Zbiór danych ilościowych:

1 Na każdą "bazę danych" składa się zanonimizowany zbiór danych ilościowych zebranych w badaniu oraz opis jego struktury (codebook).

2 Zarówno zanonimizowany zbiór danych ilościowych, jak i opis jego struktury powinny mieć format csv:

2.1 separatorem pól powinien być średnik (kod ASCII 59);

2.2 separatorem tekstu powinien być cudzysłów (kod ASCII 34);

2.3 cudzysłów w tekście oznacza się przez następujące po sobie dwa znaki cudzysłowów (kod ASCII 34) – w takim kontekście są one interpretowane jako pojedynczy cudzysłów, a nie dwa znaki separatora tekstu;

2.4 znaki narodowe powinny być zakodowane w stronie kodowej CP-1250.

3 Opis struktury zanonimizowanego zbioru danych ilościowych (codebook) powinien zawierać:

3.1 W pierwszym wierszu nagłówki kolumn (patrz punkt 3. poniżej).

3.2 W kolejnych wierszach opis poszczególnych zmiennych zanonimizowanego zbioru danych ilościowych (patrz opis kolumn)

3.2.1 N-ty wiersz (licząc z pominięciem wiersza zawierającego nagłówki kolumn) musi opisywać N-tą zmienną zanonimizowanego zbioru danych ilościowych (a więc zmienną, której wartości znajdują się w N-tej kolumnie zanonimizowanego zbioru danych ilościowych).

3.3 W kolejnych kolumnach opis cech poszczególnych zmiennych zanonimizowanego zbioru danych ilościowych (dla każdej zmiennej muszą zostać podane wszystkie stosowne dla danego typu zmiennej informacje):

3.3.1 nazwa zmiennej;

3.3.1.1 musi być unikalna w obrębie zbioru danych;

3.3.1.2 powinna oddawać hierarchię zmiennych, tzn. składać się z kodów sekcji/kategorii/grup zmiennych rozdzielonych znakiem podkreślenia. Np.: M_12_a – popunkt a pytania 12 w sekcji M ankiety;

3.3.1.2.1 każdy kod sekcji/kategorii/grupy zmiennych, który zostanie użyty w nazwie jakiejkolwiek zmiennej musi posiadać w opisie struktury zanonimizowanego zbioru danych osobowych zmienną typu kategoria, która opisuje daną sekcję/kategorię/grupę, np. jeśli zbiór zawiera zmienne S_1, S_2, M_1, M_2, to w opisie struktury zanomizowanego zbioru danych osobowych muszą pojawić się definicje zmiennych typu kategoria o nazwach S i M, których opis zawierać będzie nazwę danej sekcji/kategorii/grupy zmiennych, np. M – metryczka, S – opinie o sąsiadach (patrz też 3.3.5.1).

3.3.1.3 dla danych pochodzących z ankiet kwestionariuszowych nazwa zmiennej powinna odpowiadać numerowi pytania w kwestionariuszu ankiety; w wypadku gdy jednemu pytaniu kwestionariusza odpowiada kilka zmiennych (np. pytanie z możliwością wielokrotnego wyboru), nazwy zmiennych odpowiadających danemu pytaniu powinny się składać z numeru pytania oraz sufiksu _N, gdzie N to kolejne liczby naturalne;

3.3.1.4 nie może zawierać białych znaków (w szczególności spacji);

3.3.2 krótki opis zmiennej, zwięźle (w maksymalnie 70 znakach) opisujący czego dana zmienna dotyczy;

3.3.3 opis zmiennej, w jasny sposób definiujący, czego dana zmienna dotyczy;

3.3.3.1 dla danych pochodzących z ankiet kwestionariuszowych powinna ona zawierać pytanie odczytywane respondentowi w ankiecie odpowiadające danej zmiennej;

3.3.3.2 jeśli jest to zmienna wywiedziona z innych zmiennych, w tym polu musi zostać opisany sposób, w jaki została wywiedziona;

3.3.4 lista słów kluczowych powiązanych ze znaczeniem zmiennej. Poszczególne wyrażenia oddzielone od siebie znakiem przecinka;

3.3.5 typ zmiennej: liczba całkowita/liczba rzeczywista/tekst/data/TERC / kategoria /waga ;

3.3.5.1 kategoria to specjalny typ, który opisuje jedynie strukturę narzędzia badawczego (np. kwestionariusza), a nie informację gromadzoną od respondenta, np.:

załóżmy, że kwestionariusz składa się z następujących bloków pytań: M – pytania metryczkowe, A – Pytania o społeczność lokalną, B – Pytania o kapitał społeczny – przełożyłoby się to na trzy zmienne typu kategoria:

- zmienna o nazwie M z opisem Metryczka;

- zmienna o nazwie A z opisem Pytania o społeczność lokalną;

- zmienna o nazwie B z opisem Pytania o kapitał społeczny.

