成果を公開する活動を行ってこられた研究者，②人文学・社会科学分野の研究者を目指している大学院生など若手の人たちのうち，データを作成する際に手引きを必要とする人々のための最初の手がかりです。さらに，人文学には様々な分野がありますが，ここではそのなかでも特に歴史的文献資料の研究に際して作成されるデータについての情報提供を行います。

なお，ほとんどの分野がそうであるように，研究に用いたデータの共有は，研究成果を公表した後に行うべきことです。成果の先取権は研究者にとって極めて重要なものであり，データの共有は，それを放棄するということではありません。しかし，成果公開の後にデータを共有するとなると，公開した後には次の研究に移るべきところを，共有のためのデータ整理という作業をすることになります。その手間は少しでも少なくした方がよいので，データ作成の時点から，なるべくデータ共有を意識した形式で作成することをおすすめします。

3 データ管理計画

3.1 データ管理計画とデータのライフサイクル

データ管理計画（Data Management Plan : DMP）とは研究プロジェクトなどにおけるデータをいかに管理するかを定めたものです。具体的にはデータの種類，フォーマット，アクセスおよび共有のための方針，研究成果やデータが誰の責任のもとでどこに保管するかなどに関する計画について記載されているものを意味します。

研究助成への申請，申請された研究計画にもとづくデータの収集・分析，データアーカイブへの寄託，寄託されたデータの二次利用による新たな知的価値の生成までをデータのライフサイクルとよびます。データのライフサイクルを通して常に考慮されるべきは寄託されたデータの二次利用の利便性と客観性の担保です。以下に各段階における考慮事項の概略を述べます。

（1）申請書作成とデータ管理計画

研究助成を獲得するための申請書作成の段階から，寄託するデータアーカイブを念頭にその仕様などの概略を理解すること，そして，研究終了後も長期にわたってそのデータが第三者に利用されやすいようにデータの管理計画を立てることが必要です。したがってデータ管理計画は研究計画の中に組み込まれるべき必須の構成要素と考えることができます。

研究助成が採択された後は目的とするデータが確実に収集できるよう，予備調査・試行調査などを実施し，あわせてデータ寄託を念頭に第三者に理解されやすい文書の構成や内容の方向性を決定しておくことが必要です。なお，収集されるデータがそもそもデータアーカイブへの寄託に見合う価値を持つか否かを判断し，研究費には寄託のためのコストを計上しておくことも重要です。

なお，国内資金配分機関においても，「公的資金により行われる研究開発から生じるデータなどは国民共通の知的資産でもあり，現状では把握できていないデータの所在などを把握し，データの収集，質の確保，意味づけ，保存と活用などが適切かつ公正に行われるよう推進する役割がある」[1]とのデータ管理計画の意義を認め，たとえば，国立研究開発法人日本医療研究開発機構（AMED）や国立研究開発法人科学技術振興機構（JST），経済産業省産業技術環境局など，データ管理計画書の提出を要求するところが増えてきています。

（2）データ収集・作成と説明文書の準備

データとともにそれを説明する情報を一括してファイルに生成するために以下の注意が必要です。データの収集・作成にあたってはデータセット全体の論理的整合性をとることが必要で，その際，十分に考慮すべき諸要素としては変数名やラベルの付け方，変数の分類，コーディングの際の明確なルール，欠測データの定義，ファイル形式の選定などがあります。また，データの説明のための文書作成に必要な参考文献，引用文献，資料などの準備もあわせて進めておくことも必要です。

（3）データ分析

データ分析を遂行するにあたってはマスターのデータセットと作業過程で派生するワークファイル類とが錯綜混乱しないようデータ管理を確実に行うこと，ファイル間に適切で理解しやすい論理

構造をもたせること，事故に備えてデータや文書のバックアップを必ずとりながら作業をすすめることが必要です。

(4) データを共有するための準備

研究参加者（被験者・調査回答者など）の秘匿情報が漏洩する危険性を最小化すること，匿名化に十分配慮すること，他の利用者が読めるように，寄託するデータセットのファイル形式を決定すること，寄託するデータアーカイブに事前相談をして細部にわたって検討し，不整合やトラブルの防止に最善を尽くすことが必要です。

(5) データの寄託

データの寄託に際しては，データアーカイブが定める様式すべてを適合させること，データの頒布規格とフォーマットについてもデータアーカイブの指定に従うことが必要です。

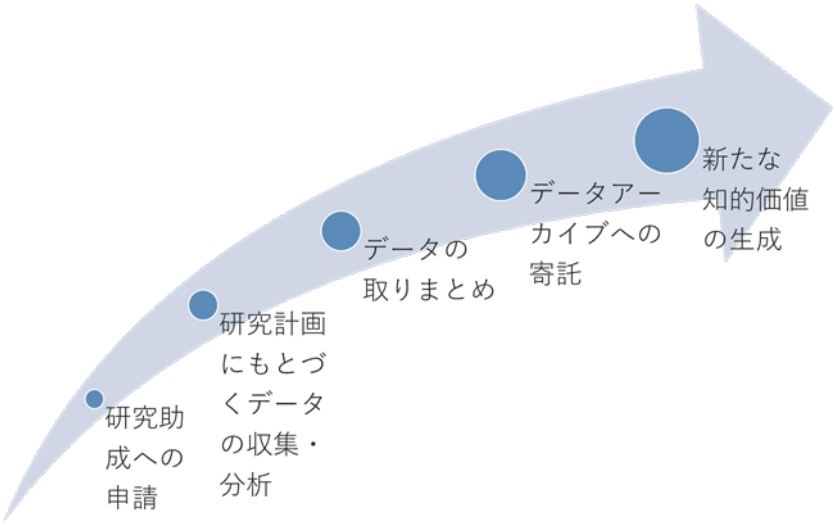

図　データのライフサイクル

1　国立研究開発法人日本医療研究開発機構，2020，「データマネジメントプランの提出について」，国立研究開発法人日本医療研究開発機構ホームページ，（2020 年 10 月 19 日取得，https://www.amed.go.jp/content/000061339.pdf）.

3.2 役割と責任

データ管理には研究者だけが関わるわけではありません。研究プロセスにはさまざまな役割と責任を担った関係者が関与しており，その役割と責任に応じてデータの管理と共有についても何らかの役割を担う場合があります。その場合には各関係者に明示的にその役割と責任を割り当てる必要があります。データの管理と共有の関係者としては以下のようなカテゴリーが例としてあげられます。

- 研究プロジェクトの責任者
- データの管理責任者
- データの収集，処理，分析の担当者
- データの収集，処理，分析の外部請負業者

- 研究事務および会計業務などの支援スタッフ
- データ保存などのサービスを提供する組織内 IT サービススタッフ
- データの共有を支援する外部スタッフ

3.3 データ管理計画

(1) 研究計画の一部としてのデータ管理計画

データ管理計画とは研究プロジェクトなどにおけるデータの取り扱いを定めるものであり，具体的にはデータの種類，フォーマット，アクセスおよび共有のための指針，研究成果の保管に関する計画などについて記載されるものです。データ管理計画は研究計画の一部を構成します。すなわちデータ管理計画は，

1) データがどのような方法で収集されるのか
2) 研究プロジェクトが進行する中でそのデータがどのように利用され保管されていくのか
3) 研究プロジェクト終了後は第三者がそのデータを再利用するために，長期間にわたっていかにその利便性を担保するのか

を明確にして立案されることが求められます。

そのために，研究プロジェクトの基本原則と進め方をデータ管理計画の作成前に確定しておくことが重要です。このようにデータを実際に収集する前にデータのライフサイクルの各段階におけるデータ管理計画を入念に準備しておくことによって，各段階相互に重複する無駄なデータ管理の労力やコストを削減し，情報の損失を防ぐことができます。もちろん研究プロジェクトが進行する途中で研究計画自体の修正が必要な場合も生じます。そのような時にはデータ管理計画を調整・変更することは可能です。

(2) データ管理計画の目的

データ管理計画は，優れた科学的実践としての研究で得られたデータが第三者によって再利用されることで，オリジナルの研究には無かった新たな科学的知見が得られる可能性や，オリジナルの研究自体にさらなる科学的担保を与える可能性を開く意味でも重要です。その目的を果たすためには，第三者によって再利用されやすいようにデータを長期にわたって安全に保護し，かつデータの共有可能性を最大限保証する必要があります。また，優れたデータ管理計画のもとに蓄積されたデータの再利用は経済的であり，リソースの節約にもなります。重複したデータ収集を回避することにも役立ち，研究方法の改善，教育と学習のための重要なリソースを提供することにもつながります。さらにはデータをデータアーカイブに寄託する姿勢自体が研究者のアカデミックな信用を熟成することにもつながり，アーカイブされたデータを再利用して公刊される研究への潜在的な信用も増すことになります。

(3) データ管理計画立案段階

データ管理計画を立案するにあたっては，

1) 研究倫理に関するレビューをすること
2) 必要な場合には研究倫理審査を受審すること
3) 当該研究に関する情報へのアクセス手段とデータの保存方法を明確にすること
4) 第三者がデータを再利用するための寄託先へのアクセス方法を指定すること

などにも配慮することが必要です。

(4) データの種類

管理対象となるデータには数量的データの他にもさまざまな種類のデータや資料，またいくつもの媒体手段（復元・再生手段の指定を含む）が存在します。技術発展によってそれらには新しい種類や手段が次々に追加されていくと見込まれます。そのため，管理計画作成にあたっては長期の再利用と共有可能性を担保する視点をおろそかにすることはできません。

(5) データの収集方法

データの収集方法は主にリサーチ・クエスチョンによって規定されます。質問紙調査，インタビュー[2]，ドキュメント渉猟，歴史的資産物の収集，Web 調査等々様々なものがあります。大規模な研究プロジェクトでは，実査を調査会社が担当するなど，再委託先が存在することもあります。その場合にはデータ管理計画の中に再委託先およびその役割を明記し記録しておくことが必要です。

(6) 著作権などの権利義務関係

著作権などの権利義務関係が研究に関連することもあります。たとえば，中間生成物の使用権，最終成果物の使用権，知的財産権が研究担当者の誰にあるいはどの機関に所属するかなどの場合です。さらにはデータの管理責任が誰にあり，データへのアクセス権は誰がどこまで持つのか，そのアクセスの許可権限は誰に有るかなどです。また，そうした権利義務関係を研究プロジェクトの参加メンバーに周知する手順もデータ管理計画の中に記述しておく必要があります。著作権法などの法律関係の知識をもった専門家の助言が必要になることもあります。なお，著作権についての詳細は，「第 9 章　データに関する著作権」を参照してください。

(7) その他

調査対象者情報の秘匿に細心の注意を払いながら，寄託するデータの中に，調査の際の依頼状や研究倫理審査結果なども含めておくことも忘れてはならない点です。

2　情報を収集するために行われる聞き取りや会話のこと。その構造化の度合いから，構造化インタビュー，半構造化インタビュー，非構造化インタビューが区別されます。特に断りなく単独で使われる場合，この手引きでは，非構造化インタビューを指します。

【人文学向けコラム 2（人文学におけるデータ管理計画）】

人文学のデータにおいても，データ管理計画は重要です。この手引きを参考にしつつ，データを作り始める前に計画を立てることをお勧めします。大規模なプロジェクトだけでなく，個人やごく少人数で行われる場合にもデータ作成における各段階の考慮すべき事項は同様ですので，一人が複数の役割を担うことを前提として検討することになります。また，人文学の場合，本コラムでは今回は扱いませんが，画像や動画など，テキストデータの保存に比べると比較的容量の大きなデータを扱うこともあり，その場合，比較的単純な作業であっても数日もしくはそれ以上かかることもありますので，そういったデータを含む計画の際にはその点も見込んでおく必要があります。

4 メタデータ

〔社会科学編〕

4.1 メタデータとは

メタデータは「data about data」とも定義され，データセットのコンテンツ，コンテキスト，出所などに関するデータです。データカタログ[1]で用いられるメタデータ（以下，単に「メタデータ」と記述します）には，データセット作成の目的，データの出所，対象となる期間，地理的な対象範囲，作成者，アクセス条件，使用条件などが記述され，ユーザーが既存のデータセットを見つけたり，特定のデータセットが研究目的に適しているかどうかを判断したり，データを引用するための書誌レコードを提供したりする際に用いられます。

また，一般にデータカタログで用いられるメタデータは標準化，構造化されたスキームによって制御されており，機械による読み取りが可能です。機械可読性の高いメタデータを作成することによって，データセットが持つ様々な切り口から検索が可能になり，精度の高いデータ検出が実現できます。

ここでは，データセットをカタログで検索するために重要なメタデータと，データセットを入手後に利用者がデータの内容を理解するために必要となるメタデータに分けて説明します。

1　データについて，そのメタデータを集めて一覧できるようにしたもの。データの収集，整理，保存，提供などに用いられます。

4.2 データの記述

本節では,社会科学分野におけるデータの概要を記述し,カタログ上の検索で重要となるメタデータ項目について紹介します。

・タイトル

調査名，あるいはデータセット名。研究費申請段階の研究題目と研究者が通称として用いる調査名とが異なることもありますが，知名度の高い名称を採用するほうがよいでしょう。

・作成者

作成者は，データの内容を担当した個人であり，多くの場合，研究プロジェクトの統括者（ディレクター）／調整者（コーディネーター）を指します。

・バージョン情報

データセットのバージョン情報。アーカイブから提供するデータについては，アーカイブ側で共有するデータのバージョンを付与します。共有されるバージョンと，研究者が寄託した段階のバージョンとの異同は，アーカイブ側で記録を保存します。

・トピック

研究関心に合うデータセットの発見を容易にするために，キーワードを付与します。自由にキーワードを選ぶ場合もあれば，データアーカイブから用語の一覧が提供され，そこから選択する場合もあります。

・概要

データが収集された研究，理論的枠組み，研究中の概念の運用に関する情報。多くの場合，データを探している研究者は概要を読んで，そのデータが自身の研究にとって有用であるかを判断します。

・対象時期

収集されたデータが対象とする時期。

・対象地域

収集されたデータが対象とする地理的範囲。

・観察単位／分析単位

観察単位はデータ収集や調査の基礎となる単位であり，たとえば，個人，組織，世帯，イベント／プロセス，地理的単位（行政区域上の区分），またはテキスト単位（新聞記事など）があり得ます。仮に個人に対してインタビューが行われたとしても個人が分析の単位でない場合もあります。たとえば，個人がインタビューされたとしても，分析単位は個人が所属する組織となる可能性があります。

・母集団

研究の対象となる，少なくともひとつの共通特性をもつ人，物，事象などの全体集合。標本調査の場合，そこから抽出された標本の観測値にもとづいて研究者が推測し一般化しようとしている対象のこと。

・サンプリング方法

母集団を代表する研究対象者の抽出に使用される，サンプリングの方法とデザイン。サンプルサイズおよび回収率への参照を含むことがあります。必要に応じて，サンプリングの手順説明を含みます。

・調査方法

データの収集方法（たとえば，個別面接法，電話法：コンピュータ支援（CATI），Web ベースのインタビュー，音声記録，視聴覚記録，または個人的な経験についての書き込みなど）を記録します。

・アクセス権

データセットのアクセス状態（利用制限など）に関する情報。一般には，データの性質に応じてデータアーカイブ側の方針により設定されます。

・権利情報

データセットの利用に関する権利情報を記入します。利用者の属性（大学院生，研究者など）による利用の可否，目的（教育／研究）による利用の可否などの条件のほか，知的所有権や著作権などに関する法的な制約情報を含みます。

・研究助成機関

データの収集またはデータが収集されたプロジェクトに資金を提供した財団や資金配分機関。必要に応じて，研究費番号も記録します。

・関連文献（データに基づく文献）

データを使用した，あるいは説明した文献・出版物のリスト。

・データソース

データが調査やインタビューを通じて収集されず，既存のデータソース（たとえば，書籍，記事，登録データ，ブログや Twitter など）に基づいている場合は，情報源を記録します。

・背景情報

調査概要，調査主体／調査代表者，研究助成機関，母集団，調査方法に関して，これらのメタデータ項目には直接記述されないものの，データ取得時に影響があったと考えられる背景情報を記録します。

【データおよびメタデータに付与する識別子】

識別子とは，様々な対象から特定の一つを識別，同定するために用いられる名前や符号，数字などを指します。ここでは，学術論文やデータに付与する識別子として，学術コミュニティにおいて広く用いられている DOI（デジタルオブジェクト識別子）を紹介します。

DOI はデジタルネットワーク上でコンテンツへのアクセスを管理するために用いられる国際的な識別子「デジタルオブジェクト識別子（Digital Object Identifier）」の頭字語です。https://doi.org/ に続けて DOI をブラウザに入力することで，自動的にコンテンツの所在情報（URL）に変換されるサービス名称でもあり，登録機関が DOI に紐づく URL のメンテナンスを行うことで，利用者からの恒久的なアクセスが実現されます。

たとえば，https://doi.org/10.1109/5.771073　と入力すると，
https://ieeexplore.ieee.org/document/771073
という URL に変換されます。

もともとは出版社が発行する論文に対して付与され始めたものですが，現在では論文に留まらず様々な学術コンテンツに付与されています。詳細については，以下のページを参照してください。

参考：DOI ハンドブック
https://www.doi.org/doi_handbook/translations/japanese/hb.html

4.3 データファイルの記述

前節までの情報は，データカタログを経由してデータを探す段階で利用者にとって重要となるものです。本節以降では，データの長期保存の観点および二次分析を行うものがデータを利用する際に重要となる情報を説明します。これらの情報は伝統的には報告書，あるいはコードブック[2]という冊子形態で準備されてきました。

まず，本節では，データファイルの記述について紹介します。将来的なファイルフォーマットの変換に備え，ファイル内のすべてのプロパティをメタデータとして記述する必要があります。次の項目を記述するとよいでしょう：

- ファイル名
- ファイルのパス（URL）
- サイズ
- フォーマット
- ファイル作成に用いたソフトウェア
- 作成日
- 作成者
- バージョン情報
- アクセス権限
- テキストデータの場合，採用した文字コード・文字エンコーディング[3]

各項目の詳細な記述方法については，「第 5 章　データのフォーマット」を参照してください。また，ファイルプロパティの具体的な確認方法はお使いの OS によって異なります。たとえば，Windows の場合は対象のファイルを右クリックし，「プロパティ」を選択することで上述した情報の多くを確認できます。

2　データを利用する際に必要な情報をまとめたもの。ファイル上のどの位置の記号（通常，数字）がなにを意味するかを示します。

3　コンピュータ上で文字をデータとして効率的に扱えるようにすべく，文字集合を策定して個々の文字に番号を割り当てた表を含む対応規則をまとめたものが文字コードです。日本語用の文字コードを規定する JIS X 0208 やその次の規格である JIS X 0213，世界中の文字を対象とした ISO/IEC 10646 など，標準化団体が定めているものが広く用いられています。Unicode は民間団体である Unicode 協会が定めていますが，現在は ISO/IEC 10646 に追従しておりほぼ完全な互換性があります。Unicode の符号化方式としては，Web では UTF-8 がよく用いられますが，UTF-16 も健在です。一方，JIS X 0208 や JIS X 0213 に対応する符号化形式として Shift JIS，EUC-JP，ISO-2022-JP などがあり，Unicode 以前の日本語データを扱う際にはほとんどがこれらのいずれかであり，現在も多少使われています。

