THE DESCRIPTION OP STATISTICAL TABLES: A PROBLEM IN DATA DOCUMENTATION

Dagobert Soergel and Karl Furmaniak
DATUM e.V. Bonn-Bad
Godesberg, Germany

Abstract

Reference to and direct storage of primary data is of growing importance. This explains the interest in the description of statistical tables. The brute-force approach to give each stratifier (that is, the list of all row captions or the list of all column captions) in full has its economical limits. A method is described which makes the description of a large number of large tables economically feasible. The basic idea is to replace the full listing of a stratifier (e.g. US counties) by a reference to a "prefabricated" stratifier available in a classification scheme (e.g. US geographical division) taking into account the hierarchical structure of such a scheme. Modification of prefabricated stratifiers is also possible in order to adapt to small variations occurring in actual tables. The method has also applications in thesaurus building.

0. Introduction

With the appearance and growing importance of material such as the US Census Summary Tapes, the need arises for the description of tables. This description must fulfill one or several of the following purposes:

(1) Describe precisely the contents of the table for the prospective user.

(2) Enable the retrieval of the table in response to a search request.

(3) Enable the inclusion of the data given in the table into a data bank.

The last point needs a short comment: We envision a data bank system which would store data in the form of tables. The system would consist of two parts: In part one, the table descriptions would be stored. In part two the data in the tables would be stored in a purely formal way as multi-dimensional arrays. If a certain data element is being searched for, one would first retrieve in part one the appropriate table and the appropriate cell of this table (as defined by row and column in the two-dimensional case). Using the table identification, the row-number, and the column-number one would then get from part two the data element searched for. This type of data bank organization is a basic feature of the SPAN-System (1) and also of the SEDAS-System (2).

In the following, the basic ideas of an economically feasible format for the description of tables are explained by means of examples. A formal and detailed description of the rules may be found in (3). The format to be explained forms a part of a larger scheme for the documentation of sets of primary or basic data (such as survey and polls material, census material and all kinds of administrative files). People interested in details are invited to contact the authors of this paper.

1. The Elements Needed for the Description of a Table

Pig. la gives a sample table, compare (4). Prom this we may see the elements needed to describe a table precisely; these are listed in fig. lb, left column. In the right column, this description format is applied to the sample table of fig. la.

Following the usage of SDC, see (5) , the term "stratifier" is used as a general term for both "stub" and "box-head". This terminology is particularly useful for the generalization to the more-than-two-dimensional case.

In fig. lc, the sample table is modified: a third stratifier "by year" is added. This brings us to a point which is often overlooked and which we have excluded up to now in order to simplify the discussion. ±n the table given in fig. la, there are involved actually three dimensions. But the third stratifier, namely "by year", consists of one element only, namely "1958". To say it the other way round: the two-dimensional table of fig. la is a subtable of the three-dimensional table of fig. lc, where the element "1958" in the third dimension is kept constant. Stratifiers consisting of one element only are called "qualifiers", but treated in exactly the same way as "normal" stratifiers. That is, a third stratifier is added to the table description in fig. lb. This point is of some importance for the organisation of retrieval, as may be seen from the example given: Somebody asking for data on employment "by state, by industry, by year" should retrieve the table given in fig. la and the corresponding tables for other years, since, in

Fig. la: Sample table

Title

Percent distribution by industry division of employees in non-agricultural establishments, by states: 1958

Boxhead

State	Mining and Manu-facturing	Electr., Gas and Water	Con-struction	Whole-sale and Retail Trade	Transport Storage, and Communic.	Finance, Insurance and Real Estate	Community Social and Personal Services
US				cell			
New England:							
Maine New	cell	cell	cell	cell	cell	cell	cell
Hampshire				cell			
Vermont				cell			
Massachusetts				cell			
Rhode Island				cell			
Connecticut							
Middle Atlantic:							
New York Mew	40.7	4.2	4.7	19.1	3.7		23.7
Jersey							
Pennsylvania							
North Central:							

Stub

Field

Fig. lb: Formal description of tables

format	example (compare fig.la)
Table number (as assigned in the system)	
Title of table	=T35
A description of the contents of the cells of the table (the meaning of the figures given in the table)	Percent distribution by industry division of employees in non-agricultural...
A listing of each stratifier (stub and boxhead)	Employees (percent, based on row sum)
Additional information (source of data; where the table is published; storage space needed in a data bank; etc.)	Stratifier 1 (stub): by state Stratifier 2 (boxhead): by industry div,.
	Statistical abstracts of the United States, 1960, table no.273, page 213
	+The listing of all the stratifier elements (e.g. states) has been omitted in this example for reasons of space only (comp. fig.2)

Fig. lc: Sample table, three "normal" stratifiers

Percent distribution by industry division of employees in non-agricultural establishments, by state, by year: 1957-1959

State	Mining and Manufacturing			Electricity, Gas and Water			Construction			Wholesale and Retail Trade			Transport, Storage and Communication			Finance, Insurance and Real Estate			Community, Social and Pers.Services		
	1957	1958	1959	1957	1958	1959	1957	1958	1959	1957	1958	1959	1957	1958	1959	1957	1958	1959	1957	1958	1959
US																					
N.E.																					
Maine																					

combination, they are equivalent to a three-dimensional table of the type given in fig. lc. (The term "qualifier" is taken from SDC (5); however, in the SDC-system qualifiers are not included into the stratifiers and are treated in another way.)

