COPYRIGHT

Luxembourg: Office for Official Publications of the European Communities, 2002

ISSN 1023-098X
Cat. No. KS-38-01-002-EN-N

© European Communities, 2002

# RESEARCH IN OFFICIAL STATISTICS

2■2001

eurostat

An international journal for research in official statistics

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server (http://europa.eu.int).

Luxembourg: Office for Official Publications of the European Communities, 2002

ISSN 1023-098X

© European Communities, 2002

*Printed in France*

PRINTED ON WHITE CHLORINE-FREE PAPER

# Research in Official Statistics

***ROS*** **—** An international journal for research in official statistics

***ROS –*** Volume 4 – Number 2 – 2001


## Contents

# Letter from the editors

The launch of the new European Research Programme is now imminent so it is timely to stress the important role of the ROS Journal in stimulating communication among researchers in official statistics.
The statistical part of the 5th Framework Programme—EPROS (Methods, Tools and Indicators)—financed numerous projects. Although the project selection phase of this R&D Programme is now ending, work on these projects continues and results will continue to flow for some time.
For this reason, priority will be given in the next issues of ROS to dissemination of results from these projects—to papers presenting findings and developments.
2003 will also be the year for first proposals under the new programme (6th Framework Programme) which with its new structure asks for more initiative from larger consortia and networks.
We are confident that the European Research in Statistics Community has by now gained enough experience and maturity to be able to handle the new challenges set by the 6th Framework Programme.

*Photis Nanopoulos* *Daniel Defays*

# A system to monitor the quality of automated coding of textual answers to open questions

Stefania Macchia ([1]) and Marcello D'Orazio ([2])

*ISTAT*
*Methodological Studies Department*
*Via C. Balbo, 16, I-00184, Rome*

**Keywords:** ACTR, coding precision, labour force survey, stratified sampling

**Abstract**

*The Italian National Statistical Institute (ISTAT) carried out some tests of automated coding of textual answers regarding occupation, education level, industry, etc., using the automated coding by text recognition (ACTR) system. The good results obtained led ISTAT to perform a further analysis of a large sample of textual data, in order to define a standardised procedure for reference when using ACTR instead of manual coding during a survey. The analysis shown in this paper aims at building up a system to monitor the quality of the results of automated coding and at verifying the improvements which can be achieved using the results of the monitoring activity to integrate the automated coding environment.*

## 1. Introduction

Coding written-in answers to open questions of statistical surveys typically requires dealing with the problem of their variability, depending on the cultural background of the respondents, on their ways of speaking and, finally, on how interested they are in cooperating. The written answers are often generic or ambiguous, since respondents are not expert in the classifications and so they respond without thinking that their answers have to be coded. The same may happen with interviewers (even if trained in advance) who are not always used to obtaining answers suitable to be easily coded.

Automated coding can help in solving the specific problems of costs, time and quality connected with the coding activity. In fact, manual coding implies high costs due to hiring, training and supervising coding personnel. It requires a long time, especially for complex questions such as industry or occupation, and an attempt to reduce time can negatively affect the quality of results. Finally, manual coding does not ensure any standardisation of the process (it is not sure that two different people assign the same correct code to the same textual description). The process is strongly influenced by factors related to knowledge of the classifications, skill and conscientiousness of coding clerks.

([1]) E-mail: macchia@istat.it; tel. (39) 06 46 73 21 57; fax (39) 06 47 88 80 69.
([2]) E-mail: madorazi@istat.it; tel. (39) 06 46 73 22 78; fax (39) 06 47 88 80 69.

For these reasons, various countries, for instance France, the United States, Canada and the United Kingdom, have developed and are successfully using automated coding systems. In France, Lorigny (1988) developed QUID (questionnaires d'identification), which was used in a number of socioeconomic surveys. Later, the Sicore system (*système informatique de codage des réponses aux enquêtes*) was designed to code different variables (Rivière, 1994).

In the United States, automated coding has been deeply studied and investigated. First papers appeared in the 1970s (O'Reagan, 1972; Corbett, 1972); other interesting papers are those of Hellerman (1982) and Appel and Hellermann (1983). For the 1990 census, the US Census Bureau developed the automated industry and occupation coding system (AIOCS); this system was adopted in the current population survey and also in the survey of income and programme participation (Lyberg and Dean, 1992). During the 1990s, the Census Bureau continued research by considering other possible techniques and systems for automated coding of industry and occupation (Creecy et al., 1990, 1992; Gillman and Appel, 1994, 1999). Still in the United States, the Centre for Health Statistics developed CLIO (classification of industry and occupation), a system derived directly from that used to code cause of death (Harris and Chamblee, 1994). Research by Statistics Canada led to the release of the automated coding by text recognition (ACTR) system, which went into production in 1986 (Wenzowski, 1988). Actually, an updated release of ACTR is used to code different variables for the census of population and the labour force survey. The United Kingdom currently uses precision data coder (PDC), which is a language-specific software, initially designed for industry coding. However, to code the different textual variables observed in the (UK's) current census of population, it was decided to adopt a more generalised software program such as ACTR.

Considering all these experiences, in 1998 ISTAT decided to test an automated coding system. Instead of developing new software, it was decided to use the third release of the ACTR system supplied by Statistics Canada. ACTR was chosen since it was language-independent and seemed easily adaptable to the Italian language, unlike other systems that were language-specific, such as, for instance, PDC. Moreover, as already mentioned, ACTR is a generalised system, so it can be used for more than one coding application. In addition, it has already been successfully used by other national statistical institutes (Tourigny and Moloney, 1995).

## 2. The ACTR system

ACTR's philosophy is based on methods originally developed at the US Census Bureau (Hellerman, 1982), but uses matching algorithms developed at Statistics Canada (Wenzowski, 1988). The coding activity follows a quite complex phase of text standardisation, called **parsing**, that provides 14 different functions such as character mapping, deletion of trivial words, definition of synonyms, removal of suffixes (these functions are completely managed by the users). The **parsing** aims at removing grammatical or syntactical differences so as to make equal two different descriptions with

the same semantic content. The parsed answer to be coded is then compared with the parsed descriptions of the dictionary, the so-called **reference file**. If this search returns a perfect match, called **direct match**, a unique code is assigned, otherwise the software uses an algorithm to find the best suitable partial (or fuzzy) matches, giving an **indirect match**. In practice, in the latter case, the software takes out of the **reference file** all the descriptions that have at least one parsed word in common with the one given by the respondent and assigns them a score. This score, standardised between 0 and 10 (10 corresponds to a perfect match), is computed as a function of the weight given to each single word in common, which is in inverse relation to its frequency of occurrence in the dictionary.

The system orders by decreasing score ($S_1 \geq S_2 \geq \ldots \geq S_n$) the descriptions selected from the **reference file** and compares them with three user-defined thresholds: the lower limit ($S_{\min}$), the upper limit ($S_{\max}$) and the minimum score difference ($\Delta S$). If $S_1 \geq S_{\max}$ and $(S_1 - S_2) \geq \Delta S$ the description with the score $S_1$ is said to be a **unique** winner and a unique code is assigned to it. If the first two (or more) descriptions are greater or equal to $S_{\max}$ ($S_1 \geq S_{\max}$ and $S_2 \geq S_{\max}$) but their difference is less than the minimum score difference ($S_1 - S_2 < \Delta S$), the system returns both as winners (**multiple** winners). The same happens if $S_{\min} \leq S_2 \leq S_1 \leq S_{\max}$; notice that in this case the similarity between the description to be coded and those selected from the **reference file** is lower than in the previous case. Finally, there are no winners if all the scores are less than $S_{\min}$ ($S_1 < S_{\min}$) and the system returns a **failed** message.

For **unique** winners, no human intervention is required, while all the other cases need to be evaluated by expert coders to choose which of the **multiples** will be the right one or whether to code at all the **failed** matches.

The following example, concerning occupation, clarifies how the indirect match works. The description *esercente di art. di abbigliamento di vario genere (esclusi i pellami)* ('trader of clothes art. of various kinds (with the exception of leather)'), after the **parsing** process we defined, becomes *abbigliament commerciant* ('clothes dealer', suffixes removed) and matches with the sentence of the **reference file** (actually used): *esercente di negozio di abbigliamento* ('shop trader of clothes'). In practice, the **parsing** first operates on strings, eliminating certain clauses (*esclusi i pellami*), deleting non-informative strings (*di vario genere*), replacing strings with synonyms and so on. It then operates on words, replacing words with synonyms (*esercente* becomes *commerciante*), deleting non-informative words (*di*, *i*) and removing suffixes from all words that do not have to be treated as exceptions. As the two sentences are similar but not identical, there is an indirect match with a score of 9.33; this score is greater than the threshold $S_{\max} = 8.0$ and, given that $(S_1 - S_2) > \Delta S = 0.2$, a unique code is assigned to the starting description.

Unfortunately, the indirect matching mechanism can produce errors. For example, consider the description *addetto ai servizi ausiliari* ('assigned to auxiliary services'); it would match

(with the actual **reference file**) with *addetto ai servizi ausiliari del reattore* ('assigned to auxiliary services of the reactor') and, having a high score, it would be uniquely coded. But, as it can be seen, the original description does not refer to any reactor; instead, it should match with the code corresponding to the description *personale inserviente negli uffici* ('office attendant').

Hence, when an automatic coding system is in production, the quality of its results has to be monitored and coding errors have to be used to update the application environment so as to prevent further errors of the same kind.

## 3. The construction of the automatic coding environment

Using ACTR requires a phase of **training**, which involves building the environment of the coding system. The first step of the training phase requires the construction of coding dictionaries (lists of texts with the corresponding codes). Afterwards the system has to be adapted to the language and to each classification; and, finally, it has to be tested.

The building of coding dictionaries **(reference files)** is the heaviest activity, as their quality and size deeply affect the performance of automated coding. Basically, it involves: (i) re-elaborating the textual descriptions used in classification manuals in order to make them simple, analytical and unambiguous; and (ii) integrating the classification dictionaries with information based on expert knowledge, with descriptions coming from other related official classifications and with empirical response patterns taken from previous surveys (in order to reproduce the respondents' natural language as closely as possible).

The already-mentioned **parsing** functions, which are managed through **parsing files**, allow the user to adapt the system to the language and to the classification. The implementation of these **parsing files** is very easy and does not require the user to be a computer expert.

As far as the adaptation to Italian is concerned, in all the applications we built, we decided to define as irrelevant the articles, conjunctions and prepositions, and we removed suffixes which determine singular and plural. Only in considering occupations was it necessary to remove the gender suffixes too. On the other hand, the definition of synonyms, both at string and at word level, is a job that requires more effort, since the classification is complex and answers can vary in their 'wording'. In order to clarify this aspect with figures, well over 2 000 synonyms were necessary when occupation type was considered, whereas just 287 have been defined for education level.

Up to now, we have trained the system to work with three variables: occupation, industry and education level. Each variable shows a different level of complexity, due to the corresponding classification complexity and to the expected variability in the 'wording' of answers (both these aspects influence the results of automated coding, as confirmed by the experiences in other countries).

The benchmark files which we used to train the system for the three mentioned variables were a sample of 9 000 households drawn to perform a quality survey on the 1991 population census and a sample drawn from the intermediate census of industry and services ('short-form survey'). To train ACTR, we ran it repeatedly on these samples, selecting each time the empirical answers to be added to the dictionaries and, at the same time, improving the **parsing** process until the highest possible number of correct unique matches was reached. The rates of matching (answer phrase: single code) obtained at the end of the runs were: 72.5 % for occupation; 86.6 % for education level; 54.5 and 73.0 % respectively for industry on the first and second sample (this difference is due to households' difficulty in answering this question). Hence they were in line with the results obtained by other countries (Lyberg and Dean, 1992).