4.4 変数の記述

データセットは 1 つ以上のファイルで構成され，定量的なデータを扱うファイルには，通常数十〜数百の変数が含まれます。一方，多くの定性的なデータにおいて，ファイルに含まれるデータ単位は 1 つだけです。以下では，定量的なデータセットの変数に関する記述例を挙げます。

変数の記述は，基本的には質問文，選択肢，選択肢に割り振られた数値，それらの頻度を示すことにより成り立ちます。質問文は変数ラベル，選択肢は値ラベルに対応しますが，字数制限などの

関係から，質問文や選択肢をそのままラベルにできないことがあります。また，欠測値についてもその頻度を示すことは重要です。

＜実例＞
Q7 あなたは政治にどの程度関心がありますか。次の選択肢から最も近いものをお選びください。

変数ラベル　　問 7　政治への関心度

	値	値ラベル	カウント
有効値	1	とても関心がある	263
	2	やや関心がある	858
	3	あまり関心がない	481
	4	まったく関心がない	82
	6	調査票記入漏れ	1
	7	答えたくない	3
	8	わからない	0

従来の統計ソフトウェアでは日本語が扱えなかったり，字数制限が厳しかったりしましたが，最近は SPSS や Stata などの統計ソフトウェアを用いてデータに変数ラベルや値ラベルを貼れば，そこから変数ラベルを記述するメタデータを自動的に作成することもできるようになっています。

なお，複数の変数から合成した尺度を寄託データに含める場合は，変数間の関係を明示しておく必要があります。

・分類体系に関する情報

使用される分類に関する情報を記録します。たとえば，「〇〇の主なカテゴリーには職業分類を使用した」「国名コード（3 桁）は ISO 3166-1 に従っている」などが挙げられます。

4.5 調査過程で作成される様々な文書と情報の保存

調査の過程では様々な文書が作成されます。標本調査を実施する場合は，研究費申請書・研究計画書に始まり，サンプリング・デザインを記述した文書，実地調査を調査会社に委託した場合の契約書，質問票，回答票，自治体への選挙人名簿あるいは住民基本台帳閲覧願，調査対象者への依頼状，実際に聞き取りを行う調査員への指示書などがあります。研究計画書をそのまま保存することはまれだと思われますが，調査過程に関する他の文書は，いずれも調査の背景やデータの構造を理解する上で重要となりますので，データと同様に保存することが重要になります。そしてこれらの文書は，できることならば，報告書やコードブックの補遺として二次利用者の閲覧に供することが望まれます。実際の質問紙がないコンピュータ支援調査の場合，質問が表示された画面のイメージを画像として保存すると同時に，質問と回答の選択肢，およびそれらが提示された順序をテキストファイルとして保存しましょう。

調査票サンプル事例：

https://ssjda.iss.u-tokyo.ac.jp/chosa-hyo/PH010c.html

【メタデータスキーマ（社会科学編）】

ここでは，メタデータの機械可読性を高めるための仕組みについて紹介します。

メタデータの記述項目や形式，語彙の定義，項目間の階層構造などを定義したものをメタデータスキーマと呼びます。事前に定義されたメタデータスキーマを用いることで，異なる機械間でもメタデータとして記述された項目の意味が一意に定まるため，精度の高い横断検索が可能になります。メタデータスキーマは，実現したいサービスの目的やコミュニティのニーズに合わせて設計され，国際標準や業界標準といった形で公開されています。ここでは，社会科学分野で使用される主なメタデータスキーマを紹介します。

◇社会科学分野で使用されるメタデータスキーマの事例

Dublin Core（ダブリンコア）メタデータスキーマは，オンラインリソースを記述するための基礎となる 15 のフィールド（Dublin Core Metadata Element Set）で構成されています。データセットのタイトル，作成者，主題，識別子，参照関係，時間，場所，利用条件などを表現することができます。

https://www.dublincore.org/resources/userguide/creating_metadata/

DataCite（データサイト）メタデータスキーマは，引用および取得のために，リソースを正確かつ一貫して識別できるよう選択された基礎的なメタデータプロパティです。Dublin Core の情報に加え，より統制されたリソース種別や地理的情報を表現することができます。

http://schema.datacite.org/

DDI（データ・ドキュメンテーション・イニシアティブ）は，社会科学，行動科学，経済科学，および健康科学の調査やその他の観察方法によって生成されたデータを記述するための国際標準です。 XML で表現された DDI メタデータ仕様は，データのライフサイクル全体をサポートしています。

https://DDIalliance.org/explore-documentation

JPCOAR（ジェイピーコア）スキーマは，オープンアクセスリポジトリ推進協会（JPCOAR）が策定した新しいメタデータ規格です。日本の機関リポジトリのメタデータの国際的な相互運用性を向上させ，日本の学術的成果の円滑な流通を図ることを目的としています。

https://schema.irdb.nii.ac.jp/ja

その他，統計分野におけるデータ交換の国際標準を定めた SDMX（Statistical Data and Metadata eXchange），デジタルライブラリの資料に関するメタデータ標準の METS（Metadata Encoding and Transmission Standard），ネットワーク環境におけるアーカイブ資料の記述を定めた EAD（Encoded Archival Description）など，目的に応じて様々なメタデータスキーマが用いられています。

本コラムの冒頭で触れたように，メタデータスキーマは，メタデータ自体の機械的な解釈可能性を担保する機能を持っています。作成したメタデータを長期的に維持するために，分野の国際標準に合致したメタデータスキーマを選ぶとともに，メタデータスキーマ自体の後方互換性にも留意するとよいでしょう。

〔人文学編〕

4.6 メタデータとは

メタデータは「data about data」とも定義され，資料の書誌・目録情報のことを指すために用いられる一方で，それらを何らかの範囲でまとめたデータセットを説明する場合にも用いられます。データセットの場合には，コンテンツ，コンテキスト，出所などを表現する手段です。データカタログで用いられるメタデータ（以下，「メタデータ」といいます）には，データセット作成の目的，データの出所，対象となる期間，地理的な対象範囲，作成者，アクセス条件，使用条件などが記述され，ユーザーが既存のデータリソースを見つけたり，特定のデータセットが研究目的に適しているかどうかを判断したり，データを引用するための書誌レコードを提供したりする際に用いられます。

また，一般にデータカタログで用いられるメタデータは標準化，構造化されたスキームによって制御されており，機械による読み取りが可能です。機械可読性の高いメタデータを作成することによって，データセットが持つ様々な切り口から検索が可能になり，精度の高いデータ検出が実現できます。

したがってここでは，データセットをカタログで検索するために用いるメタデータ（4.7）と，データセットを入手後に利用者がデータの適切な利用方法を理解するために必要となるメタデータ（4.8），それに加えて，研究対象となる資料の書誌・目録情報としてのメタデータ（4.9）を分けて説明します。（なお，デジタル化資料を扱う場合に用いられる「メタデータ」という言葉は，多くの場合，研究対象となる資料の書誌・目録情報を指しますので，その点はご注意ください。）

4.7 データカタログにおけるメタデータ

本事業におけるデータカタログでのメタデータ項目は JDCat メタデータスキーマとして設定されています。以下にその一部を見てみます。

・タイトル

データセットの名称。

・作成者

作成者は，データの内容に責任を持つ個人であり，多くの場合，研究プロジェクトの統括者（ディレクター，研究代表者，PI など）を指します。必要に応じて，データ収集者，データ入力者，そ

の他のデータ作成作業処理者（たとえば，データ構造設計者，データ校正担当者，データ編集作業者），などもその役割とともに記録します。

・バージョン情報

データセットのバージョンを記述します。これにより，アップデートがあった場合に利用者が差異について判断できるようにします。

・トピック

データセットの発見を容易にするためのキーワードとなるもの。利用者が想定しやすいものとしておく必要があるため，本事業では，国内の図書館での図書の分類に広く用いられている日本十進分類法（NDC）第 10 版の第 2 次区分表，第 3 次区分表の一部（700 番台，900 番台）を用いることとしました。

・概要

データが収集された研究，理論的枠組み，研究中の概念の運用に関する情報。

・対象時期

収集されたデータの時間的範囲に関する情報。

・対象地域

収集されたデータの地域に関する情報。

・データタイプ

収集されたデータのタイプ。

・アクセス権

データセットのアクセス状態（利用制限など）に関する情報。

・権利情報

データセットの利用に関する権利情報を記入します。利用者の属性（大学院生，研究者など）による利用可否，目的（教育／研究）による利用可否などの条件のほか，知的所有権や著作権などに関する法的な制約情報を含みます。

・データの言語

データに含まれる言語の情報。資料には様々な言語が混在することも多く，完全な記述は難しい場合もあるので，必ずしもそれを目指す必要はありません。

4.8 データの説明文書

「第 5 章　データのフォーマット」,「第 6 章　データの保管」の人文学向けコラムに登場する「データの説明文書」は，資料の目録・書誌情報や内容の情報ではなく，データを共有する際に，利用者がデータの扱い方を理解するために必要になるものです。データの持続可能性に配慮するなら，少なくとも同じ分野で常識的な IT リテラシーを持つ利用者が理解できる程度にデータの性質や

成り立ちを説明する必要があります。データセットを共有する際にはこの「データの説明文書」を同梱しておかなければなりません。このファイルに書いておくべき項目には，以下のようなものがあります。

- ファイル名およびファイルのパス（URL があればそれも含む）
- サイズ
- 作成日
- 作成者
- バージョン情報
- アクセス制限（利用条件）
- ファイルのチェックサム
- フォーマット
- ファイル作成に用いたソフトウェア
- データの元になった資料についての情報
- テキストデータの場合，採用した文字コード・文字エンコーディング
- 構造化テキストデータの場合，構造化の仕方についての説明
- その他，データを作成した際に配慮した事項

このデータは CSV，タブ区切りなどの機械可読性の高い形式で記述しておくことが望ましいです。より機械可読性を高めるには，TEI ガイドライン[4] に準拠して記述する方法があります。（具体的な記述方法については付録参照）

4 TEI（Text Encoding Initiative）ガイドラインは，人文学資料のデータを構造化するために策定されたガイドラインです。データのモデルと記述方法の両方を提示しており，様々なモデルに柔軟に対応できます。記述方式は XML に準拠しつつ，古典籍や古文書，碑文のメタデータや，古文書・古典籍・近現代小説・戯曲・演説・韻文詩・コーパス・辞書など，様々な資料を様々な観点から構造化して記述する規定や，固有名詞や外字をマークアップするための規定などが提示されています。

4.9 資料の書誌・目録情報の記述

人文学においては，資料の書誌情報や目録情報をデータとして作成したり，自分が作成したデータの中にそういった情報を含む場合があります。それらもメタデータと呼ばれます。その種のメタデータの作成に際しては，資料の性質に応じて，それぞれの専門家コミュニティにより書誌・目録情報の記述のモデルが定められています。そして，データとして記述する際には，それぞれのモデルに沿った記述が可能な記述方法に沿って記述することになります。それによって，異なるデータ作成者の間でデータに互換性を持たせ，データを横断的に広く活用することが可能になります。

代表的なものを挙げるなら，モデルに関しては，歴史的文献資料であれば，国際公文書館会議（International Council on Archives）が定める ISAD(G)があり，博物館・美術館資料として扱うものの場合には，国際博物館会議（International Council of Museums）が定める CIDOC CRM という参照モデルがあります。

記述方法については，ISAD (G) の場合は，米国議会図書館が策定する EAD3（Encoded Archival Description 3）に準拠して XML で記述する方法が広まっている一方で，AtoM などの ISAD(G)準拠の Web システムを用いることで Web のフォームからのデータ登録もできます。CIDOC CRM の場

合も様々な Web システムが存在します。いずれの場合も，モデルとして準拠できていれば Excel や Google Spreadsheet での記述も選択肢になります。

なお，前出の TEI ガイドラインは人文学資料全般に関するモデルと記述方法の両方を提供しており，ISAD(G)や CIDOC-CRM の記述にも柔軟に対応できます。

また，このような文化資料に特化されたモデルだけでなく，以下の「メタデータスキーマ」の項で挙げるルールを積極的に採用することでデータの相互運用性をより高めることができます。

4.10 変数の記述

人文学データにおいては，この項目は，構造化されたデータの要素名・属性名・属性値についての記述と考えることができます。歴史学において記述すべきものは多様であり，すべてを共通化・標準化することは困難です。一方で，他の人のデータとなるべく共通の手法で扱えるようにしないことには，データ共有の意義を十分に達成することができません。データ作成においては，その両者のバランスを取りながら行うことになります。それを踏まえた上で，コラム（メタデータスキーマ）に挙げるように，既存のスキーマにおいて色々なものが用意されていますので，関連するデータを作成する場合には検討対象としてみてください。

また，特にデータの設計・作成の段階で留意しておきたい点として，著者名，書名などの典拠情報を提供する VIAF（http://viaf.org/）や学術情報の永続的識別子を提供する DOI（https://www.doi.org/）など，URI を持つ国際的な典拠的情報へリンクすることでデータの利便性を高めるという方法があります。特に著者名については，一意に定められない場合や典拠情報が完備していない場合もありますので，全面的に採用できるとは限らず，また，どのような意図でリンクしたかの説明が必要な場合もありますが，いずれにしても，何らかの方法でリンクすることで，Web の知識情報の世界と接続しやすくなり，データの有用性・信頼性を高めることができます。

【メタデータスキーマ（人文学編）】

人文学でも，社会科学編における Dublin Core，DataCite，JPCOAR スキーマ，METS は有用であり，それぞれに利用されています。

また，デジタル人文学のデータを効率的に共有するための語彙として TaDiRAH: Taxonomy of Digital Research Activities in the Humanities がコミュニティベースで策定され，徐々に採用が広がりつつあります。

https://vocabs.dariah.eu/tadirah2/en/

また，上述のように，資料の目録・書誌情報としてのメタデータのスキーマに関しては，各資料の専門コミュニティが定めたものがあるので適宜参照してください。

さらに，データの内容についても構造化を行うことでより効率的効果的に資料を扱えるようにするためのスキーマもまた様々に策定されており，人文学全体としては，TEI ガイドラインが国際的には広く用いられています。

TEI（Text Encoding Initiative）協会による TEI ガイドライン：人文学全般（辞書，書簡，財務史料，手稿，校訂本，戯曲，韻文，貴重書書誌情報，外字，図像など）向け

https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html（国際版）

https://github.com/tei-eaj/jp_guidelines/wiki（日本版によるガイド）

TEI ガイドラインは人文学の様々な分野に対応できる内容を含んでいますが，分野によっては，この TEI ガイドラインと他のメタデータスキーマを組み合わせ，自らの必要性に応じたメタデータスキーマを作成して用いることもあります[5]。そして，研究者コミュニティによるものの多くは，TEI を通じて互換性を持たせることで応用性を高める形になります。特に汎用性が高いものは，十分な議論の後に TEI ガイドラインに統合されることもあります。

---

5 たとえば，碑文や貨幣に関するメタデータスキーマは，この種のものの典型として参考になるでしょう。碑文：EpiDoc: Epigraphic Documents in TEI XML https://sourceforge.net/p/epidoc/wiki/Home/ 貨幣：Numismatic Description Schema（NUDS）https://www.greekcoinage.org/numismatic-description-standard-nuds.html 既存のものに従わなければならないというわけではありませんが，相互運用性や持続可能性を高めるためには互換性を持たせることを念頭に置いてメタデータを設定することも重要になります。自分の分野に関わるものとしてこういったものがないかどうか確認してみてください。

5 データのフォーマット

量的調査を実施した後，そのデータを分析するためには，その情報を電子化し，データフォーマットに落とし込む必要があります。このプロセスはデータの質に直接的に影響を与えるため，どのようにデータファイル（データを格納した電子的なファイル）の形にするかについては事前に検討し，その計画を立てなければなりません。ここでは，データのフォーマットについて知っておくべきことについて解説を行います。

5.1 統計ソフトウェア

データファイルを構築するとき，まずはどのデータフォーマット[1]で保存するかを決める必要があります。どの統計ソフトウェア（統計パッケージともいう）を使うかは個々の研究者の嗜好にも左右されますが，同時に研究分野の慣習に倣うことが多いでしょう。というのも，当該研究分野で広く使われている統計ソフトウェアに合わせた形式であれば，そのデータがその後，広く使われる可能性が高まるからです。

統計ソフトウェアには様々なものがあり，代表的なものに SPSS，Stata，SAS，R，Python があります。SPSS，Stata，SAS はそれぞれ商用ソフトウェアで，その利用にはライセンスを購入する必要があります。また，それぞれには固有のデータフォーマットが存在するため，分析を行うにはそのデータフォーマットに変換する必要があることがあります。他方で，R はフリーの言語であり，分析に際しては CSV（Comma Separated Values）ファイル[2]が用いられることが多いです。

SPSS や Stata のような統計ソフトウェアは，頻繁にバージョンが更新されていくことにも注意が必要です。そのため，数年単位であればデータファイル保存について問題が起きることはありませんが，場合によっては，大幅な更新によって保存形式が変わることもありえます。そのため，たとえ分析の際に特定の商用ソフトウェアしか使わないとしても，長期的な保存を視野に入れれば，CSV ファイルのようなメタデータを含まない形式でもデータファイルを保存しておくことが重要です。

また，日本の量的調査データファイル構築で問題になるのは，日本語の文字コードに関する問題です。それぞれの統計ソフトウェアはアルファベットでの入力をベースにして開発されていますので，日本語での文字入力の際には注意が必要です。CSV ファイルを作成する際には，そのデータファイルがどの文字コード（ASCII，UTF-8，Shift JIS など）によって入力されているかを確認しておく必要があります。また，文字化けを引き起こす可能性がありますので，特殊文字や機種依存文字の使用は避けるべきです。特殊文字や機種依存文字の使用がどうしても必要な場合は，量的調査データファイルそのものではなく，調査に関するメモを別途テキスト形式のファイルとして保存する必要があります（6.1 も参照）。

1　データを格納するファイルの形式のこと。商用の統計ソフトウェアでは固有の形式が存在します。

2　CSV（Comma Separated Value）形式は，テーブルの各行に含まれる値を文字列として表し，それをコンマやタブコードで区切り，行区切りに改行コードを用いて表全体を表す形式です。タブコードを使ったものは TSV（Tab Separated

Value）とも呼ばれますが，ここではまとめて CSV 形式と呼んでいます。CSV 形式は，レイアウトや文字フォントの情報といったものは含まないファイル形式で，これはプレインテキストと呼ばれます。

