Podręcznik archiwizacji danych społecznych

Bogdan Cichomski, Tomasz Jerzyński, Marcin Zieliński

Zespół Ośrodka Badań Socjologicznych
Instytutu Studiów Społecznych
Uniwersytetu Warszawskiego

wersja 2004-09-21, 15:40
najnowszą wersję Podręcznika można znaleźć pod adresem:
http://www.ads.org.pl/pdf/podrecznikADS.pdf

Spis treści

1. Wstęp.

Archiwizacja danych pochodzących z badań nie jest pomysłem nowym. Potrzebę istnienia tego typu instytucji badacze uświadamiali sobie od dawna. Z przekazywaniem a następnie udostępnianiem zainteresowanym zbiorów danych związane są takie korzyści jak choćby realizacja postulatu otwartości warsztatu badawczego, co niewątpliwie przyczynia się do wypracowania i standaryzacji stosowanych procedur badawczych czy wzbogacenia zasobu wiedzy w dziedzinie metodologii badań społecznych.

Zbieranie a następnie udostępnianie danych za pośrednictwem jednej instytucji otwiera drogę do upowszechniania standardów konstrukcji narzędzi pomiarowych, w tym także wskaźników służących do pomiaru zjawisk społecznych. Pozwala na ich intersubiektywną kontrolę i dostarcza nie tylko gotowe wzorce ale również inspiruje do własnych poszukiwań na bazie zastanych już projektów badawczych. Zbieranie w jednym miejscu informacji na temat projektów już zrealizowanych lub pozostających w trakcie realizacji umożliwia wreszcie pozyskanie informacji i źródeł danych dla przeprowadzania własnych analiz z wykorzystaniem danych zastanych i bazując na projektach zrealizowanych przez innych. Jest to o tyle istotne, że często może pozwolić z jednej strony na uniknięcie powtarzania badań a z drugiej na ich wzbogacenie.
Często jest też tak, że badacze ograniczają stawiane przez siebie hipotezy do celów, jakie stawiali sobie projektując swoje badanie. Zastosowania gotowych już zbiorów danych mogą tymczasem znacznie wykraczać poza cele, jakie badaniu stawiano podczas tworzenia jego projektu. Udostępnienie danych osobom niezwiązanym pośrednio czy bezpośrednio z danym projektem otwiera drogę do nowych odkryć i weryfikacji hipotez innych niż pierwotnie stawiane za cel badania. Badacze powinni więc liczyć się z taką właśnie możliwością wykorzystania ich danych. Jednocześnie prowadzi to jednak do wzbogacenia zasobu informacji na temat zjawisk społecznych a zatem poszerza wiedzę na temat życia społecznego, co jest przecież pierwotnym i podstawowym celem każdego badania.

Gromadzenie danych i ich upowszechnianie poprzez redystrybucję jest wreszcie znakomitym materiałem pomocniczym, wykorzystywanym do celów dydaktyki. Różnorodność tematyki badawczej, koncepcji metodologicznych czy wreszcie osiągniętych rezultatów badania, z jaką można się zetknąć korzystając z zasobów zgromadzonych w ADS, jest niewątpliwie wartościowym materiałem służącym upowszechnianiu wiedzy na temat nauk społecznych zarówno dla osób zaawansowanych, jak i dla tych, którzy dopiero wkraczają w świat badań społecznych.

Oddając do Państwa dyspozycji Archiwum Danych Społecznych mamy nadzieję na przyczynienie się wspólnymi siłami do realizacji tych postulatów, zdając sobie sprawę jednocześnie, że powodzenie tego przedsięwzięcia zależy w równym stopniu od pracy osób odpowiedzialnych za organizację i funkcjonowanie Archiwum jak i samych badaczy, którzy deponując dane winni dbać o ich jakość i zgodność z międzynarodowymi standardami przygotowania i obróbki techniczno-metodologicznej zbiorów i dokumentacji.

Podręcznik ten zawiera informacje na temat wymogów stawianych zbiorom deponowanym w ADS a także standardowych procedur ich przygotowania, oraz informacji jakie powinna zawierać dołączana do nich dokumentacja techniczno-metodologiczna.

2. Korzyści archiwizacji danych.

Gromadzenie zbiorów danych w jednym miejscu z zachowaniem jednolitych kryteriów ich standaryzacji jest zadaniem złożonym i kosztownym, z którego płyną jednak istotne korzyści dla całego środowiska badaczy zjawisk społecznych.
S.E. Fienberg zwraca uwagę na takie korzyści z upubliczniania danych, jak:

- ułatwienie realizacji postulatu otwartości warsztatu badawczego i tym samym wdrożenie idei intersubiektywnej kontroli procesu badawczego.
- inspirowanie do mnożenia analiz i hipotez z wykorzystaniem zebranych i dostępnych już danych.
- promowanie nowych badań i umożliwienie testowania nowych lub alternatywnych metod weryfikacji postawionych już bądź stawianych hipotez.
- usprawnianie metod zbierania danych i konstrukcji pomiarów. Ogólnodostępne archiwum danych otwiera naukowej społeczności możliwość wypracowywania standardów metodologicznych
- dostęp do danych z badań już zrealizowanych może się przyczynić do poszerzenia zakresu dokonywanych analiz bez konieczność powtórnego przeprowadzania badań. Nie istniałaby ogromna liczba publikacji i artykułów powstałych na bazie takich badań jak np. Polskie Generalne Sondaże Społeczne, gdyby ich autorzy musieli sami zbierać tego typu dane.
- Archiwum jest wreszcie nieporównywalnym z żadnym innym źródłem danych wykorzystywanych dla celów dydaktycznych. Dostarcza zarówno wykładowcom jak i studentom dane o najwyższym standardzie metodologicznym.

Współpraca jest warunkiem sine qua non sprawnego działania Archiwum i skutecznego spełniania funkcji, jaka została mu powierzona. Oznacza to, że badacze powinni:

- Informować o odkrytych przez siebie błędach, a także reagować na błędy we własnych materiałach przekazywanych do Archiwum a wykrytych przez innych.
- Liczyć się z możliwością krytycznej oceny własnej pracy będącej skutkiem przeprowadzania analiz przez innych badaczy na bazie tych samych, dostępnych za pośrednictwem Archiwum danych.
- Liczyć się z możliwością dokonywania odkryć przez innych badaczy na bazie zastanych zbiorów danych.
- Pogodzić się z wykorzystaniem swoich danych przez inne osoby i w innych niż założone przez autora celach.

Często zdarza się, że użytkownicy danych wykrywają błędy popełnione przez autorów badań. Autorzy ci winni przychylnie odnosić się do takich osób i rezultatów ich pracy gdyż przyczynia się to do realizacji warunku intersubiektywnej kontroli procesu badawczego przez co możliwe jest takie odnoszenie się do wyniku, które uwzględnia popełnione przez badacza błędy. Z drugiej zaś strony przyczynia się do usprawnienia procesu planowania i realizacji przyszłych badań, z korzyścią dla całego środowiska badaczy.
Innowacje wnoszone przez użytkowników danych mogą także przyczynić się do budowania nowych i wzbogacenia już istniejących perspektyw analitycznych.
Archiwizacja danych poprzez upublicznienie procedur badawczych i uzyskanych wyników sprzyja ich ciągłemu doskonaleniu i umożliwia ich pełną kontrolę, co jest istotne zwłaszcza w przypadku badań finansowanych ze środków publicznych.
Archiwizacja danych wreszcie powinna stać się integralną częścią nauk społecznych, a ich udostępnianie - warunkiem otrzymywania grantów.

Zadaniem Archiwum jest archiwizacja deponowanych w nim zbiorów danych oraz dokumentacji techniczno-metodologicznej zgodnie z przyjętym standardem archiwizacyjnym. Podręcznik ten, poprzez opis wymogów standaryzacyjnych, ma na celu dopomożenie w takim przygotowaniu danych i dokumentacji aby były one zgodne z wymaganiami stawianymi badaczowi przez Archiwum.

Zakładamy, że czytelnik jest zaznajomiony z podstawowymi pojęciami związanymi z plikami danych, takimi jak zmienne, pola, kody itp. Dla osób, którym pojęcia te są jednak obce, na końcu dokumentu umieściliśmy słownik, który może okazać się pomocny w zrozumieniu znaczenia niektórych technicznych terminów występujących w tekście.
Standardowym formatem akceptowanych przez Archiwum zbiorów danych, jest – ze względu na swą powszechność – SPSS , stąd też podręcznik ten opisuje wymagania w odniesieniu do tego właśnie pakietu statystycznego. Zachęcamy zatem do korzystania z dokumentacji i podręczników dołączanych do samego programu.

Prosimy o kontakt w przypadku jakichkolwiek wątpliwości powstałych w trakcie czytania tego podręcznika. Dzięki Państwa sugestiom będzie możliwe jego udoskonalenie i lepsze dostosowanie do Państwa potrzeb.