This is all very simple and obvious. However, economic problems arise if you intend to document a large number of tables, because you then have to record a lot of stratifiers, which may consist of numerous elements (to illustrate: think of a table giving data on all counties of the United States). In the following, we shall describe a way out of this difficulty.

The basic idea is as follows: There are a lot of "prefabricated" stratifiers contained in classification schemes such as the geographic division of the United States or the "International Standard Industrial Classification (ISIC)" (7). If a stratifier in a table to be documented corresponds to one of these prefabricated stratifiers, a single reference will save the writing down of large numbers of stratifier elements. If the correspondence is not 100%, but, say,only 95%, it suffices to give a single reference to the prefabricated stratifier together with an indication of the *5%* modifications. Only if there is no prefabricated stratifier which at least nearly matches the table stratifier, a full listing of the stratifier elements has to be given in the description of the table. However, if it is to be expected that the same stratifier occurs in further tables, too, it is added to the list of prefabricated stratifiers and thus available at each further occurence.

In order to present the realization of this idea in an orderly fashion, we first describe in which way a full listing of stratifier elements should be constructed (section2) ; we then proceed to the treatment of references to prefabricated stratifiers (where the existence of multiple hierarchical level will be the major problem) (section 3) ; finally, we deal with the methods of modification of prefabricated stratifiers (section 4).

2. The Presentation of Stratifiers by a Full Listing of their Elements (Fig.2)

3. The Method of Referencing to Pre fabricated Stratifiers

Fig. 3b shows an extract of the regional subdivision of the United States as used by the US Bureau of the Census (comp.(8) ; for ease of reference, we have added our own systematic notation). Now it is easy to see, that the stratifier described in fig. 2 by the full listing of its elements is embedded in this regional division. However, three levels of this division, namely regions, state economic areas (SEA's) and counties are wholly omitted and there are no data lines for the divisions in the table. To represent this information, we write the references to the prefabricated stratifier as shown in fig. 3a, which is self-explaining. Note that one single line replaces a listing of all the counties of the United States.

Fig.2: Presentation of a stratifier as a full listing of its elements

(stub of the table given in fig. la)

0	US
1*	New England
3	Maine
	New Hampshire
5	Vermont
6	Massachusetts
7	Rhode Island
8	Connecticut
9*	Middle Atlantic
10	New York
11	New Jersey
12	Pennsylvania
13	East North Central
14	Ohio

* No corresponding data line (e.g. no data on the region of New England).

Fig. 3a: References to prefabricated stratifiers: examples

Symbols used:

1 All elements of the hierarchical level are present and have a data line

♦ All elements of the hierarchical level are present but only as headings without data line

0 No element of the hierarchical level is present (right zero's are omitted)

Actual examples:

Y25T=10*1	Data on US as a whole and states; divisions as headings
Y25T=100101	Data on US as a whole, states and counties
Y25.123T=101	Data on Pennsylvania and counties (of Pennsylvania)

Fig. 3b: Geographic subdivision of US

```
Y25 USA
  Y25.1 Northeastern States
    Y25.ll New England
      Y25.111 Maine
      Y25.112 New Hampshire
      Y25.113 Vermont • • •
    Y25.12 Middle Atlantic
      Y25.121 New York
      Y25.122 New Jersey
      Y25.123 Pennsylvania
        Y25.123.1 SEA (Standard Economic
                                   Area)1
          •••
        Y25.123.2 SEA 2
          •••
        Y25.123.3 SEA 3
          •••
        Y25.123.4 SEA 4
          •••
        Y25.123.5 SEA 5
          •••
        Y25.123.6 SEA 6
          •••
        Y25.123.7 SEA 7

        Y25.123.A SEA A
          •••
        Y25.123.B SEA B
          Y25.123. Bl Bucks
          Y25.123. B2Chester
          Y25.123. B3Delaware
          Y25.123. B4Montgomery
          Y25.123. B5Philadelphia
        Y25.123.C SEA C
          Y25.123. Cl Lackawana
        Y25.123.D SEA D
          Y25.123. Dl Allegheny
          Y25.123. D2Beaver
          Y25.123. D3Washington
          Y25.123. D4Westmoreland
        Y25.123.E SEA E



        Y25.123.M SEA M
          Y25.123.M1 Lehigh
          Y25.123.M2 Northampton
  Y25.2 North Central States
    Y25.21 East North Central
      Y25.211 Ohio
```

For later reference, we show another standard classification, the ISIC scheme (7) in fig, 3c.

