Technological change and the provision of documentation for time-series datasets

by Hilary Beedham[1]
ESRC Data Archive
University of Essex,

Acknowledgements.
I would like thank my colleague, Paul Child who supervised the technical aspects of the work to which this paper relates and who contributed generously to the technical content of this paper.

I would also like to thank the U.K. Central Statistical Office, not only for making the Family Expenditure Survey data available to the academic community through the ESRC Data Archive but also for their continuing interest and support of the Archive's work in making the data and documentation more easily available to secondary analysts.

Background.
The Family Expenditure Survey (FES) is a survey of household spending which originated from a recommendation of the Cost of Living Advisory Committee (now the Retail Prices Index Advisory Committee) that such an inquiry should take place as a source for the weighting pattern of the Index of Retail Prices - commonly known as the Retail Price Index (RPI). The first such survey was carried out in 1953/4 and the survey began as a continuous survey in 1957.

The Data Archive at Essex University is now the only and most complete source for these data which are used extensively for secondary analysis by researchers in a varied number of disciplines.

The data collected include not only details of household expenditure such as spending on rent, rates, transport, building maintenance and fuel expenditure but also a substantial amount of demographic information and information collected from a two week diary which is kept by all adults in the household.

The earliest data have either been lost or are not machine-readable. The schedules for the 1953/4 survey are held in their original non-anonymised form at the Public Record Office and are not available for conversion into a machine-readable file. The datasets from 1957 to 1960 inclusive were held on punch cards and are, unfortunately deemed to have been lost. The original data collectors have been generous in support of the Archive's attempts to find the cards but they have never been recovered and further efforts are now thought not to be worthwhile.

The datasets are well documented for almost all of the years for which they are preserved. There is an insuperable problem with the data for the years 1964 to 1967 inclusive as the data layout documents are inadequate for the accurate interpretation of the mixed binary/character data files. The Archive has done extensive work on this over the years but we have been unable to make an accurate interpretation of the files. One researcher with a particular interest has managed to read a few specific variables based on matching consistent codes (such as region) across years but the reliability of even these few variables must remain in doubt. For one of these years there is an additional problem in that only three of the four seasonal quarters has been preserved.

The Problem.
There is nevertheless a considerable amount of data and the FES is available for academic research continuously since 1968. Data are also readily available for the years 1961 to 1963. The type and variety of information included in the survey along with the length of time for which the data are available mean that the FES is one of the most heavily used surveys held in the Archive.

The data have a complex hierarchical structure: they are subject to changes in definitions resulting from, for example, changes in benefits available to the public; methodological changes; changes in coding and content; and changes to the structure of the databases supplied to the Data Archive. The documentation is essential to anyone who wants to use these data and with the support of the depositors the Data Archive strongly recommends that the data should only be used with the full set of documentation for each year in use and no longer offers substantive support to users who fail to purchase the documentation.

This policy is not without problems, however, since although we are funded to provide the data free of charge to most academic users, we do have to charge for the documentation on a cost recovery basis. The total cost of the documentation for the entire set of FES documentation is approximately £775.00 (US$ 1250, approx).

The effects of this on researchers are threefold:

1. Some researchers choose not to use the data because of these costs.

2. Others curtail their research and use only a few years worth of data when they would have preferred to use all.

3. Others insist on ordering all the data but with documentation for only one or two years.

All of these effects are unwelcome and act as a deterrent to good quality academic research despite the use of a very rich source of information.

The Solution.
By 1991 the Archive was becoming increasingly interested in the possibility of providing documentation for on-line use. The provision of FES documentation in this way would mean that the problems associated with the cost and quantity of the documentation to users could be almost eliminated and we began to consider a feasibility study into whether or not it would be possible to convert all of the documentation for the FES into machine-readable files.

Appendix 1 gives an indication of the quantity of documentation which required conversion by information type. The amount varied by year with the earliest years having the least. For 1961, for example, there are only 44 pages which include the data layout, the schedules and a note on the validation tests carried out on the data. In contrast, the most recent dataset (1993) is much more fully documented and has a total of 1324 pages of documentation.

If this could be achieved, we could solve the problems of researchers' access to the documentation by including the files on the same medium as the data, thus charging users only for the medium on which they are sent rather than the copious amounts of paper as at present. This solution also has the potential of reducing staff time required in preparing documentation orders since it removes the need to photocopy the documents as they were required.