3. Znaczenie planowania zarządzaniem danymi

Punktem krytycznym z punktu widzenia przyszłej archiwizacji danych zgodnie z wymaganiami stawianymi przez Archiwum tak, aby standaryzacja zbioru wymagała najmniej wysiłku ze strony badacza, jest optymalizacja planu realizacji badania.
Stąd też zachęcamy do wcześniejszego zapoznania się i wdrożenia proponowanych rozwiązań, jeszcze na etapie planowania samego badania.
Część z opisywanych w Podręczniku archiwizacji danych społecznych rozwiązań jest obligatoryjna dla osób czy instytucji deponujących dane w Archiwum. Niektóre z nich są zalecane, jako warte wdrożenia ze względu na późniejszą prostotę posługiwania się tak przygotowanymi zbiorami przez ich użytkowników.

3.1. Dokumentacja jako część planu projektu

Dokumentacja powinna być prowadzona na każdym etapie realizacji badania i powinna obejmować także wszystkie operacje jakim poddawany był produkt finalny w postaci zbioru. Już na etapie planowania badania należy ustalić:

- Strukturę zbioru. *Jak powinien wyglądać i być zorganizowany zbiór danych? Co jest jednostką analizy? Czy zbiór będzie zawierał jeden długi rekord, czy też będzie podzielony na kilka krótszych?*

- Konwencję nazewnictwa. *Jak nazwane będą pliki danych oraz wg jakich kryteriów nadawane będą nazwy zmiennym?*

- Sposoby rozwiązania problem kontroli spójności danych. *Sposób, w jaki dane zostaną wprowadzone do komputerowego zbioru danych, jakie testy kontrolne zostaną przeprowadzone na okoliczność występowania niedozwolonych wartości (wartości spoza skali), odpowiedzi sprzecznych logicznie, niekompletnych zapisów, etc.?*

- Sposób przygotowania książki kodów. *W jaki sposób książka kodów (codebook) będzie zorganizowana i jaką będzie posiadała formę oraz jakie informacje zawierała?*

- Zasady tworzenia zmiennych wtórnych. *Które ze zmiennych są zmiennymi utworzonymi na podstawie zbioru oryginalnych danych pochodzących z badania? Jak powstały te zmienne? Jak będzie dokumentowany sposób ich tworzenia?*

- Strukturę dokumentacji. *Jak i gdzie podane zostaną informacje o zbieraniu danych, kodowaniu, tworzeniu zmiennych itp.?*

3.2. Oprogramowanie

Obróbka danych
Program wczytujący dane powinien kontrolować filtry i spójność logiczną wprowadzanych danych. Możliwe jest także przeprowadzanie kontroli spójności logicznej zbioru po zakończeniu etapu wprowadzania danych a zatem na gotowym już zbiorze.
Wszelka obróbka danych już wprowadzonych powinna być dokonywana wyłącznie za pomocą oprogramowania statystycznego. Używanie w tych celach powszechnie dostępnych arkuszy kalkulacyjnych, pozbawionych narzędzi kontrolujących, może powodować generowanie błędów.
Do przygotowywania zbiorów danych przeznaczonych do archiwizacji oraz dokumentacji technicznej zaleca się używanie profesjonalnych pakietów statystycznych (SPSS lub Statistica).
Archiwum będzie przyjmowało wyłącznie zbiory danych zapisane w formacie portable, SPSS lub ASCII wraz ze słownikiem definiującym dane (*data definition statements*).

Dokumentacja
Do przygotowywania dokumentacji może być stosowane dowolne oprogramowanie. Należy jednak uwzględnić fakt, że Archiwum będzie przyjmowało wyłącznie pliki w formacie ASCII lub PDF. Przekazywanie dokumentacji w innych formatach, jako zależnych od używanej platformy systemowej, jest niedozwolone.

4. Kontrola poprawności danych w zbiorze i jego spójności

Droga od kwestionariusza do zbioru danych jest złożona i może nastręczać wiele problemów. Nieuniknione są błędy w postaci np. pięcioletnich respondentów posiadających szóstkę dzieci. Zadanie badacza polega na ich znalezieniu i korekcji --- choć idealnym byłoby niedopuszczenie do ich powstania.

Oto kilka wskazówek, które mogą być przydatne podczas procesu wczytywania danych i przygotowywania zbioru do archiwizacji:

- Jak wskazaliśmy powyżej, należy używać dedykowanego oprogramowania do wprowadzania danych. Powinno ono umożliwiać kontrolę wprowadzanych wartości, zawierać zabezpieczenia przed ,,dzikimi kodami'' (niedozwolonymi wartościami spoza skali) i utrzymywać - na podstawowym poziomie - spójność logiczną rekordu.

- Należy rozważyć możliwość dwukrotnego wprowadzania danych. Wykryte w ten sposób rozbieżności mogą być na bieżąco korygowane przez osoby wpisujące dane.

- Konieczne jest całkowite odseparowanie procesu kodowania od wprowadzania danych. Osoba wprowadzająca dane nie powinna jednocześnie dokonywać kategoryzacji.

- Proces przygotowywania danych należy poprowadzić tak, aby poszczególne jego części --- takie jak np. kodowanie zawodów --- prowadzone były od początku do końca przez jedną, przygotowaną do tego osobę lub też zespół takich osób.

- Wszelkich przekodowań czy tworzenia zmiennych wtórnych należy dokonywać wyłącznie za pomocą pakietów statystycznych. Pozwala to wyeliminować pomyłki i --- w razie sprzeczności logicznych czy programistycznych --- łatwo wprowadzać poprawki do procedury rekodującej.

4.1. Kontrola błędów w zbiorze

Występujące w zbiorach błędy mogą powstawać już na etapie projektowania narzędzia badania. Mogą też być skutkiem nierzetelnej pracy ankietera, pomyłek koderów czy wreszcie pomyłek samego badacza podczas obróbki i przygotowania zbioru do postaci dystrybucyjno-archiwizacyjnej.
Z perspektywy poprawności i integralności – a tym samym jakości przygotowanych przez badacza a dystrybuowanych przez Archiwum - zbiorów danych można wyróżnić dwa najczęściej spotykane w zbiorach danych typy błędów.

4.1.1. „Dzikie kody”

Tzw. „dzikie kody” są wartościami wykraczającymi poza założony przez badacza zakres zmienności danej zmiennej. Są one typowym błędem osób wczytujących dane.
Np. w pytaniu:

8. PŁEĆ RESPONDENTA (WPISAĆ BEZ ZADAWANIA PYTANIA):

MĘŻCZYZNA 1

KOBIETA 2

W zbiorze danych uzyskano następujący rozkład odpowiedzi:

PLEC RESPONDENTA: 1=M, 2=KOB

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	4949	44.2	44.2	44.2
	2	6238	55.7	55.7	100.0
	3	2	.0	.0	100.0
	4	3	.0	.0	100.0
	Total	11192	100.0	100.0	

Kody 3 i 4 są w tym przypadku kodami spoza dopuszczalnego zakresu zmienności.
Sprawdzenie zbioru danych na okoliczność występowania tego typu nieudokumentowanych kodów należy do obowiązków osoby/instytucji deponującej dane.

Zbiory zawierające nieudokumentowane wartości nie będą przyjmowane przez Archiwum do archiwizacji.

4.1.2. „Outliers”

Obok klasycznego przypadku występowania w zbiorze nieudokumentowanych kodów, jak w przykładzie powyżej, możliwa jest również sytuacja, w rozkładzie zmiennej występują kody wprawdzie mieszczące się w założonej skali, jednak na tyle nieprawdopodobne, że takie odpowiedzi respondenta wymagają sprawdzenia i potwierdzenia zgodności zbioru z wypełnioną ankietą/kwestionariuszem wywiadu.

Np. w odpowiedzi na pytanie:

12. A. Ile miał Pan(i) braci i sióstr? Prosimy uwzględnić żywo urodzonych -- żyjących obecnie, oraz już nie żyjących . Prosimy także uwzględnić braci i siostry przyrodnie i dzieci adoptowane przez Pan(i) rodziców.

LICZBA BRACI I SIÓSTR:..... ☐☐

NIE MIAŁ(A)M BRACI I SIÓSTR **00**

Uzyskano następujący rozkład odpowiedzi:

Q12A LICZBA BRACI I SIOSTR

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	124	6.3	6.3	6.3
	1	586	29.7	29.7	36.0
	2	394	20.0	20.0	56.0
	3	342	17.3	17.3	73.3
	4	215	10.9	10.9	84.2
	5	147	7.5	7.5	91.7
	6	91	4.6	4.6	96.3
	7	72	3.7	3.7	99.9
	17	1	.1	.1	100.0
	Total	1972	100.0	100.0	

Kodu „17”, oznaczającego że respondent miał siedemnaścioro rodzeństwa nie można traktować jako wartości nie mieszczącej się w założonej przez badacza skali odpowiedzi na pytanie o liczbę braci i sióstr. Sytuacja taka jest możliwa aczkolwiek może się wydawać nieprawdopodobna.

