Basic Architecture of Bibliographical Database on Statistical Information for More Efficient Use of Statistical Data

Yoshiro Matsuda (Hitotsubashi University, Tokyo)
Setsuo Suoh (Kansai University, Osaka: Speaker)
Tsuneharu Ohkubo (Hitotsubashi University, Tokyo)

1. INTRODUCTION

This paper aims to clarify the characteristics of statistical data from the view point of information retrieval and to propose a bibliographical database system on Japanese statistics, although its basic architecture could be applicable for foreign statistics.

The statistical data differ from the other ordinary fact data or bibliographical data: most of the fact data are constructed on a well-defined conceptual scheme, and information specialists may concentrate on constructing a retrieval system suitable for themselves, but the contents of data do not change because of the users' level. The statistical data, however, can be transformed in terms of numerical figures or tabulation forms upon the request of its user's needs: e.g., population of a certain country is one type of statistical data, but one may want it for two opposite sexes, or for age and each of the sexes. Such figures sometimes can be obtained from a single source of data or statistical report, but in other cases they can be found in more than one separately bound report, or need to be calculated by compiling different statistics from other sources of data. Even such transformations do not change an intrinsic nature of the statistical data.

The bibliographical data on statistical reports differ from that of ordinary periodicals in two respects: one is the characteristics, as briefly mentioned in the above, that all statistical reports include self-contained numerical data but that it is desirable for each of them to be connected with each other if they come from the same single statistical survey, and the other is that in the case of periodical surveys, various kinds of statistical reports on the same but time-different surveys published under different titles must be linked together from the view point of continuation of a time-series of the same statistical survey.

The statistics can be divided into three categories(Matsuda 1980): i) survey statistics, ii) administrative records, and iii) compiled statistics such as GNP or price index, and in this present paper we will focus on our two projects related to the first category, i.e., survey statistics.

Since government statistics and even some private statistics are regulated by the Statistical Act or the Statistical Report Regulation Act, advance information on the characteristics of the surveys are available such as survey procedures, sample size and sampling methods. Based on such information, we can organise a database on advance information concerning those statistical surveys that were already or will be conducted. In our first project, we have been compiling test data for this purpose. The statistical analysis of the test file for 1983 indicates that 77 percent of 879 surveys (i.e., 680 surveys) are designed to make disclosure of their results; 257 surveys are conducted irregularly, and the rest (i.e., 423 surveys) are periodical surveys with varieties of survey cycles from daily to every ten years. More results are shown in Section 2.

The over-all results taken from the first project have resulted in our second project which started a few years ago to construct a bibliographical database on statistical information on Japanese statistics called STATIONS

(STAtistical informaTION System). We have so far compiled 3.568 statistical reports covering sixteen censuses and large-scale sampling surveys conducted after World War II. The resulting database. although in the seminal stage at the moment. demonstrates two crucial points to be improved as soon as possible. The first point is the necessity of constructing authority files not only for authors or surveyors. but also for survey names that are ignored by usual library routine work. The second is to construct longitudinal files among the statistical reports on periodical surveys. An innumerable number of statistical reports continue to be published after every periodical survey. Although conducted regularly. the publication of its results is not necessarily regular. and sometimes even unknown. Moreover. survey cycles differ from survey to survey. As a result. continuation of statistical data between two consecutive surveys tend to become much more difficult to grasp than that of ordinary types of periodicals such as academic journals.

In the final section we will sketch the basic architecture in constructing STATIONS.

2. Statistical Surveys and their Reports

As far as descriptive information on current statistical surveys and their reports is concerned. there are two books available: i) "Comprehensive Indices of Statistical Information" edited by Bureau of Statistics. Prime Minister's Office (now Bureau of Statistics. Management and Coordinate Agency). and ii) "Directory of Censuses and Statistical Surveys in Japan" edited by Dept. of Statistical Standards. Statistics Bureau. Management and Coordinate Agency (former Director of Statistical Standards. Administrative Management Bureau. Administrative Management Agency).