5.2 ファイルの変換

量的調査を実施したら，データフォーマットを選択してデータ入力を行います。ここでは，データフォーマット選択の考え方について紹介し，さらに具体的な手続きについても簡潔に紹介します。

どのデータフォーマットを選ぶかは，前述の通り，その研究分野の慣習にも大きく依存しますが，大きく「メタデータ」を含むか（例 SAV ファイル；SPSS 形式のデータファイル），ほとんど含まないか（CSV ファイル）の 2 種類に分けることができます。ここでいうメタデータとは，各変数のラベル（変数に関する説明，変数名（5.5 参照）とは区別される）や値ラベル（各数値が何を表すかの説明），欠測値（後述）の指定などをいいます。

保存されたデータフォーマットとは異なる統計ソフトウェアで分析を行う場合には，データを変換する必要が出てきます。たとえば，SPSS の SAV ファイル形式の量的調査データファイルを Stata で分析したい場合，そのデータファイルを Stata のデータ形式である DTA ファイルに変換しなければなりません。その際には，Stat/Transfer のようなデータ形式変換の商用ソフトウェアを用いて変換を行うか，SPSS で DTA 形式にエクスポートをするという作業が必要になります。しかし，Stat/Transfer や SPSS のライセンスを持っていない場合にはこれらの方法を使えません。その場合には，R を用いて一旦 CSV ファイルに変換した上で，Stata で CSV ファイルを読み込むという作業，あるいは，R のパッケージ（foreign など）を用いて SAV ファイルから DTA ファイルに変換する作業が必要になります。

メタデータを含まない形式の利点は，いずれの統計ソフトウェアのファイルであっても変換が容易にできるという点にあり，また，OS や統計ソフトウェアのアップグレードに対して脆弱ではないという点にあります。そのため，データファイルの共有を可能にし，長期的な保存可能性を高めるのに適しているといえます。なお，東京大学社会科学研究所附属社会調査・データアーカイブ研究センターでは，個票データ[3]は原則として SAV ファイルと CSV 形式のファイル（タブ区切りのテキストファイル，拡張子は.dat）で提供されています。

いずれの方法であっても，ファイルのフォーマットを変換したときには，データの中の情報が抜け落ちていないか，抜け落ちていた場合はどのような情報が抜け落ちたかを確認をする必要があります。たとえば，メタデータを含む SAV 形式から CSV 形式に変換したときは，変数ラベルや値ラベルなどの情報は失われます。Excel ファイル（xlsx データ形式）でいえば，セル内の改行（Alt+Enter で改行可能）が他のデータフォーマットで引き継がれるとはかぎりません。また，統計ソフトウェアによっては変数名の文字数や文字の種類にも制限があるケースや，統計ソフトウェアによってブランクの扱いが異なる可能性があるので，その点も確認をしておく必要があります。

3　社会調査データや公的統計の作成過程で出てくる，集計前の個票（記入済みの個々の調査票）に含まれる個体情報から成るレコード群。

5.3 ファイルの作成

それでは，そもそも量的調査のデータファイルはどのように作成されるでしょうか。データファイル作成作業はそのままデータの品質に直結するため，非常に重要です。データの品質管理という

観点からも，このプロセスについては慎重に検討をして実施する必要があります。一般的には以下のように進めます。

1) 量的調査において，研究参加者（調査協力者，調査回答者など）の回答は様々な形で記録されていますが，それをデジタルデータファイルとして保存します。
2) 研究参加者に ID を与える必要があります。入力の際には，個票と紐付いた ID を用います。そうでないと，入力したデータファイルの確認をするときにどの個票を見るか，混乱する可能性があるためです。ただし，データ入力が終了した場合には，研究参加者の特定につながる情報をできるだけ削除するために，乱数を用いて ID を振り直すことが望まれます。
3) データファイルには，個票の変数だけでなく，調査地域情報や調査日などの背景情報を追加する場合があります。ただし，調査地域情報は研究参加者の特定にもつながりうる情報なので，個人が識別されないように配慮してどの程度の情報をどの範囲まで共有するかを検討する必要があります。たとえば，調査主体は記録のために市町村の情報を有しているかもしれませんが，データファイルを公開する際には，特定が避けられるように 5 段階程度の都市規模カテゴリーに留めることも検討すべきです。
4) 質問紙調査[4]の場合，まずは質問ごとに変数を作成し，どの回答にはどの数値を割り当てるかのコーディング（調査項目における回答ごとに数値を割り振る作業）のルールを定め，そのコーディング・ルールに従って入力をします。単数回答の質問に対しては 1 つの変数のみを作成しますが，複数回答の質問の場合は選択肢ごとに作成し，それに加えて無回答のケースのための変数も作成する必要があるでしょう。特に注意すべきことは，一貫性のあるコーディングを行うことです。つまり，「わからない」や「無回答」「非該当」（後述）のようなカテゴリーについては変数ごとに数値を変えるのではなく，共通した数字を付すことを検討すべきでしょう。
5) 質問票によっては，自由回答質問を含んでいる場合があります。自由回答質問とは，あらかじめ提示された選択肢から回答する形式ではなく（数値ではなく），言葉で回答する形式の質問です。この形式の質問では数値ではなく文字入力でデジタル化し，データとします。その際には，その回答記録をそのまま正確に書き写します。質問紙調査で回答者に文字や言葉の間違いがあったとしても，まずはそのまま間違い通りに入力し，記録します。この入力の際の文字コードにも注意が必要です。なお，自由回答には，個人が識別できるような情報が含まれる可能性がありますので，どの程度の情報を共有するかについては検討する必要があります。
6) コーディングが完成したら，データ入力作業を始めます。コーディングとデータ入力作業は分けることが重要です。データ入力をしながらコーディングを行うと，コーディング・ルールの適用が一貫性を欠くことになり,誤った入力がなされる可能性が高まるためです。また，データ入力の際には，すべての個票についてではないとしても，異なる入力者によって同じ個票を入力する二重入力の実施を検討すべきです。さらに，最初の 5%程度の入力が終わったあとに，データファイルを詳細に検討し，問題がないかを確認した方がよいでしょう。
7) データ入力が完了したら，度数分布表を作成し，コーディング・ルールに存在しない数値がないかを確認します。また，論理的な一貫性のない値がないかも確認します。たとえば，ある質問に対する回答次第で次の質問が変わるような分岐式質問形式のとき，本来であれば数値が入っているべきところで数値がなかったり，数値が入っているべきではないところで数値がある場合などをチェックし，入力の間違いがないかを確認します。また，そもそも調査

の段階で，誤った遷移をしている場合については，どのように処理するかを別途検討し，その処理についてはテキスト形式で記録をしておく必要があります。

4　質問紙（調査票）を用いて行う調査。調査票調査ともいいます。

5.4 ファイル名

ファイルはマスターファイルと，その後の更新ファイルを区別する必要があります。同じファイルを同じ名前のまま上書きして保存していくと，その途中段階での作業に誤りがあった場合に，その時点に戻って確認や修正をすることができなくなります。したがって，データの更新をしていく際には,そのファイルがどの段階のファイルかが分かるように別途保存をしていく必要があります。

ファイルを管理する際には，適切なファイル名を作成することに注意をすべきです。ファイル名は，その名前を見ただけで何についてのファイルか分かるようにし，かつ，簡潔な名前にすることが重要です。どのような環境でも表示できるように，できるだけ日本語は使わず，アルファベットを用いるべきでしょう。ハイフン（ - ）やアンダーバー（ _ ）の使用は問題ありませんが，スペースや特殊文字については避けるべきです。

ファイル名に含みうる要素としては，プロジェクトの名前や内容，ファイルタイプ，日付，作成者，バージョンを示す数字，現状（draft や final）が挙げられます。ただし，これらがすべて含まれると非常に長い名前になり，ファイル・マネージメントの際に視認性が悪くなるので，どの要素を含めるかはプロジェクトの中で統一したルールで作成し，それに従ってファイル名を作成すべきです。

たとえば，ABC という名前のプロジェクトでファイル名を作成するとき，マスターファイルに ABC_master.sav と名前をつけ，そこから更新がされていった場合には，ABC_20200212.sav のように日付で管理したり，場合によっては，ABC_ver2.sav のようにバージョン番号で管理したりします。ABC　データ.sav と日本語混じりであったりスペースを挟んだりする名前を付すことや，あるいは 1 つのファイル名のまま更新を繰り返すことは避けるべきです。

5.5 変数名

変数の名前についても，一定のルールのもとに作成をする必要があります。変数名の付け方にはいくつかの方法が考えられます。標準的な付け方は，変数 1，変数 2 とあったときに，(1) それぞれ v1，v2 のように通し番号を付す方法（v は variable のことを意味する），あるいは，(2) Q1，Q2 のように質問の番号をそのまま付す方法もあります。複数回答質問の場合には v1a，v1b や Q1a，Q1b のように付け，分岐式質問で下位質問がある場合には，Q1s1 のように付けることが多いです。どのように付けるにせよ，一定のルールの下に変数名を付けることでその変数がどのような構造の下に生成されたかを一目でわかるようにすべきです。

変数名の付け方としては，(3) その変数の実質的な内容をそのままつけたり，それを短縮した名前をつけたりすることもありえます。たとえば，年齢についての回答を age とし，教育程度については educ とします。この方法の利点は，異なる調査でこのように名前をつけておけば，同じ内容の変数名をすぐに見つけられる点にあります。したがって，単独の調査というよりも，同一枠組み

で複数回実施した調査の場合に検討されるべき方法といえるでしょう。なお，統計ソフトウェアによっては変数名の長さに制限がある場合があり，あまりに長い名前は避けた方がいいでしょう。

さらに，メタデータを含むデータ形式の場合には，変数ラベルを付ける場合もあります。変数ラベルには，調査票における質問番号，変数の内容，（もし複数の変数から合成された変数ならば）合成された変数であることの明示などを含めます。同様に，入力された数値について変数ごとに値ラベルを付けることもあります。値ラベルでは，その回答選択肢の内容を明示します。なお，日本語でラベルを付ける場合には，文字コードの問題にも注意を払いましょう。

5.6 欠測値

量的調査の場合，欠測値（欠損値ともいう）が付きまといます。欠測値データとは，調査項目において回答が欠落したデータのことを指します。量的調査の分析では，欠測値をどのように処理するかをまず決めないといけません。というのも，欠測値が正しく指定されていないとその分析も正しくできないからです。くわえて，その欠測値のパターン自体も調査の重要な情報となりえます。

欠測値には様々な種類があります（わからない，拒否・無回答，データ処理の誤り，非該当）。たとえば，質問に対して「わからない」と答える場合と，その質問への回答を拒否すること（質問紙調査での無回答も含む）は，異なる行動ですので，両者は区別してコーディングする必要があります。調査実施中の誤りやデータ処理の誤りもこれらとは区別されるべきです。また，分岐式の質問の場合，下位質問に答える必要がない場合や該当しない場合には「非該当」と分類されます。

欠測値データについては，ブランク（空白）として扱うのではなく，必ず何らかの数字を与えるべきです。発生する欠測値の種別は調査ごとに異なりますので，すべての調査で共通するルールを確立することは困難です。しかし，欠測値のコードについては，そのデータファイルの中では一貫したルールを採用することが必須です。たとえば，非該当は 77，わからないは 88，回答拒否・無回答は 99 のように，一貫したコーディング・ルールを作成しておくべきです。質問ごとに異なる欠測値コードをつけることは分析の際に間違いが起きる原因となりますので，避けるべきです。そのコードのルールについては文書に残して記録をしておく必要があります。また，場合によっては欠測値について補完したデータファイルを作成することもありますが，その作成情報を残し，もともとの欠測値データとは別に提供する必要があります。

上記のように，回答をデジタル化し，データファイルを作成する過程では，どの質問にどの変数名を振り，どの背景情報（調査地点，調査日など）を含め，欠測値としてどのような種類があるか確定し，それらをどのようにコードするかなど様々な決定がなされます。これらの決定は後々重要な情報となりますので，コードブックやメタデータファイルとして記録をしていく必要があります。この点については，「第 4 章　メタデータ」，「第 6 章　データの保管」も参照してください。

【人文学向けコラム 3（人文学におけるデータのフォーマット）】

人文学においては，文化的な活動を記録したデータであれば，なんであれ研究対象となり得るため，データのフォーマットとしてはあらゆるものを想定する必要がありますが，ここでは主に，歴史研究に関わる文献資料を採り上げます。

歴史研究に関わる文献資料のデータを作成時の状況に着目して区別すると，以下のように分類することができます。

（1）デジタル撮影画像など，資料（ここでは，歴史研究に関わる文献資料を指す）をデジタル複製したデータ
（2）典拠を直接確認して作成したデータ
（3）既存の二次資料・データを用いて作成したデータ
この分類を踏まえつつ，データのフォーマットについて見ていきましょう。

1．資料をデジタル複製したデータ

（1）では，画像データとして作成されることになります。撮影時には，スケールやカラーチャートを写し込むなどして，元の資料の状況を可能な限り再現できるような工夫が求められます。具体的な撮影方法については「国立国会図書館資料デジタル化の手引」（https://www.ndl.go.jp/jp/preservation/digitization/guide.html）が参考になるでしょう。撮影した画像データは，操作しやすいように圧縮してファイルサイズを小さくすることが通例ですが，一度圧縮すると撮影時点の画質に戻せなくなる場合があるので注意が必要です。通常は，撮影時点の画質に戻せるような圧縮方式（可逆圧縮）の画像（JPEG 圧縮をしていない TIFF 形式画像や JPEG2000 など）か，もしくは無圧縮の画像や RAW 形式のものを保存用画像として保管しておいて，普段閲覧したり Web 公開したりする際には，JPEG，PNG，GIF などの形式で非可逆圧縮した小さなファイルサイズのものを元画像から作成して使用することになります。画像の圧縮には様々なツールが商用でもフリーでも提供されており，大量の画像を一括処理できるものもありますので，用途に応じて効率のよいものを探してみるとよいでしょう。たくさんの画像を一つのフォルダに入れてしまうと処理に時間がかかってしまいがちですので，冊や箱など，把握しやすい単位でフォルダごとにわけて保存しておくと効率的です。フォルダや画像のファイル名は，他のものと重ならないように，ユニークな名前にしておくと後々混乱が生じにくくなります。なお，ファイル・フォルダの名前は半角英数字で桁数（文字数）をそろえておくとソートなどの操作がしやすくなります。

さらに，（1）については，ただ画像を作成するだけでは，どの画像が何であるかがわからなくなってしまうため，画像の目録データを作成する必要があります。これは Excel や Google Spreadsheet で表形式のものを作成するだけでも十分です。1 行を 1 資料もしくは 1 画像として，画像のファイル名と，それについての何らかの情報を記載しておきます。この際には，現物と画像が対応するようにしておくことも重要です。そして，タイトルなどが付された何らかのまとまりを持った複数の画像がある場合には，それも記します。入力の際には，後の検索やソートなどを効率的に行うため，「〃」などの記号は使わずに，1 行だけを取り出しても情報として通用するようにします。これらのデータはタブ区切りやカンマ区切りのテキストデータ（TSV，CSV）といった互換性の高いフォーマットに変換して保存することもできるため，データ共有の際には TSV や CSV で保存しておくと有用性が高まります。なお，より高度なデータ形式としたい場合は，前出の記述モデルやメタデータスキーマを参考にしてください。

2．翻刻データや校合テキストデータ

（2）には，（1）で述べた目録データのようなものから，翻刻[5]データ，校合[6]テキストデータ，座標データなど，歴史研究関連のデータだけを見ても様々なものが含まれます。（3）も，データのフォーマットとしては同様のものとなるため，ここではまとめて扱います。

目録データについては上に述べましたので，ここでは翻刻データや校合テキストデータなどについて見てみます。まず，使用する文字を新字体・旧字体をどちらかに寄せるか，仮名遣いをそ

のままにしたか現代に改めるか，外字フォントを使用するか，といったような，表記の扱いの方針を一つのデータの中で意識的に統一しておく必要があります。現在は Unicode を使用して，さらに IVS[7]（Ideographic Variation Sequence/Selector）も用いれば極めて多くの文字の形を表現することができますが，それでも足りないという場合には外字フォントなどを利用することになります。そして，それも含めて，データのフォーマットがどのようになっているか，ということを記述する「データの説明文書」を作成します。

それを踏まえつつ，ここでは，以下の 3 通りについて，それぞれのデータフォーマットを見てみます。

（a）Word や一太郎などのワープロソフトでレイアウト情報も含めて作成・保存する方法

（b）資料に記載された文字を入力・保存してテキストデータとして保存する方法

（c）（b）の派生として，構造の情報をマークアップして保存する方法

まず，（a）の Word や一太郎は，使い慣れている人にとっては有用なものであり，機能も豊富で，見た目の整った文書を作るには有益です。しかしながら，データ作成という観点では，データの構造に基づいて自動的に処理して分析するような，データとしての用途に適したデータを作成するにはあまり向いていないという点には留意してください。また，データ作成時のソフトウェアとデータを開く際のソフトウェアのバージョンが異なるとレイアウトが崩れることがあるため，レイアウトをそのまま保存できる PDF 形式でも保存しておくべきです。

（b）については，全文検索を簡易に行うために作成することがあります。この場合，文字コード／文字エンコーディングに配慮する必要があります。検索用途のために新字体に寄せることはよく行われますが，日本史文献の場合，東京大学史料編纂所で「異体字同定一覧」を公表していますので参考にしてください。また，検索システムを工夫すればテキスト作成の段階でどちらかに寄せなくとも曖昧検索をかけることもできます。一度作成してしまうと，あとで修正するには同じ手間をもう一度かけることになってしまいますので，入力時の手間がそれほど大きくなければ，資料で使われている字体をなるべくそのまま再現するという選択肢も有用です。また，資料に記載されたテキストしか入力されていないデータを作ると，そのテキストデータがどういうものであるか，という情報を確認することが難しくなります。解決策としては，目録データがあるならそこにこのファイル名も入れておく，あるいは，何らかの記述ルールに従ってタイトルや書誌情報などを記述しておく，といった方法があります。また，データを一定の単位で取りだして処理しやすくするために Excel や Google Spreadsheet を使用して一行・一文・一段落などの単位でデータ入力する方法もあります。また，文字が本文のものか注記なのか，宛先か奥付か，といったような情報がないと扱いが難しくなるため，その資料が何であるか，資料の中に書かれている文字列は資料の中でどういう位置づけになっているのか，といった情報を処理しやすい形で書き込む方法として次の（c）があります。

（c）構造の情報をマークアップして保存する方法は，本文やその他のテキスト中の要素に関して様々な情報を書き込む場合に用います。前出の TEI ガイドラインに準拠して，XML（Extensible Markup Language）のタグを付与する方法が国際的には広く普及しています。タグをつける操作は一見すると難しそうに思えますが，XML 用のエディタを用いることで比較的容易になります。また，より簡易な記述方法としては，青空文庫形式や，それを元にした koji という手法もあります。あるいは，XML ではなく，マークダウンと呼ばれるより簡便な記法を用いる手法も人文学でも時折見られます。一方，校合テキストデータを記述する場合には，TEI ガイドラインで充実した記法が提供されており，処理ツールも複数提供されていますので，それに準拠しておくのが後々効率的です。いずれにしても，このような場合に肝心なのは，他の研究者が理解し再利用し