Często zdarza się, że ich występowanie w zbiorze jest skutkiem pomyłek osób wczytujących dane, choć oczywiście mogą to być także odpowiedzi prawdziwe.

Przed przekazaniem zbioru Archiwum zalecane jest sprawdzenie zbioru na okoliczność występowania wartości typu Outliers, czyli odpowiedzi w znaczący sposób odbiegających od odpowiedzi udzielanych przez innych respondentów.

4.2. Kontrola spójności logicznej

Kontrola spójności logicznej odpowiedzi na pytania jest jednym z najtrudniejszych elementów przygotowywania zbioru do dystrybucji. Wymaga doskonałej znajomości struktury kwestionariusza. Polega ona na wychwytywaniu niespójnych logicznie odpowiedzi na dwa lub więcej pytań zadanych respondentowi.

4.2.1 Pytania filtrujące

Typowymi przykładami błędów logicznych wynikających z pominięcia reguł filtrujących może być deklaracja bezdzietności i jednoczesna odpowiedź na pytanie o zadowolenie z dzieci lub odpowiedź przecząca na pytanie „Czy miewa Pan(i) okazje, kiedy pija Pan(i) napoje alkoholowe, takie jak wódka, wino czy piwo, czy też jest Pan(i) całkowitym(-tą) abstynentem(-ką) i jednoczesne twierdzenie, że nie zdarza się respondentowi czasami wypić więcej niż sądzi, że powinien.
W tego typu sytuacji dochodzi do niespójności logicznej między odpowiedziami na pytanie o deklarację dzietności a kolejnymi pytaniami, których zadanie winno być konsekwencją odpowiedzi udzielonej przez respondenta na pytanie wcześniejsze.

Np. pytanie 146B zadawane winno być jedynie osobom, które zadeklarowały, że piją alkohol. Ustaleniu tego faktu służy pytanie 146A.

146. A. Czy miewa Pan(i) okazje, kiedy pija Pan(i) napoje alkoholowe, takie jak wódka, wino czy piwo, czy też jest Pan(i) całkowitym(-tą) abstynentem(-ką)?

Pijam alkohol 1

Jestem całkowitym(-tą) abstynentem(-tką) ..... 2 → PRZEJŚĆ DO PYT. **147 A**

B. Czy zdarza się Panu(-i) czasami wypić więcej, niż sądzi Pan(i), że powinien (powinna)?

Tak 1
Nie.....2
NIE WIEM.....8

Błędem wynikającym z niezastosowania się ankietera do zasady określonej w pytaniu filtrującym byłoby zatem zadanie pytania 146B osobom, które w pytaniu 146A stwierdziły, że są całkowitymi abstynentami.

4.2.2 Instrukcje przed pytaniem

Błędami logicznymi mającymi konsekwencje dla spójności i zgodności przeprowadzonego wywiadu z logiką kwestionariusza są też błędy wynikające z niezastosowania się ankietera do instrukcji umieszczanych przed niektórymi z pytań.

Np. w pytaniu 94 pytano respondentów o ocenę własnego małżeństwa/związku. Pytanie jednak należało zadawać wyłącznie osobom, które obecnie są żonate/zamężne lub też pozostają w związku konkubinackim.

PYTANIE 94 DOTYCZY TYLKO RESPONDENTÓW OBECNIE ŻONATYCH / ZAMĘŻNYCH/ W KONKUBINACIE: (POR. PYT. 41A I 41C). POZOSTALI PRZEJŚĆ DO PYT. 95.

94. Biorąc wszystko pod uwagę, jak określił(a)by Pan(i) swoje małżeństwo /związek? Czy uważa je Pan(i) za bardzo szczęśliwe, raczej szczęśliwe, czy niezbyt szczęśliwe?

Bardzo szczęśliwe.....1
Raczej szczęśliwe.....2
Niezbyt szczęśliwe.....3
NIE WIEM.....8

Błędem ankietera będącym skutkiem niezastosowania się do instrukcji umieszczonej przed pytaniem byłoby w tym przypadku zadanie tego pytania także osobom nie będącym żonatymi bądź niezamężnymi oraz nie pozostającym w związkach konkubinackich.
Instrukcje umieszczane przed pytaniem są – podobnie jak pytania filtrujące - także instrukcjami wyłączającymi pewne grupy respondentów z zadawania danego pytania lub bloku pytań.

4.2.3 Daty i przedziały czasu

Innym, równie ważnym typem błędów logicznych wymagających kontroli przed przekazaniem zbioru do Archiwum, jest sprawdzenie spójności podawanych przez respondenta dat i przedziałów czasu. W zbiorach danych przed kontrolą logiczną często możemy spotkać odpowiedzi świadczące np. o dacie rozpoczęcia pracy po okresie jej zakończenia czy np. dłuższego okresu pobierania nauki niż wynosi dotychczasowa długość życia respondenta.

Np. w pytaniu o daty rozpoczęcia i zakończenia pierwszej w życiu pracy zawodowej respondenta, uzyskano następującą informację:

28. Chciał(a)bym teraz porozmawiać o Pana(i) pierwszej w życiu stałej pracy, przynoszącej Panu(i) dochód i którą wykonywał(a) Pan(i) bez przerwy przez co najmniej rok.

A. W którym roku rozpoczął(-ęła) Pan(i) pracę w tym zakładzie

W ROKU: **1 9 6 2**

B. W którym roku przestał Pan(i) wykonywać tę pracę?

W ROKU: **1 9 5 8**

Oczywistym jest, że informacja zawarta w tej odpowiedzi jest informacją błędną. Respondent nie mógł bowiem zakończyć wykonywania pierwszej pracy przed datą jej rozpoczęcia. W przypadku tego typu niespójności logicznej między pytaniami, nieprawdziwa informacja może być zawarta w pierwszej, podanej przez respondenta dacie, drugiej lub też obie nie identyfikują poprawnie czasu rozpoczęcia i zakończenia pierwszej pracy zarobkowej.

Błędy logiczne mogą być skutkiem udzielania przez respondentów mylnych czy nieprawdziwych informacji, pomyłek ankietera podczas notowania odpowiedzi czy wreszcie pomyłek osób wczytujących dane.

Badacze deponujący zbiory w Archiwum winni dbać o poprawność i spójność logiczną deponowanych danych. Przeprowadzanie kontroli spójności logicznej deponowanych danych nie jest wymogiem dopuszczenia zbioru do archiwizacji. Fakt poddania bądź nie danego zbioru procesowi kontroli spójności logicznej, będzie jednak odnotowywany i podany użytkownikom.
Archiwum zaleca przeprowadzanie kontroli spójności logicznej przed przekazaniem danych do archiwizacji.

4.3. Kontrola zbiorów o strukturach hierarchicznych.

Strukturę hierarchiczną posiadają na ogół zbiory danych uzyskane z badań panelowych. Ich cechą charakterystyczną jest przypisanie jednej jednostce badania więcej niż jednego rekordu w zbiorze.

Zbiory danych o strukturze hierarchicznej winny być kontrolowane na okoliczność występowania niedozwolonych wartości (dzikich kodów), wartości w znaczący sposób odbiegających od odpowiedzi innych respondentów (outliers) jak i pod względem logicznej spójności dla każdego rekordu z osobna oraz nawzajem pomiędzy rekordami.

5. Przygotowanie dokumentacji techniczno-metodologicznej badania

5.1. Książka kodów (*codebook*)

Książka kodów (*codebook*) jest niezbędnym elementem dokumentującym zarówno samo badanie jak i powstały w jego wyniku zbiór danych.
Książka kodów dostępna wraz ze zbiorem danych powinna zawierać informacje na temat:

1. Dane o kierowniku projektu i informacje o jego afiliacji w czasie wykonywania badań oraz informacje na temat instytucji realizującej badanie.

2. Oficjalny tytuł badań.

3. Wzór cytowania danych

4. Źródła finansowania, numer grantu oraz opis innych, związanych z badaniem, sposobów finansowania.

5. Skład osobowy zespołu przetwarzającego dane.

6. Opis projektu, w tym: opis celu badań, krótka historia projektu z opisem ważniejszych problemów i decyzji podejmowanych w trakcie jego realizacji.

