JOSS2021（2021年6月18日）「FAIRなデータキュレーションの実践」

社会科学における データキュレーション

三輪　哲

（東京大学社会科学研究所）

社会科学における研究データ

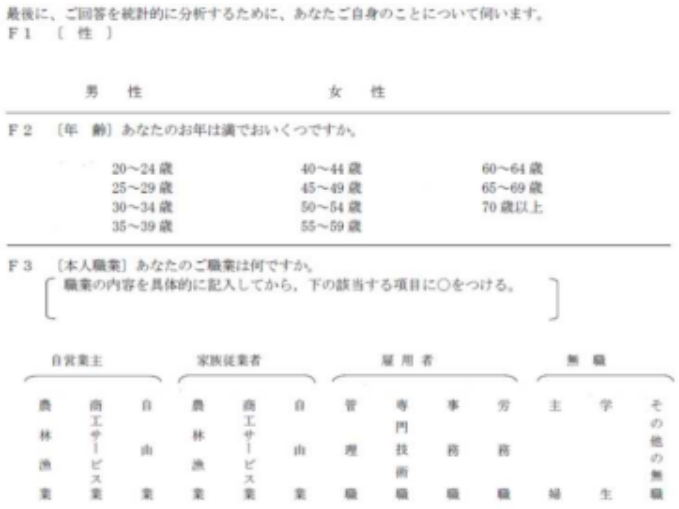
- 社会科学　Social Sciences
  - 社会の中での人間行動や社会現象を対象とする科学の総称
    - 法学、政治学、経済学、経営学、社会学　など・・・

- 研究データは主に社会調査によって収集
  - 定性的データ
    - インタビュー、フィールドワーク、文書記録　など
  - 定量的データ
    - 行政管理データ、集計データ、個票データ　など

質問紙調査（survey research）

■何を聞くか？

質問紙への回答



最後に、ご回答を統計的に分析するために、あなたご自身のことについて伺います。
F１ 〔 性 〕

男 性　　女 性

F２ 〔年 齢〕あなたのお年は満でおいくつですか。

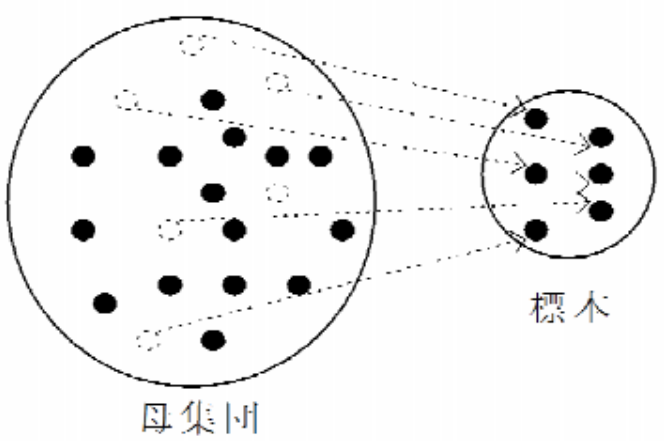
20～24 歳　40～44 歳　60～64 歳
25～29 歳　45～49 歳　65～69 歳
30～34 歳　50～54 歳　70 歳以上
35～39 歳　55～59 歳

F３ 〔本人職業〕あなたのご職業は何ですか。
〔職業の内容を具体的に記入してから、下の該当する項目に○をつける。〕

自営業主：農林漁業／商工サービス業／自由業
家族従業者：農林漁業／商工サービス業／自由業
雇用者：管理職／専門技術職／事務職／労務職
無職：主婦／学生／その他の無職

■誰に聞くか？

母集団と、無作為抽出による標本



■どのように聞くか？

訪問面接法，郵送法，電話法，ウェブ調査

質問紙による社会調査による個票データ

	ZQ07A	ZQ07B	ZQ07C	ZQ07D	ZQ07E	ZQ07F	ZQ07G
1	2	3	1	2	2	4	4
2	2	2	2	4	1	2	3
3	2	3	3	2	4	2	4
4	1	2	2	2	1	2	4
5	3	2	3	4	3	1	4
6	2	2	2	3	2	4	3
7	2	3	4	2	2	2	3
8	2	4	5	4	1	1	4
9	2	3	4	1	2	2	3
10	2	2	5	3	2	2	4
11	2	2	5	3	1	3	3
12	3	2	2	2	2	3	1
13	2	4	4	4	3	3	4
14	2	3	4	2	3	4	4
15	1	3	5	2	2	1	4