Obviously this method is applicable only in those cases where for each hierarchical level the following holds: Either none or all of the elements of the hierarchical level are present. Fortunately this is a condition which often holds. If it does not hold, one may try to use the procedures described in the following section for the modification of stratifiers.

4. The Modification of Stratifiers

The first method to be described here consists in a combination of the full-listing method presented in 2 and the reference method described in 3, as may be explained by means of the following example (fig. 4a). For some reason or other, the author of the table choose to give data on those entities. Obviously, the reference method of section 3 is not applicable in this case. However, we could replace two large blocks within the stratifier by a short reference; this would give us the short description of the stratifier shown in fig. 4b.

We can formulate the principle behind this reduction as follows: A block is a contiguous sequence of lines (including the case, where the "sequence" consists of one line only). A block may be represented by a reference to a prefabricated block as described in section

Fig. 3c: ISIC scheme

```
N89G ISIC
  N89G.1 Agriculture, hunting
         forestry and fishing
    N89.1.1 Agriculture and hunting
      N89.1.1.1 Agricultural and live-
      stock production
        ...
  N89G.2 Mining and quarrying
  N89G.3 Manufacturing
  N89G.4 Electricity, gas and water
  N89G.5 Construction
  N89G.6 Wholesale and retail trade
  N89G.7 Transport, storage and
         communication
  N89G.8 Financing, insurance, real
         estate and business services
  N89G.9 Community, social and
         personal services
  N89G.1O Activities not adequately
  |    |   defined
  |    |
  |    └─ Notation used within ISIC
  |
  └──── Code-number assigned to ISIC in the
        framework of a larger scheme
```

3. A stratifier may be represented as an arrangement of blocks, taking into account the appropriate hierarchical levels.

This method gives more flexibility, but still does not allow for the modification of prefabricated blocks. These problems are solved using the following conventions and notation: To any block, as represented by a reference, a modification description may be added. A modification description consists of several modification statements. A modification may be either (1) the insertion of new subblock or (2) the replacement of a subblock by a new subblock or (3) the deletion of a subblock. It follows, that a modification statement must specify (a) the location, where the new subblock is to be inserted or the old subblock to be replaced or deleted, respectively, and (b) the new subblock (to be inserted or replacing an old subblock) or the deletion operator, as the case may be. The new block (to be inserted or replacing an old block) may be represented either as a full listing of its elements or as a reference to a prefabricated block, which may in turn be modified by a modification description. Examples are given in fig. 4c and 4d.

5.

As a pre-requisite for the application of this method for the economical description of tables, relevant classification schemes must be collected, and a code symbol must be assigned to each of them. This is a minor requirement, however, compared with the savings to be achieved by the method. The rules are based on a few basic principles so that their application is easy. This is also an advantage for computerizing the interpretation of the stratifier format.

6. Application of the Method to Thesaurus-Building

At the end of this paper we may take the liberty to mention an interesting application of the format described in quite another area: Imagine that you would be given the task of constructing a specific classification scheme to be applied in a specific institution having specific requirements as to the more or less detailed treatment of the different subjects, the priority and sequential arrangement of the subjects and the use of very specific descriptors, so that no existing scheme could be used. You could consult different sources, but in essence you would have to build a totally new classification scheme, with all the effort necessary for such an undertaking.

Now, imagine that somebody has created a storehouse of information on

Fig. 4a: Stratifier, where reference method not directly applicable

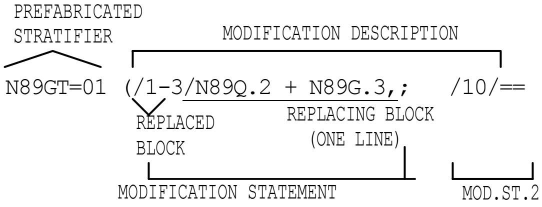
```
Y25.123 Pennsylvania
  Y25.123.B SEA B Y25.123.B1
    Bucks Y25.123.B2 Chester
    Y25.123.B3 Delaware
    Y25.123.B4 Montgomery
    Y25.123.B5 Philadelphia
  Y25.123.C SEA C
  Y25.123.D SEA D
    Y25.123.D1 Allegheny
    Y25.123.D2 Beaver
    Y25.123.D3 Washington
    Y25.123.D4 Westmoreland
```

Fig. 4b: Arrangement of several blocks (short representation of the stratifier listed in fig. 4a)

```
Y25.123 Pennsylvania
  Y25.123.BT=11 SEA B    (includ. counties)
  Y25.123.C     SEA C    (without counties)
  Y25.123.DT=11 SEA D    (includ. counties)
```

Data on: Pennsylvania as a whole, SEA's B, C, D, and the counties of B and D (but not of C)

Fig. 4c: Modification description (representation of boxhead of the table given in fig.4a using the ISIC scheme shown in fig.3)



That means: Industrial division according to ISIC, but the subblock consisting of lines 1-3 is replaced by a new line (there is no sum column corresponding to N89G itself), and line 10 is deleted.