In our first project. we used the latter together with its magnetic-tape version, which contains the outline concerning almost all statistical surveys registered to the agency: e.g., surveyor, survey cycle, title of report, date of publication , publisher/editor, etc. It is compiled through the routine administrative procedures imposed upon government statistical surveys such as designated statistics, approved statistics and notified statistics, which means that its data is advance information on statistical surveys currently conducted. The analysis of the data(Ōhkubo 1986) reveals that about one sixth of all surveys do not produce their result as a report for general use. There are 467 periodical surveys that have a clearly defined survey cycle and that are designed to make disclosure of their results. The comparison between the periodical surveys and their publication cycle or frequency indicates that out of 467 periodical surveys, reports on 275 surveys are published in the same cycle as respective survey cycles, 57 less frequently than their own survey cycle, and 135 more frequently than their own survey cycle. Note that these numbers are not actual numbers of surveys, but the same survey can be counted repeatedly, because one single survey can produce more than one report.

In our second project, we have more precisely shown the complexities of bibliographical data on sixteen censuses and large-scale sampling surveys(Table 2.1). As part of working procedure for STATIONS, we have inputted not only their bibliographical information necessary for library house-keeping but also extra information relevant to the two characteristics of statistical data mentioned in the previous section: i.e., i) connection among the same series of reports produced from one single survey, whether or not periodical, and ii) continuation of reports in terms of a time-series of the same statistical survey conducted repeatedly in a regular or irregular cycle.

For the latter, we introduced a 'connector' that is given to each bibliographical unit, that is, each statistical report, as a kind of self-identifying code by hyphenating a published year and numbers, so that the way of hyphenation of connectors can explicitly indicate the relations of the first characteristic among various reports. For the second characteristic, each bibliographical unit is given an appropriate number of pairs of a connector of itself and that of those reports that are published from the following survey and that the contents of each other are continuous as time-series data. By having this one-way relationship in the initial input data, we can also compute a backward relationship.

Figure 2.1 shows the complexities of such relations in the case of Census of Agriculture and Forestry as an example. Each connector enclosed by a pair of brackets '< >' denotes a connector of each report, and underneath connector(s) with '. .' are those report(s) to be linked as time-series data published from the following survey, and the lines are drawn according to the underneath connectors for visual understanding.

3. Basic Architecture of STATIONS

STATIONS has been designed and constructed to computerize and improve the library service system at the Documentation Centre of Japanese Economic Statistics, Institute of Economic Research, Hitotsubashi University. The present STATIONS is only part of what we are eventually aiming at. Its total file organisation is shown in Figure 3.1, and we are now working on the 'bibliographical data' as our second project. Our first project deals with the 'authority file for statistical survey names'. Both of the project teams also have started working on the 'authority file for authors' from different viewpoints. We also have been working on the library house-keeping file for future computerization of the Documentation Centre. We have a plan for organising the 'contents file', but it will be left over until other files are completed.

The present bibliographical data for STATIONS have been organised and constructed as shown in Figure 3.2. When designing the initial input file, much emphasis was placed upon efficient data coding and punching, and less error-free file. This is very important particularly for those large surveys such as Population Census that produces many reports with similar titles. To cope with it, we adopted a file with hierarchical structure(Figure 3.3). Since the number of letters (or characters) of titles of statistical reports is much longer and more complicated than that of ordinary books or periodicals, and it often occurs that the first part of two titles is identical and only the second part is different, we separated such titles into half, or sometimes into three parts, as shown below using Backus Notation.

```
<common part of title> ::= record with '¥' as a tag
<different part of title> ::= record with 'B' as a tag
<title (type 1)> ::= <common part of title> /
                     <title (type 1)><different part of title>
<title (type 2)> ::= <different part of title>
<complete title> ::= <title (type 1)> / <title (type 2 )>
```

Thus when two reports with a common part of title come into the initial input file physically next to each other, the second occurrence of the record of common part is not inputted. The similar way of abbreviation of input record occurs in the case of a editor record and surveyed year record: e.g., when in Figure 3.3 (3), shadowed records are identical to those with 'X' respectively, the latter are not included in the initial file, as in Figure 3.3 (4).