問7.　現在（または直近）のお仕事に関して、次にあげるＡからＧのそれぞれについて、あてはまる程度をお答えください。（○はそれぞれにつき１つ）

	かなりあてはまる	ある程度あてはまる	あまりあてはまらない	あてはまらない	部下はいない
A. 自分の仕事のペースを、自分で決めたり変えたりすることができる	1	2	3	4	
B. 職場の仕事のやり方を、自分で決めたり変えたりすることができる	1	2	3	4	
C. 部下の仕事のやり方を、自分が決めている	1	2	3	4	5
D. 教育訓練を受ける機会がある	1	2	3	4	
E. 仕事を通じて職業能力を高める機会がある	1	2	3	4	
F. 子育て・家事・勉強など自分の生活の必要にあわせて、時間を短くしたり休みを取るなど、仕事を調整しやすい職場である	1	2	3	4	
G. 今後１年間に失業（倒産を含む）をする可能性がある	1	2	3	4	

あなたのふだんの生活についておうかがいします。

問8.　あなたはどのくらいの頻度で以下のことをしていますか。（○はそれぞれにつき１つ）

	毎日	週に5～6日	週に3～4日	週に1～2日	月に1～3日	ほとんどしない
A. 運動（ウォーキング・ジョギング・エアロビクス・水泳・テニスなど）	1	2	3	4	5	6
B. １日に３食を食べる	1	2	3	4	5	6
C. 栄養バランスの取れた食事を取る	1	2	3	4	5	6
D. カップ麺やファーストフードを食べる	1	2	3	4	5	6
E. 食事の用意	1	2	3	4	5	6
F. 洗濯	1	2	3	4	5	6
G. 家の掃除	1	2	3	4	5	6
H. 日用品・食料品の買い物	1	2	3	4	5	6
I. 友人・恋人（配偶者は除く）と食事をする	1	2	3	4	5	6
J. 友人・恋人（配偶者は除く）と話をする	1	2	3	4	5	6
K. インターネットを利用する（仕事以外で）	1	2	3	4	5	6

社会科学におけるデータキュレーション

- 調査（実査）が実施されてから、以下3段階でおこなわれる

1. 質問紙回収直後の段階
   - エディティング、コーディング、クリーニング
2. 調査報告の段階
   - 調査概要、コードブック、変数リスト
3. データ共有の段階
   - データフォーマット、データ確認と処理、メタデータ

※あくまで、質問紙調査による個票データの話に限定

１．質問紙回収直後の段階

- 首尾よく質問紙が回収できた、とする
- しかし、回収された質問紙はそのまま入力することはできない
- なぜか？　→回答がしばしばキレイではないから
  - 記入にミスがある、文字が読みにくい、空白になっている箇所がある
- そこで、必要となるのが、***エディティング***
  - エディティングとは、回収された調査票の点検と修正

    （１）「白紙」、いいかげんな回収票、不正記入票を見つけ出し、取り除く

    （２）読みにくい字、誤字があれば、修正する

    （３）不完全な回答を見つけ、できるなら修正する、できないなら無回答扱いとする

    （４）欠損値を割り当て、数値を記入する

    ⇒これを通して、有効票を確定させる

問１　あなたの性別とお生まれになった年をお答えください。

① 男　2 女	昭和 ~~72~~ 47 年	現在 37 歳

問２ａ　現在のあなたの居住形態について教えてください。

1 家族と同居	2 家族と一時的に別居	③ ひとり暮らし	4 その他

問２ｂ　問２ａで１に○をつけた方にお聞きします。同居されているご家族とあなたとの続柄について、あてはまるものすべてに○をつけてお答えください。また、あなた自身を含めて同居されているご家族は何人でしょうか。以下の回答欄にご記入ください。

				人数
1 母親（義母を含む）	4 きょうだい（義理のきょうだいを含む）	⑦ 子ども		2 人
2 父親（義父を含む）	5 祖父			
③ 配偶者（夫または妻）	6 祖母			

単身赴任か？
判断保留

問３ｃ　お勤め先の業種は、次のうちどれにあてはまりますか。

1	農林漁業	7	運輸業・郵便業	13	教育・学習支援業
2	鉱業	8	卸売業・小売業	14	医療・福祉業
3	建設業	9	金融業・保険業	15	その他サービス業
4	製造業	10	不動産業	16	公務
5	電気・ガス・熱供給・水道業	(11)	学術研究・専門・技術サービス業	17	その他
6	情報通信業	12	宿泊・飲食サービス業		（具体的に：　　）