やすいようなフォーマットにしておくことです。そのためには「データの説明文書」を十全に記述しなければなりません。その内容を一からすべて説明するのは，時間も手間もかかりしばしばかなり困難ですが，TEI などの既存の枠組みを適切に利用することで省力化できます。

なお，すでに（b）のところでも少し触れましたが，（a）～（c）のような内容のテキストデータは，もし画像データも一緒に共有できる場合には，画像データのファイル名，もしくは URL と紐付けられるようなリンク情報も記述しておきます。このリンク情報は，目録データの一部として記述することもできますし，テキストデータのファイル名を画像ファイル名と関連付けられるようなものにするという方法もあります。両方に対応できればより望ましいです。さらに，テキストの任意の箇所と対応する画像上の一部分を対応づける手法として，TEI 以外にも，Web 画像の相互運用の枠組みである IIIF[8]（International Image Interoperability Framework, http://iiif.io/）による Web アノテーションや，画像上のレイアウトとテキストデータを紐付ける ALTO/METS（Analyzed Layout and Text Object with Metadata Encoding and Transmission Schema, https://www.loc.gov/standards/alto/）などが広く利用されていますので関心がある方は挑戦してみてください。

3．資料を離れた二次的なデータ

人文学のデータには，辞書や索引，目録，人名辞書，年表，地理情報など，単なる資料のデジタル複製ではなく，研究に必要なデータとして抽出され取りまとめられたデータも多く存在します。目録の作り方についてはすでに上で触れたとおりです。それ以外のデータについても，基本的には目録と同様の考え方が適用できます。データを作成する際には，すでに広く流通している参照 ID が存在する場合には，その情報とリンクしておくことで共有した際により広く活用されるようになります。たとえば，論文情報なら DOI，人名辞書なら VIAF，典籍の目録なら国文学研究資料館のデータベースにおける「統一書名」や大正新脩大蔵経のテキスト番号，芸術関連なら Getty Vocabularies などはよく用いられます。他にも分野ごとに様々なものが作られつつありますので，自分のデータを作成する際には確認してみることを強くおすすめします。

4．既存の二次資料・データを用いて作成したデータ

このコラムの冒頭に挙げた 3 つのうちの（3）について配慮すべき点を見ておくと，前出の「データの説明文書」をきちんと記述しておくことが必要です。すでに作られたデータを元にしてさらに新たなデータを作成するという取り組みは，デジタルデータの特長を活かすものであり，今後広く展開されていくことが期待されますが，一方で，データの正確性についての責任の所在は元データの作成者に依拠することになりますので，元データの作成者の貢献を明らかにするという意味も含め，どのデータにどう依拠したかをなるべく正確に記述しておくことが重要になります。

5　写本・版本などを，原本どおりに活字に起こしたりテキストデータにするなどして新たに出版すること。特に後者を指してデジタル翻刻ということがあります。

6　複数の写本や木版本などを比べ合わせて，本文の異同を確かめたり誤りを正したりすること。また，それらの異同や修正も含めてデータ化したものを校合テキストデータといいます。

7　IVS は，Unicode においては包摂もしくは統合と判定されるにも関わらず字形が異なる文字を，枝番号を付することで表示できるようにする仕組みです。これにより，細かな字形差にも対応できるようになりました。これを国際的に相互運用するために，字形のセットをデータベース（http://www.unicode.org/ivd/）にすることになっています。“Adobe-Japan1”や“Hanyo-Denshi”をはじめ，2021 年 3 月現在，7 件の字形セットが登録されています。

8　Web サイトで公開された画像などのマルチメディアコンテンツを相互運用することを目指す技術仕様。これに準拠して公開されたコンテンツはサイトの内外で自在にアノテーション可能となります。人文学では Web 画像への注釈

や部分切り出しといった用途で利用が広がっています。人文学研究者にも利用しやすいシステムの一つに IIIF Curation Platform があります。http://codh.rois.ac.jp/icp/

6 データの保管

研究で作り上げたデータを長期間に渡り利用可能な状態で保管するにはいろいろな面からの配慮が必要です。ここでは，データを作成し，蓄積するメディアについてソフトウェア，ハードウェアの両面からデータ保管の際に留意すべき点について述べます。

6.1 保管対象データについて

研究過程において，質問紙調査を行い，結果を整理しながら調査結果の分析のためのソフトウェアを利用してデータを蓄積し，それを利用して集計表を作り，さらに分析結果をまとめるといった作業が行われます。また，そうしたソフトウェアを利用して作ったデータは，データベース[1]として利用されます。また，データを利用する際に必要な情報をまとめたコードブックが作られます。データベースはもちろんのこと，コードブックや関連する記録もデジタルデータとして保存する対象ととらえることができます。これらに加えて，データの作成過程やデータ間の関係などに関する記録を保存することも重要です。

一般に，保管対象データは以下のように分類することができます。

（a）SPSS などの調査データ分析用のソフトウェアを使って作られるひとまとまりのデータ

（b）Excel などの汎用のソフトウェアを使って作られるひとまとまりのデータ

（c）MySQL などの汎用のデータベース管理システム，あるいはそれらを使って開発したサービスを用いて作られるデータ

（d）コードブック，データの構成規則などに関する説明文書などの文書データ

（e）調査の過程で得られた資料や記録などをデジタル化して作られたデータ（質問紙をデジタル化して作ったデータなど），もともとデジタル形式で作られた記録や資料（インタビューの記録，オンラインの入力データなど）

上のような様々なデータを適切に保管するには，保管すべきデータがどのような実体でありどのような特性を持つものであるかということと，保管対象となるデータの間の関係を把握して記述することが求められます。そのため，

（f）調査において作られるデータ間の関係，データ作成や維持管理に関する記録，その他データの利用のために必要な情報などをまとめた文書データ

も保管すべきデータです。なお，（f）は研究過程で作られるデータに関するデータであるので二次的データと呼ぶこともできます[2]。この二次的データは，調査において作られるデータをひとまとまりのものとして保管する上で必要になるものです。たとえば，ひとまとまりのデータを一つのフォルダに収め，そのフォルダの中のデータの説明を書いた文書（ReadMe といった名前をつけられるケースが多くあります）は，この二次的データの例です。そのため，このデータは，保管したデータを将来に渡って利用し続けることができるようにする上でとても重要な役割を持ちます。また，権利情報や秘匿性などの管理要件も記録しておく必要があります。後述するように，データの長期保管の過程においては，データの新しい環境への移行や新しいソフトウェアバージョンへの変

換といったことを想定しなければなりません。そのため，データの管理状況の変遷，いわばデータの来歴情報を記録していくことも求められます。

1　データを目的に応じて構造化し，データへのアクセスと利用を効率化したもの。

2　「第 4 章　メタデータ」に定義されているように，「データに関するデータ」はメタデータと呼ばれます。メタデータは，多くの場合，標準規格を参考にしながら，記述項目とその値の対の集まりからなる構造を持つデータとして作られます。そのため，メタデータとは「データに関する構造を持つデータ」と理解されることも多くあります。その一方，データベースやフォルダの内容説明の場合，作者や作成日，内容説明といった構造的な表現をする場合もありますが，そうした構造を持たない文章表現として作られることも多くあります。ここで「二次的データ」と呼ぶものはメタデータですが，本章では，混乱を避けるため，「二次的データ」と表記しています。

6.2 データのタイプと保管のための維持管理

保管対象となるデータのタイプは大きく分けて二つあります。

データベース型データ：　調査データを表形式にまとめ，それを用いた集計，表の内容の検索，並べ替え，結合などの機能を利用することを目的として作られたデータ

文書型データ：　調査結果をまとめた文書や調査時に作られるいろいろな記録など，利用者が読むことを目的として作られたデータ

大雑把に言って，6.1 に示した（a），(b），(c）は前者，(d），(e），(f）は後者です。ここでは前者をデータベース型データ，後者を文書型データと呼ぶことにします。いずれのタイプであっても，データをコンピュータ上で利用するためにはデータごとに適したソフトウェアを使う必要があります。たとえば，SPSS などの調査分析用ソフト，Word のようなワープロソフトや PDF 文書の表示閲覧ソフト，Excel などの表形式データを扱うソフトなどです。こうしたソフトウェアは，バージョンによってデータの表示形式などが異なることもあるため，保管されたデータの利用には適切なバージョンのソフトウェアを利用することが求められます。

データベース型，文書型いずれのデータであっても，その作成時に用いたソフトウェアが利用可能である間は，データの保管に大きな問題は生じません。しかしながら，多くの場合，ソフトウェアのバージョンが進んだり，ソフトウェアの提供がされなくなったりすることがあるため，データの長期保管の際には，データごとに必要なソフトウェアが使えるかどうかを，できるだけ定期的にチェックする必要があります。同じソフトウェアであっても，バージョンが変わることで表示のレイアウトが異なったり，ソフトウェアが提供する機能が変更になったりすることもあるので，データを利用できるかどうかのチェックは重要です。また，データを古いバージョンのまま残しておくと，将来適合するソフトウェアが利用できなくなることも考えられるので，新しいバージョンに適合するようにデータを更新あるいは移行することも必要になります。ただし，こうした更新や移行による影響が，元のデータの内容や利用性を損なわないことを確認する必要があります。

データベース型のデータの場合，データの加工や表示のための手続きの記述（スクリプトやマクロ）を合わせて保存する場合があります。こうした手続きはソフトウェアおよびそのバージョンに依存することが多く，長期利用のための維持管理の視点からは特に注意が必要です。

データベース型のデータの保存の際に，CSV（Comma Separated Values）形式が用いられることがあります。CSV 形式はプレインテキストであるため，ソフトウェアに依存することはなく，長期間の安定したデータ保存や異なるソフトウェア間でのデータの移行には向いています。その一方，表示レイアウトやマクロなどのソフトウェア固有の機能を利用したデータ作成時の設定を保存することはできません。また，Excel などのソフトウェアで作ったオリジナルのデータ中に，改行

コードなどの CSV 形式でのテキスト保存に適さないものが含まれていると正しい CSV 形式が作り出されないことがあるので，保存用の CSV 形式データが正しく再ロードできるかどうかといったチェックは必要です。

文書型データの場合，Word などの広く使われているファイル形式の場合は安定的に利用できると思われますが，異なるバージョン間でデータを移行した際のレイアウトのずれといった問題があります。固定された文書データとして保存することを考えた場合 PDF 形式（特に，アーカイブ向けである PDF/A）で保存することが望まれます。その一方，調査データの継続的な利用に応じて文書編集を行うことが期待される場合には，Word などのファイル形式で残さざるを得ません。その場合，編集の記録を残すことが求められます。

最近は，データ中で用いられる文字コードの違いによる問題が起きることは少なくなってきましたが，データベース型，文書型，いずれの場合でも，データを長期に渡って保管することを考慮すると，前述の二次的データとして利用している文字コードに関する記述を含めること，システム依存の文字は使わないことといったことに留意することが求められます。

【人文学向けコラム 4（人文学におけるデータの保管）】

データの保管に関しては，基本的な考え方については社会科学分野のデータと人文学分野のデータの間には全体として大きな違いはありません。データの種類の違いに伴って若干異なる部分があるので，ここではそれについて述べておきます。

保管対象データについては，特定企業の製品や特定のソフトウェア・ハードウェアに依存しなければデータを利用できない形になってしまうことを避ける必要があります。その上で，可能な限り汎用的なフォーマットに変換した上で保存しておくことが望ましいです。そして，フォーマットに関することは「データの説明文書」に，少なくとも同程度にこの種の事に詳しい同業者が読んで理解できるような形で記述しておきます。

前章を踏まえつつ，歴史研究に関わる文献資料データの主なタイプを以下に列挙した上で，個々に検討してみます。

（1）画像データ
（2）Word や一太郎などのデータ
（3）PDF ファイル
（4）テキストデータ
（5）Excel やファイルメーカーのデータ
（6）XML などの構造化データ
（7）Wordpress，Drupal，Omeka などの CMS に保存されたデータ
（8）データベースシステムに保存されたデータ
（9）企業や研究者が開発した専用システムに保存されたデータ

（1）の画像データについては，TIFF や RAW，JPEG2000 形式など，撮影時点での画質をなるべく維持できるフォーマットにする必要があります。圧縮方式は時折進歩することがあり，扱える画像のサイズも徐々に大きくなっていきますので，パソコンやネットワークがある程度進歩したら元の画像から圧縮をしなおすことで，低コストでシステムを改良することができます。逆に，JPEG，GIF などの非可逆圧縮で保存してしまうと，システムが進歩した場合に撮影をしなおさないとデータの改良が行えないため，費用と人手がかかるだけでなく，資料を再度持ち出して撮影しなおさなければなりませんので，資料への負荷という観点からも好ましくありません。

（2）に関しては，頁内のどの位置にどの文字があるか，といったことまで正確に残したければ，むしろ（3）PDF ファイルとして保管するべきです。ただし，PDF ファイルでは再編集が難しいことがあるため，その可能性を考慮して（2）も同時に保管しておくことが有用です。

（4）の保管に関しては，タグ付きテキストデータである（6）と同様に考えることができます。この場合，文字コード・文字エンコーディングについて特記すべき事項があれば「データの説明文書」に記載しておくことが大切です。また，外字フォントを使用した場合には外字フォントも同時に保管しておきます。ただし，フォント自体にも利用条件があり，データ共有の際には確認が必要です。

（5），（7），（8），（9）に関しては，ソフトウェアやシステムを丸ごとそのまま保管したいところですが，OS のアップデートに伴い，やがて使えなくなってしまいます。多くのシステムはその事態を見越してテキストデータ，CSV などに出力できるようになっていますので，システムのデータとともに，出力したテキストデータも保管しておくとよいでしょう。また，文書のレイアウトも維持する必要がある場合には PDF 形式で保存しましょう。（9）に関して，テキストデータや CSV などに出力できないという場合には，いずれ訪れるシステム移行の際に困難が生じますので，システム導入時にそれが可能であることを必ず確認してください。また，文書のレイアウトも維持する必要がある場合には PDF 形式で保存しましょう。

（9）に関して，テキストデータに出力できないという場合には，いずれ訪れるシステム移行の際にも困難が生じますので，出力できるように改良を依頼するとよいでしょう。

6.3 保管のためのシステム環境

データを収集蓄積し，分析するシステム環境は，利用者ごとに様々です。その一方，どのようなシステム環境であれ，その環境に応じて貴重なデータを保存し，保管する必要があります。システム環境を大別すると，以下のように考えられます，

- パソコンなどにインストールした個人向け環境
- 情報センターや研究所などが提供する環境
- インターネット上のサービスとして提供される環境

パソコン環境では，ソフトウェアのインストールからデータの保存管理まで，基本的に研究者自身が行うことになります。情報センターのシステムやインターネット上のサービスを用いる場合，研究者は自身のデータの保存は行いますが，ソフトウェアやシステムの管理を行う必要はありません。

いずれのケースにおいても必要とされるのは研究者が作るデータの「適切な保存」です。適切な保存には，事故などによるデータ喪失が起きないようにするためのバックアップデータづくりとその管理，研究過程において作られるデータのバージョン管理といった観点が含まれます。さらに，できあがったデータの長期利用の観点からの保存も含まれます。作成中におけるバックアップは，パソコンなどのシステムに準備されたバックアップ機能を用いることができます。ただし，こうしたシステム側でのバックアップの場合，バックアップ時の環境に依存する可能性があります。他方，完成したデータの場合には作成過程の延長としてバックアップを取っておくこともありますが，一般にデータが作成者の手から離れることを想定する必要があるため，完成したデータを作成に用いたシステムと切り離して保存することを考えておく必要があります。

研究過程におけるデータ管理と完成したデータでは管理の仕方が異なりますが，いずれの場合もデータのバージョン情報やファイル形式などの情報，そしてデータを利用するためのソフトウェアなどに関する情報を適切に管理しておくことは必要不可欠です。完成データの保存の場合には，データベース型のデータの他にコードブックなどの関連データ，そしてそうしたデータの間の関係やデータの利用方法などを書いた二次的データも一緒に保存する必要があります。

他方，セキュリティ対策やアクセス権限などのデータ管理のために，ファイルやフォルダに対してパスワードなどによる保護設定を行うことがあります。長期保存の過程でパスワードが失われてファイルが開けなくなることなどの事態が起きないように保護設定に関する情報を適切に管理しなければなりません。

6.4 データの保管メディア

前述のように，研究者が置かれているシステム環境によって，データの保管に適した記録メディアは異なります。情報センターのような環境でデータ作成を進めている場合は，保管のための記録メディアの選択や維持管理は情報センターに任せることになると思われます。ただし，その場合でも，情報センターが提供する保存サービスの範囲を正確に知ることと，保存しているデータの利用可能性を維持するために必要な情報の記録とその記録の適切な保存を行うことが研究者には求められます。

パソコン上のツールで作成したデータの場合，研究者自身によるデータ保存が求められることになります。保存対象となるデータはいくつものファイルに分かれることが想定されます。そのため，(1) に示した二次的データを含めて対象データのファイルをひとまとめにして保存する必要があります。「ひとまとまりのデータ」を一つのフォルダにまとめることや，いくつかのファイルでひとまとまりのデータを構成していることの記述をすること，そしてそうしたフォルダや記述のありかをわかりやすく残していくことが必要です。

データの安全な保管の視点からは，複数のコピーを作り，異なる物理媒体に保存し，異なる場所で保管することが推奨されます。その際，DVD などの持ち運び可能な記録メディア（可搬記録メディア）を利用することが想定されます。他方，複数のコピーの管理において矛盾が起きないようにするための注意が必要です。たとえば，マスターデータとバックアップデータの対応関係が失われないようにすること，データへのアクセス権限に関する情報に矛盾が生じないようにすること，データの紛失などが起きた際の復元方法を明確化しておくことなどが求められます。

可搬記録メディアには DVD，Blu-ray などの光ディスク，外付けのハードディスクや半導体ディスク（SSD），さらに USB メモリやメモリカードなどがあります。高密度に集積する必要がある情報センターなどでは磁気テープが保存用に用いられることもあります。個人や小規模な研究室では，コンパクトディスクや外付けディスクが扱いやすいであろうと思います。コンパクトディスクには 10 年以上の耐用年数があるといわれますが，保存の環境に依存することもあるので注意が必要です。メディア自体の耐久性の限界やメディアを利用するために必要なハードウェア機器が旧式化（陳腐化）し手に入らなくなるため，どのようなメディアであれ何十年にも渡って使い続けることができるものではありません。それに加えて，前述のようにソフトウェアの陳腐化やバージョンの違いによる互換性の問題といった問題もあります。そのため，数年に一度と想定されるパソコンやシステムのリプレースなどの機会を利用した利用可能性のチェックが強く推奨されます。事情が許せば，信頼性の高いデータセンター，クラウドサービスを利用してデータを保存し，可搬記録メ