7. Informacje na temat próby i procedury jej doboru. Sekcja ta powinna opisywać badaną populację i metodę użytą do wyłonienia z niej próby. Powinna tu także znaleźć się odpowiedź na pytanie, czy stosowanie - przy tak wylosowanej próbie - standardowych procedur wyliczania błędów estymatorów jest możliwe. Jeśli nie, należy zamieścić procedury szacowania błędów estymatorów dostosowane do danego modelu losowania próby. Jeśli zbiór danych zawiera wagę, musi być szczegółowo opisana procedura jej tworzenia i sposób użycia. Niezbędną informacją w tej sekcji jest dokładne określenie liczebności próby założonej (wylosowanej) oraz liczebności próby zrealizowanej. Informacje te są niezbędne do poprawnego stosowania procedur szacowania błędów estymatorów. Nie mniej ważne jest dokładne określenie okresu w którym przeprowadzane było badanie terenowe.

8. Obszar geograficzny, na którym badanie zostało zrealizowane oraz termin jego realizacji

9. Źródła danych, tj. informacje na temat technik użytych do zbierania danych

10. Informacje na temat tego, co było jednostką badania

11. Zestawienie rozkładów liczebności i częstości jednej zmiennej (wraz z używanymi kodami specjalnymi na oznaczenie braków danych) dla wszystkich zmiennych występujących w zbiorze wraz z :

a. Pełnym, identycznym z zamieszczonym w kwestionariuszu brzmieniem pytania. Jeśli pytanie było zaczerpnięte z innych badań, konieczne jest podanie źródła.

b. Jednoznacznym odniesieniem danej zmiennej do jej odpowiedniego miejsca w kwestionariuszu.

c. Określeniem grupy będącej adresatem pytania a zatem uwzględnienie przyczyn i zastosowania zasady filtrowania określonych grup respondentów.

d. Pełnym zestawieniem wartości kodów specjalnych wraz z ich opisem.

e. Jeśli zmienna nie jest zmienną pierwotną czyli powstała w wyniku obliczeń na zmiennych pierwotnych, lub jeśli zmienna została poddana jakiejkolwiek innej edycji, należy zamieścić dokładny opis użytych procedur.

f. Należy opisać wszystkie wartości wszystkich zmiennych, których rozkłady są prezentowane w książce kodów.
Jedynymi akceptowanym odstępstwem od tej zasady jest zamieszczenie dokładnego opisu wartości w postaci aneksu dołączanego do książki kodów w przypadku niektórych zmiennych (jak np. klasyfikacje zawodowe) lub – w przypadku zmiennych ciągłych – opisanie jedynie najniższej i najwyższej wartości danej zmiennej z podaniem jednostki w jakiej została wyrażona.

g. Kody typu *BRAK DANYCH* i *NIE DOTYCZY* powinny zostać wyłączone z podstawy oprocentowania. W książce kodów i zbiorze wartości zadeklarowane jako wyłączane z podstawy oprocentowania (typu *MISSING VALUES*) powinny być identyczne.

h. Struktura, układ oraz kolejność zmiennych w książce kodów i dokumentacji powinna odpowiadać kolejności zmiennych w zbiorze danych. W przypadku jakichkolwiek rozbieżności należy załączyć opis niezgodności i podać jej przyczynę.

i. Dokumentacja badania musi zawierać kopię narzędzia badania, które posłużyło do zbierania danych. Umożliwia to np. śledzenie kontekstu zadanego pytania.

j. Pomocnym w pracy ze zbiorem danych i jego dokumentacją jest umieszczenie w książce kodów alfabetycznej listy zmiennych wraz z odnośnikami do stron, na których zmienne te zostały opisane.

k. W złożonych projektach, gdzie wykorzystywana jest większa liczba narzędzi badawczych wskazane jest rozrysowanie ich struktury w postaci grafu przepływu informacji.

l. Dokładny opis stosowanych skrótów i konwencji stosowanych w nazewnictwie zmiennych oraz standardów przydzielania kodów specjalnych.

11.1. Zestawienie rozkładów liczebności i częstości jednej zmiennej może zostać przedstawione w postaci zestawienia rozkładów liczonych osobno dla każdego roku badania (w przypadku badań powtarzanych) lub zestawień rozkładów dla każdej fali badania (w przypadku badań panelowych). Podstawową zasadą, jaką powinien kierować się badacz konstruujący zestawienia tabelaryczne jest wygoda użytkownika korzystającego z tych zestawień.

Przykładowe zestawienie tabelaryczne z książki kodów (*codebook*) badania powtarzanego wraz z opisem poszczególnych elementów zawiera Aneks A.

5.2. Rozkłady częstości (*frequencies*)

Do każdego deponowanego zbioru danych winny zostać dołączone nieważone rozkłady częstości jednej zmiennej, zapisane w pliku ASCII. Dołączane rozkłady winny zawierać nazwę zmiennej z jej etykietą, jej wartości wraz z kodami, rozkład liczebności, częstości oraz częstości skumulowanej.
Dołączane rozkłady powinny zostać skonstruowane w oparciu o ostateczną wersje zbioru – tą, która ma zostać zdeponowana w Archiwum. Niedopuszczalne są jakiekolwiek rozbieżności między załączonymi rozkładami częstości a rozkładami, które można uzyskać z załączonego zbioru danych.

Zbiory danych bez dołączonych rozkładów częstości nie będą przyjmowane do archiwizacji.

5.3. Statystyki opisowe (*descriptive statistics*)

Do deponowanego zbioru danych należy również dołączyć zestawienie statystyk opisowych dla wszystkich zmiennych znajdujących się w deponowanym zbiorze danych obliczone dla danych nieważonych. Przez statystyki opisowe należy tu rozumieć zestawienie liczebności, minimum, maksimum, średnią oraz odchylenie standardowe wyliczone dla każdej zmiennej występującej w zbiorze wraz z nazwą, pod jaką występuje i jej etykietą.

Zbiory bez dołączonych zestawień statystyk opisowych nie będą przyjmowane do archiwizacji.

6. Przygotowanie zbiorów do archiwizacji

6.1. Akceptowane formaty zbiorów danych

Archiwum będzie przyjmować zbiory danych w następujących formatach:

1. Format ASCII wraz ze słownikiem definiującym (*data definition statements*), zawierającym informacje na temat szerokości kolumn dla każdej zmiennej, ich etykiety, etykiety wartości zmiennych i definicje kodów specjalnych. Przez format ASCII należy tu rozumieć dane w postaci tekstowej bez separacji kolumn. Szczegółowych informacji na temat przygotowywania słownika definiującego dane dostarcza pkt. 6.2.1.i Aneks B Podręcznika.
2. Systemowy plik pakietu statystycznego SPSS (SAV)
3. Przenośny (portable) plik systemowy (POR)

Z archiwizacją plików w takich właśnie formatach wiążą się następujące korzyści:

1. Pliki ASCII są niezależne od platformy systemowej. Korzystać z nich można w dowolnym systemie używając dowolnego oprogramowania statystycznego. Zadaniem użytkownika jest jedynie odpowiednia preparacja słownika definiującego (*data definition statements*) w języku zrozumiałym przez pakiet statystyczny, którego używa.
Korzystanie z plików ASCII umożliwia także dostosowanie zbioru do własnych potrzeb jeszcze przed załadowaniem go do programu, co jest szczególnie istotne w przypadku zbiorów wymagających dużej przestrzeni dyskowej.

2**.** Zbiory danych typu portable oraz systemowe pliki zapisane w formacie SPSS są już preformatowane i gotowe do użycia dla osób posiadających i używających odpowiedniego oprogramowania. Brak ingerencji użytkownika w obszar definicyjny zbioru eliminuje możliwość powstawania błędów. Tego typu zbiory są bezpieczne i szybkie w użyciu. Jednak uciążliwe jest, a często niemożliwe otworzenie zbioru w programie innym niż ten, w którym zbiór został zapisany. Ten typ formatu zbioru danych zawęża grono jego użytkowników do użytkowników oprogramowania, w którym został zapisany. Pomimo wszystko systemowe pliki programów statystycznych są bardzo atrakcyjne dla użytkowników gdyż pozwalają na natychmiastowy, wolny od pomyłek dostęp do danych.

Idealnym zatem rozwiązaniem jest przekazywanie do archiwizacji zbiorów danych w dwóch formatach: ASCII wraz ze słownikiem definiującym (*data definition statements*) oraz pliku systemowego SPSS lub portable. Dopuszczalnym jest jednak przekazywanie zbioru w postaci pliku jednego z ww. rodzajów.

6.2. Techniczne aspekty zbiorów danych

6.2.1. Słownik definiujący dane (*data definition statements*)

Dołączenie słownika definiującego dane (*data definition statements*) jest niezbędne w przypadku dostarczania Archiwum zbioru danych zapisanego w formacie ASCII. Jego zadaniem jest określenie miejsca (poprzez podanie numeru kolumny), w którym program statystyczny ma odnaleźć daną zmienną a następnie określić jej właściwości, takie jak jej nazwę (*variable name*), etykietę (*variable label*), etykiety wartości (*value labels*) i w końcu wartości, które nie będą brane pod uwagę podczas analiz statystycznych (*missing values*).