## 4. First results of automated coding

After training the system, it needs to be tested in order to verify if the application environment, built using small samples, is suitable to be used for data sets of bigger size. For this purpose, the quality of automated coding has to be measured in terms of **recall**, i.e. the percentage of codes automatically assigned, as well as in terms of **precision**, i.e. the percentage of correct codes automatically assigned.

Table 1 shows the results obtained in terms of **recall** on data collected in the 1994 health survey, the 1998 labour force survey (four quarters collected and already manually coded), the 1999 labour force pilot survey and the 1998 intermediate census of industry and services ('long-form survey'). These results are consistent with those obtained during the system training.

*Table 1: Some results on recall of automatic coding*

| Source of texts | Occupation | | Industry | |
|---|---|---|---|---|
| | Number of texts | Recall (%) | Number of texts | Recall (%) |
| 1994 health survey | 33 735 | 72.3 | — | — |
| 1998 labour force survey | 356 231 | 71.9 | — | — |
| 1999 labour force pilot survey | 1 307 | 67.6 | 1 252 | 44.6 |
| Long-form survey | — | — | 37 161 | 63.0 |

As far as **precision** is concerned, with the aid of expert coders who analysed all the automatically assigned codes, it was possible to achieve the results shown in the following table.

*Table 2: Precision of automatic coding* ([1])

| Source of texts | Occupation | | Industry | |
|---|---|---|---|---|
| | Uniquely coded | Precision (%) | Uniquely coded | Precision (%) |
| 1994 health survey | 24 404 | 97.0 | — | — |
| 1999 labour force pilot survey | 884 | 99.0 | 558 | 86.0 |

([1])Evaluation of **precision** for the long-form survey is still in progress.

It was not possible to do the same thing in the labour force survey, due to its great amount of texts (256 748 texts coded as **unique**). Here, the **precision** can be evaluated only on a sample basis; hence, the need to build a system to monitor the quality of automatic coding, which can determine the extraction of samples of texts that have to be submitted to expert coders (see Section 5).

## 5. Monitoring and enhancing the quality of automatic coding of a great amount of texts

We analysed the textual answers for the 1998 labour force survey (four quarters collected) with the purpose of: (i) thoroughly evaluating the performance of the automatic coding; (ii) building up a quality monitoring system; and (iii) doing further training of the coding environment, the main purpose of which is to enrich the dictionary with new texts.

As a first step, we quantified how many 'different' texts existed in the original file and defined some frequency classes, so as to evaluate the performance of the system, class by class. To identify the 'different' texts, we performed a kind of 'raw standardisation' with only a few **parsing** functions, so as to delete from descriptions the articles, conjunctions, prepositions and suffixes (in practice, all the elements that determine the gender of words, the singular/plural, etc.). As can be seen in Table 3, the initial 356 231 texts were reduced to 59 561 different ways of describing the occupation. On the other hand, 74 % of these descriptions occurred only once in the original file, thus proving a high variance in wording of answers, if compared with the 6 319 official elementary definitions derived from just 599 occupations listed in the classification manual.

*Table 3: Distribution of 'different' texts by classes of occurrence*

| Original texts | 'Different' texts | Occurrence | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3–10 | 11–50 | 51–1 000 | 1 001–10 000 |
| 356 231 | 59 561 | 43 349 | 7 344 | 6 404 | 1 783 | 640 | 41 |
| | (100.00 %) | (73.78 %) | (12.33 %) | (10.75 %) | (2.99 %) | (1.07 %) | (0.07 %) |

## 5.1. Evaluating the performance of the automatic codingenvironment

A primary indicator of the performance of the automatic coding environment is achieved by comparing its **recall** on the original data set (the one with all non-parsed texts) with that of 'different' texts. Obviously the system **recall** on this latter file is lower, as can be seen in Table 4.

*Table 4: ACTR results on 'different' texts: recall*

| ACTR output | Recall | |
|---|---|---|
| | Number of texts | % |
| Unique | 19 404 | 32.5 |
| Multiple | 20 537 | 34.5 |
| Failed | 19 620 | 33.0 |
| Total | 59 561 | 100.0 |

**Recall** grows as frequency class becomes higher (Table 5). In particular, for 'different' texts occurring only once, ACTR assigned a unique code in 27 % of cases, while for texts occurring more than 100 times, this rate goes beyond 79 %. This means that the actual **reference file** already includes most of the occupation descriptions that occur frequently in common speaking.

*Table 5: ACTR results on frequency classes of 'different' texts: recall*

| ACTR output | Occurrence | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2–10 | | 11–100 | | 101–1 000 | | 1 001–10 000 | |
| | Number | % | Number | % | Number | % | Number | % | Number | % |
| Unique | 11 786 | 27.2 | 5 869 | 42.7 | 1 437 | 69.0 | 273 | 79.6 | 39 | 95.1 |
| Multiple | 15 735 | 36.3 | 4 303 | 31.3 | 431 | 20.8 | 66 | 19.2 | 2 | 4.9 |
| Failed | 15 828 | 36.5 | 3 576 | 26.0 | 212 | 10.2 | 4 | 1.2 | 0 | 0.0 |
| Total | 43 349 | 100.0 | 13 748 | 100.0 | 2 080 | 100.0 | 343 | 100.0 | 41 | 100.0 |

## 5.2. Lack of standardisation in the manual coding process

The quality of automated coding can be further evaluated by comparing it with the level of standardisation in the manual coding process.

As the labour force data were previously manually coded, we could quantify the different codes assigned by manual coders to the same text (Table 6).

*Table 6: Lack of standardisation in manual coding*

| Texts frequency classes | Different codes assigned to 'equal' texts | | | |
|---|---|---|---|---|
| | Maximum number | Mean | Median | Mode |
| 2 | 2 | 1.27 | 2 | 1 |
| 3–5 | 5 | 1.84 | 3 | 1 |
| 6–10 | 10 | 2.68 | 3 | 1 |
| 11–50 | 33 | 4.65 | 4 | 2 |
| 51–100 | 42 | 10.05 | 8 | 4 |
| 101–1 000 | 119 | 18.65 | 14 | 7 |
| 1 001–10 000 | 389 | 67.46 | 51 | 33 |

The results in this table show how low the level of standardisation of manual coding is. The discrepancy between codes assigned by different operators can usually be ascribed to different interpretations of the response text, to different knowledge of the classification and to misunderstandings. On the other hand, there is surely a percentage of texts (which we could not quantify) to which operators assigned different codes in view of some other information taken from other correlated questions in the questionnaire (for instance, industry).

## 5.3. The system to monitor the quality

Given the characteristics of ACTR, the sample of $n$ 'different' texts to be checked has to be drawn from those uniquely coded with a score of less than 10 ($N = 13,821$). In fact, a text coded with a score of 10, corresponding to a direct match, has a correct code (unless there are some mistakes in the **reference file**).

We decided to use a stratified random sampling design to draw the sample. In practice, texts were first stratified according to their frequency of occurrence $M_j$; then, within each stratum, a simple random sample (without replacement) of texts was selected. The strata coincided with the previously defined classes of occurrences with exception of the '1 001–10 000' one, given that all its 39 'different' texts had a coding score equal to 10, i.e. they were all correctly coded.

In deciding the sample size, it is possible to choose between two different strategies: (i) to compute the overall sample size and then allocate it between the strata; or (ii) to compute the sample size independently for each stratum, according to the precision of estimates required in each of them.

With the first strategy, the overall optimal sample size can be approximately computed by using **Neyman allocation** (see, for example, Cochran, 1977, p. 105). In this circumstance, it is important to decide a priori how the sample should be allocated between strata. For

example, with proportional allocation, the sample is allocated according to the relative size of each stratum $W_h = N_h / N$. However, with the problem at hand, a better approach could be that of allocating the sample according to the relative sum of frequencies of 'different' texts in the same class, so as to sample more 'different' texts with higher importance.

The advantage of this procedure is that of computing directly the overall sample size, given an allocation criterion. The disadvantage is that for some strata the optimal sample size may be greater than the entire stratum size ($N_h$); here, one has to revise the allocation following Cochran (1977, p. 104).

The alternative strategy avoids this last problem; it involves deciding the optimal sample size independently from stratum to stratum using each time the following expression (see Cochran, 1997, pp. 75–76):

$$n_h^* = \frac{n_{0h}}{1 + (n_{0h} - 1)/N_h}$$

with

$$n_{0h} = \frac{z_{1-\alpha/2}^2 \tilde{\pi}_h (1 - \tilde{\pi}_h)}{d_h^2}, \qquad h = 1, 2, \ldots, L .$$

In this expression $\tilde{\pi}_h$ is the hypothesised **precision** of automated coding for texts belonging to class *h*; $d_h$ represents the overall margin of error allowed in estimating the unknown **precision**, $\pi_h$, of automated coding and *z* is the percentile of standardised normal distribution such that $\Pr(|\hat{\pi}_h - \pi_h| \geq d_h) = \alpha$.

Then, the overall sample size is achieved by summing up the so obtained optimal sample sizes: $n = n_1^* + n_2^* + \ldots + n_L^*$. The problem with this procedure is that *n* may easily explode if some $n_h^*$ values are too large.

We used this latter strategy in deciding the size of the sample of text to submit to expert coders. An equal **precision** rate of automated coding in each class, $\tilde{\pi}_h = 0.75$, was hypothesised, while the margin of error *d* was progressively reduced (fourth column of Table 7) in higher classes of occurrences of 'different' texts; this guaranteed estimates with a higher **precision** for heaviest 'different' texts. The approximate sample size computed for various classes with $\alpha = 0.05$ ($z_{0.975} = 1.96$) can be found in the fifth column.

*Table 7: Optimal sample sizes in the strata*

| Classes of occurrences | Number of different texts ($N_h$) | Hypothesised precision of automated coding ($\tilde{\pi}_h$) (%) | Margin of error ($d_h$) (%) | Approximate optimal sample size ($n_h^*$) | Sampling fraction ($f_h = n_h^*/N_h$) (%) |
|---|---|---|---|---|---|
| 1 | 10 007 | 75.0 | ± 5.0 | 148 | 1.48 |
| 2 | 1 756 | 75.0 | ± 5.0 | 138 | 7.86 |
| 3–5 | 1 187 | 75.0 | ± 4.5 | 160 | 13.48 |
| 6–10 | 473 | 75.0 | ± 3.0 | 222 | 46.93 |
| 11–50 | 349 | 75.0 | ± 2.5 | 221 | 63.32 |
| 51–100 | 33 | 75.0 | ± 1.0 | 33 | 100.00 |
| 101–1 000 | 16 | 75.0 | ± 1.0 | 16 | 100.00 |
| Total | 13 821 | | | 938 | 6.79 |

The sample of 938 texts was then submitted to expert coders, in order to evaluate if ACTR had assigned correct codes. In this way, it was possible to estimate **precision** for each class of occurrences and hence for all the 13 821 'different' texts. The estimates, computed using the theory of stratified random sampling (see Cochran, 1977, pp. 90–96), can be found in Table 8, with the corresponding values useful to derive the 95 % confidence interval (last column of the table).