There were, however, a number of problems which had to be considered before even a feasibility study could begin:

1. Collation of the documentation.
Key documentation, the schedules and the data layout tables with coding information attached, had always been made available to users via the Archive and was filed in such a way as to make it readily and easily available for both internal and external use. Between 1968 and 1986, other, more comprehensive documentation had been available to users directly from the depositors. In 1990/91, the depositors sent the remaining copies of these 'Information packs' to the Archive for dissemination as required. These were not available for every years and some did not contain all the documentation indicated by their contents pages. Some of these gaps might be filled with the documentation already held in the Archive and some of the missing information could be acquired from the annual reports for the surveys. Another problem was that some of the documents overlapped with the schedules and coding notes we already held. In order to collate the information we had to list the contents of each pack, compare this with the contents list and the contents of each other pack for the same year and then check whether any known gaps could be filled either from the documentation already in the Archive or from the depositors.

2. Quality of the documentation.
As might be expected the quality of the paper and the typeface of the documents varied over the years and both the typeface and the paper quality varied between the earliest and the latest year. Many of the early documents were photocopies of type-written originals whilst later documents are copies of either word-processed files or of output from computer printout. The latter did not all present problems since we can generate equivalent machine-readable files from the data files we hold because they were created using the SIR software. This is not, however, the case for all such documentation because the complex documentation processes in place at the Employment Department (the Government department with responsibility for this aspect of work on the FES) could not all be reproduced with the files we archive. We could not re-create any documentation prior to that for 1986 by this means.

3. Current scanning technology.
We needed to spend a significant amount of time deciding which of the documents could be scanned and which would have to be typed in. From the outset it was agreed that the schedules could not be scanned using existing scanning technology so any feasibility study would have to include provision for typing these in. There were other documents which lay in a 'gray area' with respect to the scanning/typing decision and the only solution was to include comparative time tests for the same documents - scanning versus typing.

4. Which output format should we choose?
A decision also had to be made as to whether the paper should be converted into Optical Character Recognition (OCR) or image files. This decision was relatively easily made on the basis that very large amounts of storage space which would be required if the documentation were to be converted into image files. Also, experience suggested that users would have less problems reading OCR files than image files and we did not wish to embark on such a large project only to create new problems for users.

5. Equipment.
The scanner which was then available to the Archive was a Onescanner attached to an Apple Macintosh II ci computer. The software used was Omnipage. This is a fairly dated set of equipment with a flatbed rather than document feeder and it was clear that we would need a more recent machine if the full project were to be undertaken efficiently. We were aware that any feasibility study using this equipment would necessarily suggest considerably more resources than would be needed if we had a more up to date set-up. However, we were fortunate in learning of a much better scanner in another department and staff there generously allowed us use of the machine during the Summer vacation. There were

restrictions on the use of this but it offered a much better opportunity to produce realistic figures on the resources which would be needed to convert the entire set of documentation into machine-readable files. One important outcome of the project has been to demonstrate the Archive's need for funding for a better scanner.

The feasibility study.
The feasibility study was conducted in a number of stages with some aspects of the work running in parallel.

Stage 1: Preliminary work.
Stage 1 involved both the collection of documentation and the practical testing of the Archive's flatbed scanner. The latter simply demonstrated the impracticality of using such a machine for such a large project. At this stage a list was compiled of what documents were expected to exist for each year (Appendix 1). Having itemised all the documentation, a document typology was developed to serve as a framework in which to apply different OCR settings and to create Recognition Training files (RT files) within the scanning software. Five different types of text were described within this typology, shown on Table 1:

Table 1: A document typology for use with OCR Scanning.

Typology	Description
Simple text	Text written from left to right of a whole page without non-grammatical breaks. Structural objects such as headings and formatting objects such as indents may be included.
Formatted Text	Includes simple text, lists and ASCII tables i.e. Tables without graphic lines, the columns are separated with spaces or tabs.
Columnar text	Simple or formatted text in a newspaper style layout.
Tables	Columns of information with graphic lines**.**
Complex Text	Includes simple text, formatted text and graphic lines. Unlike formatted text there may or may not be a relationship between lines of text. A good example of this is schedules.

Table 2 gives the estimated proportion of each type of text within the documents.

During this period, after the typology had been created, some of the documents were given to experienced typists in the Archive to compare typing with scanning and determine approximate costs if scanning proved not to be feasible.

Stage 2: assessment.
With the information gained in stage 1 and with access to a Kutzweil 6000 scanner attached to a 486 PC, using Textbridge software, we were able to employ a clerical assistant for a few weeks to take the project forward. During the first week, the documents were carefully collated and many gaps were filled. As so many of the documents proved to be unique, some time was also spent in photocopying the originals: it has long been Archive policy that any extensive internal work is not carried out on original documents as a preservation measure.

Also during the first week, time was spent in familiarising both the clerical assistant and the responsible staff member with the new machine. The learning curve proved to be steep and a number of false starts were made.