Słownik definiujący dane zapisane w formacie ASCII winien więc zawierać:

1. Informację na temat lokalizacji wszystkich zmiennych (nr kolumny) wraz z ich nazwami
2. Zdefiniowane etykiety tych zmiennych nadane zgodnie z pkt. 6.2.3. Podręcznika
3. Opis etykiet wszystkich wartości tej zmiennej nadany zgodnie z pkt. 6.2.5. Podręcznika
4. Definicję wartości wyłączanych z analizy typu *BRAK DANYCH* i *NIE DOTYCZY,* nadane zgodnie z pkt. 6.2.6. Podręcznika

Przykładowy słownik definiujący dane dla pierwszych 5 zmiennych badania PGSS został przedstawiony w Aneksie B.

W przypadku dostarczenia zbioru danych zapisanego w formacie ASCII, słownik definiujący dane winien zostać przygotowany w języku poleceń programu SPSS tak, aby możliwe było wczytanie zbioru do pakietu SPSS bez konieczności interwencji użytkownika w składnię słownika.

6.2.2. Nazwy zmiennych (*variable names*)

W nadawaniu nazw zmiennym stosuje się następujące konwencje:

1. Numerowanie proste. W tym systemie każda zmienna numerowana jest od 1 do n, gdzie n jest liczbą zmiennych w zbiorze danych. Większość programów nie pozwala na zapisywanie nazw zmiennych zaczynających się od cyfry. W tym przypadku numery zmiennych można poprzedzić literą np. V1,V2,V3,...,Vn. Pomimo iż większość programów obsługuje etykiety zmiennych i możemy odczytać, że zmienna V13 to tak naprawdę ,,Q3b Wykształcenie matki" to system ten jest niewygodny w użyciu i bardzo podatny na błędy.
   W przypadku decyzji o zastosowaniu takiego właśnie systemu nadawania nazw zmiennym, koniecznym jest naniesienie w kwestionariuszu tak nadanych nazw odpowiadającym im pytaniom.
   Zalecane jest także umieszczanie numeru pytania, do którego odnosi się dana zmienna w opisie zmiennej (tj. etykiecie zmiennej).

2. Numerowanie według numerów pytań w kwestionariuszu. System ten jest podobny do poprzedniego. Nazwy zmiennych są tu jednak numerowane zgodnie z numerami pytań w kwestionariuszu np. Q1, Q2a itd. Podejście to ułatwia pracę ze zbiorem danych w oparciu o oryginalny kwestionariusz. Przy dużych zbiorach danych trudno jest jednak zapamiętywać tak nadane zmiennym nazwy, choć posługując się kwestionariuszem nie ma żadnych problemów z ich identyfikacją.

3. Nadawanie zmiennym nazw mnemonicznych, ułatwiających zapamiętywanie. Nazwy mnemoniczne są skrótami mającymi odnosić użytkownika do faktycznej treści pytania i zawartości zmiennej. Jest to najłatwiejszy do zapamiętania i rozpoznania systemu danych. Jednak należy pamiętać, że dla osoby, która stworzyła skrót jest on oczywisty, dla innych osób tak być nie musi. Tym bardziej, że skróty te mogą mieć nie więcej niż osiem liter – większość pakietów statystycznych nie akceptuje dłuższych nazw zmiennych.

   > Systematyzacja nazw mnemonicznych jest pomocną w zapamiętywaniu techniką nadawania zmiennym nazw. Polega ona na podziale nazwy zmiennej na przedrostek, rdzeń i przyrostek. Zmiennym odnoszącym się do danego, konkretnego przedmiotu badania nadawany jest identyczny przedrostek lub przyrostek, co ułatwia identyfikację tych zmiennych.

Np.: Zmiennym odnoszącym się do wykształcenia przypisano przyrostek ED. Zgodnie z tą zasadą, wykształcenie matki posiada nazwę mnemoniczną MAEDUC, ojca PAEDUC, współmałżonka SPEDUC a samego respondenta EDUC.
Korzystanie z takiego systemu nazewnictwa wymaga od badacza zaplanowania standardowych dwu- lub trzy- literowych skrótów, dla przedrostków, rdzeni i przyrostków.

Należy pamiętać, że nazwy zmiennych są jednym z najczęściej używanych elementów zbioru danych, stąd nie mogą zawierać nieprawidłowych informacji o zawartości zmiennych. Idealnie byłoby, gdyby niosły jak najwięcej informacji na temat zmiennej, do której się odnoszą.

Zalecane jest przekazywanie do archiwizacji zbiorów z nazwami zmiennych nadawanymi wg numerów pytań w kwestionariuszu, przez co stosunkowo łatwe jest odnalezienie żądanego pytania w zbiorze posługując się kwestionariuszem i pytania w kwestionariuszu, gdy użytkownik posługuje się w danym momencie zbiorem. W przypadku zastosowania innego rodzaju numeracji zmiennych koniecznym jest naniesienie nadanych im nazw w kwestionariuszu przy pytaniach, do których się odnoszą.
Dopuszczalne jest także przekazywanie dwóch zbiorów, z których każdy posiada nazwy zmiennych nadane wg. odmiennego klucza. Warunkiem jest jednak, że jest to jedyna różnica między tymi zbiorami.

6.2.3. Etykiety zmiennych (*variable labels*)

Większość programów statystycznych pozwala przypisywać do zmiennych ich etykiety. Są one bardzo ważne. Powinny dostarczać przynajmniej trzech informacji:

1. numer pytania z kwestionariusza (o ile sama nazwa zmiennej nie zawiera tej informacji),
2. jasną informację o zawartości zmiennej, treści pytania,
3. informację czy jest to zmienna pierwotna, czy wtórna. Jeśli liczba znaków w etykiecie jest ograniczona, dobrze jest zawczasu przygotować odpowiedni zestaw skrótów.

Wymogiem archiwizacyjnym jest nadanie etykiet wszystkim zmiennym występującym w zbiorze danych. W przypadku stosowania numeracji zmiennych nie zawierającej informacji bezpośrednio odnoszącej daną zmienną do pytania w kwestionariuszu, zalecane jest umieszczenie tej informacji w jej etykiecie.

6.2.4. Wartości zmiennych (*values*)

Zanim dane sondażowe będą mogły być analizowane, werbalne odpowiedzi ankietowanych należy zamienić na ich numeryczne reprezentacje. W tej sekcji wyjaśnimy konwencje kodowania wartości zmiennych tak aby zapewnić zgodność danych z każdym oprogramowaniem statystycznym, oraz zmaksymalizować porównywalność i użyteczność stosowanych miar.

1. Na początku rekordu zawsze należy zarezerwować miejsce dla wszelkich zmiennych służących identyfikacji przypadku. Najczęściej będzie to numer badania i numer respondenta. Są one niezbędne aby zapewnić kontrolę unikalności każdego z rekordów.
2. Kategorie kodowe powinny być wzajemnie rozłączne, wyczerpujące, i dokładnie zdefiniowane. Każda odpowiedź powinna posiadać tylko jedną wartość. Wszelkich kategoryzacji należy dokonywać już po przygotowaniu całego zbioru danych a nie w trakcie jego tworzenia.
3. Kodowanie powinno zachować tak dużo szczegółów odpowiedzi respondenta jak to jest możliwe. Należy pamiętać o tym, że pozostawienie takich zmiennych jak wiek czy dochód w postaci zmiennych ciągłych jest znacznie bardziej użyteczne niż ich agregowanie i tworzenie na ich podstawie skal przedziałowych.
   Pozostawienie maksymalnie uszczegółowionych lub nie agregowanych danych, umożliwi użytkownikowi użycie silniejszych procedur statystycznych i pozwoli na dostosowanie danych do własnych potrzeb.
4. Pytania zamknięte należy kodować zgodnie z numeracją odpowiedzi w kwestionariuszu. Pozwoli to uniknąć błędów i dezorientacji użytkownika.
5. W przypadku pytań otwartych, wyróżniamy dwa główne podejścia do kodowania. Pierwszym jest odgórne przygotowanie przez badacza klucza kodowego. Drugim - przestudiowanie pojawiających się odpowiedzi i tworzenie klucza kodowego na ich podstawie. Wybór metody zależy w tym przypadku od istoty badania i celów badacza.
6. Gdy w grę wchodzą zmienne klasyfikujące dane zjawisko należy stosować hierarchiczny zapis wartości. W tym wypadku pierwsza cyfra kodu oznacza największą kategorię klasyfikacyjną, kolejne zaś następne kategorie, coraz bardziej uszczegóławiające.
   Tego typu klasyfikacją jest np. *Międzynarodowa Klasyfikacja Zawodów i Specjalności ISCO-88*.