3.3.5.2 typu waga należy użyć, gdy dana zmienna przechowuje wagę obserwacji;

3.3.5.3 typ TERC odpowiada kodowi GUS jednostki terytorialnej, w której leżą poszczególne obserwacje (jeśli dane takie są gromadzone w badaniu);

3.3.6 skala zmiennej: dychotomiczna/nominalna/porządkowa/interwałowa/ilorazowa;

3.3.7 rozmiar zmiennej:

3.3.7.1 dla liczb i wag – maksymalna liczba cyfr, z jakich składa się liczba (w wypadku liczb rzeczywistych łącznie cyfr dziesiętnych i ułamkowych);

3.3.7.2 dla tekstów – maksymalna liczba znaków;

3.3.7.3 dla zmiennej typu TERC – 2, 4 lub 6 w zależności czy przechowywane będą identyfikatory, odpowiednio, województw, powiatów czy gmin;

3.3.7.4 dla dat – nie dotyczy (data ma stały rozmiar 19 znaków)

3.3.8 dokładność zmiennej

3.3.8.1 dla zmiennych rzeczywistych – liczba cyfr ułamkowych w reprezentacji liczby (pusta, jeśli nie dotyczy);

3.3.8.2 dla zmiennych typu TERC – rok, z którego pochodzą kody województw/powiatów/gmin;

3.3.9 wartość minimalna – tylko dla liczb i dat – minimalna poprawna wartość zmiennej (pusta, jeśli nie dotyczy);

3.3.10 wartość maksymalna – tylko dla liczb i dat – maksymalna poprawna wartość zmiennej (pusta, jeśli nie dotyczy);

3.3.11 etykiety wartości – tylko dla zmiennych tekstowych zakodowanych liczbowo – tekst, który w kolejnych liniach zawiera pary kod:etykieta, gdzie kod to kod liczbowy przyporządkowany danej etykiecie (pusta, jeśli nie dotyczy).

3.3.11.1 dla danych pochodzących z ankiet kwestionariuszowych:

3.3.11.1.1 etykiety wartości powinny być takie same, jak teksty odpowiadających im punktów kafeterii danego pytania kwestionariusza;

3.3.11.1.2 opis etykiet wartości musi pokrywać całą kafeterię danego pytania kwestionariusza;

3.3.12 warunek posiadania przez zmienną wartości – tylko dla zmiennych podlegających filtrom – zdanie logiczne opisujące to, kiedy zmienna powinna mieć wartość, o strukturze:

3.3.12.1 w każdym wierszu po jednym warunku logicznym;

3.3.12.2 każdy warunek logiczny postaci: [i/lub] nazwaZmiennej operator wartość, gdzie:

3.3.12.2.1 i/lub na początku warunku oznacza sposób łączenia się danego warunku z warunkami w poprzednich wierszach, przy czym:

3.3.12.2.1.1 obowiązuje matematyczny priorytet warunków i/lub;

3.3.12.2.1.2 człon ten nie jest obecny w pierwszym warunku logicznym;

3.3.12.2.2 dostępne operatory to: =, !=, <, >, <=, >= (równe, różne od, mniejsze, większe, mniejsze bądź równe, większe bądź równe);

np.

A1 = 3

i B2 > 5

lub B3 = 9

oznacza warunek: (jeśli na pytanie A1 udzielono odpowiedzi o kodzie 3 oraz na pytanie B2 odpowiedzi o kodzie większym od 5) lub na pytanie B3 udzielono odpowiedzi o kodzie 9

4 Opis struktury zanonimizowanego zbioru danych ilościowych (codebook) powinien opisywać wszystkie zmienne znajdujące się w zanonimizowanym zbiorze danych ilościowych.

5 Zanonimizowany zbiór danych ilościowych powinien zawierać:

5.1 W pierwszym wierszu nazwy zmiennych przechowywanych w poszczególnych kolumnach zbioru (patrz opis struktury zanonimizowanego zbioru danych ilościowych).

5.2 W kolejnych wierszach wartości poszczególnych zmiennych dla kolejnych jednostek obserwacji.

5.2.1 Jeśli dana zmienna jest typu kategoria, to powinna ona mieć pustą wartość dla wszystkich jednostek obserwacji (patrz 3.3.5.1).

5.2.2 Jeśli dana zmienna jest typu data, powinna mieć format: RRRR-MM-DD GG:MM:SS, gdzie RRRR to czterocyfrowy zapis lat, MM to dwucyfrowy zapis numeru miesiąca, DD to dwucyfrowy zapis numeru dnia w miesiącu, GG to dwucyfrowy zapis godzin w formacie 24-godzinnym, MM to dwucyfrowy zapis minut, a SS to dwucyfrowy zapis sekund, np. 2010-02-04 17:06:48, co oznacza drugi lutego 2010 r., sześć minut i 48 sekund po godzinie siedemnastej. Jeśli data pozyskiwana była z mniejszą dokładnością, zapis należy uzupełnić zerami, np. jeśli data pozyskiwana była z dokładnością do dni, zapis powinien mieć postać RRRR-MM-DD 00:00:00.