問３ｄ　あなたが現在のお勤め先に入社されたのは、何歳のときですか。以下の回答欄にご記入ください。

~~40~~ 歳のとき

99 (あり得ない回答

問４　お勤め先でのあなたの役職は、大きく分けて、この中のどれにあたりますか。あてはまるもの１つに○をつけてお答えください。

(1)	役職なし	4	課長、課長相当職	~~(7) その他~~	
2	職長、班長、主任相当職	5	部長、部長相当職	（具体的に：准教授）	
3	係長、係長相当職	6	社長、役員級	9	わからない

役職とはみなさな

問 22b　それは何が原因だと思いますか。以下のＡ～Ｌについて、あてはまるもの１つに○をつけてお答えください。

	とてもあてはまる	あてはまる	あまりあてはまらない	あてはまらない
A．自分の健康について	1	2	3	4
B．家族の健康について	1	2	3	4
C．自分の生活（進学、就職、結婚など）上の問題について	1	2	3	4
D．家族の生活（進学、就職、結婚など）上の問題について	1	2	3	4
E．現在の収入や資産について	1	2	3	4
F．今後の収入や資産の見通しについて	1	2	3	4
G．老後の生活設計について	1	2	3	4
H．家族・親族間の人間関係について	1	2	3	4
I．近隣・地域間との関係について	1	2	3	4
J．勤務先での仕事や人間関係について	1	2	3	4
K．自由時間を過ごすための施設が充実していないことについて	1	2	3	4
L．事業や家業の経営上の問題について	1	2	3	4

いいかげんな回答

１．質問紙回収直後の段階

- 時に、回答は自由回答（文章を自由に記入）で得られる
  - 自由回答の利点
    - 回答があらかじめ用意した選択肢に制約されないので、意外な回答が得られる
    - 選択肢を用意できないような、詳細な情報が得られる
    - 要望や苦情などはしばしば自由回答の形式をとるしかない
    - 情報を調査者側が自由に加工できる
- そこで、必要となるのが、***コーディング***
  - コーディングとは、自由回答の文章や言葉へと数値を割り当てること

    ⇒これを通して、数量的データができあがる

１．質問紙回収直後の段階

- 地域　総務省の標準地域コードなど
  - 例）宮城県仙台市青葉区　　04101

- 大学名　日本学術振興会の大学コードなど
  - 例）東北大学　　0132

- 職業名　日本標準職業分類など
  - 例）大学教員　　198

１．質問紙回収直後の段階

- エディティングおよびコーディングを終えた後、質問紙の回答は入力される
  - テキスト形式、エクセル形式など
- しかしながら、まだデータファイルは完成していない
  ・・・データの誤りを修正できていないから
- そこで、必要となるのが、***クリーニング***

１．質問紙回収直後の段階

- クリーニングとは、データの中の「誤った回答」「おかしい回答」を、修正する作業
- エディティングとの違い
  - エディティング　・・・調査票を1票1票みる
  - クリーニング　・・・データファイルをみる

  （１）そもそも対象外の人が回答していないか
  （２）「ありえない値」が入力されていないか
  （３）論理的エラーや矛盾はないか
  ⇒これらを通して、データファイルを完成させる

2．調査報告の段階

- 調査概要

『2005年SSM日本調査コード・ブック』より

1．調査設計と回収状況

1．調査設計

2005 年 SSM 日本調査は、全国の男女個人（ただし、(2) 項参照）を母集団とする調査である。なお、調査票は、面接調査票および留置調査票から成っており、留置調査票は A 票、B 票の 2 種類ある。

(1) サンプル数

計画サンプル数は 14,140 である。

(2) 対象者の範囲

対象者は、2005 年 9 月 30 日現在で満 20 歳～69 歳の男女。
(1935 年 10 月 1 日～1985 年 9 月 30 日に生まれた者)

(3) 地点抽出

- 全国の区市町村を人口規模に応じて表の通り層化した上で、区市町村内の投票区を第一次抽出単位として層毎の系統抽出によって計 1,010 地点を抽出した。実際には、(層毎の全市区町村内の) 有権者数の長さを持つ各投票区を並べたリストから、系統抽出法により調査対象投票区（この時点で市区町村も同時に決定される）とランダムなスタート番号を決定した。地点の確率比例抽出を模した手続きとなっている。
- 各地点について 14 サンプルずつ割り当てた。