As can be seen, we estimated that 75.77 % of the 13 821 'different' texts were correctly coded by ACTR. True **precision** lies between 70.58 % ($= 75.77 - 5.19$) and 80.95 % ($= 75.77 + 5.19$) approximately with a probability of 0.95. The **precision** tends to be higher (over the 80 %) for the last classes. Notice that for the last two classes we do not have an estimate but the true **precision**, as all texts (rather than a sample) were checked. Here the coding **precision** is over the 80 % and this further proves that the system works well with more frequent descriptions.

*Table 8: Estimated precision of automatic coding of different texts*

| Classes of occurrences | 'Different' texts | Sample size | Sampling fraction (%) | Estimated precision (%) | Estimated margin of error |
|---|---|---|---|---|---|
| 1 | 10 007 | 148 | 1.48 | 74.32 | ± 6.99 |
| 2 | 1 756 | 138 | 7.86 | 81.88 | ± 6.17 |
| 3–5 | 1 187 | 160 | 13.48 | 78.13 | ± 5.96 |
| 6–10 | 473 | 222 | 46.93 | 73.42 | ± 4.23 |
| 11–50 | 349 | 221 | 63.32 | 80.09 | ± 3.19 |
| 51–100 | 33 | 33 | 100.00 | 87.88 | — |
| 101–1 000 | 16 | 16 | 100.00 | 81.25 | — |
| Total | 13 821 | 938 | 6.79 | 75.77 | ± 5.19 |

If we also consider the 6 083 ($=19,904-13,821$) 'different' texts coded with a score of 10 (all correctly coded), the overall estimated **precision** goes up to 83.17 % of 19 904 'different' texts.

The estimated **precision** of automated coding when applied to original texts can easily be derived from that of the 'different' texts, by considering the occurrences of these latter (Table 9). In practice, each 'different' text can be viewed as a cluster of original texts and the theory of cluster sampling allows us to derive the estimates of **precision** and the corresponding 95 % confidence intervals reported in the table below.

*Table 9: Estimated precision of automatic coding of original texts*

| Classes of occurrences | 'Different' texts | Original texts | Estimated precision (%) | Estimated margin of error |
|---|---|---|---|---|
| 1 | 10 007 | 10 007 | 74.32 | ± 7.01 |
| 2 | 1 756 | 3 512 | 81.88 | ± 6.19 |
| 3–5 | 1 187 | 4 337 | 78.34 | ± 6.55 |
| 6–10 | 473 | 3 492 | 73.40 | ± 4.52 |
| 11–50 | 349 | 7 320 | 86.29 | ± 5.08 |
| 51–100 | 33 | 2 214 | 87.49 | — |
| 101–1 000 | 16 | 3 731 | 81.96 | — |
| Total | 13 821 | 34 613 | 79.70 | ± 2.57 |

It is estimated that the 79.70 % (27 586 texts) of the 34 613 original texts uniquely coded with a score of less than 10 were coded correctly. The true **precision** lies between 77.13 % ($=79.70-2.57$) and 82.27 % ($=79.70+2.57$), with an approximate confidence of 0.95. Here, too, if we consider the 222 135 original texts uniquely coded with a score equal to 10, it comes out that 249 721 of the 256 748 original texts uniquely coded had a correct code (i.e. 97.26 %). This last estimate is in line with that obtained for the 1994 health survey (see Table 2).

Thus, with a small but well-designed sample (in this case 6.79 % of single texts), it was possible to evaluate the precision of automated coding results with a high confidence.

## 5.4. First results of the further training phase of the coding environment

The further training phase consists of adding new texts to the dictionary and updating the coding environment: (i) to prevent further texts being processed as coding errors found; and (ii) to increase the future **recall** rates.

To prevent further coding errors, the sample of 'different' texts for which ACTR did not assign a correct code, as determined by expert coders, needs to be analysed.

In order to increase the future **recall** rate, texts for which ACTR was not successful in assigning a single code need to be examined, including coded texts having enough informative content to be assigned a unique code (i.e. those which are not too generic, or which describe concepts which can be directly linked with single codes). In this regard, it is convenient to examine first the more frequent ones, while the analysis of texts belonging to lower frequency classes, given their minor importance, can be restricted to only a sample.

We analysed the failed matches returned by ACTR when coding the file of 'different' texts occurring more than 10 times. By analysing only 216 different texts (212 belonging to the '11–100' class of occurrences and 4 to the '101–1 000' one), we added 299 new texts to the **reference file** and 46 synonyms (at both the level of string and of the word).

The **recall** rates obtained on the original text data set after this further training activity are shown in Table 10.

*Table 10: ACTR results on original labour force survey sample after further training: recall*

| ACTR output | Recall | |
|---|---|---|
| | Number of texts | % |
| Unique | 269 485 | 75.6 |
| Multiple | 58 848 | 16.6 |
| Failed | 27 898 | 7.8 |
| Total | 356 231 | 100.0 |

As can be seen, the percentage rises from 71.9 to 75.6 % and would likely to have been even higher if we had also analysed the multiple matches.

Finally, we verified if the update of the coding environment for occupation, achieved by analysing labour force descriptions, could imply better results for other coding applications performed on data from other surveys. For this purpose, we automatically coded again the health survey texts, as it was the next biggest file we had at our disposal after the labour force one. As shown in Table 11, the **recall** rate grows from 72.3 to 75.1 %, thus confirming that the outcomes of each coding application represent precious feedback for updating the coding environment and give the chance of achieving higher **recall** rates.

*Table 11: ACTR results on health survey sample after further training: recall*

| ACTR output | Recall | |
|---|---|---|
| | Number of texts | % |
| Unique | 25 337 | 75.1 |
| Multiple | 5 827 | 17.3 |
| Failed | 2 571 | 7.6 |
| Total | 33 735 | 100.0 |

# 6. Conclusions

As mentioned in the previous sections, ISTAT spent much work and time in order to introduce automated coding of written-in answers to open questions regarding occupation, education level and industry by means of the ACTR system. Most of the work involved building the **reference files** and the corresponding **parsing files** for both occupation and industry. The first results obtained in this direction (Section 4) were encouraging, especially if compared with those of manual coding, and led us to improve further the automated coding environment, using all available sources of textual descriptions to enhance the **reference files** and to refine the **parsing** step. Alongside this activity, we thought it was necessary to introduce an evaluation procedure so as to quantify the quality of ACTR output (Section 5). This procedure was kept as general as possible in order to get a reliable idea, even if on a sample basis, of how well ACTR codes texts that do not exactly correspond to descriptions of the **reference file**. We performed this evaluation step on a large amount of texts regarding occupation (Subsection 5.3) and the results obtained were particularly satisfactory (overall coding precision was estimated to be about 97 %).

All the work invested in ACTR training and the good results obtained in the testing/evaluation phase convinced us that it can successfully be adopted for use in different surveys, even to code such complex descriptions as the occupation and industry ones, giving more consistent results than those of manual coders. Moreover, these results seem to be achievable at a lower cost; gains are likely to increase with the amount of descriptions collected. In any case, the application of ACTR should constantly be monitored in all its phases.

Despite the advantages, it has to be kept in mind that the application of ACTR still presents a problem in cases where the system fails in assigning a unique code. Different solutions are available. One, for example, is that of trying to code by making use of additional information derived from related questions of the same form. This could be achieved automatically or with the intervention of expert coders, possibly aided by an assisted coding system. If this does not work, and if the number of unsolved cases is not high, it may be necessary to consider re-contacting the respondents. Therefore, further investigation is needed in order to choose the strategy that performs best in terms of number of cases solved, costs and quality of final results.

## 7. References

[1] Appel, M. and Hellerman, E. (1983), 'Census Bureau experience with automated industry and occupation coding', *Proceedings of Section on Survey Research Methods*, American Statistical Association, pp. 32–40.

[2] Chen, B., Creecy, R. and Appel, M. (1993), 'Error control of automated industry and occupation coding', *Journal of Official Statistics*, Volume 9, pp. 729–745.

[3] Cochran, W. G. (1977), *Sampling techniques*, Third edition, Wiley, New York.

[4] Corbett, J. P. (1972), 'Encoding from free word descriptions', unpublished manuscript, US Census Bureau.

[5] Creecy, R. H., Causey, B. D. and Appel, M. V. (1990), 'A Bayesian classification approach to automated industry and occupation coding', paper presented at the American Statistical Association's joint statistical meetings, Anaheim, CA.

[6] Creecy, R. H., Masand, B. M., Smith, S. J. and Waltz, D. L. (1992), 'Trading MIPS and memory for knowledge engineering', *Communications of the ACM*, Volume 35, pp. 48–68.

[7] De Angelis, R. and Macchia, S. (1998), 'Applying automated coding to the pilot survey of next population census: a challenge', paper presented at the Conference on New Techniques and Technologies for Statistics, Sorrento, Italy, 4 to 6 November 1998, pp. 309–314.

[8] Dumicic, S. and Dumicic, K. (1994), 'Optical reading and automatic coding in the census '91 in Croatia', paper presented at the Conference of European Statisticians, work session on statistical data editing, Cork, Ireland, 17 to 20 October 1994.

[9] Gillman, D. and Appel, M. V. (1994), 'Automated coding research at the Census Bureau', *Research Report*, No 4, US Census Bureau, Washington.

[10] Gillman, D. and Appel, M. V. (1999), 'Developing an automated industry and occupation coding system for the 2000 census', paper presented at a joint statistical meeting held in Baltimore, 1999.

[11] Harris, K. W. and Chamblee, R. F. (1994), 'Evaluation of an automated industry and occupation coding system', paper presented at a joint statistical meeting held in Toronto, 1994.

[12] Hellermann, E. (1982), 'Overview of the Hellerman I&O coding system', US Census Bureau internal paper, Washington.

[13] Kalpic, D. (1994), 'Automated coding of census data', *Journal of Official Statistics*, Volume 10, pp. 449–463.

[14] Knaus, R. (1987), 'Methods and problems in coding natural language survey data', *Journal of Official Statistics*, Volume 1, pp. 45–67.

[15] Lorigny, J. (1988), 'QUID: a general automatic coding method', *Survey Methodology*, Volume 14, pp. 289–298.

[16] Lyberg, L. and Dean, P. (1992), 'Automated coding of survey responses: an international review', paper presented at the Conference of European Statisticians, work session on statistical data editing, Washington, DC, 1992.

[17] Massingham, R. (1997), 'Data capture and coding for the 2001 Great Britain census', paper presented at XIVth Annual International Symposium on Methodology Issues, 5 to 7 November 1997, Hull, Canada.

[18] O'Reagan, R. T. (1972), 'Computer-assigned codes from verbal responses', *Communications from the ACM*, Volume 15, pp. 455–459.

[19] Rivière, P. (1994), 'Le système de codification automatique Sicore', working paper, Conference des statisticiens européens, Séminaire ISIS 1994, Bratislava.

[20] Tourigny, J. Y. and Moloney, J. (1995), 'The 1991 Canadian census of population experience with automated coding', paper presented at the United Nations Statistical Commission on Statistical Data Editing.

[21] Wenzowski, M. J. (1988), 'ACTR — a generalised automated coding system', *Survey Methodology*, Volume 14, pp. 299–308.