After abbreviated information is recovered by SAS programs to create the master file, the latter is compiled and converted into basic SAS datasets.

If we took a look at Figure 2.1 and a matching list of titles and connectors, we could trace titles in time-series, but it is too cumbersome to do so. The Bibliographical Transition Retrieval System built on PC-SAS has been developed to help statistical users who want to collect time-series data. Given a statistical survey No. and report No. of the statistical report that the user know, the system displays three sets of information on the screen (Figure 3.4); i.e., various attributes of the statistical report itself to be first given, and two sets of bibliographical information about those statistical reports having time-series connection with the given report that were published from both the preceding survey and the following survey.

REFERENCES

[1] Matsuda, Y.(1980),"Survey Data vs. Compiled Data," Proc. of the 4th Japan-U.S. Forum on International Issues, the World Economy, Energy and International Relations, pp 133-137.

[2] Ohkubo, T.(1986),"Disclosure of Official Statistical Survey in Japan," Proc. of the 2nd Japan-China Symposium on Statistics, pp 194-196.

[3] Suoh, S.(1987), Computational Technique for Constructing Catalogue Data base on Periodicals with Flexible Type of Publication, (in Japanese, kasoteki Kankou keitaino Chikuji kankoubutsu Mokuroku De-ta Be-su Hensei Gihou), Tokyo: Hitotsubashi University.

	Survey Names	survey cycle	survey freq.	No. of reports
1	国勢調査 Population Census	5 years	7	1138
2	住宅統計調査 Housing Survey	5 years	8	265
3	全国消費実態調査 National Survey of Family Income and Expenditure	5 years	5	53
4	全国物価統計調査 National Survey of Prices	5 years	5	69
5	社会生活基本調査 Survey on Time Use and Leisure Activities	5 years	2	15
6	事業所統計調査 Establishment Census	3 years	13	530
7	就業構造基本調査 Employment Status Survey	3 years	10	40
8	学校基本調査 School Basic Survey	1 year	34	61
9	学校教員統計調査 School Teachers Survey	3 years	15	32
10	農林業センサス Census of Agriculture and Forestry	5 years	7	599
11	漁業センサス Census of Fisheries	5 years	7	167
12	工業統計調査 Census of Manufactures	1 year	38	221
13	商業統計調査 Census of Commerce	3 years	15	112
14	本邦鉱業のすう勢調査 Survey of Mining Trend of Japan	1 year	38	38
15	全国貨物純流動調査 Real Goods Flow Survey in Japan	5 years	3	13
16	賃金構造基本統計調査 Basic Survey of Wage Structure	1 year	37	215

total 3.568

Table 2.1 16 Censuses and Large-scale Sampling Surveys Conducted after World War II