2．調査報告の段階

- 調査概要

『2005年SSM日本調査コード・ブック』より

1．調査設計と回収状況

1．調査設計

2005 年 SSM 日本調査は、全国の男女個人（ただし、(2) 項参照）を母集団とする調査である。なお、調査票は、面接調査票および留置調査票から成っており、留置調査票は A 票、B 票の 2 種類ある。

(1) サンプル数

計画サンプル数は 14,140 である。

(2) 対象者の範囲

対象者は、2005 年 9 月 30 日現在で満 20 歳～69 歳の男女。
(1935 年 10 月 1 日～1985 年 9 月 30 日に生まれた者)

(3) 地点抽出

- 全国の区市町村を人口規模に応じて表の通り層化した上で、区市町村内の投票区を第一次抽出単位として層毎の系統抽出によって計 1,010 地点を抽出した。実際には、(層毎の全市区町村内の) 有権者数の長さを持つ各投票区を並べたリストから、系統抽出法により調査対象投票区（この時点で市区町村も同時に決定される）とランダムなスタート番号を決定した。地点の確率比例抽出を模した手続きとなっている。
- 各地点について 14 サンプルずつ割り当てた。

2．調査報告の段階

- 調査概要

奈良県『令和2年県民アンケート調査報告書』より

1. 調査の目的

身近な生活に関する事柄についての重要度・満足度のほか、「観光」や「農林業」等に関する意識やニーズを把握し、今後の県政運営の基礎資料とすることを目的に、アンケート調査を実施しました。

2. 調査の設計

■調 査 地 域　奈良県全域
■調 査 対 象　県内在住の満 20 歳以上の男女・個人
■調査標本数　5,000 人
■調査抽出法　層化二段無作為抽出法
■調 査 方 法　郵送配布・郵送回収。調査期間内に、はがきによるお礼状兼督促状の配布１回
■調 査 時 期　令和２年８月１日（土）～令和２年８月 20 日（木）

3. 調査票の配布・回収の状況

■配布件数　5,000 件
■回収件数（率）　2,891 件（57.8%）
■有効回答数（率）※1　2,809 件（56.2%）

2．調査報告の段階

- カラム・ガイド

『1995年SSM日本調査
コード・ブック』より

質問項目	カラム	内容	DK・NA	非該当
	〔カード番号 ＝ 1〕			
地点番号	1 - 3		—	—
個番	4 - 5		—	—
調査票タイプ	6	＝1	—	—
カード番号	7 - 8	＝1	—	—
問1　性・年齢				
(1)性別	9		—	—
(2)出生年	10 - 11	西暦下2ケタ	—	—
出生月	12 - 13		—	—
満年齢	14 - 15	調査時点の年齢	—	—
問2　兄弟姉妹				
兄弟姉妹数	16 - 17		99	—
全体順位	18 - 19			
同性順位	20 - 21			
問3　15歳時耐久消費財				
1(ア)持家	22	あり＝ 1 なし＝ 0	9	—
2(イ)自家風呂	23			
3(ウ)ラジオ	24			
4(エ)テレビ	25			
5(オ)冷蔵庫	26			
6(カ)自転車	27			
7(キ)自動車	28			
8(ク)ピアノ	29			
9(ケ)電話	30			
10(コ)応接セット	31			
11(サ)文学全集・図鑑	32			
12(シ)株券または債券	33			
13(ス)美術品・骨董品	34			
14(セ)別荘	35			
15　どれもない	36	ない＝ 1	9	—
問4　本人現職				
a 従業上の地位	37 - 38		99	—
c 従業先産業	39 - 40	産業コード	99	98
d 従業員数	41 - 42		99	98
e 仕事の内容（職業）	43 - 45	職業コード	999	—
f 役職名	46		9	8