There were also problems which were due to the machine being on loan: we were unable to alter basic settings and the software had not been fully installed so that some of the features which would have improved certain aspects of recognition were not available. At one point the system crashed and the machine was reconfigured differently but we were not in a position to rectify this.

Nevertheless, significant progress was made and there were clear advantages to using the more recent equipment:

Table 2: The estimated proportion of each document type contained in each document

Document Typology	Estimated proportions per document
Simple text	20%
Formatted text	33%
Tables	22%
Columnar text	7%
Complex Formatted text	18%

1. Time was saved because of the document feeder;

2. Verification was more efficient since once a correction has been made during scanning, the software 'remembers' it and can apply it elsewhere in the file during further scanning. Verification was slower but more extensive than with Omnipage but the actual scanning was considerably quicker.

3. The Textbridge software allows for zoning where the operator can isolate different parts of a page and apply different settings to each part. This was not fully exploited but is potentially very useful where distinct document types exist on the same page.

4. The software allows the operator to load a dictionary of terms which are specific to the area to which the document relates and also has the option to set lexical and grammatical rules depending on the type of document being scanned. For instance, when tables are being scanned, grammar rules can be turned off.

5. Extensive tests were not run using the available delayed processing facility, partly because of the need to ensure that we did not impinge on the work of the department which loaned us the machine. However, Textbridge allows image files to be created in Tiff format on which normal processing can be undertaken at a later time. Combined with the zoning facility, groups of documents can be processed outside normal working hours resulting in substantial efficiency gains.

Results.
Using the more advanced scanner, the clerical assistant succeeded in scanning the documentation for a full year; deliberately one of the years with the most documentation. This may seem limited, given that he was employed for 3 weeks in total but as has been explained, most of the first week was spent in collating, checking and photocopying documentation. Over a week was spent in familiarisation with the equipment and with running test documents through the system to create the RT files and only a few days of the final week were actually spent on systematic scanning of the documents.

We are confident that the bulk of the FES documentation can be scanned. That which cannot, will have to be typed in and this is expected to be costly because of the complexity of the schedules although it may be worthwhile to examine the possibility of reformatting the schedules so that information is not lost. We would only do this in consultation with the depositors and would not release altered schedules without their approval.

There are two key elements to the successful completion of this project: ready access to a state of the art scanner; and resources to employ a staff member who can make maximum use of this. The clerical assistant who worked on the feasibility study made it quite clear that the scanner was still 'learning' even after completing the work on one set of documentation.

As a result of this work, the Archive has been awarded funds jointly with another ESRC centre at Essex for the purchase of an extremely powerful scanner. It is hoped that resources can now be found for a staff member to work on this and take the project to its conclusion.

References
Family Spending 1993: A report on the 1993 Family Expenditure Survey. Ed. John King. London: HMSO 1994.

Family Expenditure Survey Handbook. WFF Kemsley, RU Redpath & M Holmes. London: HMSO 1980.

1 Paper presented at the IASSIST Conference, Quebec City, May 1995.

APPENDIX 1

Availability matrix of documents within the information packs, year by title.

	Intro	Annex A	Coding Notes	Annex B	Notes on Coll/ dp	Sched's	II & S	V.S	ABTD	cf of codes	Prob's notes
1986	¸	ü	ü	ü	ü	ü	ü	ü		ü	ü
1985	ü		ü		ü	ü	ü	N/A	ü		ü
1984	ü		ü		ü	ü	ü	N/A	ü	ü	ü
1983			ü		ü	ü	ü	N/A	ü	ü	ü
1982	ü	ü	ü	ü	ü	ü	ü	N/A	ü	ü	ü
1981	ü	ü	ü	ü	ü		ü	N/A	ü	ü	
1980	ü		ü			ü	ü	N/A	ü	ü	ü
1979	ü		ü			ü	ü	N/A	ü	ü	ü
1978	ü		ü			ü	ü	N/A	ü	ü	ü
1977	ü		ü			ü	ü	N/A	ü	ü	ü
1976	ü		ü			ü	ü	N/A	ü		
1975	ü		ü			ü	ü	N/A	ü		
1974						ü		N/A	ü		
1973			ü			ü		N/A	ü		
1972			ü			ü		N/A	ü		
1971			ü			ü		N/A	ü		
1970			ü			ü		N/A	ü		
1969			ü			ü		N/A	ü		
1968			ü			ü		N/A	ü		

KEY:

Intro:-	Introduction.
Notes on Coll/dp:-	Notes on collection and data processing.
Sched's:-	Schedules.
II and S:-	Interviewers instructions and sampling.
V.S:-	Variable Schedules.
ABTD:-	Annual base tape documents. (Data layout files with some coding information)
cf of codes:-	Comparison of codes between current and previous years.
Prob's/notes:-	Problems and notes.