6.2.5. Etykiety wartości (*value labels*)

Etykiety wartości zmiennej są elementem ich rozróżniania bez konieczności odwoływania się za każdym razem do narzędzia badania. Większość programów statystycznych pozwala na ich przypisanie zmiennej. Etykiety wartości powinny dostarczać informacji na temat:

1. treści odpowiedzi, jaką dana wartość reprezentuje

2. znaczenia wartości, w przypadku kodów specjalnych

3. Zalecane jest stosowanie konwencji zapisywania etykiet wartości niedostępnych respondentowi (w tym kodów specjalnych) wielkimi literami, zaś tych, które są mu dostępne – małymi. W przypadku zastosowania jednolitej wielkości liter, koniecznym jest zamieszczenie w dokumentacji informacji na temat zastosowanej konwencji dostępności respondentowi poszczególnych odpowiedzi z kafeterii.

4. Etykiety wartości należy nadać wszystkim wartościom wszystkich zmiennych występujących w zbiorze. Każda wartość zmiennej musi być opisana, poprzez nadanie jej etykiety.
   W niektórych przypadkach, np. zmiennych zawierających klasyfikacje zawodowe, wartości poszczególnych kodów mogą zostać opisane w aneksie dołączanym do dokumentacji metodologicznej (książki kodów). Innym, dopuszczanym odstępstwem od wymogu opisu każdej wartości zmiennej jest przypadek zmiennych ciągłych o jednolitej jednostce kodowej (np. zarobki respondenta). W tym przypadku opisana winna być co najmniej najniższa i najwyższa wartość danej zmiennej wyrażonej w danej jednostce.

Archiwum będzie przyjmować do archiwizacji wyłącznie zbiory, w których etykiety zostały nadane wszystkim wartościom danej zmiennej wraz z kodami specjalnymi. Jedynym dopuszczalnym odstępstwem od tej reguły jest opisanie znaczenia poszczególnych wartości w Aneksie dołączanym do dokumentacji metodologicznej oraz opisanie jedynie najniższej i najwyższej wartości skali zmiennej ciągłej.

6.2.6. Braki danych i kody specjalne (*missing values*)

1. Zdecydowanie nie zaleca się stosowania nadawania wartości ,,0'' lub wartości z wewnątrz przedziału zmienności zmiennej jakiemukolwiek z kodów specjalnych.

2. Niedopuszczalne jest pozostawianie w zbiorze danych pustych komórek, bez przypisania im kodów specjalnych (*system missing*).

3. Braki danych oraz kody specjalne powinny być kodowane jako wartości liczbowe z poza przedziału zmienności danej zmiennej.

Wyróżnia się co najmniej pięć typów braków danych, każdy z nich powinien posiadać odrębny kod.

1. BRAK DANYCH / ODMOWA ODPOWIEDZI. Respondent odmówił odpowiedzi na pytanie. Brak odpowiedzi na pytanie w sytuacji, której dane pytanie nie zostało wyłączone ze względu na wcześniej zadane pytanie filtrujące bądź instrukcję zamieszczoną przed pytaniem.

2. NIE WIEM. Respondent nie potrafił odpowiedzieć na pytanie, nie miał opinii w danej kwestii.

3. BŁĄD PRZETWARZANIA DANYCH. Zdarza się, że w zbiorze danych brakuje wartości, chociaż respondent takiej udzielił. Może to być błąd ankietera, kodowania, wczytywania, itp.

4. NIE DOTYCZY. Występują wtedy, gdy respondent nie odpowiadał na pytanie ze względu na strukturę wywiadu i kwestionariusza. Brak danych jest w tym przypadku poprawnie zakodowaną wartością jako konsekwencja struktury kwestionariusza.

5. NIE ZGODNE. Tego typu błędy mogą powstają gdy dane pochodzą z różnych źródeł i informacja z jednego z nich nie może zostać zidentyfikowana.

Nie będą przyjmowane do archiwizacji zbiory danych zawierające systemowe braki danych (SYSMIS – *system missings*) oraz nieopisane kody specjalne. Każda wartość, również kodu specjalnego, przed przekazaniem zbioru Archiwum musi zostać opisana.

6.2.6.1 Wybór kodów specjalnych

Większość badaczy stosuję zasadę ustawiania kodów specjalnych poza prawym marginesem zmienności zmiennej. Może to być kłopotliwe gdyż czasami powoduje konieczność zwiększania liczby znaków potrzebnych do zakodowania danej zmiennej. Podobny problem pojawia się gdy stosuje się kody ujemne gdyż trzeba przeznaczyć miejsce na znak ,,minus".

Archiwum zaleca stosowanie następującej notacji dla oznaczania braków danych w zbiorach:

KOD*	Opis kodu
-2	*NIE DOTYCZY* ze względu na nie zadanie danego pytania w danym badaniu. Np. w sytuacji, w której w badaniu corocznie powtarzanym na różnych grupach respondentów lub badaniu panelowym nie zadano danego pytania w danym roku badania lub danej fali badania.
-1	*NIE DOTYCZY* ze względu na wcześniejsze pytanie filtrujące lub instrukcję umieszczoną przed pytaniem. Np.: w pytaniu o zarobki z pracy w przypadku osób bezrobotnych należy przypisać wartość –1. Stosowane w przypadku wyłączania grup respondentów z odpowiedzi na dane pytanie.
9x9	*BRAK DANYCH* ze względu na nie udzielenie przez respondenta odpowiedzi na pytanie, na które zgodnie ze strukturą kwestionariusza odpowiedzi winien był udzielić. Stosowany także na oznaczenie pomyłkowego nie zadania danego pytania przez ankietera.
9x8	odpowiedź typu *nie wiem* gdy respondent nie potrafił odpowiedzieć na dane pytanie, nie miał zdania w badanej kwestii.

* gdzie x oznacza cyfrę 9 lub jej wielokrotność w zależności od ogólnej liczby cyfr wykorzystanych do opisania wartości znaczących.
Np. w przypadku pytania o zarobki, najwyższą podaną przez respondentów kwotą było 25000 PLN. W tym przypadku, kod 9, 99, 999, 9999 są wartościami umieszczonymi w środku skali a zatem brakowi danych należy przypisać kolejny, wolny kod spoza niej, czyli 99999, zaś odpowiedzi typu *NIE WIEM* – 99998.

Stosowanie takich właśnie kodów specjalnych jest przez Archiwum zalecane aczkolwiek nie stanowi wymogu decydującego o przyjęciu bądź odrzuceniu zbioru do archiwizacji, pod warunkiem jednak spełniania wymogów opisanych w pkt. 6.2.6. Podręcznika.

7. Zarządzanie danymi

Gotowy zbiór danych powinien być przechowywany na niezapisywalnym nośniku. Nie więcej niż jedna, bądź dwie osoby w zespole powinny mieć prawo i być odpowiedzialne za wprowadzanie do gotowego zbioru danych zmian. Wersje zbioru danych powinny być w przejrzysty sposób numerowane. Nie powinno budzić wątpliwości która z dostępnych wersji jest starsza. Dokumentacja powinna odnosić się do danej wersji zbioru danych.

Do kontaktów z Archiwum w sprawach związanych z deponowaniem zbioru danych winna zostać wyznaczona jedna lub najwyżej dwie osoby z zespołu realizującego dane badanie. Pracownicy Archiwum wyłącznie za pośrednictwem tej osoby/osób będą kontaktować się z zespołem realizującym badanie w celach uzupełnienia bądź poprawienia informacji zawartych np. w Formularzu deponowania danych bądź dokumentacji techniczno-metodologicznej.

8. Kopie zapasowe

Każdy zbiór danych powinien być, niezależnie od kopii przechowywanej w Archiwum, archiwizowany przez zespół realizujący badanie. Najlepszym rozwiązaniem jest archiwizacja na płytach CD. Zawsze powinny istnieć co najmniej dwie kopie zapasowe gotowego zbioru danych oraz dokumentacji techniczno-metodologicznej, przechowywane przez zespół realizujący badanie. Nie należy przechowywać danych wyłącznie na dyskach komputerów, w szczególności jeśli są one podłączone do sieci komputerowej.
Ze swej strony Archiwum zobowiązuje się do niezależnego przechowywania i udostępniania zbiorów wg. zasad określonych przez deponenta w Formularzu deponowania danych oraz Formularzu rejestracyjnym zbioru a także do przechowywania ich kopii zapasowych.

9. Anonimizacja

Zgodnie z obowiązującym prawem (por. Ustawa o ochronie danych osobowych z dnia 29 sierpnia 1997 r., Dz. U. z dnia 29 października 1997 r.) zbiory danych zawierające informacje mogące być podstawą identyfikacji respondenta wymagają rejestracji w Generalnym Inspektoracie Ochrony Danych Osobowych.

Przed przekazaniem zbiorów danych do archiwizacji deponenci zobowiązują się do anonimizacji zbiorów celem uniemożliwienia identyfikacji respondenta, o ile zachodzi taka potrzeba i podpisują oświadczenie o jego dokonaniu.