ディアでの保管をバックアップデータとすることで，ハードウェア的な維持管理の手間を減じることは可能です。ただし，秘匿性など，一定のデータ管理基準が求められるデータの保管の場合，研究機関などが定める管理基準を満たしたサービスを選定する必要があります。また，可搬記録メディアのみを用いる場合であっても，複数のメディアに複製を作りそれを別の場所で保存することで，個別メディアによる寿命の違いや災害や事故によるデータの損失に対する備えとすることができます。いずれの場合でも，ソフトウェアの陳腐化や互換性の欠如によってデータが使えなくなることが起きないようにするための定期的なチェックを行うといった維持管理は必要です。

コンパクトディスクなどの可搬記録メディアによる保存の場合は，物品としての記録メディア管理が求められます。すなわち，可搬記録メディアを保存する場所でのメディアの物理的な管理とそこでの配架情報の管理が求められます。災害対策や盗難，データ漏洩対策などを含めたセキュリティの観点から，重要なデータ資源は安全な環境で保管することが求められます。個人や研究室などの小規模な環境での保管の場合であっても，物品として管理のしやすい環境を整える必要があります。

【人文学向けコラム 5（人文学におけるデータ保管のためのシステム環境と保管メディア）】

この点における社会科学分野のデータと人文学分野のデータの考え方にも大きな違いはありません。ただし，文献資料のデジタル複製で扱うことになる画像データの場合にはデータ容量が社会科学分野に比べてかなり大きくなりますので，それに伴う実務面での違いが出てきます。可逆圧縮形式で高精細デジタル画像のデータを保存すると 1 枚あたり数百 MB になりますので，それをまとまった数で保管できる環境とメディアが必要になります。保管用画像は一度保管したら書き換えを行うことはありませんので，あとは適正に保管されているか，必要な時に取り出せるかを確認することになります。数がそれほど多くなければ，Blu-ray などの光ディスクで保管することができます。ただし，あまりに多い場合には，保管や内容チェックのためのディスクの入れ替えの手間が大きくなってしまうため，光ディスク複数枚をカートリッジに入れて自動的に入れ替えてくれるタイプの機器や，磁気テープ（LTO8 規格で 1 本あたり 12TB）などを利用するのがおすすめです。ただし，費用はあまり安くないため，ある程度大きなプロジェクトでの選択肢と考えておくのがよいでしょう。また，最近は，コールド・ストレージと呼ばれる，書き換え頻度を極端に落とすことを前提とした契約内容による非常に安価なクラウドストレージが大手各社から提供されるようになりましたので，データ復元に多少時間がかかってもよく，かつ，データセキュリティ上の問題がない場合は選択肢として検討するとよいでしょう。

なお，いずれにしても，メディアとしての寿命やデバイス・規格としての寿命がありますので，永続的にそのままで保存できることはありません。長期的な保管に際しては一定期間で保存媒体を更新する必要があることを前提として保管計画を立ててください。

6.5 データの受け渡しと廃棄

研究で作成したデータの研究者間，あるいは研究者とデータアーカイブの間での受け渡しには，可搬型メディアを用いる，ダウンロード・アップロードする，あるいはデータ交換のためのオンラインストレージを用いることが想定されます。いずれの場合も，データを受け渡しする両者の間で

の取り決めを明確に決めることが重要です。また，データの内容が安全かつ完全に受け渡しされるデータを利用するために必要なシステム要件などの情報も正確に伝える必要があります。

データの受け渡しの際には，ウィルスチェックやパスワードによるロックの状態チェックなど，データを受け取る側での混乱を避けるようにしなければなりません。受け渡しされるデータの一覧，データ間の関係，データ毎の作成担当者などを記述した文書データを受け渡す必要があります。また，6.1 で示した「データ間の関係やデータ作成に関する記録（来歴情報）をまとめた記録文書データ」がないと，他者が「利用できないデータ」を移管することになるので注意が必要です。

データの廃棄は，データの管理要件にもよりますが，秘匿すべき情報が含まれるデータが蓄積された可搬型の記録メディアは物理的に破壊することが確実な廃棄方法です。パソコンの固定ディスクの場合はパソコン上での削除，専用ソフトを用いたディスクの内容の消去，物理的なディスクの破壊などを行う必要があります。大学などの機関が決めるデータの廃棄ルールを知ることが重要です。

6.6 まとめ

最近では，すでにふれたように，研究にあたってデータ管理計画（DMP）を作ることが求められる場合が増えてきました。そうした場合，データ管理計画に従ってデータを管理することが求められますが，その一方，研究終了後のデータ管理は忘れられがちになります。気が付いた時には必要なデータが使えなくなっていたということにならないように，データの利用性を保ちながら保管することに注意を払わねばなりません。

データの利用可能性を保証した長期保管に関する万能薬はなく，保管されたデータを利用可能な状態に保つ維持管理を続けることが必要です。頻繁に使われるデータは利用者によるチェックがはいるので，データ保管の面からは安全な側にあると言えます。その一方，すべてのデータが頻繁に使われると考えることはできず，かつ，全てのデータを自動チェック可能なシステム環境に置くことも現実的とは言えません。そのため，上に述べたいろいろな観点に基づきデータの保管環境を整えることが重要です。

7 データ共有に関する倫理的側面

ここでは，人を対象とする研究において得られたデータ（人体から取得された試料より得られたデータも含む）を共有する可能性がある研究について，倫理面での留意点について述べます。

7.1 人を対象とする研究における倫理原則

国際的には，「人を対象とする研究」のなかに，医学や生命科学の研究として行われる臨床試験や実験だけでなく，社会科学の研究として実施されるインタビュー調査や質問紙調査，フィールドワークなども含まれています。人を対象とする研究において，多くの国々で共有されている倫理原則は，人格の尊重（研究参加者の人格を尊重し，尊厳が保たれること），善行（科学的社会的に妥当性が高く，危険性に勝る利益が期待できること），無危害（科学的社会的な意義が高い研究であっても，研究参加者への危害が上回ってはならないこと），正義（研究参加者を公正に選定し，社会正義に反することのない研究であること）が挙げられています[1]。

人を対象とする研究を実施する研究者は，これらの原則が，研究の最初から最後まで貫かれるように配慮して，必要な手順を踏んだ上で研究を実施しなければなりません。これらの原則が守られた研究のみが実施されるようにするため，人文・社会科学系の研究計画の倫理審査を行う研究機関も増えています。また，学術誌によっては，査読の際に倫理審査の実施やインフォームド・コンセントの手続きの説明を求められます。そのため，事前の研究計画の審査，インフォームド・コンセントの取得などの手順を前提とした研究計画の立案が必要になってきました。

人を対象とする研究計画を実施してよいかどうかは，研究から得られる利益（学術的な成果や人々に還元できる知見など）の見込みと，侵襲（精神的なものを含む）や介入（人の意識や行動に影響を与える要因の有無または程度を制御する行為）の程度に応じた研究参加者への負担とリスクのバランスによって決定される必要があります。素晴らしい目的をもち，その成果が学術的にも社会的にも強く期待される研究であったとしても，その達成のために研究参加者への負担やリスクが上回ることは許されません。そのため，研究者は研究参加者の立場に立って，様々な配慮をする必要があります。

たとえば，質問紙調査の場合，設問文や実施方法について回答者がどう受け止めるかはわかりません。回答者にとって許容しがたい不快感や負担感の発生，トラウマの惹起などの被害を事前に予測することは難しいものです。そのため，設問を厳選して分量の軽減に努める，休憩をはさみながら答えられる環境をつくる，設問の意図について断り書きを入れる，機微に触れる質問項目を最小限にする，回答拒否の機会を保障するなどの工夫をする必要があります。

また，研究対象者は，これから自分が経験する研究の目的や内容，負担やリスクについて十分に説明を受け，研究に協力するかどうかを自由に意思決定する権利があります。一度，同意をしたとしても，理由の如何を問わず，その意思を撤回する権利もあります。

こうした考え方は，データ共有においても同様です。

---

1 Jonsen, Albert R., 1998, *The Birth of Bioethics*, Oxford, Oxford University Press.（細見博志訳，2009，『生命倫理学の誕生』勁草書房.）

7.2 事前の研究計画の審査

研究者が立案する研究計画には，研究が進展した後のデータ共有の可能性やその目的，意義，リスク，データ共有の方法，データの格納場所についても含める必要があります。

研究倫理審査委員会は，研究機関に設置され，研究計画の科学的倫理的妥当性について，様々な立場から構成される第三者によって判断するための仕組みです。研究を開始するときに加え，研究計画を変更する必要が生じるとき（研究開始当初はデータ共有を予定していなかった場合など）に，審査が行われます。

現時点の日本では，人を対象とする研究計画の科学的倫理的妥当性について，事前に研究倫理審査委員会で審査を受けることを必須とするかどうかは研究領域によって異なり，人文・社会科学系の研究では必ずしも必須ではありません。

ただし，人文・社会科学系研究であっても，論文を投稿する学術誌の編集部または査読者より，研究倫理審査委員会の承認を受けて行った研究かどうかを確認される事例が増加しているほか，論文採択の条件としてデータ共有を求められる事例が増えています。

さらに，研究から得られたデータの寄託を受けるデータアーカイブでは，調査の回答者がデータ共有についても説明を受けて同意を与えたデータの寄託を受けたいと考えており，その確認の責任を倫理審査委員会に求めている場合もあります。そのため，研究倫理審査委員会による許可を受けた研究計画かどうかは，データを寄託しようとしているデータアーカイブからも確認される場合があるので注意してください。

7.3 データ共有に伴って研究参加者に与えるリスク

データの共有に伴って研究参加者に与えるリスクには，一般的に以下のようなものが挙げられます。

- 個人の身元が推測されること，または判明することによる不快や衝撃の付与
- 個人の身元情報が悪用され，社会的な不利益を受ける可能性やその恐怖の付与
- 個人の家族，所属する集団や地域に対して，社会的な不利益やスティグマが付与される可能性やその恐怖の付与

こうしたリスクを軽減する手段として，共有前の段階でデータから研究参加者に関する情報を匿名化[2]することが挙げられます。匿名化は，研究参加者の身元を特定できるような情報，たとえば，氏名，居住地，所属機関，職業などを記号に置き換えたり，マスキングしたりすることで可能となります。

ただし，研究領域によっては，匿名化が進みすぎることで研究の意義が損なわれ，二次利用の意義が失われる場合があります。どの程度の匿名化を行うかは，事前に与える説明の中で説明する必要があります。第三者から見て，本人の身元を特定しうる状態でデータ共有を行う場合には，その旨とリスクについて同意を得る必要があります。

2　データに含まれる情報が具体的にどの調査対象に関するものかデータの利用者には分からないように加工を施すこと。

7.4 インフォームド・コンセント

人を対象とする研究では，原則として，研究参加者への負担や予測されるリスク，利益を総合的に評価し，事前の十分な説明および自由意思による同意を取得した上で，研究を実施しなければなりません。

ただし，事前の説明および同意取得が研究目的の達成を阻害するような研究の場合には，代わりに事後の研究参加者への謝罪と丁寧な説明を加えることによって，例外的に許容される場合があります（デブリーフィング）。

社会的に弱い立場にある者を研究対象とする場合には，研究のあらゆる側面で特別な配慮が必要となります。説明を理解し同意を与える能力がない，または制限されている者（①16 歳未満の者，②成年であって，インフォームド・コンセント[3] を与える能力を欠くと客観的に判断される者，③死者であって，研究の実施が生前の明示的な意思に反していない者）を研究対象とする場合には，例外的に，その人の意思や利益を代弁できると考えられる人による代諾の可能性を検討してよいです。

先住民族を対象とした研究においては，研究目的や実施方法について，その代表者と事前の協議が必要であり，データの寄託についても知らせておくべきです[4]。 その他，学校や地域住民，職場などを対象とした調査や研究でも，その代表者との事前協議は重要です。特に，社会的少数者集団（マイノリティグループ）を対象にした場合には，公表された研究結果やデータの寄託によって研究参加者が不利益を被らないよう，特段の配慮が必要となります。

研究が進展した後のデータ共有の方針については，事前に説明すべき事項に含まれます。データ共有に関して，説明すべき内容は，一般的に以下の通りです。

- データの匿名化の程度
- データの保管場所
- データの二次利用の目的
- データの二次利用申請を認める手続き
- データの二次利用の状況を知る手段
- データの二次利用開始後のデータ削除要望への対応

同意の取得段階で，データ共有方針が具体的に決定していない場合には，追加的な情報提供と同意の取得をすることが認められます。また，研究参加者の身元を推測できるような状況でデータ共有を行う場合には，その旨とリスクも説明して同意を得る必要があります。

なお，国の「人を対象とする生命科学・医学系研究に関する倫理指針」では，データ共有について事前に説明しておらず，同意を取得していない場合には，研究倫理審査委員会で研究参加者に与える不利益の可能性を検討した上で，その可否判断を受け，①再同意の取得，②通知又は公開と拒否の機会の保障，③通知又は公開，のいずれかの対応を取ることになっているので，参考にしてください。

---

3 研究対象者などが，調査や研究の目的，方法，研究対象者に生じる負担，個人情報やデータの取扱などについて十分な説明を受け，自由意思に基づいて研究者に与える，当該研究の実施に関する同意のこと。個人情報を取得しない調査では，回答をもって協力の同意があったとみなすことができます。

4 「先住民族の権利に関する国際連合宣言」，「生物多様性条約」第 8 条（j）項の先住民族関連条項，「同条約戦略計画の目標 4」および「これからのアイヌ人骨・副葬品に係る調査研究の在り方に関するラウンドテーブル報告書」で保障されているものです。

7.5 同意撤回の申出への対応

研究参加者からの同意の撤回の申出があった場合には，随時，その意思を尊重する必要があります。同意の撤回の申出方法や，その場合に対応できること（撤回の効果）や対応できないことについては，事前に説明しておく必要があります。そして，もしそのような申出があった場合には，研究参加者と丁寧なコミュニケーションをはかることが不可欠です。

同意の撤回の申出に対してどのような対応をとれるかは，調査や研究の進展状況によって異なります。たとえば，定期的な調査協力を依頼している場合は，今後の調査協力の依頼を停止し，それまでのデータは使用の許可を得られるかを相談する方法があります。既にデータ収集が終わり，分析の過程にある場合には，申出者のデータを削除した分析ができないこともあるため，継続使用の許可を得られるかを相談し，データ寄託時には削除する方法があります。

データの寄託後や二次分析が進んでいる段階でのデータの廃棄や不使用の希望については応じることができないため，事前に説明しておく必要があります。しかし，事前にそのように説明し，同意を得ていたとしても，研究参加者や代諾者からデータの削除を要望される可能性はあります。たとえば，高齢になったから，調査に協力していた家族が亡くなったから，個人情報漏えいや研究不正などの報道を契機として個人情報保護や学術研究に対する不信を持ったから，といった実例があります。まず，研究参加者などがどのような事情からデータ削除を申し出てきたのかを確認する必要があります。そして，インフォームド・コンセントで同意を得ていた内容を，丁寧に説明しましょう。研究参加者などが事前に説明したことを覚えていないことも多く，あらためて説明をすることで理解が得られることもあります。しかし，それでも理解が得られない場合には，データアーカイブ側に相談し，二次利用をいったん停止して，データを削除したデータセットを共有できる可能性があるかどうかを検討してみましょう。

【人文学向けコラム 6（人文学におけるデータ共有に関する倫理的側面）】

人文学におけるデータに関する倫理的側面は，社会科学において要求されている事柄については同様に考えるべきです。そして，紙媒体において機微情報として配慮されてきた事柄は，データにおいても少なくとも同等かあるいはそれ以上の配慮が必要となります。とりわけ，注意しておきたいのは，データの場合，デジタル情報としてインターネットを通じて急速に拡散してしまうことがありますので，データ共有をする際には，内容のみならず共有範囲についても十分に検討する必要があります。それにあっては，アクセス制限などの技術的な対応だけでなく，適切な運用のためにデータ利用のガイドラインを作成することも有用です。

また，著作権保護期間は終了しているものの資料所蔵者への配慮が必要なものがあり，それをデータ化した場合にも同様の配慮が必要です。本来は，パブリックドメインと位置づけられ，特に平面画像としてデジタル化された複製物の流通を制限することは基本的にできません。しかしながら，所蔵者の意向に反する状況になると資料の閲覧自体が困難になることもあります。特に貴重な資料の場合には，著作権の概念が成立する以前，数百年から千年以上に渡り保管の労を執っている所蔵者も少なくなく，そうした貴重な営為をないがしろにしてしまうことがないようによく注意する必要があります。また，所蔵者が自由な利用条件での公開に前向きになることが

ありますが，そのような場合には，いずれ訪れる責任者の交代に備えて書面で利用条件に関する覚え書きなどを交わしておくことが大切です。

8 個人情報と匿名化について

日本には様々な個人情報保護法制や条例があります。学術研究の目的のために個人情報を利用する研究者は，個人情報取扱事業者としての適用を受けない場合もありますが（個人情報保護法第 76 条第 3 項），個人情報の基本的な取り扱い方を知っておくことは大切です。ここでは，個人情報の考え方と，寄託するデータに個人情報を保護するために行う匿名化についての留意点について述べます。

8.1 個人情報の定義

個人情報保護法[1]における「個人情報」とは，「生存する個人に関する情報であって，当該情報に含まれる氏名，生年月日その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ，それにより特定の個人を識別することができることとなるものを含みます。）」をいいます（個人情報保護法第 2 条第 1 項）。

ただし，地方公共団体の条例では，死者の個人情報も定義に含んでいる場合があるので，特定の地域を対象にして調査を行っている場合には注意が必要です。

個人情報としての取り扱いが求められるものに，個人識別符号があります。個人識別符号とは，「特定の個人の身体の一部の特徴を電子計算機の用に供するために変換した文字，番号，記号その他の符号であって，当該特定の個人を識別することができるもの」とされています（個人情報保護法第 2 条第 2 項第 1 号）。また，「個人に提供される役務の利用若しくは個人に販売される商品の購入に関し割り当てられ，又は個人に発行されるカードその他の書類に記載され，若しくは電磁的方式により記録された文字，番号，記号その他の符号であって，その利用者若しくは購入者又は発行を受ける者ごとに異なるものとなるように割り当てられ，又は記載され，若しくは記録されることにより，特定の利用者若しくは購入者又は発行を受ける者を識別することができるもの」も個人識別符号に該当します（個人情報保護法第 2 条第 2 項第 2 号）。それぞれについて具体的には，モーションキャプチャデータや指紋認証データのように，個人の特性を示しうるデータ（個人情報保護法施行令第 1 条第 1 号），さらに旅券番号，基礎年金番号のような個人に割り当てられた記号や番号が政令などで指定されています（個人情報保護法施行令第 1 条第 2 号〜第 8 号，同施行規則第 3 条および第 4 条）。個人識別符号は，個人情報としての保護が必要であり，利用目的を特定した同意を取得する必要があります。また，原則として，変更前の利用目的との関連性を有すると合理的に認められる範囲を超えた変更はできません。