Fig. 4d: Modification description (representation of a stratifier using the geographical subdivis-ion shown in fig. 3b)

```
Y25.123T=11 (/6/==; /9/Y25.123.BT=11;
            /ll/25.123.DT=ll(/la,2/
            Y25.123.D1.4 Pittsburgh))
```

That means: Data on Pennsylvania as a whole and the SEA's of Pennsylvania.But: Line 6 deleted: No data on SEA 6. Lines 9 and 11 replaced by blocks: Data on the counties of SEA's B and D. The block replacing line 11 is in turn modified: One line is inserted after line 1 on the hierarchical level 2 (relative to the inserted block): Data on Pittsburgh (part of Allegheny county).

concepts and their interrelationships in the form of a large classification scheme, universal in scope and detailed in the treatment of the different subjects, stored in a computer! You could not apply this scheme directly for the reasons mentioned. But you could do the following: Using the method described in sections 2-4, you could write a list of references to prefabricated blocks contained in the universal scheme, and modify these blocks as necessary. This is all you would have to do. The rest is done automatically: The computer program would prepare an extract of the universal scheme modified and supplemented in accordance with the specific requirements to be met, and create an alphabetical index (which would also contain synonyms of the terms chosen). This seems to be a realistic way out of the alternative universal vs. specific classification schemes, which has caused much controversy.

Appendix:

A formal description of the format described

The following are recursive definitions written in a Backus-like notation:

```
<Element ::= <descriptor>/<data-set-specific
               characteristic>/<deletion operator>

<Block>::= <element>/<block, to which another block
             has been added at the end>/ <block
             modified by another block>

<subblock>::= <any block contained in another block>

Modification of a block by another block>::=
               <insertion of another block as sub-block
               of the resulting new block>/ <replacement
               of one subblock by another block>
```

Further notes

Descriptors are elements of a classification scheme and referred to by notations. Data-set specific characteristics are defined in the description of the data set to which the table belongs and referred to by numbers assigned within that description. The inclusion of the deletion operator into the set of elements means that deletion of a subblock is equivalent to replacement by another block, the other block being the deletion operator.

Addition of a block at the end and insertion of a block could be identified in the formal definitions. This has not been done in order to reflect the different ways the two operations are performed in practice.

Some blocks may be referred to by short references as has been shown. Modification is useful only for these blocks, of course.

A block may be used as stratifier in the description of a table. The blocks so used are not distinguished formally from other blocks.

References

1. Systems Development Corporation, SPAN/36O Capabilities for Census Use Applications. SPAN System Overview, SDC, Santa Monica, Calif., 1970, 9p. (internal memo available on request)

2. Manuals for SEDAS are in preparation at Infas, 53 Bonn-Bad Godesberg, Germany

3. DATUM e.V., Erfassungsschema fur Dateien und Umfragen: Auf Grund praktischer Erfahrungen revidierte Endfassung, DATUM e.v., Bonn-Bad Godesberg, 1970 (Fundortkatalog fur Daten des Arbeitsmarktes. Projekt-Report Nr. 2(3))

4. Bruce L. Jenkinson, Bureau of the Census Manual of Tabular Presentation, GPO, Washington, D.C., 1960, 266 p., p.10-11

5. J.B. Irwin, A Systematic Approach to Describing 1970 Census Summary Tape Contents, Systems Development Corp., Santa Monica, Calif., 1968 April, 5 p., 4 app. (SDC-TM(L)-3917/OOO/OO)

6. F. Lachmann, An Automated System for Census Summary Tape Contents, Systems Development Corp., Santa Monica, Calif., 1968 August, 84 p. (SDC-TM-4073/OOO/OO)

7. United Nations Statistical Office, International Standard Industrial Classification of all Economic Activities, United Nations, New York, JOT, 1968, 48 p. (UN/STAT/SER.M/4/Rev. 2)

8. US Bureau of the Census, County and City Data Book (A Statistical Abstract Supplement), GPO, Washington, D.C., 1962, 669 p.

9. US Bureau of the Census, Statistical Abstract of the United States, GPO, Washington, D.C., annually