# Refining electronic data-collection instruments and data-dissemination tools through usability testing

Elizabeth Murphy, Kent Marquis,
Elizabeth Nichols, Kristina Kennedy and David Mingay

*Statistical Research Division*
*US Census Bureau* ([1])
*Washington, DC 20233, United States of America*
*E-mail: elizabeth.d.murphy@census.gov*

**Keywords:** Online questionnaires, information retrieval, user-interface evaluation

**Abstract**

*Electronic systems for collecting and disseminating data need usable interfaces if they are to be accepted by the general public. Ease of use cannot be assumed without conducting usability testing. Such testing identifies problems tha t actual users are likely to have in completing computer -based tasks. Methods of usability testing can be tailored to the available software development products. Some short case studies on usability testing at the US Census Bureau illustrate its value i n improving user interfaces for public use.*

## 1. Introduction

As government statistical agencies turn to electronic collection and dissemination of data, the role of usability engineering in user-interface design becomes increasingly important. Usability engineers seek to design computer-based systems for effectiveness, efficiency, and user satisfaction [1]. Exploratory testing conducted at the US Census Bureau indicates that electronic questionnaires and data-dissemination tools can benefit in many ways from iterative usability evaluations using relatively few test users.

### 1.1. Exploratory usability testing during system design and development

The purpose of exploratory usability testing is to identify problems that end-users are likely to have because of usability violations in the design of the user interface. Such violations may include misuse of colour, complexity of visual layouts, and lack of navigational cues. The earlier testing begins, the more likely it is that changes can be made to improve usability.

([1]) This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

Low-fidelity exploratory testing, often using paper mock-ups or screen shots, can occur during the early design phase of a user-interface design effort, before any code is written. Usability testing of software prototypes typically begins and continues throughout the development cycle. Testing right before a scheduled release is not encouraged unless project managers agree to correct serious usability violations. For each major class of end-users, at least three test participants are needed to identify most usability problems [2].

### 1.2. Usability issues for electronic data-collection and dissemination

Usability issues for electronic data collection include the challenges of going from a paper questionnaire to an interactive medium and the ease of navigating through forms (see [3], [4]). Usability issues for electronic data dissemination (and the user's retrieval of information) include the design's support for user tasks and mismatches between actual system behaviour and the user's expectations for system operation (see [4], [5]). We present examples from both perspectives to illustrate the value of usability testing for government statistical agencies.

## 2. Usability testing of electronic questionnaires

We have tested various Census Bureau data-collection web sites and computer self-administered questionnaires (CSAQs) for usability. Here, we focus on selected issues identified during exploratory testing of the electronic forms for the 2000 census of population and housing and the electronic forms for the US 2002 economic census.

### 2.1. Respondent's task: access the 2000 census Internet form

Usability testing of the prototype web site began about one year before the scheduled release date for the Internet form. For the first round of exploratory testing, we recruited 22 demographically diverse participants who were familiar with the Internet. Participants came from inside and outside the Census Bureau. We used three variations of form type, gave some participants incorrect identification (ID) numbers and/or invalid forms, and turned off JavaScript to disable the browser for some participants. The purpose of these manipulations was to determine whether respondents would be able to recover from unfavourable conditions and still supply their census data, either online or on paper.

Of particular concern was the issue of confirming an online respondent's identity. To prevent fraud and maintain confidentiality, Internet respondents were required to enter a 22-digit housing unit identification number printed on the paper version of the questionnaire (e.g. 00224-2571156-28-120-112-93). During the first round of exploratory usability testing, 55 % of participants had trouble finding their official ID number. Once they had found it, 90 % of participants had no trouble entering the ID number. This high success rate belied some expectations that entering the ID number would be too difficult.

No changes were recommended in the design of the entry fields, but we recommended revising the instructions to tell respondents where to look for the mailing label containing the ID number. The developer added these textual instructions along with a graphic illustrating where to find the ID number (Figures 1 and 2).

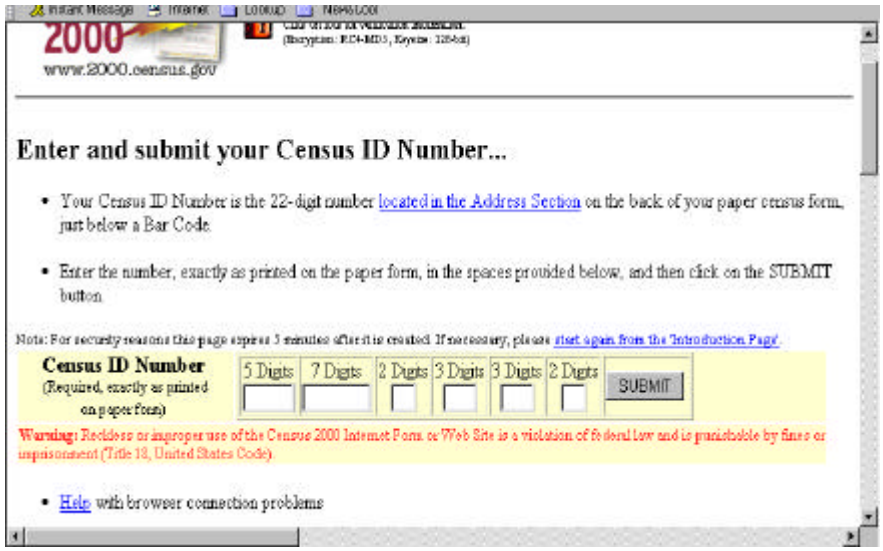



*Figure 1: Textual instructions and fields for finding an d entering the 2000 census ID number*

The graphic shown in Figure 2 appeared when the respondent clicked on the link embedded in the instructions ('located in the Address Section'), as shown in Figure 1.

For the second round of usability testing, we could not use outside participants because of heightened concern for the security and integrity of the census 2000 Internet form. We conducted an exploratory usability test with 12 census employees. The data and comments they submitted were merged with data and comments collected from 268 additional participants in load-and-usability testing that was open to all employees of the Census Bureau. Results of laboratory testing indicated that the instructions shown in Figure 1 were enough to help the participants find their census ID numbers. None of these participants clicked on the link to the ID-location graphic. Several commented on other aspects of the design of the census ID entry area or on the difficulty of entering the long ID number, but all laboratory participants who received a valid paper form were able to enter their ID numbers successfully once they had found it.

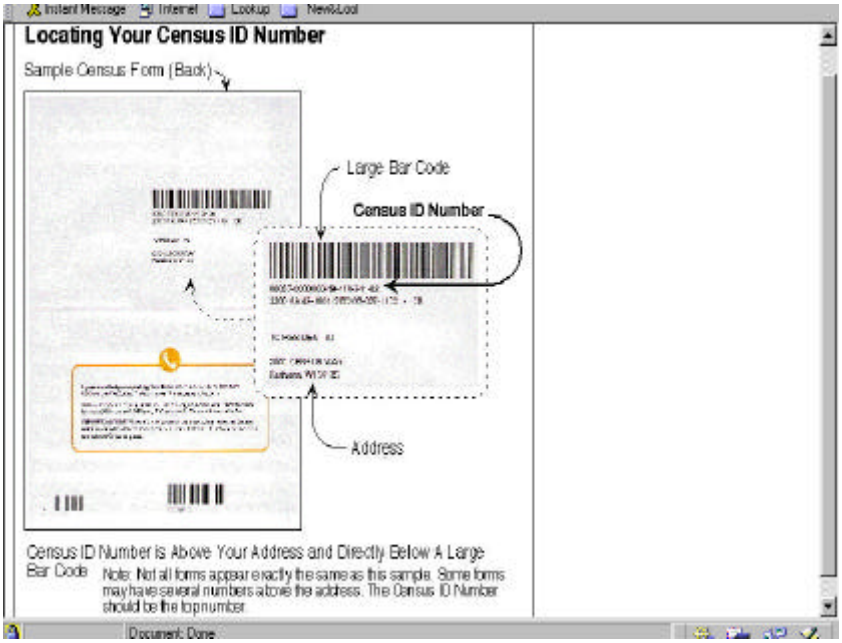



*Figure 2: A graphic to help respondents find the census 2000 ID number for their housing unit*

In the load-and-usability testing, participants were given a list of valid census ID numbers from which to pick a number. Thus, they did not need to locate their number on the back of a paper form. None of these participants reported being unable to enter their number, although some mentioned problems associated with the census ID number, such as the sheer length of the number and no indication of entering an incorrect number. Suggesting that actual respondents might rarely access 'Help', only two people commented on the graphic help for locating the census ID number.

The second round of testing identified several serious usability violations. Among these was an error message that appeared when some of the participants tried to submit their census ID numbers. This problem was observed in the laboratory testing and reported by many participants in the load-and-usability testing. Although the solution was to click again on the Submit button, our concern was that some respondents might give up if they received an error message on their first try. We also discovered by observation that pressing the Enter key after entering data would take the respondent out of the form and cause the loss of any entered data. We recommended investigating and resolving these problems. The development team recognised these as high-priority issues. These problems were resolved before the site went online.

In the actual census 2000 environment, about 83 % of all attempted ID submissions were successful [6]. Possible reasons for failed attempts included not only keying errors, but also lower-than-required levels of encryption and forgetting to check the disclaimer box indicating acceptance of the security warning. Attempts to misuse the system, for example by entering numbers at random, could also have contributed to failed ID submissions [6].

Despite the absence of advertising, over 66 000 unique Internet forms were processed [6]. Server logs recorded 7 017 visits to the ID-location graphic (Figure 2) during March and April 2000. Although visits cannot be equated with the number of respondents ([2]), the number of visits to the ID-location graphic suggests that it helped some respondents find their census ID numbers.

By identifying problems that the test respondents had, both rounds of usability testing contributed to the success of the US census 2000 Internet form. The recommendations implemented to resolve these problems presumably improved the design of the web site's user interface, making it easier for respondents to submit their census data over the Internet. This example illustrates the usefulness of usability testing of all aspects of a web application, including the log-in process. It also shows that expert judgments about usability can be wrong, as was the expectation that respondents would not be able to enter a 22-digit identification number. It is always best to conduct usability tests, as well as to solicit and use expert design advice.

## 2.2. Respondent's task: act on electronic edits in the 2002 economic census

Respondents to the 2002 economic census will have an electronic reporting option, in the form of a downloadable questionnaire. In the electronic mode, an edit message will be presented to respondents when entered data fail an automated checking function. Several cognitive and usability issues are associated with automatic edit messages. Cognitive issues largely involve the potential effects of wording in conveying the message to the respondent, including the nature of the problem and guidance on how to solve the problem. Usability issues centre largely on the method of presenting the messages, for example whether to use dialogue boxes and whether to present messages immediately when edit-check failures occur, or in a list at the end of the form (deferred presentation). Some combination of immediate and deferred edit messages may be best.

In March 2001, we tested various types of edit messages using wording drafted by a style-guide development group at the US Census Bureau. To test the messages in a realistic context, we developed a paper questionnaire modelled after an online mortgage questionnaire. In order to easily create situations that would fail edits, we prepared a partially completed questionnaire containing some discrepancies.

([2]) The number of visits to a web page is not necessarily equal to a comparable number of unique visitors (D. Coon, e-mail, 12 April 2001).

Participants were asked to pretend that they had completed it and, later, to make corrections. We used low-fidelity paper mock-ups, since software prototypes were not yet available.

Three messages were the result of failing preventive edits — that is, the messages were presented immediately upon an edit-check failure. We presented these edits to participants as pop-up dialogue boxes on paper. For example, as the participant proceeded through the questionnaire, reading the questions and pre-filled answers, he or she came to the following item.