さらに，個人情報のなかでも，一段高い保護を求められるカテゴリーとして，「要配慮個人情報」があります。「要配慮個人情報」とは，「本人の人種，信条，社会的身分，病歴，犯罪の経歴，犯罪により害を被った事実その他本人に対する不当な差別，偏見その他の不利益が生じないようにその取り扱いに特に配慮を要するものとして政令で定める記述などが含まれる個人情報」をいい，詳しくは政令などで指定されています。（個人情報保護法第 2 条第 3 項，同施行令第 2 条，同施行規則第 5 条）。「要配慮個人情報」は，本人の同意がない取得や第三者への提供が原則として禁止されています。

1　本節では，現行の個人情報保護法に基づいて記述をしています。2020 年 6 月に「令和 2 年 改正個人情報保護法」が国会で成立しましたが，それに伴う個人情報の法的な取り扱いに関しては，改正後の個人情報保護法および関連する法令などの条文をご参照ください。

8.2 個人情報の該当性の要件

その情報が個人情報に該当するかどうかは機械的に決められるものではなく，個人特定性，個人識別性，容易照合性という観点から確認する必要があります。個人特定性とは，その個人の氏名にたどりつけるかどうかを意味しています。個人識別性とは，個人特定性がなくても，複数の識別子を使えば一定の集団の中からある特徴をもつ個人を識別できるかどうかを意味しています。容易照合性とは，一見，個人特定性がないように見えても，識別子をたどれば個人に紐づけられる可能性があるかどうかを意味しています。

社会調査のデータを含む研究で用いるデータにおいて，個人識別性があるだけでなく，個人特定性を有する具体的なケースとしては，以下の 2 つが考えられます。

①データの保存の観点から，研究で用いるデータセットだけでなく，名前などの直接的な識別子を含む調査対象者のリストも調査実施者以外の第三者に提供することによって，データセットと調査対象者のリストとの連結が可能になり，個人特定性を有する場合

②研究で用いるデータに名前などの直接的な識別子は含まれないものの，(1) 特定の調査対象者がそのデータに含まれていることを知っている，(2) 詳細な地域情報（たとえば地点情報など）がデータに含まれている，(3) 外観識別性[2] を持つ属性がデータに含まれるといった理由で，個人識別性があるだけでなく，データが容易に個人に紐づけられる可能性があるため，個人特定性を有する場合

こうした事情から，研究で用いるデータのうち，一見，直接的に個人情報とは見えない情報も，個人情報としての取り扱いが求められる場合があります。

その上で匿名化に関する方針を検討する必要があります。その際に，データの利用目的，利用場所，利用対象，利用の仕方について検討し，どのような二次利用を許容するか，どのような二次分析が行われるかを想定しましょう。

①利用目的　　学術・研究目的，教育目的などの利用目的の設定

②利用場所　　研究室といった学術研究機関における指定された場所あるいは技術的な安全管理措置が施されたセキュアなオンサイト施設などの指定

③利用対象　　特定の利用者，特定の分野の研究者，「データ提供者が認めた者」など

④利用の仕方　分析に最低限必要な変数の指定，分析に使用する集計表やモデルの指定，分析結果のチェック（output checking），オンデマンド集計，リモートアクセスによる個票データの提供，プログラム送付型のリモートエグゼキューションなど

2　外観識別性とは，外から見た場合に個人の特定に結び付く可能性があることを意味します。外観識別性を有する属性の例としては，個人に関する性別や国籍，住宅における建物の大きさや構造などが考えられます。

【公的統計における調査票情報について】

統計作成部局は，公表の対象となる統計表を作成するために，統計調査票を用いて調査対象者に統計調査を実施することによって，記入済みの調査票を収集します。記入済みの調査票には調

査対象者一人一人に関する情報が含まれますが，この記入済みの調査票に基づいて統計表の元になる調査票情報（個票データ）が生成されます。

調査票情報は，「統計調査によって集められた情報のうち，文書，図画又は電磁的記録（電子的方式，磁気的方式その他人の知覚によっては認識することができない方式で作られた記録をいう。）に記録されているもの」と統計法（平成 19 年法律第 53 号）第 2 条第 11 項で規定されており，法的には，公的統計[3]の調査票情報に含まれる個人情報（行政機関個人情報保護法などで定義されている個人情報）は，行政機関個人情報保護法などの規定の適用を受けません（統計法第 52 条）。また，調査票情報は，複数の調査項目（属性)と調査項目の回答値（属性値）の属性群から構成されるレコード群を表していますが，名前や住所のような個別主体の直接的な識別子が含まれていない個票データに関しては，非識別データ（deidentified data）だということができます。

なお，「調査票情報等の管理及び情報漏えい等の対策に関するガイドライン」（平成 21 年 2 月 6 日総務省政策統括官（統計基準担当）決定，平成 31 年 4 月 19 日最終改正，以下，「ガイドライン」といいます）によれば，基幹統計調査の場合，記入済みの調査票については，調査規則が定める期間中は，調査実施者などによって保存することが求められています。調査票情報についても同様に，調査規則で規定されている期間は保存する必要がありますが，調査票情報の場合，基本的には永年保存することが「ガイドライン」で明記されています。

3　国の行政機関，地方公共団体又は独立行政法人などが作成する統計（統計法第 2 条第 3 項）。官庁統計・政府統計と言われることもあります。

8.3 安全な二次利用を目指した匿名化

次に，匿名化にあたっては，データセットにどのような識別子が含まれているかを確認しなければなりません。

その情報単体で個人を特定できる情報のことを，直接識別子（あるいは識別子）といいます。具体的には，氏名，社員番号，会員番号などが挙げられます。これらは，一般的には，二次利用や二次分析に不必要であり，削除する必要があります。ただし，直接識別子を削除することによって二次利用や二次分析の意義を失わせることにつながるため，新たに付与した別の数字やカテゴリーに置き換えるなどの方法によって生かすことは可能です。また，削除しないことについて研究参加者（調査回答者）から同意を得ている場合は例外となります。

また，稀少性の高いデータセットなどの場合には，直接識別子のみを削除しただけでは，匿名化が不十分となる場合があります。そのため，間接識別子（準識別子）と呼ばれる，他の情報と容易に照合することができ，それにより特定の個人を識別することができることとなるものについても考慮が必要です。間接識別子の例としては，性別，年齢，居住地情報，学歴などが挙げられますが，間接識別子は，研究にとって重要なデータでもあるため，一般化したり，再グルーピングしたりすることによって，二次利用に生かせるような処理を検討する必要があります。

たとえば，人口の少ない地域の住民から得たデータなどでは，年齢と性別だけで個人が特定しうる場合があります。そのため，安全に二次利用に生かすことを目的として，他の地域と同一カテゴリーに置き換える，年齢区分を大きなグループに区切るといった処理を施してみます。そうした処理が困難な場合には，データセットからのさらなる匿名化を検討します。

【公的統計における匿名化について】

公的統計における匿名化措置は，(1) 個体情報の秘密保護に関する統計法制度的な措置と (2) 技術的な匿名化手法の適用に類別されます。個体情報の秘密保護に関する法的制度的な措置としては，①ミクロデータ提供に関する法的な規定（適正管理，守秘義務などの規定），②ミクロデータの提供に関する審査機関の存在（各府省・統計センター），③チェックリストに基づく露見リスク（disclosure risk）の評価（オンサイト利用における提供審査）があります。

一方，技術的な匿名化手法の適用に関しては，統計作成部局によって様々な技法が追究されています。公的統計ミクロデータに適用される匿名化技法は，①非攪乱的（non-perturbative）な手法と②攪乱的な（perturbative）な手法に大別されます。非攪乱的手法には，リコーディング（recoding），データの削除（suppression），トップ（ボトム)・コーディング，サンプリング（リサンプリング）といった手法があります。また，攪乱的な手法に関しては，ノイズの付加，スワッピング（data swapping）といった技法が含まれます。秘匿性の強度や有用性（情報量損失）の程度だけでなくて，統計実務面も考慮した上で，適用可能な技法の組み合わせが統計作成部局によって選択されます。

わが国の統計法制度の下では，基幹統計を対象に，公的統計の調査票情報に対して各種の匿名化技法を適用することによって，「匿名データ」（統計法第 2 条第 12 項で定義）が作成・提供されています（法第 35 条「匿名データの作成」，法第 36 条「匿名データの提供」）。匿名データの作成においては，主としてリサンプリング，リコーディング，トップ（ボトム)・コーディングといった非攪乱的な手法が用いられてきており，わが国で攪乱的手法（パータベーション）が採用されることは少ないようです。ただし，国勢調査の匿名データの作成においてスワッピングが適用されたという事例が存在します。

8.4 匿名化処理の安全性の確認

匿名化処理の後，適切に処理がなされたかどうかを確認する必要があります。確認するポイントとしては，以下の通りです。

- 対応表を残している場合，元情報や他の個人情報などと容易に照合できないですか？
- 共有・公開した先（提供先機関）において特定の個人を再識別化されることはないですか？
- 公表した場合に，個人を再識別化されることはないですか？（再識別化テストによる検証など）

8.5 データ共有後に匿名化が不十分であったことに気づいた場合

匿名化処理を適切に行ったつもりでも，データ寄託後に匿名化が不十分であったことに気づいたり，二次利用した研究者から指摘を受けたりすることがあります。

その場合には，直ちにデータを寄託したデータアーカイブに連絡し，再度の匿名化処理を行ったデータに差し替えられるか，それまでの間は新たな二次利用の休止ができるかどうかを相談しましょう。

9 データに関する著作権

9.1 著作権についての一般的な考え方

著作権は，著作物を生み出した著作者に対して自動的に与えられる知的財産権であり，これにより，元の著作物が著作者の許諾を得ないままに複製されたり，発行されたりするのを防ぐ役割を果たしています。

著作物とは，「思想又は感情を創作的に表現したものであつて，文芸，学術，美術又は音楽の範囲に属するもの」（著作権法第 2 条第 1 項第 1 号）であり，「思想又は感情」，「表現したもの」，「創作的」，「文芸，学術，美術又は音楽の範囲」の要件から，それぞれ，単なるデータ，アイディア，他人の作品の単なる模倣，工業製品などが除かれます。

著作者とは，「著作物を創作する者」（第 2 条第 1 項第 2 号）です。

著作者が有する権利は，財産権としての著作権と，人格権としての著作者人格権に分けられます（第 17 条）。財産権としての著作権については，複製権，公衆送信権，翻案権など（第 21 条以下）の種々の権利が含まれる一方，私的使用のための複製，引用（第 30 条以下）など一定の制限があります。

著作権の保護期間は，原則として著作者の死後 70 年です（第 51 条以下）。

9.2 データの著作権

(1) 著作物としての保護が与えられる可能性のあるデータについて

一般的に，単なるデータ[1]は著作物にあたりませんが，創作的な表現にあたる場合は，著作物としての保護を受け，データを作成する研究者が各自のデータの著作権を有します（データに限らず，研究者が生み出す研究成果のすべては，創作的な表現にあたることを条件に，著作権で保護されます）。したがって，研究者 A が作成したデータ α の著作権は A に帰属するため，別の研究者 B が α を利用する際には，A から許諾を得るのが原則です。

なお，データの作成者は自動的に当該データの最初の著作権者になりますが，例外的に，異なる著作権の割り当てを定める契約が有効である場合や，著作権者の署名入りの著作権譲渡書面がある場合には，異なる取り扱いをすることができます。

著作権による保護の対象となりうるデータには，どのようなものが含まれるでしょうか。たとえば，数値などの単なる事実の提示は，「思想又は感情」を含まないため保護の対象となりません（たとえば，富士山の標高など）。また，誰が行っても同じ結果が出てくる計算データも保護の対象になりません。他方で，それらのデータを収集し，一定の考え方のもとに複数の事実を組み合わせて整理したものなど，結果の表現方法において創意工夫が見られるものは，創作性があるものとして，保護の対象となります。

データは，創作的な表現にあたることを条件にそれとして著作権が認められる場合のほか，「編集著作物」（第 12 条），「データベースの著作物」（第 12 条の 2 第 1 項），「図形の著作物」（第 10 条第 1 項第 6 号）として保護の対象になる場合があると考えられます。

(2) データベースの著作物

データベースの著作物とは，「論文，数値，図形その他の情報の集合物であって，それらの情報を電子計算機を用いて検索することができるように体系的に構成したもの」（第 2 条第 1 項第 10 号の 3）であり，「データベースでその情報の選択又は体系的な構成によって創作性を有するものは，著作物として保護」（第 12 条の 2 第 1 項）されます。

したがって，データベース内にある数値などのデータそのものは著作物ではありませんが，それらの情報の集合物であるデータベースは，情報の選択や体系的な構成に創作性があれば著作物として保護されます（例として，学術論文の書誌情報や全文を蓄積したデータベースなど）。

編集著作物との相違点は，「素材の配列」ではなく「情報の体系的な構成」に著作物としての重要な要素を認めている点であり，「体系的な構成」とは，コンピュータで検索するためのコード，個々の情報の属性（数値，文字など），情報の文字数や桁数などを設定し，それに従って情報を整理し，組み立てることです [2]。

---

1 本章においては，主に社会科学の量的データを対象にしているため，論文，写真，画像，映像を研究目的で利用する場合の著作権については，別途議論がありえます。

2 文化庁，2020，「著作権制度に関する情報」，文化庁ホームページ，（2020 年 2 月 27 日取得，https://www.bunka.go.jp/seisaku/chosakuken/seidokaisetsu/index.html）.

9.3 データ共有上の法的問題

(1) 著作権を有する主体

（ア）共同研究などの場合

データが複数の研究者・研究組織による共同研究の成果に基づく場合，または，データが多様なデータおよび資料に基づく場合には，元となる各データの著作権がそれぞれの作成者にどのように割り当てられているかを確認する必要があります。

（イ）職務著作などの場合

研究機関に雇用されている研究者が雇用中に作成したデータの著作権者は，雇用者であるということも考えられますが（＊），実際には，多くの研究機関は，データ，研究資料，それに基づく発行物などの著作権を研究者に付与しています。研究者は，所属している研究機関が著作権をどのように割り当てているかを確認する必要があります。

（＊）職務著作・法人著作について

「法人その他使用者（...）の発意に基づきその法人等の業務に従事する者が職務上作成する著作物（...）で，その法人等が自己の著作の名義の下に公表するものの著作者は，その作成の時における契約，勤務規則その他に別段の定めがない限り，その法人等とする。」（著作権法第 15 条）

(2) データの二次利用

（ア）データアーカイブを介さない場合

データの二次利用者は，データを複製する前に，当該データの著作権保持者から許諾を得なければなりません。研究倫理の問題として，データの二次利用者は，研究成果の謝辞において，利用するデータの出典，データの提供者，著作権保持者の貢献を明示する必要があります [3]。

（イ）データアーカイブを介する場合

研究者はデータをデータアーカイブに寄託することができます。寄託とは，収集したデータなどをデータアーカイブやリポジトリに預けることをいいます。データアーカイブを介してデータが共有される場合においても，データ作成者が当該データの著作権を保持することに変わりはなく，データアーカイブに対してデータの処理および提供を許諾する形をとります。すべての著作権保持者が特定され，かつ，各々の著作権保持者が各データの保存と提供を許可した場合に，データアーカイブはデータの寄託を受けることができます（もちろん，データアーカイブに寄託されるデータは，著作権の保護が与えられるデータに限りません）。データの二次利用を考える研究者は，データアーカイブが定める条件にしたがい，データアーカイブとの間でデータの利用に関する契約を結ぶことになります。研究倫理の問題として，データの二次利用者は，研究成果の謝辞において，利用するデータの出典，データの提供者，著作権保持者の貢献を明示する必要があります。

なお，公的統計データにおける統計表（集計結果表）などを対象にした著作権の考え方および取り扱いについては，コラム「公的統計における統計表の著作権について」を参照してください。

3 Eynden, Veerle Van den, Louise Corti, Matthew Woollard, Libby Bishop and Laurence Horton, 2011, *Managing and Sharing Data: Best Practice for Researchers*, Colchester, Essex: UK Data Archive University of Essex, p.29（東京大学社会科学研究所附属社会調査・データアーカイブ研究センター訳，2013，『データの管理と共有――研究者向け最良事例』東京大学社会科学研究所附属社会調査・データアーカイブ研究センター．29 頁）によると，イギリスにおいては，非営利の教育または研究目的の場合，データの著作権者に対する謝辞が掲載されることを前提に，公正な取り扱いという考えの下で，著作権を侵害することなくデータを複製することができるとされています。

【人文学向けコラム 7（人文学のデータに関する著作権）】

著作権は，「思想又は感情を創作的に表現したものであって，文芸，学術，美術又は音楽の範囲に属するもの」に対して発生する権利です。ここで述べられている点は人文学でも十分に踏まえるべき事柄ですが，さらに，人文学における学術的なデータの中には，翻刻データや校合テキストのデータなど，作成に高度な知識を必要とし，学術の発展に大きく貢献するにもかかわらず，創作性を見出しがたいために著作物とは言い切れないデータを作成する場合があります。また，前コラム（人文学向けコラム 6）で述べたように，著作権保護期間が終了した資料であっても倫理的観点からその利用について配慮する必要が生じる場合もあります。

このような課題はデジタル化において世界的に顕在化しており，特に前者については，文学分野における米国現代語学文学協会（Modern Language Association：MLA）や歴史学分野における米国歴史協会（American Historical Association：AHA）のように，研究に資するデジタル成果物を研究コミュニティへの貢献として評価するためのガイドラインを策定している学会も出てきています。（AHA のガイドライン日本語訳: https://www.jadh.org/guidelines-for-the-evaluation-of-digital-scholarship-in-history）

今のところは米国中心に活動する学会が主ですが，日本でも今後対応していくことが期待されます。

【データの二次利用とクリエイティブ・コモンズ・ライセンス】

現在，普及しつつある「クリエイティブ・コモンズ・ライセンス」（CC ライセンス）を学術の世界で活用することができれば，データの著作権を有する研究者が，データの著作権を保持したまま，データを自由に流通させることができ，データの受け手もライセンス条件の範囲内でデー

タの再配布や結合を行うことが可能になります（「クリエイティブ・コモンズ・ジャパン」Web サイト参照）。
https://creativecommons.jp/licenses/