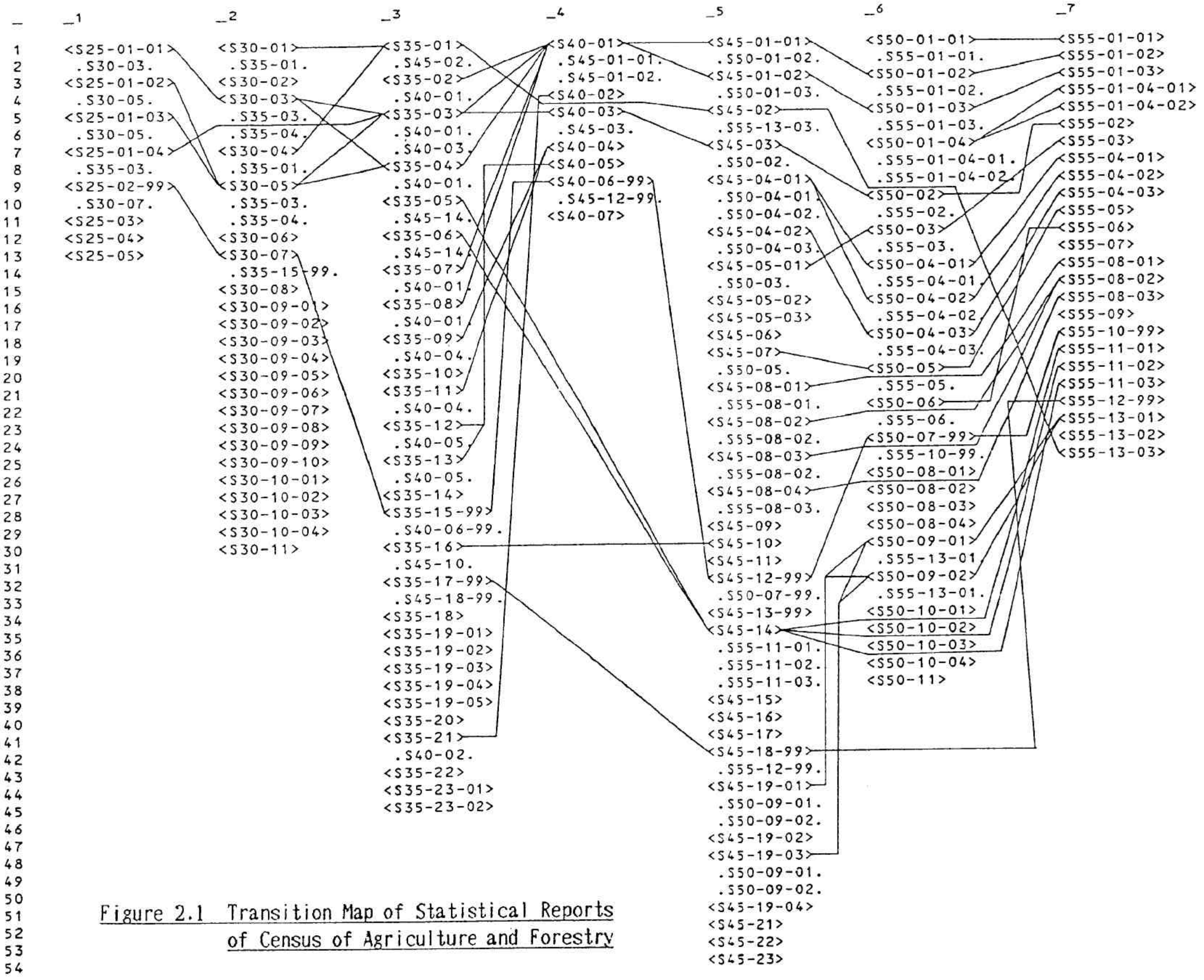


Figure 2.1 Transition Map of Statistical Reports of Census of Agriculture and Forestry