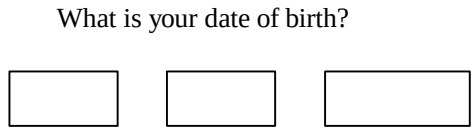



*Figure 3: Sample question used in testing electronic edit messages*

The partially completed questionnaire had no value in any of the date fields and did not give a format for data entry. The test administrator said that the respondent's birthday occurred in June, and the participant entered a 'J' for June. Then, the test administrator, acting as the computer, presented the following edit message on paper.

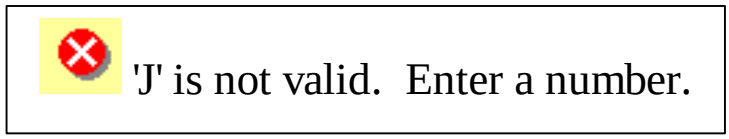



*Figure 4: Example of a preventive edit in low-fidelity usability testing of electronic edit messages for the 2002 economic census*

All participants understood this message. Most claimed they would enter either a 6 or 06 for the month, but they did not know which format to use. They wanted the format to be specified on the questionnaire. We recommended indicating the required format for dates on the questionnaire (e.g. MM/DD/YYYY). The design practice of displaying the format for the user follows long-standing user-interface design guidelines (e.g. [7], [8]).

We recognise the limitations to generalisation of the results of the exploratory methodology used in testing the edit messages: (i) the questionnaire tested was not an economic census form; (ii) the participants were not actual, corporate respondents; (iii) the

situation was manipulated; and (iv) neither the form nor the edits were electronic. However, there is ample precedent for the use of low-fidelity methods early in the design cycle (e.g. [9], [10], [11]). Invaluable information can be gained from early testing, using low-fidelity, paper-based prototypes.

## 3. Usability testing of web-based data-dissemination tools

On the data-dissemination side, we have used exploratory usability-testing methods to evaluate several Census Bureau web sites for usability. Here, we focus on selected issues associated with the American FactFinder and the project management repository, a Census Bureau intranet site.

### 3.1. User's task: select a geographical area of the United States in formulating a data request

The American FactFinder (AFF) is a major Census Bureau data-dissemination tool, available at http://www.census.gov. In mid-1999, prior to public release, the AFF user interface required users to perform a series of steps in a precise order and used hierarchical lists, or tree widgets. One use of a tree widget was to provide users with a list of geographical areas for selection. In formulating a query, the user needed to select a geographical area (e.g. one of the states) and then request further data for that area. Figure 5 shows a partial set of tree-buttons used for selecting one state.


*Figure 5: Use of a tree widget in an American FactFinder prototype*

In the exploratory usability testing, most of the 10 test participants understood how the tree-buttons functioned. Some, however, had never worked with a tree and needed an

explanation of the pluses and minuses. The six participants who took the AFF tutorial prior to testing received an explanation there, along with explanations of many other features. It was questionable whether they remembered these explanations later. The four who did not take the tutorial had to figure out how to expand and close the trees on their own, if they were not already familiar with this feature. Instructions about expanding and contracting trees were confusing. Many test participants had trouble making the appropriate geographical selection. We suggested adding explicit instructions to guide users in making geographical selections and noted that navigation needed to be simplified. The current online version of AFF does not use hierarchical lists. Instead, users can select a geographical area from a list box, as shown in Figure 6.

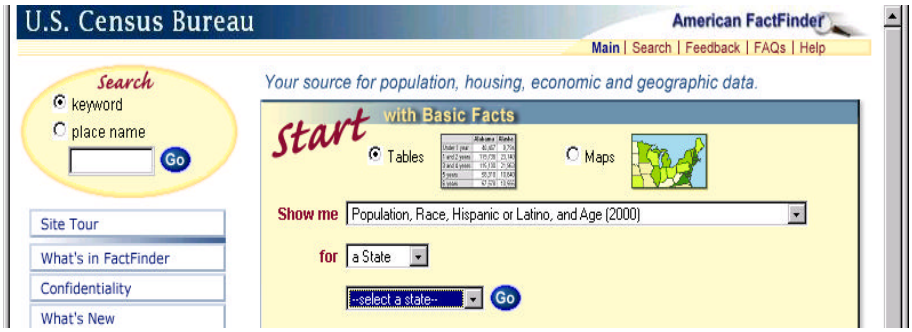



*Figure 6: Current method for selecting a geographical area in the American FactFinder*

When the user clicks on the Go button next to the 'select a state' list box, an alphabetical list of the states is displayed. The user can then select one state by clicking on the state's name. This is a more straightforward way for the user to inform AFF about the geographical area of interest.

Although some usability problems remain, this version of the AFF user interface simplifies the task by giving the user a list from which to pick, instead of a tree hierarchy. The value of simplifying the user interface cannot be overstated.

## 3.2. User's task: find a document on project management

Usability testing of the project management repository (PMR) site took place in early 2001. The purpose of this census intranet site is to provide ready access to a collection of documents on project management, as well as a private, collaborative workspace for a project team. Typical users are trained in project management and are familiar with the documents posted at the PMR site. The evaluation method was exploratory, scenario-based usability testing, with participants thinking aloud — i.e. verbalising what they were doing and why they were doing it, as well as describing any difficulties they

may have had.

The PMR site provides the user with both 'basic' and 'advanced' functions for searching the database of documents. Because the basic search function was not working during the testing period, participants were told to use the advanced search, as shown in Figure 7. Participants could have entered a search term in the initial search field, accepted the other defaults and initiated a search, but they all explored the other categories and choices presented to them.

When asked about the difference between 'contains all' and 'contains any' as a search qualifier, some participants misinterpreted their meanings. The participant who was the most familiar with the Internet thought that 'contains all' was a broad query and that 'contains any' would be broader. The reverse is actually true in a Boolean context. The growing literature on search behaviour at web sites indicates that users in general are not familiar with Boolean logic and find it difficult to use in constructing queries (e.g. [12]).

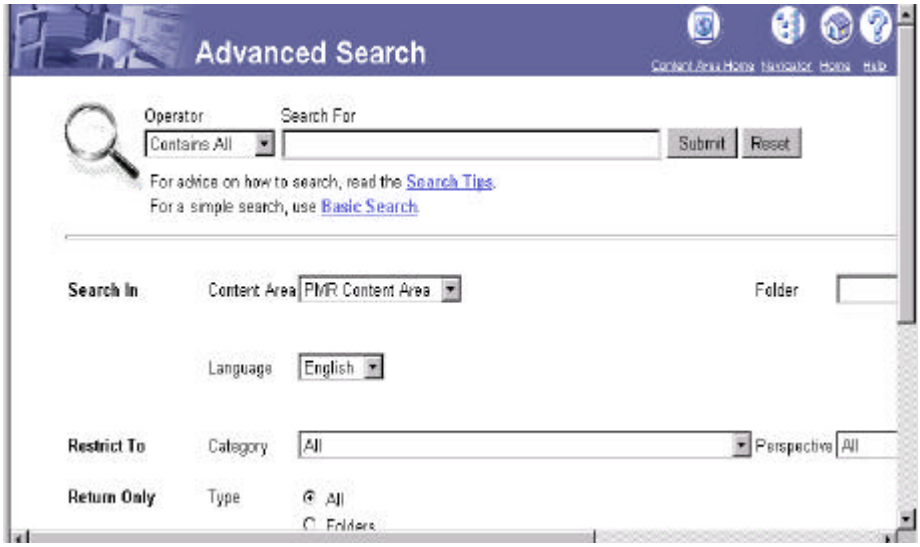



*Figure 7: Beginning of the advanced search in the project management repository*

We recommended a redesign of the advanced search function to make it more consistent with the way people think about finding a document. At the time of writing, the basic search function is operational, but plans call for a redesign of the advanced search, with usability testing to be repeated during the summer of 2001.

We plan to conduct further research to improve the usability of the search features used in data-dissemination tools.

## 4. Conclusions

Our experiences with exploratory usability testing at the US Census Bureau reflect the need for user-centred, not technology-centred, design and development of user interfaces. We have seen that providing even minimal guidance to potential respondents (e.g. where to find their ID number) makes a difference to the respondents' ability to provide important demographic data over the Internet. We have demonstrated the benefits of using paper prototypes in testing even before any software prototypes become available. If nothing else, we have shown the value in keeping user interfaces as simple as possible so that the user does not become confused or disoriented.

Our experience underscores the importance of focusing on the end-user early in the design process. Early and iterative usability testing can identify problems that can be designed out of the user interface before release of the electronic questionnaire or information-retrieval tool. When confirmatory (acceptance) testing is conducted near the end of the development cycle, usability goals are more likely to be met if design has been user-centred and usability has been a concern from the start.

## 5. References

[1] Nielsen, J., *Usability engineering*, Academic Press, New York, 1993.

[2] Dumas, J. S. and Redish, J. C., *A practical guide to usability testing*, Ablex, Norwood, NJ, 1993.

[3] Nichols, E. and Sedivi, B., 'Economic data collection via the web: a Census Bureau case study', *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 1998, pp. 360–365.

[4] Murphy, E., Marquis, K., Hoffman, R. III, Saner, L., Tedesco, H., Harris, C. and Roske-Hofstrand, R., 'Improving electronic data collection and dissemination through usability testing', *Proceedings of the Federal Committee on Statistical Methodology Research Conference* (Statistical Policy Working Paper 29, Part 5 of 5), 1999, pp. 117–126. Available from NTIS (PB99-166795) and at http://www.bts.gov/fcsm.

[5] Tedesco, H., Brady, M. and Ciochetto, S., 'Data-centric or user-centric? A case study of the design of an Internet data-dissemination system', *Proceedings of the Third ASC International Conference*, Edinburgh, Association for Survey Computing, Chesham, Bucks, UK, 1999, pp. 383–392.

[6] Coon, D., 'Implementing a web-based reporting option for the 2000 US census of population and housing,' presentation slides prepared for Fedcasic 2001, US Census Bureau, Computer-Assisted Survey Research Office (CASRO), Washington, DC, 2001.

[7] International Organisation for Standardisation (ISO), *Ergonomic requirements for office work with visual display terminals (VDTs) — Part 10: Dialogue principles* (ISO 9241-10), ISO, Geneva, 1996.

[8] Smith, S. L. and Mosier, J. N., *Guidelines for designing user-interface software* (ESD-TR-86-278/MTR 10090), MITRE, Bedford, MA, 1986.

[9] Fleming, J., *Web navigation: designing the user experience*, O'Reilly, Sebastopol, CA, 1998.

[10] Hix, D. and Hartson, H. R., *Developing user interfaces: ensuring usability through product and process*, Wiley, New York, 1993.

[11] Weinschenk, S., Jamar, P. and Yeo, S. C., *GUI design essentials*, Wiley, New York, 1997.

[12] Spink, A., Milchak, S., Sollenberger, M. and Hurson, A. R., 'Elicitation queries to the Excite web search engine', *Proceedings of the Ninth International Conference on Information Knowledge*, 2000.