【公的統計における統計表の著作権について】

公的統計の統計表（集計結果表）は，単位（統計単位），属性（集計事項，調査事項），場所（地域）と時間を規定することによって作成されます。『統計学辞典』[4] によれば，統計表は，「いくつかの変数値区分の組み合わせの各セルごとに，対応する観測単位数，あるいは観測単位のもつ値の平均値や比率などの統計量を表示した，表形式の集計データ」と定義されています。わが国の統計法の下では，統計ないしは統計表の定義は明文化されていませんが，一般に公表されている統計表は，調査票から集められた個体情報（統計単位の情報）に基づいて，統計作成部局によって選定された調査事項をクロスさせた集計表であるとみなされます。

公的統計において公表される統計表は，集計計画において予め定められます。集計計画では，統計表の表頭（列）と表側（行）に含まれる集計事項，および集計事項の分類区分が設定されています。これらの統計表の表形式そのものに関しては，統計法上の規定はないものの，その著作権は集計計画の策定者である統計作成部局に帰属すると考えることもできます。しかしながら，統計調査によって収集された記入済みの調査票から集計された統計表は，許諾を要することなく，誰でも Web からアクセス可能な形で入手できるオープンデータです。その理由は，公的統計において公表されている統計表さらには統計表に含まれる結果数値は，「国民の共有財産とし，広く活用する」とされているからです。したがって，公的統計における統計表については，著作権という概念は当てはまりません。ただし，総務省統計局のように，統計作成部局によっては，政府標準利用規約に基づき，統計調査の統計データを引用・転載する場合に出典の表記を求めていることがあります。なお，公的統計データを分析に利用した際の分析プログラムには，著作権があるとされています。さらに，公的統計データが用いられる白書類には著作権はありますが，出所を明らかにすれば，白書類を刊行する官公庁などに対して許諾を得る必要はないと考えられます。

4　竹内啓編，1989，『統計学辞典』東洋経済新報社．

10 データアーカイブの役割

10.1 データアーカイブとは

データアーカイブとは，主に研究・教育を目的として，幅広いデータを収集・保存し，提供するサービスを行う組織・機関です。社会科学分野の代表的なデータアーカイブとしては，米国のICPSR，ヨーロッパ諸国の GESIS（ドイツ）や UKDA（イギリス），DANS（オランダ）などがあります。過去の調査で得られたデータがデータアーカイブに寄託されることで，新たな角度からの分析や比較研究の実施，再調査のコスト節約といった効果が期待でき，ひいては人文学・社会科学の発展に寄与します。

また，研究者の立場からデータアーカイブの重要性を見ると，データの寄託は通常，調査研究プロジェクトの終了後に行われ，データを保存する責任は，アーカイブに特化した部門が引き受けます[1]。保有するデータのスムーズな引き渡しにあたっては，データの保存と物理的ストレージに伴う基本的な知識が必要です。本章では，データアーカイブを有効に活用するために必要な知識について紹介します。

1　日本学術振興会の「人文学・社会科学データインフラストラクチャー構築推進事業」では，以下の機関に「拠点機関におけるデータ共有基盤の構築・強化委託業務」の業務委託を行っています。
東京大学社会科学研究所附属社会調査・データアーカイブ研究センター (SSJDA): https://csrda.iss.u-tokyo.ac.jp/
慶應義塾大学経済学部附属経済研究所パネルデータ設計・解析センター: https://www.pdrc.keio.ac.jp/
一橋大学経済研究所: http://www.ier.hit-u.ac.jp/Japanese/index.html
大阪商業大学 JGSS 研究センター: https://jgss.daishodai.ac.jp/
東京大学史料編纂所:https://www.hi.u-tokyo.ac.jp/

【人文学向けコラム 8（人文学におけるデータアーカイブの役割）】

人文学分野では，資料の種類とそれに基づくデータがそれぞれに多様であるのと同様に，データのアーカイブも多様です。代表的なものとしては，ERIC（European Research Infrastructure Consortium）の傘下で活動する CLARIN（https://www.clarin.eu/）が，欧州各国の歴史学も含めた人文学全般のコーパスや辞書，ツールなどを集約したサイトとして運営されています。また，資料を所蔵する機関による書誌・目録データの公開も世界各地で行われており，各国の図書館・博物館・公文書館が目録の検索や画像データの公開など様々な形でデジタルデータを Web 公開しています。そして，米国国立公文書館では市民アーキビスト向けの参加型公文書翻刻サイトを提供したり，英国図書館では芝居のビラのクラウドソーシング翻刻サイトを運用するなど，広く市民が基礎データの構築に参加できるようにしています。さらに，こうしたデータの中には横断検索を通じて発見性を向上させているものもあります。たとえば，欧州各国の公文書館のポータルサイトとして Archives Portal Europe が運用されており，さらに，これも含めた様々な文化機関の資料目録の横断検索ができるポータルサイトとして Europeana が提供されています。

10.2 データアーカイブの機能とサービス

データアーカイブが持つ主な機能として，データの適切な保存・管理，長期的なアクセスの提供および研究分野へのサポートがあります。これらを通じて，科学研究の透明性，蓄積，効率的な再利用を促進する役割を果たしています。以下では，データアーカイブが担う業務の概要と，データを登録する際の一般的な注意点を紹介します。

(1) データの保存・管理

データアーカイブでは，データの保存・管理のため以下の業務を行っています。

提供されたデータの受入・保管
メタデータの管理・保管
データの変更を含むデータキュレーション（匿名化などを含む）
データのバックアップの取得・保管
データの復元を実施するためのバックアップシステムの確保
上述した管理などを行うためのデータ管理システムの開発・運営（情報システム上の対応のみならず，制度・組織間のスキームを含む）

データやメタデータの受入対象や基準，準備すべき情報は，各データアーカイブが持つポリシーによって異なりますので，各データアーカイブの Web サイトをご覧ください。

(2) 長期的なアクセスの提供

データアーカイブでは，保管するデータおよびメタデータに対して長期的なアクセスを保障すべく，データカタログを公開するとともに以下の情報を提供しています。

データへのアクセス手順
データの利用条件（目的，利用期間など）
データおよびメタデータの版（バージョン），および変更履歴

幅広い研究者による二次分析を可能にする観点からは，データのアクセス手順や利用条件は可能な限り簡便かつ緩やかであることが望ましいと考えられます。また，調査回答者との個別契約により限定的な共有条件などが定められている場合は，利用上の注意を合わせて表示すると詳細がより分かりやすくなります。データアーカイブの担当者とも相談の上，簡潔な利用案内を作成しましょう。

(3) 研究分野へのサポート

データアーカイブでは，国内外の研究コミュニティとの連携を通じたデータの普及，二次分析を促進するための技術開発や研修実施といった活動を行い，データの有効活用や統計調査・社会調査の水準維持・向上に寄与しています。具体的には，以下のような効用が期待できます。

データ管理，二次分析に関する十分な訓練を受けた研究者の確保
データキュレーター，データアーキビストの専門性維持
最先端のデータ分析スキルの迅速な実装
データ分析の方法論の開発に関連するアイディアの交換を促進する環境の提供

データアーカイブは,データ管理やデータ分析に関する最新の動向がいち早く集まる場でもあり，情報収集やネットワーク作りの場としても活用できます。

10.3 データアーカイブの今後の展開

長年にわたるデータアーカイブ活動によって，各国のデータアーカイブには多数かつ多様なデータが蓄積されています。また，近年のデータ集約型科学への期待の高まり，オープンサイエンスの潮流などを受け，データアーカイブには専門分野の垣根を超えた横断検索サービス，分析サービスなどが求められつつあります。このような背景の中，データアーカイブの価値を高める取り組みとして，国内でもデータアーカイブ間の統合的なデータカタログ提供システム開発，オンライン分析システムとの連携などが試みられています。

また，データアーカイブはデータの適切な保存・管理，長期的なアクセスの提供および研究分野へのサポートを行う基盤となります。ここで，データアーカイブを支えるシステム（組織，技術，資金，ポリシーなど）の持続可能性はその根幹に関わる問題となります。データアーカイブの信頼性を高め，ユーザーと資金提供者に対してデータアーカイブの価値を示していくために，データリポジトリ[2]の信頼性を実証するための指針の策定や，データリポジトリの整備・運用に関する国際的な認証基準の開発，取得の動きが高まりつつあります。具体的には，CoreTrustSeal（https://www.coretrustseal.org/）や TRUST 原則（https://doi.org/10.1038/s41597-020-0486-7）などがあります。

データアーカイブは，今後ますます研究・教育に不可欠な基盤としての役割を担っていくことになるでしょう。

2　電子的データの保存・共有などを行うための広い意味の情報基盤であり，計算機基盤（狭義の情報基盤）のみならず，運営体制および人的基盤を含みます。

【FAIR データ原則】

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）」があります。「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ます。内閣府の検討会が定める「研究データリポジトリ整備・運用ガイドライン」においても FAIR データ原則に沿ったデータ管理の機能が提言としてまとめられており，今後のデータアーカイブにおける業務展開に影響を及ぼしていくものと考えられます。

参考 1：データ共有の基準としての FAIR 原則：

https://DOI.org/10.18908/a.2018041901

参考 2：研究データリポジトリ整備・運用ガイドライン：

https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/guideline.pdf

付録：「データの説明文書」の用例

SAT 大蔵経テキストデータベース研究会による『勝鬘經義疏』のデータ（構造化テキストデータ）を事例として三つの書き方を示します。このデータの詳細については以下の URL を参照してください。

https://21dzk.l.u-tokyo.ac.jp/SAT/sat_tei.html

なお，以下の例は，この種のものとしては比較的詳細に記述したものです。同じようなことをしようとしている同じ分野の人，個人では難しいとしても複数人で智慧を寄せ合えば資料の内容を把握して適切に利用できるようにこの文書を書いておくことは極めて重要であり，それを目指していくべきですが，一方，この種の文書をどれくらい詳細に記述するかというのは，作成者・作成グループがどれくらい時間を割くことができるかに依存しますので，あくまでも無理のない範囲ということで考えてください。

事例 1：テキストファイルや Word 文書などで項目ごとに記載する

なお，同種類の複数のデータ（たとえば，同一文書群の個々の文書データなど）について説明文書を作成する必要が生じた場合，Excel などの表計算ソフトを用いて表形式で作成するとその後の処理を効率化しやすくなるのでぜひ検討してみてください。

・ファイル名およびファイルのパス（URL があればそれも含む）

https://21dzk.l.u-tokyo.ac.jp/SAT/sat_tei/2185.xml

・サイズ（ファイルのサイズを記載する）

626KB

・作成日（作成日を記載する）

2020 年 4 月 1 日作成。

・作成者（このデータの作成者を記載する）

永崎研宣（一般財団法人人文情報学研究所主席研究員）

・バージョン情報（改訂履歴を記載する）

2020 年 10 月 31 日版　<editorialDecl>を導入し，デジタル翻刻に際しての情報をより確認しやすくした。

2020 年 6 月 22 日版　タイトル情報を ab 要素に入れて，大正蔵のヘッダと大正蔵本文を div で分けてそれぞれ@type を付与。

2020 年 4 月 3 日版　c@type を metamark@function に変更し，大正蔵と#丙の facsimile を 2 件追加。大正蔵対応の pb を追加。

・アクセス制限（利用条件）

CC BY-SA

・ファイルのチェックサム（ファイルチェックサムを取得して記載する）

1591f29ac932b5c4bf91ddac39d24c85e76c81e24b0450d037e89e380ce9194e

・フォーマット（ファイルフォーマットを可能な限り詳しく記述する）

TEI/XML

・ファイル作成に用いたソフトウェア（ファイル作成に用いたソフトウェアを記述する）

Oxygen XML Editor

・データの元になった資料についての情報（書籍や古文書などであればその書誌情報など，資料を一意に示すことのできる情報を記載する）

「勝鬘經義疏」『大正新脩大藏經』1924-1932 年, 大正一切經刊行會

・テキストデータの場合，採用した文字コード・文字エンコーディング（特に，どのように文字を正規化したか・しなかったかの方針を記述する）

文字表記に関しては，Unicode に準拠している。大正新脩大藏経の文字の形に可能な限り準拠するために IVS を一部使用しており， また，Unicode CJK Ext.F 所収の文字を含んでいるため，作成者の意図の通りに表示するにあたっては，対応するフォントを利用されたい。

・構造化テキストデータの場合，構造化の仕方についての説明

縦書きであることの記述は，<gi>body</gi>直下の<gi>div</gi>において以下のように記述した。style="writing-mode: vertical-rl"しかしながら，本書の場合は<gi>body </gi>にこの属性を付与してもよいかもしれない。

・異文情報は，double-end-point attachment method を用い，それに沿って大正蔵の脚注の内容を記述した。ただし，「下同」については大正蔵の表記をそのまま転写しているので留意されたい。

・引用に関しては，ほとんどが勝鬘經からと思われるが，確認できた範囲で，<gi>quote</gi>を用い，@source で SAT DB 2018 の行番号・文字番号を記載した。ただし，文言が若干違うものの同定できると思われるものについては@corersp で参照した。その場合の多くは，甲本・乙本であれば一致する。

・返り点は，大正蔵のものをそのまま転写した。返り点記号は Unicode のものを使用し，<metamark function="kaeriten"></metamark>とした。

・割書に関しては，暫定的に，次のようにした。：<seg type="wari">大倭國上宮王<milestone unit="wlb"/>私集非海彼本</seg>

・句読点に関しては，誤りが多いとされるものの，大正蔵のものをそのまま転写した。

・大正蔵テキストは改行・段落分けなどがないため，主に藤井教公訳（大乗仏典：中国・日本篇第 16 巻所収）（ISBN 9784124026368）を適宜参照して段落分けを行なった。

・その他，データを作成した際に配慮した事項

・SAT 大蔵経テキストデータベースで公開されている簡易な構造化データを元にして作成した。

事例 2：TEI/XML 形式での記述

TEI/XML では，この種の情報を電子テキストファイルのヘッダとして XML で記述することになっており，個々の項目にタグ付けすることで情報を抽出しやすくしています。以下に，ヘッダの例を示します。特に，<encodingDesc>以下および<revisionDesc>以下をご覧ください。注目していただきたい箇所には次のような形式でコメントを付与しています。<!-- ここにコメント-->

```
      <teiHeader>
           <fileDesc>
              <titleStmt>
                 <title>勝鬘經義疏</title>
                 <author>聖徳太子</author>
                 <editor>SAT大蔵経テキストデータベース研究会</editor>
                 <respStmt>
                    <resp>Conversion from a traditional format to TEI P5</resp>
                    <persName sameAs="#knagasaki">Kiyonori Nagasaki</persName>
                 </respStmt>
                 <respStmt xml:id="knagasaki">
                    <resp>TEI Encoding</resp>
                    <persName>Kiyonori Nagasaki</persName>
                 </respStmt>
              </titleStmt>
              <publicationStmt>
                 <publisher>
                     <orgName ref="https://21dzk.l.u-tokyo.ac.jp/SAT/">SAT大蔵経テキストデータベース研究会</orgName>
                 </publisher>
                 <pubPlace>The Internet</pubPlace>
                 <availability>
                     <licence target="https://creativecommons.org/licenses/by-sa/4.0/">
                        <p>CC BY-SA</p>
                     </licence>
                 </availability>
                 <date when="2020-04-01">2020年4月1日</date>
              </publicationStmt>
              <sourceDesc><!--この要素以下に，元になった資料の書誌情報などを記述する-->
```

```
                <bibl>
                    <title>勝鬘經義疏</title>
                    <author sameAs="#聖徳太子">聖徳太子</author>
                    <series>大正新脩大藏經</series>
                    <date>1924</date>
                    <biblScope>續經疏部</biblScope>
                    <idno type="Taisho">2185</idno>
                </bibl>
                <listWit>
                    <witness xml:id="原">寶治年間版本</witness>
                    <witness xml:id="甲">安永八年版本</witness>
                    <witness xml:id="乙">明治二十八年蕃根版本<ref
                          target="https://dl.ndl.go.jp/info:ndljp/pid/817844 https://dl.ndl.go.jp/info:ndljp/pid/817845 https://dl.ndl.go.jp/info:ndljp/pid/817846"
                        /></witness>
                    <witness xml:id="丙">大日本佛教全書<ref target="https://dl.ndl.go.jp/info:ndljp/pid/952684/4"
                        /></witness>
                </listWit>
            </sourceDesc>
        </fileDesc>
        <encodingDesc><!--この要素以下で，テキスト入力やタグ付けをした際に採用したルールについて記述している-->
            <variantEncoding method="double-end-point" location="external"/>
            <editorialDecl>
                <normalization><!-- 新旧仮名遣いや使用した漢字の範囲（JIS第二水準など），文字表記などを正規化する方針について記述している-->
                    <p> 文字表記に関しては，Unicodeに準拠している。大正新脩大蔵経の文字の形に可能な限り準拠するためにIVSを一部使用しており， また，Unicode CJK Ext.F所収の文字を含んでいるため，作成者の意図の通りに表示するにあたっては，対応するフォントを利用されたい。</p>
                    <p>
                縦書きであることの記述は，<gi>body</gi>直下の<gi>div</gi>において以下のように記述した。style="writing-mode: vertical-rl"しかしながら，本書の場合は<gi>body </gi>にこの属性を付与してもよいかもしれない。</p>
                </normalization>
                <interpretation><!-- 本文の解釈に関わる事柄について記述している。-->
                    <p> 異文情報は，double-end-point attachment methodを用い，それに沿って大正蔵の脚注の内容を記述した。ただし，「下同」については大正蔵の表記をそのまま転写しているので留意されたい。 </p>
                </interpretation>
                <quotation><!-- 引用に関する事柄について記述している。-->
```

```
                    <p>
                引用に関しては，ほとんどが勝鬘経からと思われるが，確認できた範囲で，<gi>quote</gi>を用い，@sourceでSAT DB 2018 の行番号・文字番号を記載した。ただし，文言が若干違うものの同定できると思われるものについては@corerspで参照した。その場合の多くは，甲本・乙本であれば一致する。
                </p>
                </quotation>
                <punctuation><!-- 句読点，返り点などに関する事柄について記述している。-->
                    <p><![CDATA[
                 返り点は，大正蔵のものをそのまま転写した。返り点記号はUnicodeのものを使用し，<metamark function="kaeriten"></metamark>とした。
                割書に関しては，暫定的に，次のようにした。: <seg type="wari">大倭國上宮王<milestone unit="wlb"/>私集非海彼本</seg>
                句読点に関しては，誤りが多いとされるものの，大正蔵のものをそのまま転写した。
                ]]></p>
                </punctuation>
                <segmentation><!-- 資料内の「区切り方」について記述している。-->
                    <p> 大正蔵テキストは改行・段落分けなどがないため，主に <bibl><author>藤井教公</author>訳（<series><title>大乗仏典 ：
                                中国・日本篇</title>
                            <biblScope>第16巻</biblScope>所収</series>）(ISBN <idno type="ISBN">9784124026368</idno>)</bibl>
                        を適宜参照して段落分けを行なった。 </p>
                </segmentation>
            </editorialDecl>
        </encodingDesc>
        <profileDesc>
            <textClass>
                <classCode scheme="http://ndl.go.jp/dcndl/terms/NDC8">183.9</classCode>
                <keywords>
                    <term>仏教</term>
                    <term>論部．論疏</term>
                </keywords>
            </textClass>
        </profileDesc>
        <revisionDesc>
            <change resp="#knagasaki" when="2020-10-31"
                >2020年10月31日，<gi>editorialDecl</gi>を導入し，デジタル翻刻に際しての情報をより確認しやすくした。</change>
            <change resp="#knagasaki" when="2020-06-22"
```

```
                    >タイトル情報をab要素に入れて，大正蔵のヘッダと大正蔵本文をdivで分けてそれ
ぞれ@typeを付与。</change>
               <change resp="#knagasaki" when="2020-04-03"
                  >2020年4月3日，c@typeをmetamark@functionに変更し，大正蔵と#丙のfacsim
ileを2件追加。大正蔵対応のpbを追加。</change>
            </revisionDesc>
         </teiHeader>
```

参考文献

American Historical Association, 2015, “Guidelines for the Professional Evaluation of Digital Scholarship by Historians”, (Retrieved September 3, 2020, https://www.historians.org/teaching-and-learning/digital-history-resources/evaluation-of-digital-scholarship-in-history/guidelines-for-the-professional-evaluation-of-digital-scholarship-by-historians). (菊池信彦・小風尚樹・師茂樹・後藤真・永崎研宣訳，2016，『歴史学におけるデジタル研究を評価するためのガイドライン』（2020 年 9 月 3 日取得，http://hdl.handle.net/2261/59142）.）

CESSDA, 2020, “About the CESSDA Training Working Group”, CESSDA Training, Bergen, Norway: CESSDA ERIC, (Retrieved February 27, 2020, https://www.cessda.eu/Training/About).