2．調査報告の段階

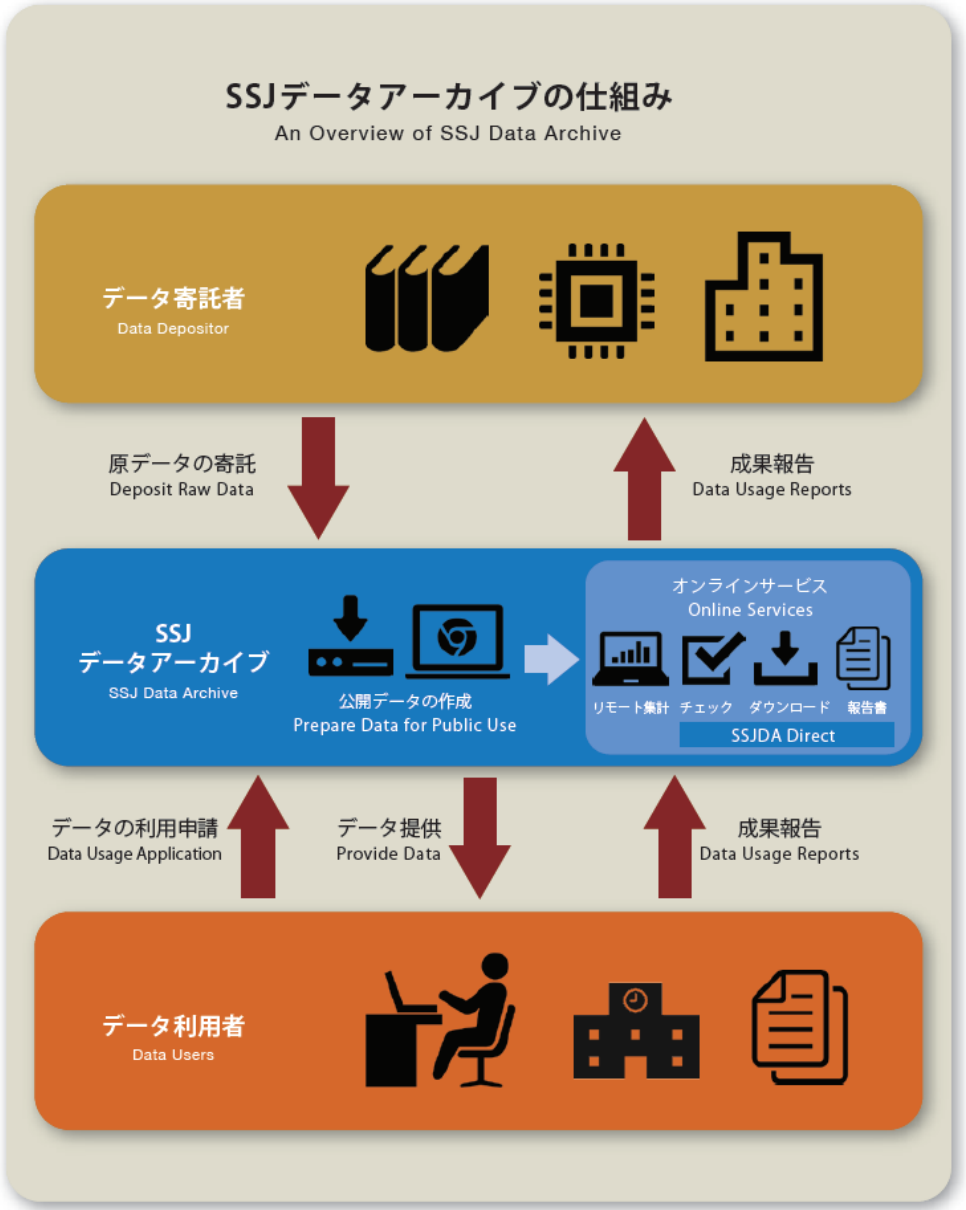
- 変数リスト

リクルートワークス研究所「全国就業実態パネル調査」資料より

聴取有無 継続サンプル	聴取有無 追加サンプル	問番号 JPSED 2016	JPSED 2017	JPSED 2018	JPSED 2019	今回 JPSED 2020	質問項目	質問カウント 138	うに聞く質問	毎年変化する項目であり、今回も同じよ	一部の人だけ変化する可能性のある質問	確認のために再度残す質問	毎年変化しないが	備考
						追加設問 / 修正設問								
属性（基本情報：回答者本人の確認のため）														
○	○	Q1	Q1	Q1	Q1	Q1	性別	0				●		
○	○	Q2	Q2	Q2	Q2	Q2	年齢	0				●		
○	○	Q3	Q3	Q3	Q3	Q3	誕生年月	1				●		
○	○	Q4	Q4	Q4	Q4	Q4	居住地	1			●			
○	○	Q12	Q5	Q5	Q5	Q5	学歴	1				●		
○	○	－	Q6	Q6	Q6	Q6	学部	1						
○	○	Q13	Q7	Q7	Q7	Q7	学年	1				●		
○	○	－	－	－	Q7-1	Q7-1	卒業予定の学部	1			●			
○	○	Q14	Q8	Q8	Q8	Q8	中退経験	1				△		
家族の状況について														
○	○	Q5	Q9	Q9	Q9	Q9	配偶者有無	1			●			
○	○	Q6	Q10	Q10	Q10	Q10	子ども有無	1						