# Central metadata system (CMS) in the statistical production and publication environment

Reinhard Karge (*) and Lars Rauch (**) ([1])

(*) *Run-Software Werkstatt GmbH*
*Koepenicker Straße 325, D-12555 Berlin*
*E-mail: reinhard.karge@run-software.com*

(**) *Statistics Sweden*
*Karlavägen 100, S-104 51 Stockholm*
*E-mail: lars.rauch@scb.se*

**Keywords:** object orientation, metadata, data exchange, statistical production, statistical publication, Bridge[NA], classification database, XML, Internet

**Abstract**

*The paper presents strategic aspects in using central metadata systems. It demonstrates the advantages of a central metadata system as a primary metadata source. The paper defines requirements on a central metadata system regarding the content to be covered and connectivity aspects. The base for practical experiences is the statistical metadata system Bridge [NA], which is based on the outcome of the IMIM project of the fourth framework programme. Bridge [NA] is used as a metadata system in some European statistical offices. First practical experiences with the implementation of Bridge [NA] for a classification database at Statistics Sweden are presented.*

## 1. Role of central metadata systems

The development of central metadata repositories is on the agenda in several national and international statistical organisations. There is no doubt today about the importance of using metadata bases in the statistics production system and, in particular, for the dissemination to users. What can often be observed is that statistical metadata bases have been developed or are under development for different purposes and main statistical applications. There are metadata bases for statistical databases, for Internet dissemination of statistics, for subject-matter-oriented projects, for documentation of surveys, etc.

One crucial problem is that these metadata bases are more or less isolated from each other or only loosely coupled. Such a situation causes problems concerning office-wide harmonisation and consistency of metadata. A well-organised central metadata resource could avoid many such problems. On the other hand, there is a certain danger that to have only **one** metadata base for the whole national statistical organisation (NSO) could be an

([1]) The paper should not be considered as the official position of Statistics Sweden. It expresses only the opinion of the authors.
L Rauch is a staff member at Statistics Sweden.

extreme solution that would not be able to serve in an efficient way all the different needs of metadata in an office. Therefore, we want to speak about a central metadata repository. A central metadata repository should have the following functions.

- It should store all metadata that are of comprehensive interest to the office. This means that metadata, which are only of interest locally for very special purposes, need not be stored and maintained centrally.

- It should serve as a common source for other distributed metadata bases in the production process.

- It should serve as a general, metadata-supported entry point for statistical users to obtain information on the content of the statistical information system and support the access to the data.

A central metadata repository may be organised in different technical ways, but it would be preferable to organise it as a database.

Figure 1, developed by Bo Sundgren (Statistics Sweden), may serve as a good guideline for the design of a metadata-supported production system. It demonstrates the statistics production system and the position and role of the metadata system.

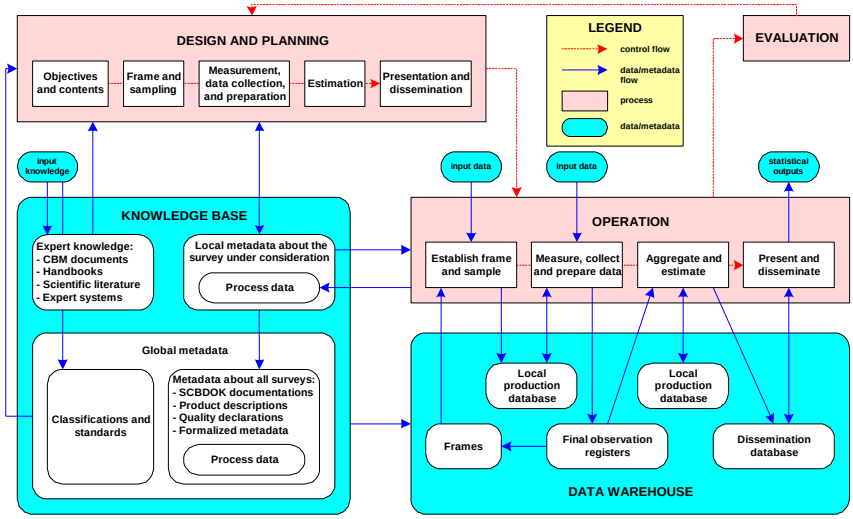



*Figure 1: Statistics production schema*

This figure shows the metadata system — the knowledge base — in the statistics production system. The knowledge base has its own internal structure. The main structure

is the division into global metadata and local metadata. This division is of highly practical importance. Local metadata are linked to special activities carried out in the statistical office. Local metadata may be organised in a different way to global metadata. But it has to be stressed that global metadata are a main source for local processes, as indicated by the link between global metadata and the design process. Global metadata should be the only resource for the output part of the production process that is presented as the data warehouse in the figure. The global metadata are the part which we want to consider as a central metadata repository.

Statistical metadata consist of very different kinds of metadata for different purposes. Statistical metadata can be classified in a number of ways. They can, for instance, be divided into:

- physical and operational (technical) metadata necessary for software processes (e.g. variable codes, position of variables in records, data types, etc.);

- metadata related to publications, as, for instance, tables, books, etc. Typical metadata for that would be footnotes and readable labels for all kinds of data (column names, value names in tables, footnotes, etc.);

- metadata related to the content of statistics (e.g. definition of variables, description of surveys — SCBDOK ([2]), data quality statements, etc.);

- metadata related to administrative issues around the statistics (e.g. the organisational structure of the NSO, responsible contact persons, administrative division of statistics into subject-matter areas, etc.

These examples should only demonstrate that metadata are very heterogeneous in their functions, representation, and usage. A central metadata repository must reflect this broad spectrum of metadata.

This central metadata repository will not only be used in the statistics production process but it should be the main entry point for external users to the statistical information system as such.

In this respect, it would be advisable to differentiate between the metadata repository as a storage vehicle and the tools to access the metadata, both for other system components and users. Such a differentiation will provide a higher degree of flexibility in the development of the metadata storage and the adoption of new user requirements to work with metadata.

([2]) SCBDOK — Statistics Sweden's documentation system for surveys, administrative register, etc.

Figure 2 demonstrates this facility from a general point of view. It should also be mentioned that it would be an advantage to have separate access tools for accessing output databases, data warehouses, etc. It is important to understand that different users will have different objectives regarding how to work with a statistical information system. The access tool in the figure is very general and includes, for instance, Internet access. In this paper, however, the metadata part will be considered in more detail.

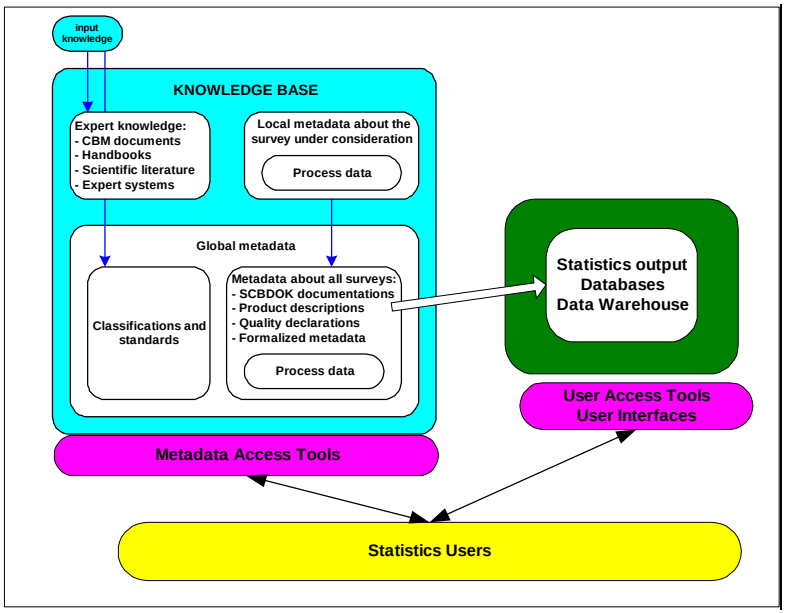



*Figure 2: User access to metadata and data*

It is essential that metadata are entered into the metadata system only once. The system should guarantee and support the use of metadata for different purposes. Moreover, as much metadata as possible should be generated from the conceptual level down to the technical level (see also Section 3). Metadata stored in a central repository have to be available for specific metadata bases. It may be advisable — for practical and efficiency reasons — to keep and maintain specific metadata bases for a public database. But that should mainly be limited to such metadata that are necessary for the physical access to the data, i.e. metadata that are used by software.

As a result, we may get a number of metadata bases linked together with the central metadata repository as a spider in the web (Figure 3).

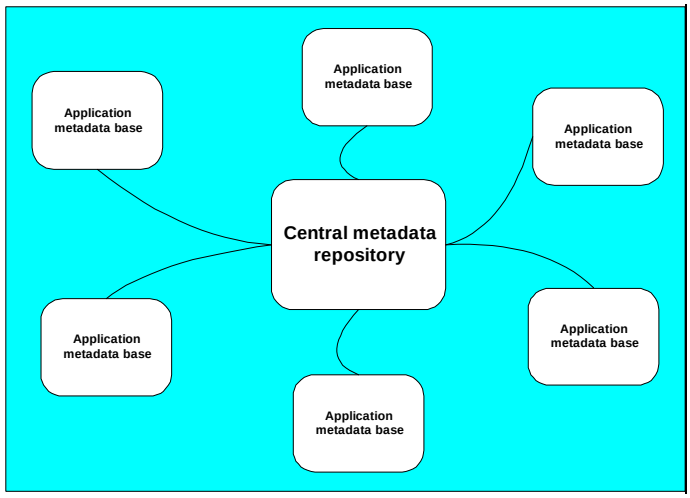



*Figure 3: Central metadata repository and distributed application -oriented metadata bases*

There is also another important aspect that has to be taken into consideration. Today, a net of NSOs and international statistical organisations is under development. The exchange of statistical data and the delivery of data to international statistical organisations for compiling international statistics require a system of metadata that is linked together. It must be possible to have mutual access and to exchange metadata between the different systems. At the end, we can imagine having a global statistical metadata base for the European region, which contains metadata that is of global interest in this respect. That is still a vision.

## 2. Online metadata

Already in a number of NSOs, the Internet is the main channel for dissemination of statistics. Without doubt, the role of the Internet will increasingly dominate the dissemination of statistics in the future. This means that the metadata system must be integrated with the web site of a statistical organisation.

It is a strategic question as to how to make all relevant statistical metadata accessible for the users. The metadata will be used to access the system and to navigate through the statistical information. The usage of metadata on the web site requires that the metadata are consistent across the whole system. To achieve this goal, it is necessary to have a good metadata system behind web site publishing of statistical information. A standardised resource for the statistical system, not only for web-publishing needs, would support this global metadata goal.

The Internet approach has completely changed the situation for both office and external

users. When printing statistical publications such as tables, booklets, and statistical yearbooks, the available statistical data are very restrictive and limited. Both the data and how they are published are in the hands of the statisticians. The statistician chooses what should be published and normally only consistent data will be disseminated. That would be more or less the same situation when publishing only fixed tables via the Internet. But the expectations and needs of users today are quite different. They want to get direct access to databases and have flexible possibilities to combine statistical data from different points of view. Users have their own technical equipment to carry out further processing of accessed data in a way that is out of statistical office control.

The only way to avoid misunderstanding and misinterpretation of data is to provide users with correct and helpful metadata. Metadata are also necessary for user navigation in the statistical information system. In many cases, it is not an easy task to find the right data. Often it is impossible for an external user, even if the data are formally available for direct access. Many users do not know what data are really available. They often are surprised when they get a complete overview of the content of the statistical information system.

It is therefore necessary to make the global metadata repository accessible via the Internet, but accessibility has to be supported by an advanced user interface that makes it easy to navigate and that is easily understood by different kinds of user.