Wyłączną odpowiedzialność za ewentualne konsekwencje wynikłe z nienależycie przeprowadzonego przez zespół realizujący badanie procesu anonimizacji ponosi osoba deponująca dany zbiór w Archiwum.

10. Sumaryczne zestawienie wymagań stawianych deponentom

10.1. Zbiór danych

Wymagane	Zalecane
Określenie w nazwie zmiennej bądź jej etykiecie jednoznacznego odniesienia do odpowiadającego jej numeru pytania w kwestionariuszu. W przypadku stosowania innej konwencji nazewnictwa zmiennych niezbędnym jest naniesienie w kwestionariuszu informacji jednoznacznie wiążącej dane pytanie ze zmienną w zbiorze (pkt. 6.2.2.).	Przyjęcie konwencji nazewnictwa zmiennych wg numerów pytań w kwestionariuszu (pkt. 6.2.2.2.).
Nadanie etykiet wszystkim zmiennym występującym w zbiorze (pkt. 6.2.3.).	
Nadanie etykiet wartości wszystkim wartościom wszystkich zmiennych (w tym kodom specjalnym) występującym w zbiorze. Wyjątki od tej zasady zostały opisane w pkt. 6.2.5.4. Podręcznika.	Zapisanie etykiet wartości odpowiedzi standardowo niedostępnych respondentowi - wielkimi literami, w odróżnieniu od zapisywanych małymi literami - odpowiedzi mu dostępnych (pkt. .6.2.5.3.)
Zastąpienie wszystkich systemowych braków danych (*system missings*) kodami specjalnymi wraz z opisem ich znaczenia (pkt. 6.2.6.).	Dla oznaczania wartości kodów specjalnych zalecane jest stosowanie wartości spoza skali zmienności danej zmiennej, zgodnie ze wskazówkami zawartymi w pkt. 6.2.6.1. Podręcznika.
Zapis przygotowanego do archiwizacji zbioru danych do pliku systemowego SPSS lub pliku typu *portable* lub pliku ASCII (pod warunkiem dołączenia do niego słownika definiującego dane - *data definition statements*) (pkt. 6.1. Podręcznika).	Przesłanie zarówno systemowego zbioru danych w formacie SPSS bądź zbioru typu *portable* jak i pliku ASCII wraz ze słownikiem definiującym dane (*data definition statements*) (pkt. 6.1. Podręcznika)
Przeprowadzenie procesu anonimizacji zbioru celem usunięcia wszystkich informacji mogących stać się podstawą identyfikacji respondenta (rozdz. 9)	

10.2. Dokumentacja

Wymagane	Zalecane
Informacja na temat kierownika badania i jego afiliacji w czasie wykonywania badań oraz informacje na temat instytucji realizującej badanie.	
Oficjalny tytuł badań.	
Wzór cytowania danych.	
Wskazanie źródeł finansowania (w tym nr grantu).	
Informacje na temat składu osobowego zespołu przetwarzającego dane.	
Opis projektu badań.	
Informacja na temat technik użytych do zbierania danych.	
Informacja na temat tego, co było jednostką badania.	
Opis zastosowanej w badaniu próby i procedury jej doboru	
Dołączenie wszystkich narzędzi, które posłużyły do zbierania danych	
Załączenie rozkładów jednej zmiennej dla wszystkich zmiennych występujących w zbiorze.	
Powyższe informacje powinny być ujęte w postaci jednego dokumentu zapisanego w formacie ASCII lub PDF. Dokładny opis wymagań, jakie musi spełniać książka kodów zawarty został w pkt. 5.1. Podręcznika.	
Niezależnie od rozkładów zamieszczonych w książce kodów (*codebook*), przygotowanie zestawienia nieważonych rozkładów jednej zmiennej (*frequencies*)dla wszystkich zmiennych występujących w zbiorze, zapisane w formacie ASCII (pkt. 5.2. Podręcznika).	
Zestawienie statystyk opisowych (*descriptive statistics*) dla wszystkich zmiennych występujących w zbiorze, przygotowanych zgodnie z instrukcją zawartą w pkt. 5.3. Podręcznika i zapisanych w formacie ASCII.	

11. Przekazywanie danych do Archiwum

11.1 Informacje ogólne

Proces deponowania danych składa się z trzech części.

1. Wypełnienie Formularza rejestracyjnego i przesłanie go na adres:

   Archiwum Danych Społecznych
   Instytut Studiów Społecznych UW
   ul. Stawki 5/7
   00-183 Warszawa

2. Wypełnienie formularza deponowania danych
3. Przesłanie zbioru danych oraz dokumentacji do Archiwum.
   a. Deponowanie danych zarchiwizowanych na płycie CD
   b. Przesyłanie danych z wykorzystaniem protokołu FTP

Państwa dane mogą zostać zdeponowane w Archiwum poprzez przesłanie ich do Archiwum bezpośrednio - z wykorzystaniem protokołu **FTP** lub też mogą zostać dostarczone na fizycznym nośniku danych (ze względu na wysoką niezawodność dane będą tą drogą przyjmowane **WYŁĄCZNIE** na płycie **CD**).

Niezależnie od tego, którą z możliwości dostarczenia zbiorów Archiwum Państwo wybiorą, **WSZYSTKIE WYSYŁANE PLIKI, MUSZĄ ZOSTAĆ UPRZEDNIO SKOMPRESOWANE DO JEDNEGO PLIKU ARCHIWIZACYJNEGO (np.: ZIP)**.

11.2. Deponowanie danych zarchiwizowanych na płycie CD

1. Wypełniony Formularz rejestracyjny zbioru:
   - http://www.ads.org.pl/pdf/FORM02.pdf
2. Wypełniony formularz deponowania danych:
   - http://www.ads.org.pl/pdf/FORM04.pdf
3. oraz płytę z nagranym plikiem archiwizacyjnym zawierającym zbiory i dokumentację

należy przesłać na adres:

Archiwum Danych Społecznych
Instytut Studiów Społecznych UW
ul. Stawki 5/7
00-183 Warszawa

WSZYSTKIE WYSYŁANE PLIKI, MUSZĄ ZOSTAĆ UPRZEDNIO SKOMPRESOWANE DO JEDNEGO PLIKU ARCHIWIZACYJNEGO (np.: ZIP).

W trosce o bezpieczeństwo Państwa danych zalecane jest przesłanie płyty z danymi przesyłką poleconą.

11.3. Przesyłanie danych z wykorzystaniem protokołu FTP

1. Wypełniony Formularz rejestracyjny zbioru:
   - http://www.ads.org.pl/pdf/FORM02.pdf

należy przesłać na adres:

Archiwum Danych Społecznych
Instytut Studiów Społecznych UW
ul. Stawki 5/7
00-183 Warszawa

2. Wypełnić formularz deponowania danych:
   - http://www.ads.org.pl/FORM04.php
3. Przesłać skompresowany plik archiwizacyjny na serwer ADS.

KONIECZNYM JEST UMIESZCZENIE WSZYSTKICH WYSYŁANYCH PLIKÓW W JEDNYM SKOMPRESOWANYM PLIKU ARCHIWIZACYJNYM (np.: ZIP).

Metoda kompresji pliku archiwizacyjnego jest dowolna, zależna jedynie od preferencji wysyłającego. Mogą to być formaty *.zip, .rar, .arj,...* i inne