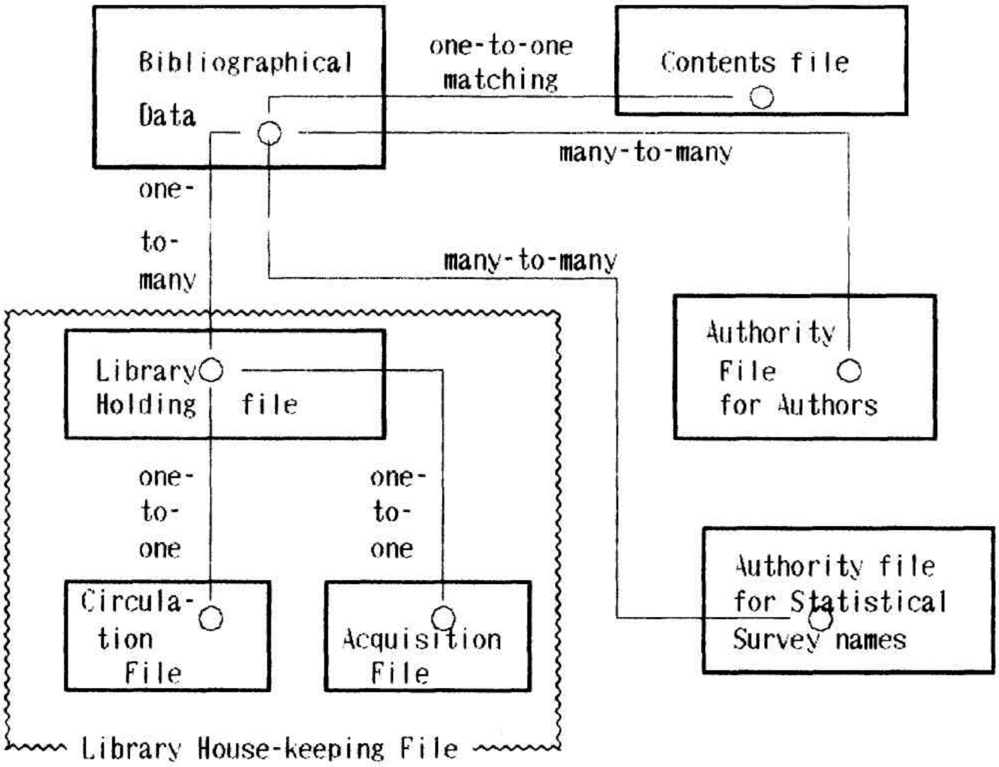
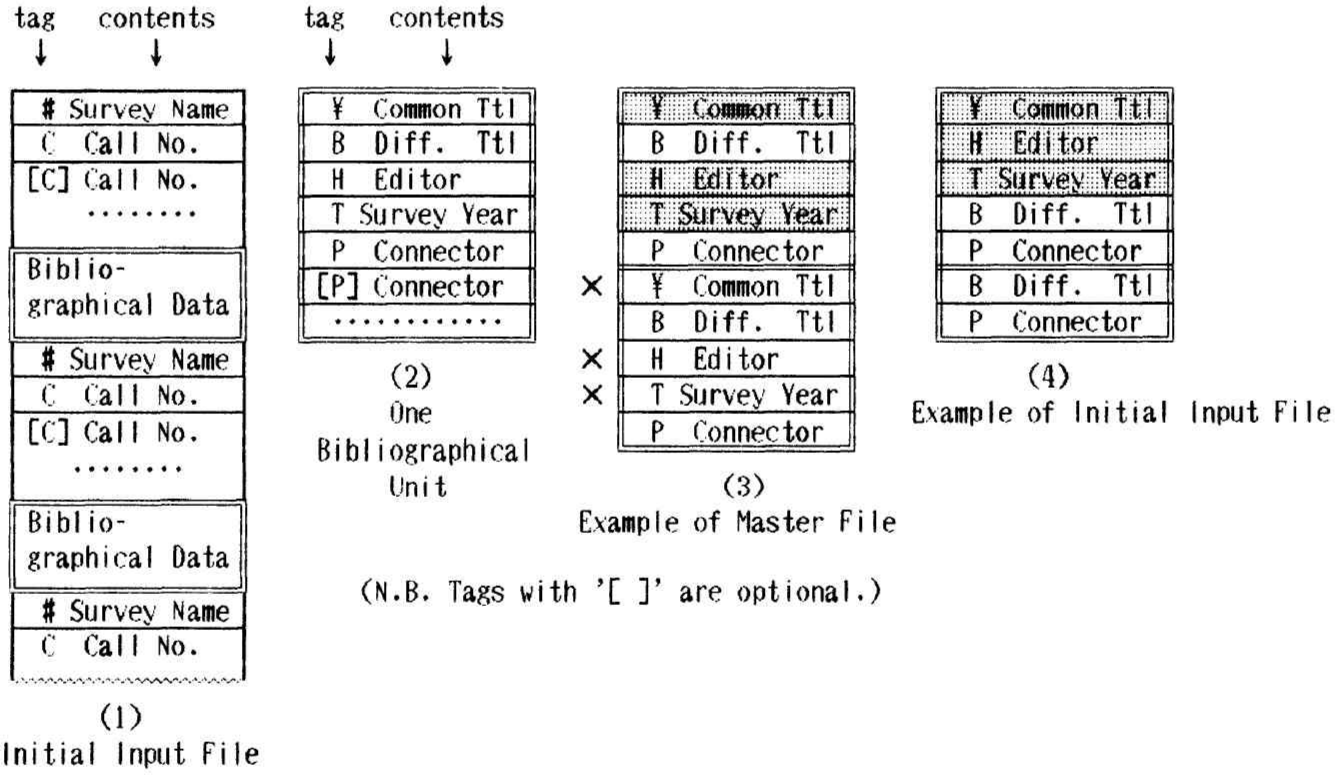


Figure 3.1 File Organisation of STATIONS



(1)
Initial Input File

(2)
One
Bibliographical
Unit

(3)
Example of Master File

(4)
Example of Initial Input File

(N.B. Tags with '[ ]' are optional.)