## 3. Requirements of a central metadata system

Several requirements result from the central role of statistical metadata. Since the central metadata repository is considered to be a knowledge base, which is accessible in the global context, it must provide standard access for any type of user. It must be possible to import metadata from other sources as well as exporting them.

Metadata must be stored for all basic statistical concepts such as classifications, variables, tables and others by appropriate metadata object types. The definition of standardised statistical concepts is the key problem for harmonisation of metadata, both within an NSO and even more between NSOs and international statistical organisations. A number of efforts are being made to achieve such standardisation; the Metanet project of the fifth framework programme is one of the most important. Special working groups have started to work on the standardisation issue in different fields. The so-called Neuchâtel Group ([3]) has explored the area of classifications. At present, there are more than 70 such basic concepts (or object types) defined and agreed between some countries; but, finally, there will be more than 100 basic concepts. All these object types must reflect different aspects of these basic concepts, such as conceptual, operational and physical aspects [1].

Since concepts in statistical production will change over time, a central metadata repository must support versions for metadata objects. Last, but not least, a central

([3]) Members of the Neuchâtel Group are Statistics Sweden, Switzerland, Norway, Denmark and Run-Software GmbH.

metadata repository must provide multilingual support, especially in the context of the European Union.

## 3.1. Standard access and data exchange

There are good technical standards for data access (COM, CORBA) and data exchange (XML), but these are still not sufficient for standard access to a central metadata repository.

When having a general function, which can provide any pieces of metadata (e.g. GetMetadataObject), we still have to refine our requirement by specifying what type of metadata object we want to see and which one:

GetMetadataObject ('ClassificationVersion','NACErev1').

This would be sufficient to get the NACE Rev. 1 specification for the classification NACE. This works, however, only if the metadata provider understands what a 'ClassificationVersion' is. This is not a big problem for people but it is for technical systems. These systems will already have problems if there is a simple spelling error but they will have even more problems if there is a request for receiving a 'Nomenclature' or a 'KeyBridge'.

Hence, an agreement is required for being able to exchange metadata in a standardised way. Such an agreement should be on a level where it is relatively easy to agree. Therefore, the base for such agreements should not be a common metadata model, because different offices are using different database technologies and have different internal requirements on their metadata models. Moreover, data models are far too detailed to be a good base for such an agreement. It is much easier to find such an agreement on a conceptual level. During the last two years, such conceptual agreements have been defined as terminology models [2], [3], [4].

A terminology model regards statistical concepts as 'classifications' and their characteristics as taken from concepts such as 'name', 'owner', etc. [1].

A **terminology model** consists of textual definitions of statistical concepts defining a name (term) and a detailed description. In contrast to usual terminology definitions, each concept definition is appended by a list of characteristics, which describe properties or attributes of a concept and references to other concepts. A terminology model is similar to a data model but not as formalised and more expert-oriented.

Since terminology models reflect a purely conceptual view, they cannot be used directly for standard metadata access. A semantic specification for a semantic interface, however, is a more technical reflection of the terminology model.

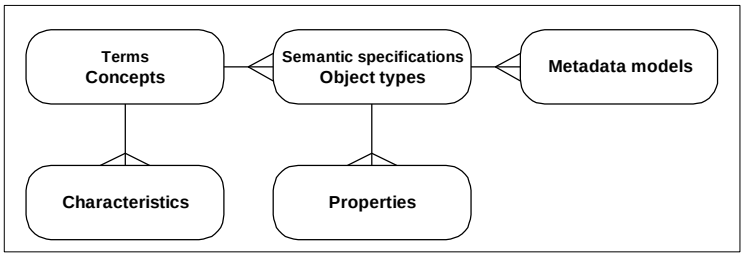



*Figure 4: Terminology model and semantic interface*

A **semantic specification** results directly from a terminology model but names are converted to technical names (which must not contain spaces) and some technical information is added, which allows presentation of the terminology model, for example, in UML. Concepts are reflected as logical object types and characteristics as properties of these object types in a semantic specification.

A semantic interface provides a defined set of functions for accessing any type of statistical metadata based on a common access mechanism (COM, CORBA). The same semantic specification can be used as an XML schema for metadata exchange. Accessing statistical metadata via a semantic interface metadata access and exchange becomes independent of the underlying metadata base. Moreover, the metadata repository might consist of different metadata sources, which are accessed via the semantic interface simultaneously. One example for an implementation of a semantic interface is ComeIn (common metadata interface, a public domain specification) [5].

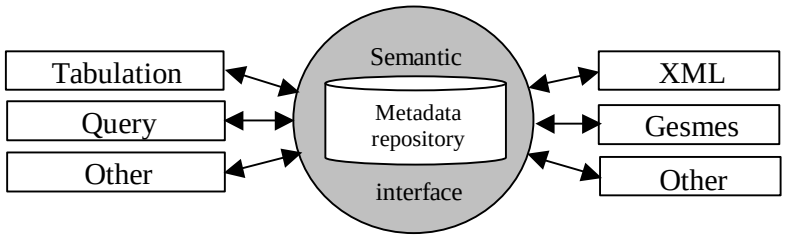



*Figure 5: Semantic interfaces*

The semantic interface specification is implementation independent and can be used as a base for XML as well as for ComeIn or other standard interfaces for special purposes. Metadata repositories can be adapted to the best database systems without changing the interface for production systems and external users.

Even though the technical problems have been solved in principle, the terminology agreement is still an open problem. There have been several groups discussing terminology models but this cannot be considered as an agreement for all statistical offices or even the offices in the EU. An important initiative to improve the situation has been taken by the Metanet project already mentioned above. Other standards as 'Dublin Core' [6] or ISO 11179 [7] are too weak to act as a semantic specification. They neither include descriptive exclusions, nor the 'includes' or 'includes also' characteristic(s) for classification items and many other items, which are defined in terminology models. Moreover, the semantic specification includes a description of metadata rules (such as, when deleting a classification item, all sub-items will be deleted as well), which is not part of the standards mentioned above and not part of the terminology model so far. Considering a metadata repository as the knowledge base for a statistical office also includes, however, a description of metadata rules, which are part of the semantic interface as well.

## 3.2. Complexity of a metadata repository

As mentioned above, several terminology groups have already defined about 70 statistical concepts (terminology objects). Modelling 70 objects in a terminology model results in more than 100 database objects or more than 300 relations. This is the experience of the IMIM project where about 120 metadata objects have been implemented in an object-oriented database. Implementing a model with the same information in a relational database would result in more than 500 relational tables.

This shows the complexity of the problem and it must be taken into account that this is just the beginning of defining a quite complex model. In future, metadata repositories might consist of 200 or even more metadata object types, which means nearly 1 000 or more relational tables.

The implementation effort for a relational and an object-oriented implementation for the Bridge system has been estimated at 3 person-years for the object-oriented approach and 10–15 person-years for the relational approach. There are no comparable experiences with maintenance resources required for such systems in a relational implementation. The Bridge model, however, has been continuously developed during the last two years and adding new object types, attributes or relationships is only a matter of days. On the other hand, it takes months or sometimes years to extend a relational metadata model in a statistical office.

The complexity is the reason that most implemented metadata models are very simplified. They have, for instance, often no, or very limited, multilingual support and, in most cases, no support for the history of metadata, i.e. taking care of the metadata changes over time (versions of metadata) etc. Such simplified solutions, which do not cover the existing complexities, can very well serve as a good basis for a while. At least they are often good and efficient starting points for the development of more advanced metadata systems.

However, it is questionable whether there are really good solutions with relational databases for applications with a high degree of complexity but with a relatively small amount of data. Object-oriented database systems could be an alternative for such cases. At least it is worth investigating the advantages and disadvantages of introducing object-oriented database systems into the statistics production process for tasks that are not very well supported by the relational approach.

It becomes obvious, when considering the metadata repository as a knowledge base, that metadata rules are an integral part of the metadata model. Object-oriented systems provide much better facilities to define rules as part of the model. Maintaining a metadata repository makes a big difference and an object-oriented implementation becomes much easier to maintain.

## 3.3. Versions, multilingual support

Conceptual metadata contain a lot of textual information. Most metadata objects have a title and a description. For example, classification items contain an 'official title' and 'general notes', as well as 'includes' or 'includes also' descriptions (exclusions). In an international environment, it is necessary to support a multilingual approach for this type of metadata. Providing unlimited multilingual support in a relational database will strongly influence the data model and could cause a lot of problems. This, again, is simpler in object-oriented databases such as ODABA2 [8], where the appropriate attributes need only be defined as 'generic attributes'. This allows the creation of language-dependent attributes as well as language-dependent indexes for fast search.

Metadata have a history, i.e. they have a time dimension that should be reflected in a system solution. Definitions of variables and the structure of classifications change over time, so it is not sufficient to keep only the current version. It should be possible to go back in the history of the whole metadata system. 'How was the situation two years ago?' could be a question to be answered. This opens another dimension in the database and will also cause several modelling problems when implementing in a consistent way (i.e. versioning of indexes and relationships between objects) in a relational database [9]. This becomes much easier when using an object-oriented system that supports a version dimension because in that case the time dimension will not influence the data model.

## 3.4. Data exchange with other sources

Some metadata are very well structured and stored in databases: for instance, variable names, code sets, etc. Other metadata may only be documented in a verbal way using text editors such as MS Word. Metadata stored in Word documents are difficult to access. Hence, it should be possible to exchange metadata between documents and the repository, i.e. breaking down the metadata in the Word document into atomic metadata items to be stored in the metadata repository and composing documents again from these atomic items in the repository.

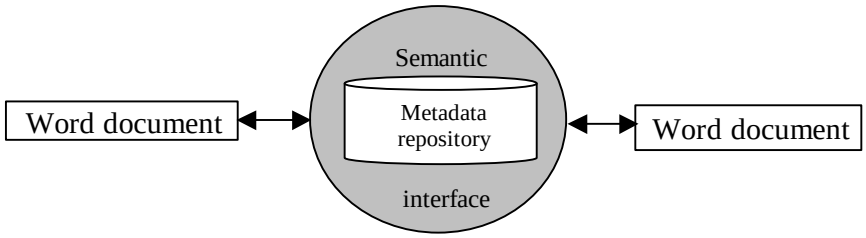



*Figure 6: Data exchange*

SCBDOK is an example of a well-structured Word document that contains textual metadata. Nevertheless, this is a first step and the structure should be more detailed to make it easier to exchange metadata with a formalised metadata repository. Data exchange with Word documents as shown in Figure 6 is only one example of exchanging metadata with other sources. Instead of Word documents, one may consider XML files or relational databases, EXCEL tables or other formats.

Similar requirements result from the situation where it might be necessary to transfer metadata from an existing metadata base to a central repository that provides more facilities. Thus, we can say in general that a metadata repository should provide data-exchange features with any type of existing metadata storage (including documents).

## 3.5. Choosing the right system

In principle, a central metadata system can be implemented in a relational database management system (RDBMS) as well as in an object-oriented one since the data models are equivalent, i.e. a relational model can be transformed into an object-oriented one and vice versa. The advantage of the object-oriented approach is, however, that many features which need explicit programming in an RDBMS are provided as built-in functions in an object-oriented environment. Thus, the degree of consistency in an object-oriented system is much higher than in a relational one.

Another problem is that the relational model requires a conceptual transformation of problems into a table presentation (i.e. one has to think about how to present, for example, the children of a person in appropriate tables). This is more natural in object-oriented models, which allows much better handling of complexity in object-oriented systems than in relational ones.