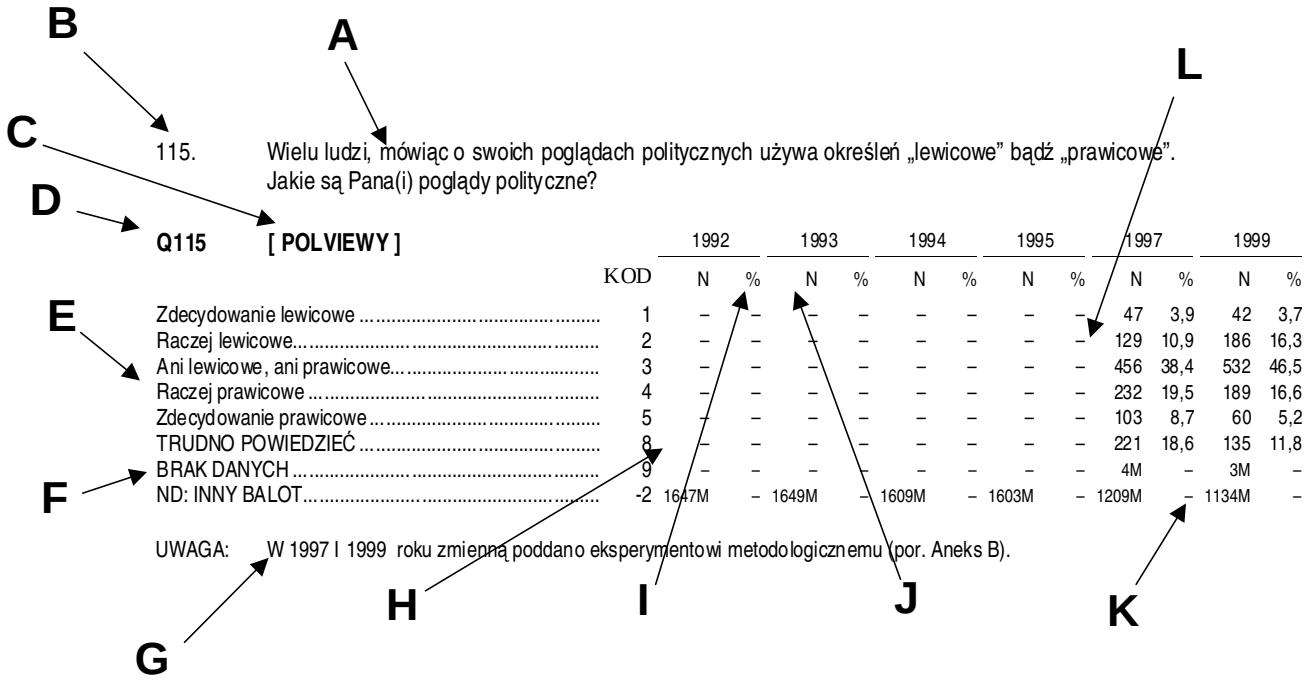
11.3.1. Konfiguracja programu FTP do pracy z serwerem ADS

Osoby/instytucje nie posiadające zawartej z ADS Umowy o współpracy proszone są o kontakt pod adresem: Marcin.Zielinski@uw.edu.pl w celu uzyskania informacji na temat loginu i hasła dostępu niezbędnych do przesłania zbioru i dokumentacji za pomocą protokołu FTP.

Większość programów służących do komunikacji za pomocą protokołu FTP ma zbliżoną zasadę działania. Do ich konfiguracji niezbędne są następujące informacje:

1. Host name (nazwa hosta): **samba.iss.uw.edu.pl**
2. Port number(numer portu): **22**
3. User name (login): tu należy wpisać swój login
4. Password (hasło): tu należy wpisać hasło dostępu

Aneks A.
Przykładowe zestawienie tabelaryczne jednej zmiennej dla badania powtarzanego (książka kodów - codebook)



	KOD	1992 N	1992 %	1993 N	1993 %	1994 N	1994 %	1995 N	1995 %	1997 N	1997 %	1999 N	1999 %
Zdecydowanie lewicowe	1	–	–	–	–	–	–	–	–	47	3,9	42	3,7
Raczej lewicowe	2	–	–	–	–	–	–	–	–	129	10,9	186	16,3
Ani lewicowe, ani prawicowe	3	–	–	–	–	–	–	–	–	456	38,4	532	46,5
Raczej prawicowe	4	–	–	–	–	–	–	–	–	232	19,5	189	16,6
Zdecydowanie prawicowe	5	–	–	–	–	–	–	–	–	103	8,7	60	5,2
TRUDNO POWIEDZIEĆ	8	–	–	–	–	–	–	–	–	221	18,6	135	11,8
BRAK DANYCH	9	–	–	–	–	–	–	–	–	4M	–	3M	–
ND: INNY BALOT	-2	1647M	–	1649M	–	1609M	–	1603M	–	1209M	–	1134M	–

UWAGA: W 1997 I 1999 roku zmienną poddano eksperymentowi metodologicznemu (por. Aneks B).

A Pełna treść pytania (identyczna z kwestionariuszem)
B Nr pytania w kwestionariuszu
C Nazwa mnemoniczna zmiennej (jeśli zbiór został przygotowany w dwóch wersjach z dwoma rodzajami nazw - jeśli nie występuje należy pominąc ten element)
D Nazwa zmiennej w zbiorze danych
E Treść odpowiedzi odczytywanej respondentowi (małe litery)
F Treść odpowiedzi nieodczytywanej respondentowi oraz oznaczenia kodów specjalnych (wielkie litery)
G Miejsce na dodatkowe uwagi (jeśli nie występują należy zostawić puste miejsce)
H Wartości zmiennej
I Kolumna z określeniem liczebności
J Kolumna z określeniem odsetek liczonych bez uwzględniania kodów typu *BRAK DANYCH* i *NIE DOTYCZY*
K Liczebności dla wartości typu brak danych i nie dotyczy (oznaczane dodatkową literą M)
L Liczebności i odsetki są podawane w postaci kresek w sytuacji, gdy dane pytanie nie było w ogóle zadane oraz gdy dana wartość zmiennej jest wyłączona z podstawy procentowania (odpowiedzi typu *BRAK DANYCH* i *NIE DOTYCZY*)

Aneks B.
Przykładowy słownik definiujący dane (*data definition statements*) dla zbioru danych zapisanego w formacie ASCII (dla pierwszych 5 zmiennych badania PGSS)

1. Informacje na temat lokalizacji zmiennych (nr kolumny) wraz z jej nazwą

```
FILE HANDLE PGSS9299
/NAME = 'p9299.dat'                  -> apostrofem ujęta nazwa pliku ASCII
/LRECL=2799.                         -> numer ostatniej kolumny w pliku ASCII
DATA LIST
  FILE=PGSS9299 RECORD=1/
    RECORDID          1-5            -> zmienna zapisana w kolumnach
    PGSSYEAR          6-9              od 1 do 5
    WEIGHT           10-31 (16)      -> zmienna zajmuje kolumny od 10 do 31.
    BALLOT           32-33             Kolumny 16 do 31 zawierają cyfry po
    PGSS_BAL         34-41             przecinku
.                                    -> opis lokalizacji wszystkich
                                        zmiennych należy zakończyć kropką.
```

2. Definicja etykiety zmiennej

```
VARIABLE LABELS                              -> komenda definiująca etykiety
    RECORDID      'NUMER KOLEJNY REKORDU'       zmiennych
    PGSSYEAR      'ROK BADANIA PGSS'
    WEIGHT        'WAGA ANALITYCZNA'         -> po nazwie zmiennej
    BALLOT        'PODPROBY LOSOWE: A/B'        apostrofami została ujęta jej
    PGSS_BAL      'PODPROBY LOSOWE: A/B/ROK'    etykieta
.                                             -> opis etykiet wszystkich
                                                 zmiennych należy
                                                 zakończyć kropką
```

3. Opis etykiet wszystkich wartości zmiennych

```
VALUE LABELS                        -> komenda definiująca etykiety wartości
    PGSSYEAR                        -> nazwa zmiennej
           1992    '1992r'          -> po wartości zmiennej apostrofami ujęto
           1993    '1993r'             etykietę danej wartości
           1994    '1994r'
           1995    '1995r'
           1997    '1997/98r'
           1999    '1999r'/         -> przejście do opisu kolejnej zmiennej
    BALLOT                             sygnalizowane jest znakiem / (slash)
             -2    'NIE BYLO BALOTU'
              1    'BALOT A'
              2    'BALOT B'/
    PGSS_BAL
           1992    'PGSS1992'
           1993    'PGSS1993'
           1994    'PGSS1994'
           1995    'PGSS1995'
          19971    'PGSS97-A'
          19972    'PGSS97-B'
          19991    'PGSS99-A'
          19992    'PGSS99-B'
.                                   -> nadanie etykiet wszystkim wartościom
                                       wszystkich zmiennych należy
                                       zakończyć kropką. Po opisie ostatniej
                                       zmiennej nie należy wstawiać znaku /
                                       (slash).
```

4. Definicje wartości typu *BRAK DANYCH* i *NIE DOTYCZY*

```
MISSING VALUE           -> komenda definiująca wartości wyłączane z analizy
   BALLOT    ( -2,    ) -> po nazwie zmiennej w nawiasach ujęto wartości
.                           wyłączane z analizy, oddzielając je przecinkami
EXECUTE.
```

Podobnie, jak w przypadku poprzednich poleceń języka SPSS, definiowanie wartości zmiennych wyłączanych z analizy należy zakończyć kropką.
Komendą nakazującą wykonanie wszystkich, w ten sposób opisanych, poleceń jest EXECUTE.

Wszystkie elementy słownika definiującego dane (data definition statements) należy zapisać w postaci jednego pliku w formacie ASCII w kolejności odpowiadającej powyższemu opisowi.

Pełnym, wzorcowym słownikiem definiującym dane jest dostępny w bazie ADS słownik dla badania Polskie Generalne Sondaże Społeczne 1992-1999.
Podczas tworzenia słownika zalecane jest także korzystanie z wbudowanego systemu pomocy dla pakietu SPSS oraz dokumentacji dołączanej do programu.