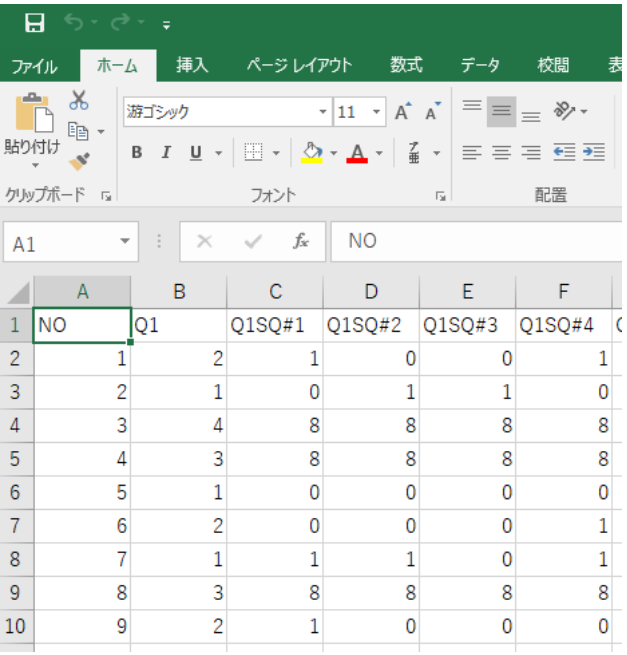
3．データ共有の段階

- データ共有は、もっぱらデータアーカイブ（データレポジトリ）が担う
- データアーカイブとは、調査データを収集・保管・公開する機関
  - 「データ版の図書館」のイメージ
  - データアーカイブの意義
    1. データの散逸を防ぐ
    2. データを見つけやすく、アクセスしやすく、相互運用しやすく、再利用しやすく
    3. 分析の再現性を担保する
    4. 新たな視角からの再分析を可能にする
    5. 過去データとの比較可能性を確保する