5.2.3 Jeśli dana zmienna nie ma wartości dla danej jednostki obserwacji z uwagi na to, że nie była ona w ogóle gromadzona dla danej jednostki obserwacji (np. wskutek filtru na pytaniu kwestionariusza lub tego, że wykonywany przez danego ucznia zeszyt ćwiczeń nie zawierał danego pytania), powinna zostać jej przypisana wartość:

5.2.3.1 dla zmiennych liczbowych, wag, zmiennych typu teryt: maksymalna wartość zmiennej o danym rozmiarze - 2, np. dla zmiennej o rozmiarze 3 cyfr 997;

5.2.3.2 dla zmiennych będących datami: 9997-00-00 00:00:00;

5.2.3.3 dla zmiennych tekstowych: ciąg znaków NIE DOTYCZY;

5.2.4 Jeśli dana zmienna nie ma wartości dla danej jednostki obserwacji z uwagi na odmowę lub nieudzielenie odpowiedzi przez respondenta / ucznia (także gdy uczeń w teście zamkniętym nie zaznaczył żadnej odpowiedzi), a także omyłkowe niezadanie pytania wywiadu kwestionariuszowego przez ankietera:

5.2.4.1 dla zmiennych liczbowych, wag, zmiennych typu teryt: maksymalna wartość zmiennej o danym rozmiarze, np. dla zmiennej o rozmiarze 3 cyfr 999;

5.2.4.2 dla zmiennych będących datami: 9999-00-00 00:00:00;

5.2.4.3 dla zmiennych tekstowych: ciąg znaków ODMOWA ODPOWIEDZI;

5.2.5 Jeśli dana zmienna nie ma wartości dla danej jednostki obserwacji z uwagi na to, że respondent / uczeń nie potrafił udzielić odpowiedzi lub w teście zamkniętym zaznaczył więcej niż jedną odpowiedź:

5.2.5.1 dla zmiennych liczbowych, wag, zmiennych typu teryt: maksymalna wartość zmiennej o danym rozmiarze - 1, np. dla zmiennej o rozmiarze 3 cyfr 998;

5.2.5.2 dla zmiennych będących datami: 9998-00-00 00:00:00;

5.2.5.3 dla zmiennych tekstowych: ciąg znaków TRUDNO POWIEDZIEĆ;

5.2.6 Długości zmiennych liczbowych muszą być dobrane w ten sposób, by opisane powyżej wartości specjalne znajdowały się poza dopuszczalnym zakresem „zwyczajnych” wartości zmiennych.

6 Zanonimizowany zbiór danych osobowych:

6.1 musi zawierać zmienną stanowiącą unikalny identyfikator obserwacji w zbiorze;

6.2 nie może zawierać zmiennych stanowiących dane osobowe (np. imiona, nazwiska, adres zamieszkania, telefon, itp.);

6.3 musi być w pełni zgodny z opisem swojej struktury.

7 Standardy kodowania wybranych zmiennych

7.1 Jeśli zanonimizowany zbiór danych ilościowych zawiera zmienną określającą gminę, powinna ona być kodowana z użyciem sześciocyfrowego kodu TERC. W opisie struktury zbioru danych taka zmienna powinna mieć typ TERC.

7.2 Jeśli zanonimizowany zbiór danych ilościowych zawiera zmienną identyfikującą podmioty posiadające numer REGON, to powinna ona zawierać numery REGON tych podmiotów.

7.3 TU BY MOŻNA JESZCZE POMYŚLEĆ, ŻEBY MIEĆ JEDNAKOWY SPOSÓB KODOWANIA TYCH SAMYCH DANYCH W RÓŻNYCH BADANIACH, NIESTETY NP. Z UCZNIAMI JEST KŁOPOT – ROZSĄDNY ZDAJE SIĘ BYĆ TYLKO PESEL (KTÓRY STANOWI DANĄ OSOBOWĄ, WIĘC POWINIEN BYĆ W ODDZIELNYM PLIKU – PATRZ PONIŻEJ)

8 Zgodność zanonimizowanego zbioru danych ilościowych oraz opisu jego struktury z OPZ potwierdza pisemnie Zamawiający. W wypadku niezgodności z OPZ Zamawiający przekaże na piśmie opis zauważonych niezgodności, uniemożliwiających przyjęcie zbioru danych i/lub opisu jego struktury.

*Autorem opisu i twórcą standardu jest Mateusz Żółtak.*