Figure 3.3 Initial Input file and Master File with Hierarchical Structure

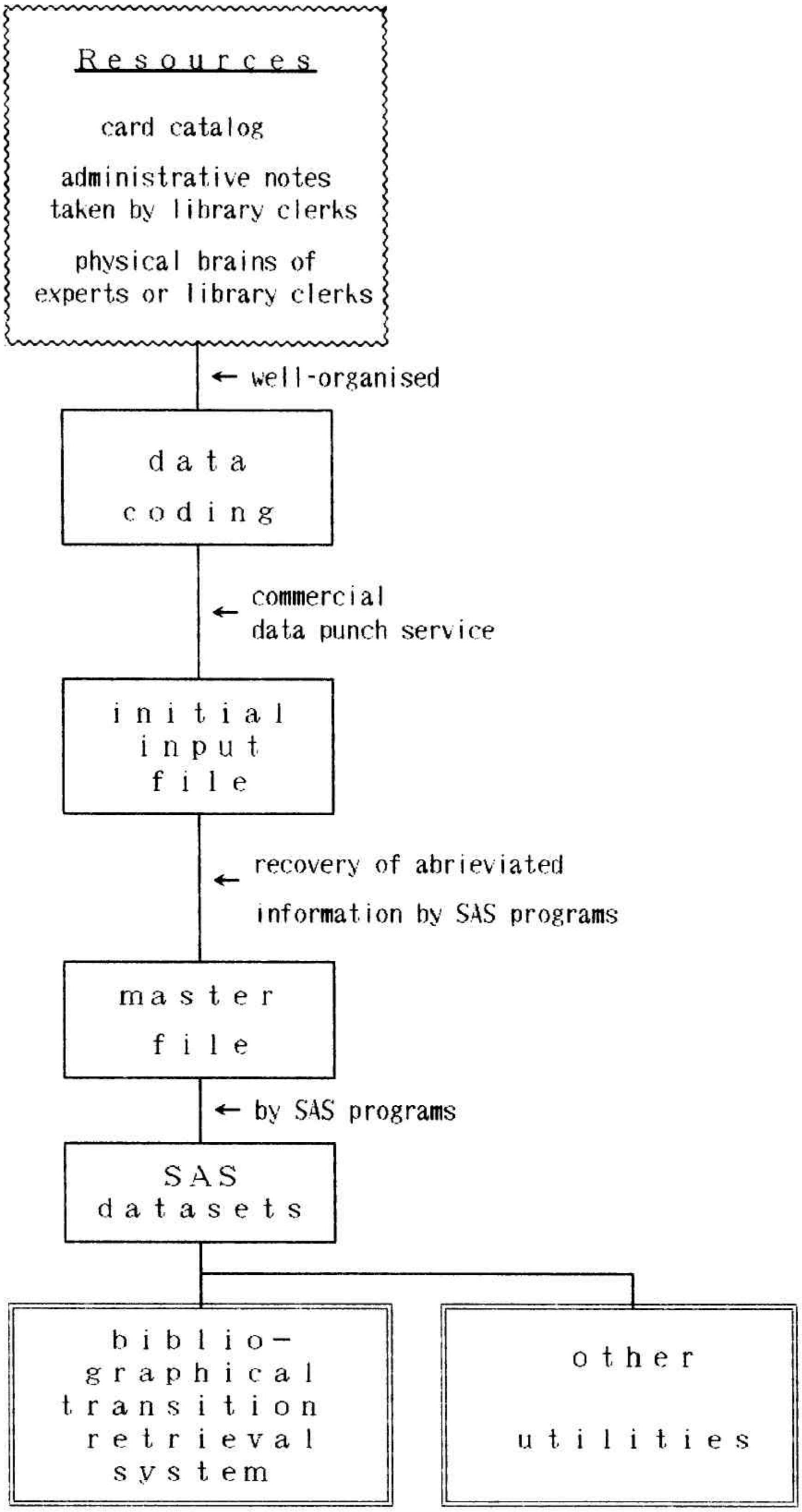

Resources
card catalog
administrative notes taken by library clerks
physical brains of experts or library clerks
← well-organised
data coding
← commercial data punch service
initial input file
← recovery of abrieviated information by SAS programs
master file
← by SAS programs
SAS datasets
biblio-graphical transition retrieval system
other utilities

Figure 3.2 Working Process to Construct STATIONS

```
《今回調査報告書》=========================================
前回　5冊:次回　2冊　　調査対象年 S40.02.01
調査#　　　　1000　　刊行年　　S42.03
報告書#　　195000　　編集形態　J
S40-01　　　　　　　　編集機関
農業センサス@１９６　農林省・農業経済局・統計調査部
５年　農家調査報告書

《前回調査報告書》-----------------------------------------
前回　0冊:次回　1冊　前回　0冊:次回　1冊　前回　0冊:次回　1冊　前回　0冊:次回　1冊　前回　0冊:次回　1冊
調査#　　　　1000　　調査#　　　　1000　　調査#　　　　1000　　調査#　　　　1000　　調査#　　　　1000
報告書#　　　78000　報告書#　　　79000　報告書#　　　80000　報告書#　　　83000　報告書#　　　84000
S35-02　　　　　　　S35-03　　　　　　　S35-04　　　　　　　S35-07　　　　　　　S35-08
世界農林業センサス@　世界農林業センサス@　世界農林業センサス@　世界農林業センサス@　世界農林業センサス@
１９６０年　農家調査　１９６０年　農家調査　１９６０年　農家調査　１９６０年　農家調査　１９６０年　農家調査
報告書　果樹編　（１　報告書　農家・人口編　報告書　生産手段編　報告書　農産物販売農　報告書　農産物販売農
９６０年世界農林業セ　（１９６０年世界農　（１９６０年世界農林　家編　１　（１９６０　家編　２　（１９６０
ンサス資料　Ｎｏ．２　林業センサス資料　Ｎ　業センサス資料　Ｎｏ　年世界農林業センサス　年世界農林業センサス
）　　　　　　　　　ｏ．３）　　　　　　．４）　　　　　　　資料　Ｎｏ．７）　　資料　Ｎｏ．８）

《次回調査報告書》-----------------------------------------