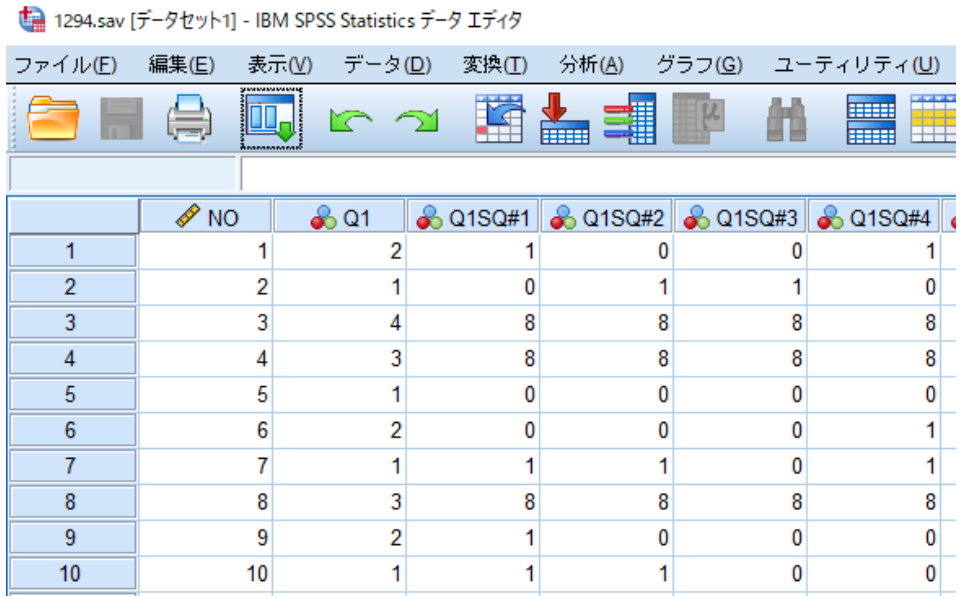


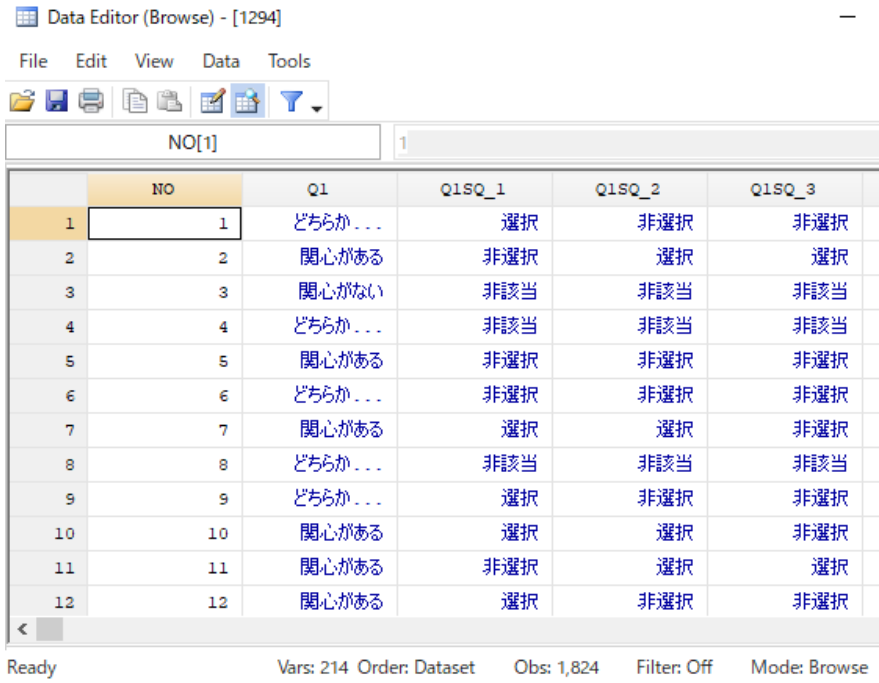
3．データ共有の段階

- データフォーマット
  1）最もシンプルな、テキスト形式（.csv）
  2）社会科学でユーザーの多い、SPSS形式（.sav）
  3）近年利用が増えつつある、STATA形式（.dta）



	NO	Q1	Q1SQ#1	Q1SQ#2	Q1SQ#3	Q1SQ#4
1	1	2	1	0	0	1
2	2	1	0	1	1	0
3	3	4	8	8	8	8
4	4	3	8	8	8	8
5	5	1	0	0	0	0
6	6	2	0	0	0	1
7	7	1	1	1	0	1
8	8	3	8	8	8	8
9	9	2	1	0	0	0
10	10	1	1	1	0	0





	NO	Q1	Q1SQ_1	Q1SQ_2	Q1SQ_3
1	1	どちらか . . .	選択	非選択	非選択
2	2	関心がある	非選択	選択	選択
3	3	関心がない	非該当	非該当	非該当
4	4	どちらか . . .	非該当	非該当	非該当
5	5	関心がある	非選択	非選択	非選択
6	6	どちらか . . .	非選択	非選択	非選択
7	7	関心がある	選択	選択	非選択
8	8	どちらか . . .	非該当	非該当	非該当
9	9	どちらか . . .	選択	非選択	非選択
10	10	関心がある	選択	選択	非選択
11	11	関心がある	非選択	選択	選択
12	12	関心がある	選択	非選択	非選択

３．データ共有の段階

- データ確認と処理
  - 報告書の集計表が再現できるかを度数分布表によりチェック
    → もしズレがあるなら、調査実施した方々へ問い合わせ
  - データの再クリーニング
    → エラーをみつけたら、修正するか、エラーがある旨を文書に残すか
  - 個人情報の処理
    → 調査の独自ID、地点情報など、個人情報とつながりうる情報は削除
  - レアケースの処理
    → データ分析の結果として、レアケースであり、それが個人を特定しかねない場合は何らかの対処
      - 分類を粗くすることでレアではないようにする
      - その質問の変数自体を削除する
      - （SSJDAではやっていないが）時には、意図的にエラーを含めることも

３．データ共有の段階

- メタデータ
  - 当該の調査のスペックを整理して、情報を示す
    - 社会科学ではDDI（Data Documentation Initiative）というメタデータ国際標準があり、それへと対応するように
  - 示すべき情報
    1. 調査の問題関心・目的などデータの実質的内容
    2. 調査実施に携わった団体・研究会や研究者など形式的側面
    3. 調査の概要が一読して分かるような重要情報
    4. 実際に用いられた質問紙
  - 調査の報告書や、前回調査・関連調査のメタデータを参照しつつ作成

SSJDAでの概要作成の各項目

●社会調査における基本的項目など15項目

- **調査の概要**
- 調査対象
- サンプルサイズ
- 調査時点
- 調査地域
- 標本抽出
- 調査方法
- 調査実施者
- 調査票・コードブック・集計表など
- 委託者（経費）
- 寄託時の関連報告書・関連論文
- **主要調査事項**
- 公開年月日
- トピック
- 特記事項