CESSDA, 2020,“Documentation and metadata”, CESSDA Training, Bergen, Norway: CESSDA ERIC, (Retrieved February 27, 2020, https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Documentation-and-metadata).

CESSDA, 2020,“Organisation of variables”, CESSDA Training, (Retrieved February 27, 2020, https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/2.-Organise-Document/Designing-a-data-file-structure/Organisation-of-variables).

Charters Encoding Initiative, 2020, “CEI - Charters Encoding Initiative”, (Retrieved September 3, 2020, https://www.cei.lmu.de/).

CLARIN, 2020, “CLARIN - European Research Infrastructure for Language Resources and Technology”, Drift: CLARIN ERIC, (Retrieved September 3, 2020, https://www.clarin.eu/).

CoreTrustSeal, 2020, “CoreTrustSeal”, (Retrieved February 27, 2020, https://www.coretrustseal.org/).

Corti, Louise, Veerle Van den Eynden, Libby Bishop and Matthew Woollard, 2014, *Managing and Sharing Research Data: A Guide to Good Practice*, Los Angeles: Sage.

Data Cite, 2020, “DataCite Metadata Schema”, Hannover: Data Cite, (Retrieved June 22, 2020, http://schema.datacite.org/).

DOI, 2020, “The DOI® System”, DOI, (Retrieved September 3, 2020, https://www.DOI.org/).

DOI, 2016，「DOI ハンドブック」，（2020 年 2 月 27 日取得， https://www.doi.org/doi_handbook/translations/japanese/hb.html）.

Eynden, Veerle Van den, Louise Corti, Matthew Woollard, Libby Bishop and Laurence Horton, 2011, *Managing and Sharing Data: Best Practice for Researchers*, Colchester, Essex: UK Data Archive University of Essex, (Retrieved February 19, 2020, https://ukdataservice.ac.uk/media/622417/managingsharing.pdf). （東京大学社会科学研究所附属社会調査・データアーカイブ研究センター訳，2013，『データの管理と共有――研究者向け最良事例』東京大学社会科学研究所附属社会調査・データアーカイブ研究センター，（2021 年 10 月 12 日取得，https://csrda.iss.u-tokyo.ac.jp/UKDAguide.pdf）.）

Finnish Social Science Data Archive, 2020, “Data Management Guidelines”, Tampere: Tampere University, (Retrieved February 27, 2020, https://www.fsd.tuni.fi/aineistonhallinta/en/).

Finnish Social Science Data Archive, 2020, “Data Description and Metadata”, Data Management Guidelines, Tampere: Tampere University, (Retrieved February 27, 2020, https://www.fsd.uta.fi/aineistonhallinta/en/data-description-and-metadata.html).

ICPSR, 2020, *Guide to Social Science Data Preparation and Archiving: The Best Practice Throughout the Data Life Cycle・6th edition*, Ann Arbor, MI: University of Michigan, (Retrieved June 22, 2020, https://www.icpsr.umich.edu/files/deposit/dataprep.pdf).

ICPSR, 2020, “Guide to Social Science Data Preparation and Archiving: Phase 3: Data Collection and File Creation”, ICPSR Start Sharing Data, Ann Arbor, MI: University of Michigan, (Retrieved February 27, 2020, https://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html#elements).

ICPSR, 2020, “Summer Program in Quantitative Methods of Social Research”, Ann Arbor, MI: University of Michigan, (Retrieved February 27, 2020, https://www.icpsr.umich.edu/icpsrweb/content/sumprog/about.html).

Jonsen, Albert R., 1998, *The Birth of Bioethics*, Oxford, Oxford University Press.（細見博志訳，2009，『生命倫理学の誕生』勁草書房.）

Library of Congress, 2020, “LCCN Permalink Frequently Asked Questions“, Washington, DC: Library of Congress, (Retrieved September 3, 2020, https://lccn.loc.gov/).

OCLC，2020，「VIAF:バーチャル国際典拠ファイル」，（2020 年 9 月 3 日取得，http://viaf.org/）.

Online Greek Coinage, 2020, “Numismatic Description Schema”, (Retrieved September 3, 2020, https://www.greekcoinage.org/numismatic-description-standard-nuds.html).

Source Forge, 2020, “EpiDoc: Epigraphic Documents in TEI XML”, San Diego, CA: Source Forge, (Retrieved October 29, 2020, https://sourceforge.net/p/epidoc/wiki/Home/).

Text Encoding Initiative, 2020, “P5: Guidelines for Electronic Text Encoding and Interchange”, TEI Consortium, (Retrieved August 19, 2020, https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html).（TEI 協会東アジア／日本語分科会訳，2020，「日本語向け TEI ガイドライン」，（2020 年 9 月 3 日取得，https://github.com/tei-eaj/jp_guidelines/wiki）.）

The DDI Alliance, 2020, “Explore Documentation”, The DDI Alliance (Retrieved June 22, 2020, https://DDIalliance.org/explore-documentation).

The Dublin Core™ Metadata Initiative, 2021, “Creating Metadata”, The Dublin Core™ Metadata Initiative, (Retrieved September 7, 2021, https://www.dublincore.org/resources/userguide/creating_metadata/).

The Text Encoding Initiative, 2020, “TEI: P5 Guidelines”, The Text Encoding Initiative, (Retrieved June 22, 2020, https://tei-c.org/guidelines/p5/).

Ubiquity Press, 2020, “The Journal of Open Humanities Data”, (Retrieved June 24, 2020, https://openhumanitiesdata.metajnl.com/).

UK Data Service, 2020, “Catalogue metadata”, Colchester, Essex: University of Essex and University of Manchester, (Retrieved February 27, 2020, https://www.ukdataservice.ac.uk/manage-data/document/metadata.aspx).

大波純一・八塚茂・信定知江・箕輪真理・三橋信孝・畠中秀樹，2018，「データ共有の基準としての FAIR 原則」，バイオサイエンスデータベースセンター，（2020 年 2 月 27 日取得，https://DOI.org/10.18908/a.2018041901）.

オープンアクセスリポジトリ推進協会，2020，「ホーム」，JPCOAR スキーマガイドラインホームページ，（2020 年 6 月 22 日取得，https://schema.irdb.nii.ac.jp/ja）.
科学技術・学術審議会 学術分科会 学術情報委員会，2016，「学術情報のオープン化の推進について（審議まとめ）」，文部科学省ホームページ，（2020 年 10 月 28 日取得，https://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu4/036/houkoku/1368803.htm）.
金甫榮，2020，「アーカイブズ資料情報システムの構築と運用── AtoM (Access to Memory)を事例に」，『アーカイブズ学研究』32: 4-29.
クリエイティブ・コモンズ・ジャパン，2020，「クリエイティブ・コモンズ・ライセンスとは」，クリエイティブ・コモンズ・ジャパンホームページ，（2020 年 2 月 27 日取得，https://creativecommons.jp/licenses/）.
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，2019，「研究データリポジトリ整備・運用ガイドライン」，内閣府ホームページ，（2020 年 2 月 27 日取得，https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/guideline.pdf）.
国立研究開発法人日本医療研究開発機構，2020，「データマネジメントプランの提出について」，国立研究開発法人日本医療研究開発機構ホームページ，（2020 年 10 月 19 日取得，https://www.amed.go.jp/content/000061339.pdf）.
国立国会図書館， 2017，「国立国会図書館資料デジタル化の手引」，国立国会図書館ホームページ，（2020 年 9 月 3 日取得， https://www.ndl.go.jp/jp/preservation/digitization/guide.html）.
後藤真・橋本雄太編，2019，「『歴史情報学の教科書』全文公開アーカイブ」，図書出版文学通信，（2020 年 9 月 3 日取得，https://bungaku-report.com/blog/cat43/）.
下田正弘・永崎研宣，2019，「下田正弘・永﨑研宣編『デジタル学術空間の作り方　仏教学から提起する次世代人文学のモデル』全文公開サイト」，図書出版文学通信，（2020 年 9 月 3 日取得，https://bungaku-report.com/sat.html）.
情報処理推進機構，2020，「IVD/IVS とは」，IPA 文字情報基盤整備事業ホームページ，（2020 年 6 月 24 日取得，https://mojikiban.ipa.go.jp/1292.html）.
総務省， 2009， 「調査票情報等の管理及び情報漏えい等の対策に関するガイドライン」，（2020 年 9 月 3 日取得， https://www.soumu.go.jp/main_content/000618528.pdf）.
総務省， 2009，「匿名データの作成・提供に係るガイドライン（抄）」，（2020 年 9 月 3 日取得，https://www.soumu.go.jp/main_content/000599961.pdf）.
竹内啓編，1989，『統計学辞典』東洋経済新報社.
東京大学社会科学研究所パネル調査プロジェクト，2009，「東大社研・高卒パネル調査（JLPS-H）wave1，2004.3　高校生調査票」，（2021 年 4 月 30 日取得，https://ssjda.iss.u-tokyo.ac.jp/chosa-hyo/PH010c.html）.
東京大学史料編纂所，2020，「史料編纂所データベース異体字同定一覧」，（2020 年 9 月 3 日取得，https://wwwap.hi.u-tokyo.ac.jp/ships/itaiji_list.jsp）.
永崎研宣，2019，京都大学人文科学研究所・共同研究班「人文学研究資料にとっての Web の可能性を再探する」『日本の文化をデジタル世界に伝える』樹村房.
日本学術振興会「科学の健全な発展のために」編集委員会，2015，『科学の健全な発展のために── 誠実な科学者の心得』丸善出版．[1]
文化庁，2020，「著作権制度に関する情報」，文化庁ホームページ，（2020 年 2 月 27 日取得，https://www.bunka.go.jp/seisaku/chosakuken/seidokaisetsu/index.html）.

北海道アイヌ協会・日本人類学会・日本考古学協会，2017，「これからのアイヌ人骨・副葬品に係る調査研究の在り方に関するラウンドテーブル報告書」，（2021 年 6 月 10 日取得，http://archaeology.jp/wp-content/uploads/2017/05/dc163de9d75c26bfb9452b3db6526dfe.pdf）.
北海道アイヌ協会・日本人類学会・日本考古学協会・日本文化人類学会，2019，「アイヌ民族に関する研究倫理指針（案）」，（2021 年 6 月 10 日取得，https://www.ainu-assn.or.jp/news/files/3b014e7a03b0c1567978f9a1da5f17b8e8813a5a.pdf）.
三輪哲・佐藤香，2018，「連載　教育研究の現在　第 11 回　データアーカイブの教育研究への活用　――世界的動向をふまえて」，『教育学研究』85(2): 206-215，（2020 年 2 月 27 日取得，https://www.jstage.jst.go.jp/article/kyoiku/85/2/85_206/_pdf/-char/ja）.
山田太造，2021，「chapter2　歴史データをつなぐこと――目録データ」，文学通信，（2021 年 4 月 30 日取得，https://bungaku-report.com/blog/2019/03/chapter-2.html）.

1 「科学の健全な発展のために」は平成 27 年発行以降改訂されていないため，ここで紹介されている法令や指針は，それぞれ既に廃止・改正されており，特に個人情報のところは現行法と異なるため，参照すべきではない部分もあります。

グロッサリー

DOI（Digital Object Identifier） デジタルネットワーク上で，コンテンツへのアクセスを管理するために用いられる国際的な識別子。https://doi.org/に続けて DOI をブラウザに入力することで，自動的にコンテンツの所在情報（URL）に変換されるサービス名称でもあり，登録機関が DOI に紐づく URL のメンテナンスを行うことで，利用者からの恒久的なアクセスが実現される。

インタビュー 情報を収集するために行われる聞き取りや会話のこと。その構造化の度合いから，構造化インタビュー，半構造化インタビュー，非構造化インタビューが区別される。特に断りなく単独で使われる場合，この手引きでは，非構造化インタビューを指す。

インフォームド・コンセント 研究対象者などが，調査や研究の目的，方法，研究対象者に生じる負担，個人情報やデータの取扱などについて十分な説明を受け，自由意思に基づいて研究者に与える，当該研究の実施に関する同意のこと。個人情報を取得しない調査では，回答をもって協力の同意があったとみなすことができる。

寄託 収集したデータなどをデータアーカイブに預けて，データセットの保存・管理と共有にかかる業務を依頼すること。

校合 複数の写本や木版本などを比べ合わせて，本文の異同を確かめたり誤りを正したりすること。

欠測値 調査項目において回答が欠落したもの。

公的統計 国の行政機関，地方公共団体又は独立行政法人などが作成する統計（統計法第 2 条第 3 項）。官庁統計・政府統計と言われることもある。

コーディング 調査項目における回答ごとに数値を割り振る作業のこと。

コードブック データを利用する際に必要な情報をまとめたもの。ファイル上のどの位置の記号（通常，数字）がなにを意味するかを示す。

個票データ（ミクロデータ） 社会調査データや公的統計の作成過程で出てくる，集計前の個票（記入済みの個々の調査票）に含まれる個体情報から成るレコード群。

識別子 様々な対象から特定の一つを識別，同定するために用いられる名前や符号，数字などを指す。

質問紙調査 質問紙（調査票）を用いて行う調査。調査票調査ともいう。

データアーカイブ 主に研究・教育を目的として，幅広いデータを収集・保存し，提供するサービスを行う組織・機関。

データカタログ データについて，そのメタデータを集めて一覧できるようにしたもの。データの収集，整理，保存，提供などに用いられる。

データ管理計画 研究プロジェクトなどにおけるデータをいかに管理するかを定めたもの。

データ共有 既存のデータを研究者など特定の条件を満たす主体に利用できるようにすること。

データ公開 無制限にだれもがデータにアクセスできるようにすること。

データセット 調査において収集された情報の総体のこと。複数のデータファイルを有することがありえる点で，データファイルとは区別される。

データファイル データを格納した電子的なファイルのこと。

データフォーマット データを格納するファイルの形式のこと。商用の統計ソフトウェアでは固有の形式が存在する。

データベース データを目的に応じて構造化し，データへのアクセスと利用を効率化したもの。

データリポジトリ（レポジトリ） 電子的データの保存・共有などを行うための広い意味の情報基盤であり，計算機基盤（狭義の情報基盤）のみならず，運営体制および人的基盤を含む。

匿名化 データに含まれる情報が具体的にどの調査対象に関するものかデータの利用者には分からないように加工を施すこと。

二次分析 すでに分析がなされている既存のデータを再利用して，新たな視点や分析方法にもとづいて分析すること。

二次利用 既存のデータを当初の調査・研究目的以外に利用すること。二次分析より広義。

翻刻 写本・版本などを，原本どおりに活字に起こしたりテキストデータにするなどして新たに出版すること。特に後者を指してデジタル翻刻ということがある。

メタデータ データセットのコンテンツ，コンテキスト，出所などに関するデータ。資料の書誌・目録情報のことを指すこともある。

メタデータスキーマ メタデータの記述項目や形式，語彙の定義，項目間の階層構造などを定義したもの。

名簿

人文学・社会科学データインフラストラクチャー構築推進事業
運営委員会作業部会（共通ガイドライン（手引き）策定）

部会長

松本　　康（まつもと　やすし）	立教大学　社会学部　元教授／データインフラストラクチャー構築推進事業運営委員会委員

専門委員

遠藤　晶久（えんどう　まさひさ）	早稲田大学　社会科学総合学術院　准教授
柴山　　直（しばやま　ただし）	東北大学　大学院教育学研究科　教授
杉本　重雄（すぎもと　しげお）	筑波大学　名誉教授
田尾　亮介（たお　りょうすけ）	首都大学東京　法学政治学研究科　准教授
永崎　研宣（ながさき　きよのり）	一般財団法人人文情報学研究所　人文情報学研究部門　主席研究員／データインフラストラクチャー構築推進事業運営委員会委員
南山　泰之（みなみやま　やすゆき）	情報・システム研究機構国立情報学研究所　オープンサイエンス基盤研究センター　特任技術専門員
武藤　香織（むとう　かおり）	東京大学　医科学研究所　教授

人文学・社会科学データインフラストラクチャー構築推進センター

センター長

廣松　　毅（ひろまつ　たけし）	東京大学　名誉教授

研究員

池内　有為（いけうち　うい）	文教大学　文学部　専任講師
伊藤　伸介（いとう　しんすけ）	中央大学　経済学部　教授
前田　幸男（まえだ　ゆきお）	東京大学　大学院情報学環　教授

（令和 3 年 3 月 31 日現在）

奥付

人文学・社会科学におけるデータ共有のための手引き
－人文学・社会科学データインフラストラクチャーの構築に向けて－
2021 年 11 月発行
発行：独立行政法人　日本学術振興会
執筆・編集：人文学・社会科学データインフラストラクチャー構築推進事業
　　運営委員会作業部会（共通ガイドライン（手引き）策定）
　　人文学・社会科学データインフラストラクチャー構築推進センター
製作：中西印刷株式会社

独立行政法人 日本学術振興会
研究事業部 研究事業課
〒 102-0083
東京都千代田区麹町 5-3-1
TEL：03-3263-4645,1106
Email：di-hs@jsps.go.jp
URL：https://www.jsps.go.jp/j-di/index.html

引用例：
人文学・社会科学データインフラストラクチャー構築推進事業運営委員会作業部会，2021，『人文学・社会科学におけるデータ共有のための手引き――人文学・社会科学データインフラストラクチャーの構築に向けて』日本学術振興会.