前回　1冊:次回　1冊　前回　1冊:次回　1冊
調査#　　　　1000　　調査#　　　　1000
報告書#　　247000　　報告書#　　248000
S45-01-01　　　　　　S45-01-02
世界農林業センサス@　世界農林業センサス@
１９７０年　農家調査　１９７０年　農家調査
報告書　農家・人口編　報告書　生産手段編
```

English Translation

```
《Survey Report to be Examined》=========================================
Previous Survey, 5 reports; Following Survey, 2 reports | Surveyed : Feb. 1 of Showa 40th Year
Survey No. 1000                                         | Published: March of Showa 42nd Year
Report No. 195000                                       | Editting Style: J
Self-identifying connector: S40-01                      | Editor: Dept. of Statistical Survey, Ministry of
Title of Report: Census of Agriculture 1965,            |         Agriculture, Forestry and Fisheries.
                 Farm Households Survey Report.         |

《Reports from Previous Survey 》-----------------------------------------

  N.B. The format is the same as the left hand side of the above.
       In this example, five reports are listed, as indicated in the above; 'Previous Survey, 5 reports'.

《Reports from Following Survey》-----------------------------------------

  N.B. The format is the same as the above.
       In this example, two reports are listed, as indicated on top; 'Following Survey, 2 reports'.
```

Figure 3.4 Output Example from the Bibliographical Transition Retrieval System:
(Census of Agriculture and Forestry)