SSJDA Direct
Social Science Japan Data Archive

English

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

概要	
調査番号	PM080
調査名	東大社研・壮年パネル調査（JLPS-M）wave1-8, 2007-2014
寄託者	東京大学社会科学研究所パネル調査プロジェクト
利用申込先・承認手続き	利用方法の詳細はこちら SSJDAが利用申請を承認したときに利用できる
教育目的(授業など)の利用	教育（授業・卒論等）も可
利用期限	研究はなし教育は一年
データ提供方法	ダウンロード
メタデータ閲覧・オンライン分析システムNesstar	利用不可
調査の概要	労働市場の構造変動，急激な少子高齢化，グローバル化の進展などにともない，日本社会における就業，結婚，家族，教育，意識，ライフスタイルのあり方は大きく変化を遂げようとしている。これからの日本社会がどのような方向に進むのかを考える上で，現在生じている変化がどのような原因によるものなのか，あるいはどこが変化してどこが変化していないのかを明確にすることはきわめて重要である。 東京大学社会科学研究所パネル調査プロジェクトは，こうした問題をパネル（追跡）調査の手法を用いることによって，実証的に解明することを研究課題とする。このため東京大学社会科学研究所では，「働き方とライフスタイルの変化に関する全国調査」（Japanese Life Course Panel Surveys－JLPS）として，若年パネル調査（JLPS-Y），壮年パネル調査（JLPS-M），高卒パネル調査（JLPS-H），中学生親子パネル調査（JLPS-J）の4つのパネル調査を実施している。 2007年1月に始められたJLPS-YとJLPS-Mは，職業，家族，教育，意識（政治的態度を含む），健康など，網羅的な質問項目を含んでおり，日本では数少ない大規模パネル調査の1つである。また，職業に関する項目は非常に詳細で，社会階層と社会移動に関する全国調査（SSM調査）に準拠する項目が尋ねられている。 このように，JLPSは，特に英米における豊富なパネル調査の経験を参考に，国際比較分析が可能になるような設計を心がけているのみならず，既存の日本の調査（クロスセクショナルなものを含む）の調査項目も参考にしている。 JLPS-Yは，2006年12月末現在で20歳から34歳のいわゆる「若年層」を，またJLPS-Mは，35歳から40歳の「壮年層」を対象としたもので，両者の質問項目は一致している。 ここに収録の調査は，上記4つの調査のうち，2007年～2014年にそれぞれ実施した「壮年パネル調査」（Japanese Life Course Panel Surveys of the Middle-aged; JLPS-M）wave1～wave8に関するものである。同調査はその後も毎年継続して実施する予定である。なお，2014年同時期に行った「若年パネル調査」はSSJDA調査番号 PY080に収録されている。 また，2007年～2013年に行った「壮年パネル調査（wave1～7）」については，以下をそれぞれ参照のこと。 PM010　PM020　PM030　PM040　PM050　PM060　PM070 また2011年度からは，長期追跡に伴う回答者の脱落問題を考慮し，継続調査と同年代の対象者（2011年に39歳から44歳）を新たに追加している。2007年からの継続調査データはPM080.sav，2011年からの追加調査データはPM080_add2.savとなっている。 なお，社会科学研究所パネル調査プロジェクトの推進にあたっては，以下の資金提供を受けている。東京大学社会科学研究所研究経費（2003年度～），独立行政法人日本学術振興会科学研究費補助金（基盤研究S：2006～2009年度，2010～2014年度），厚生労働科学研究費補助金（政策科学推進研究：2004～2006年度），奨学寄付金：株式会社アウトソーシング（代表取締役社長・土井春彦，本社・静岡市）：2006～2008年度。

背景の問題関心

プロジェクト概要

前回調査関連調査

予算など

まとめ

- 社会科学では、社会調査により研究データを得ることが多い
- そのうち質問紙調査による個票データのキュレーションの実務を、１）質問紙回収直後、２）調査報告、３）データ共有の各段階での具体的作業について述べた
- FAIRの原理に適うための仕事の多くは、データ共有段階すなわちデータアーカイブが担っているとみてよい
- しかしながら、社会科学の調査データのうち、データアーカイブから公開されているものは決して多くはない
- データアーカイブを育てるための支援を拡大すること、研究者・機関にとってデータ共有をするためのインセンティブを与えること、などがこれからの重要な課題といえる