The disadvantage of object-oriented systems is that standards such as object query language are not as well known as structured query language (SQL) for relational databases and are more difficult to understand. On the other hand, queries in complex relational systems become rather complicated and are rather slow when the nesting level increases.

The differences in object-oriented systems are essential. Some systems (e.g. O2, ODABA) provide a large amount of conceptual features. Others, such as ObjectStore or POET, concentrate on efficiency. Since metadata are more complex and of lower quantity, database management systems (DBMS) that support the features described in the previous topics should be preferred.

## 4. Metadata transformation

Modern presentations of statistical data will include statistical metadata for the user, with enhanced explanations of statistical data shown in a table, whether on the Internet or in an enhanced tabulation system. This requires metadata on a base of conceptual metadata objects. However, different components of a statistical table, such as categories (lines or columns) or variables (columns), can either be linked with metadata stored for these concepts or with metadata objects. To guarantee consistency between statistical data and metadata, statistical metadata must be linked to the whole production process of statistical data. For instance, it should be possible for a user to go back to the questionnaire to see what has really been asked to get a better understanding of certain variables. This, however, is only possible after metadata have been transformed in the production phase.

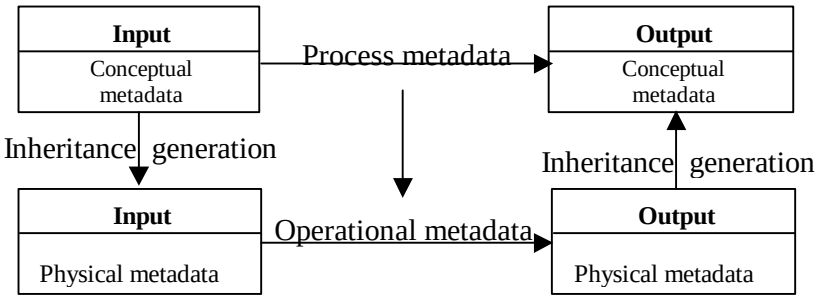



*Figure 7: Derived metadata transformation*

When considering metadata transformations, two different methods of statistical production are possible.

- Metadata-driven system
  In this case, the input and the output are defined conceptually. Physical metadata are created from the conceptual information. Operational metadata are generated from the input and output definitions and conceptual process definition.

- Production-driven system

In this case, only the input is defined conceptually. Physical metadata are created from the conceptual input metadata. Operational metadata are defined explicitly. The process generates the physical output metadata, while conceptual metadata for the output have to be generated from physical output and/or added manually.

In practice, both methods will be used, depending on the production and metadata tools which are available. Two types of metadata transformation result from the picture above. One is the transformation of metadata from the conceptual level to the operational or physical level and vice versa (vertical metadata transformation). Another results from operations on data such as aggregation or calculation of new variables (horizontal metadata transformation).

## 4.1. Deriving physical or operational metadata from conceptual

When conceptually describing an observation register, descriptions of different variables such as 'income' or 'date of birth' are required. Such an observation register might be implemented both as a relational database (e.g. Access) and as a text file, i.e. one conceptual observation register may have different physical realisations.

The definition of a physical implementation differs from the conceptual one, because data types must be added and size and precision must be defined. Nevertheless, more than 50 % of the physical and operational metadata information can be provided from the conceptual definition using different techniques.

- Inheritance or reference
  Metadata on the conceptual level, such as descriptions, can be inherited or referenced on the operational or physical level. Thus, descriptions are stored once but are visible for each implementation of a metadata concept. This guarantees consistency, because modifications are immediately visible on the physical level.

- Generating
  'Generating' means to create information on the physical or operational level, based on the conceptual definitions. This can be considered as a proposal for definition on the physical level, which can be modified later. This method provides higher flexibility than inheritance and referencing, but it does not guarantee consistency.

Usually these techniques are combined. Inheritance and references are used when the same information is referenced in all implementations. Generating of features is used when new metadata are created or when metadata might be modified.

This physical metadata can be provided both from the conceptual input and from the output and so the production process can be controlled by this metadata information.

### 4.2. Returning metadata from the production process

Most production tools perform metadata transformation on the physical level. This, however, will not return any conceptual metadata. When conceptual metadata are not available on the output side, online metadata cannot be provided for tables and other products. Hence, conceptual metadata have to be provided in production-driven systems from the physical metadata created for the output.

In a production-driven system, the conceptual metadata for the output must be provided according to the production process. This requires some support from the production system, but such support is not available at present. However, modern systems like SuperSTAR ([2]) are developing a support system to guarantee consistency between data and metadata.

## 5. Metadata-based retrieval functions

Retrieval functions usually require knowledge about the concepts of the facts we are looking for. When asking for something that is related to income, one might get many hits and it becomes quite difficult to find what we are interested in. But if we know about the concept of variables, we can start by asking for income-related variables first. Then we can select some variables that reflect the information we are looking for. Next, we can ask for the tables (assuming that we are aware of the concept of statistical tables) or cubes containing the selected variable. We can restrict the information to specific regions or age groups when we know the concept of classifications and value sets.

This shows very roughly how metadata-based retrieval functions can work and how statistical metadata can be used to provide enhanced retrieval functions. But again we will have big problems when we search for information in different environments.

Using terminology-based web interfaces allows combining of keyword search and object link techniques in different web applications, regardless of the metadata storage used in a specific environment.

A retrieval function is considered as a function that returns a number of web pages, or, more generally, a number of (meta) objects. Retrieval functions based on terminology models are more flexible than simple keyword search functions, since we can combine three search techniques in a web application [9].

- Keyword search
  Keyword search allows searching for all objects that contain a number of defined search words or that are linked to these terms. Keyword search is provided as free text

([2]) SuperSTAR is a product of Space-Time Research, Melbourne.

search (slow) and index search (fast).

- Type-oriented search
  This technique is usually combined with keyword search. By defining object type, the number of hits can be drastically reduced and the search process will be much faster. One typical way to reduce the scope is to search for objects of a given type (e.g. variables or tables referring to the search or keyword).

- Object link
  Object link techniques are based on the fact that a number of relationships exist between different (meta) objects. Thus, it is quite common in many cases to follow the links between objects, which are part of the knowledge stored in the metadata base.

Combining keyword search facilities with linked object techniques and reduced scope features allows flexible, fast and user-friendly search strategies. XML can be used as well as active server pages generating HTML directly from the metadata base.

## 6. The classification database at Statistics Sweden

Statistical metadata systems appear to be quite complicated and need substantial support by advanced software. At present, it is practically impossible to find standard software on the market, supplied by big vendors, that could solve the metadata software problem for an NSO. A number of NSOs have started in-house developments based on commercial database management systems and other basic standard software packages. Relational database systems are used almost exclusively by NSOs.

Based on the results of the IMIM ([3]) project of the fourth research framework programme, a metadata management system, Bridge$^{NA}$, has been developed. The main partners in the IMIM project were Statistics Sweden, Denmark, Norway, World Systems (Europe) Ltd, and World Systems Informatics GmbH in Berlin. The last cooperated very closely with Run-Software GmbH in Berlin. Experts from this company worked on a consultancy basis for World Systems Informatics. The main result of the IMIM project was a software package Bridge. After the project lifetime, Run-Software GmbH continued the development on a commercial basis and developed a new and improved structured package called Bridge$^{NA}$.

Bridge$^{NA}$ consists of an object-oriented metadata repository using the ODBMS ODABA2 that has also been developed by Run-Software GmbH. The company was also represented in the Neuchâtel Group (see footnote 3) that developed standards for classifications. The group has finished the present work in the field of classifications.

([3]) IMIM — integrated meta-information management.

Bridge$^{NA}$ is being used in the development of a classification database at Statistics Sweden. In the present phase, a number of standard classifications have been and are to be loaded into the system. In a further step, the classification database will be accessible inside the office and, later, will be published on the Internet. The classification database will be the central (global) source for this kind of metadata at Statistics Sweden. All maintenance work will be carried out in that environment. Other applications will download the classifications from that central resource. Perhaps that is the beginning of developing a central metadata resource on a new platform. The success of using Bridge$^{NA}$ will be very decisive in that respect.

The software was developed in a continuous way in parallel with the requirements of the Neuchâtel Group and it was flexible enough to adopt new requirements in a very short time. The software is completely new and still needs stabilisation and further development. The functionality implemented in the system is much broader than that used for the classification database. As can be seen from other developments of metadata systems at Statistics Sweden, that would not have been possible using the traditional relational approach.

Statistics Switzerland is preparing the use of the system for different metadata purposes. It would probably be very difficult for Statistics Sweden to carry out such a development in-house, since there is a lack of the necessary resources — both in amount and mainly in skills competence.

It is intended to continue this software development. Scheduled cooperation with the SuperSTAR development of Space-Time Research will broaden the functionalities and open a new perspective. There is also a clear vision to implement standards for metadata that will be developed in other metadata projects, in particular by the Metanet project.

## 7. References

[1] Karge, R., ‘Metadata, semantic interfaces and meaning in global communication’, International Conference on Survey Research Methods, May 2001.

[2] Classification terminology, Neuchâtel Group 2000 (www.run-software.com/download/ClassificationTerminology.doc).

[3] Variable terminology, Oslo Group, September 2000 (www.run-software.com/download/VariableTerminology.doc).

[4] Terminology for statistical activities, cubes and registers, November 2000 (www.run-software.com/download/ActivityTerminology.doc).

[5] ComeIn user’s guide, Run-Software, November 2000 (www.run-software.com/download/CIUsersGuide.doc).

[6] Using Dublin Core, July 2000 (dublincore.org/documents/2000/07/16/usageguide/).

[7] ISO 11179 — Specification and standardisation of data elements, Parts 1–6 (ftp://sdct-sunsrv1.ncsl.nist.gov/x3l8/11179/).

[8] ODABA2 — Technical overview (www.run-software.com/download/TECKONEN.DOC).

[9] Karge, R., 'Efficient modelling techniques for complex problems', Fifth International Conference on Information Systems Analysis and Synthesis (ISAS '99), August 1999.

# WAID 4.1: a computer program for imputation of missing values

Ton de Waal

*Statistics Netherlands*
*PO Box 4000, 2270 JM Voorburg, Netherlands*
*E-mail: twal@cbs.nl*

**Keywords:** donor imputation, imputation software, mean imputation, nearest neighbour imputation, tree-based models, weighted automatic interaction detection

**Abstract**

*One of the aims of the fourth framework programme project Autimp was the development of advanced prototype imputation software. Institutes participating in this project, which lasted from February 1999 until November 2000, were the University of Southampton, the Office for National Statistics (UK), Statistics Finland, Instituto Nacional de Estatística de Portugal and Statistics Netherlands. The imputation software that has been developed is based on automatic interaction detection (AID) trees. Because the developed algorithm gives lower weights to outliers while constructing regression trees, the technique is referred to as weighted automatic interaction detection (WAID). The developed imputation software, also called WAID, can be used to impute for missing values of both categorical and numerical variables. It is a stand-alone program, and can be used under Windows 95/98 and Windows NT. The present paper briefly describes WAID (version 4.1).*

## 1. Introduction

Autimp was a relatively small fourth framework programme project that was partly funded by the European Commission. The aims of Autimp were the evaluation of software that can be used to impute for missing data (see Chambers et al., 2001a, for several evaluation reports) and the development of advanced prototype imputation software. Autimp lasted from February 1999 until November 2000. Participating institutes were the University of Southampton, the Office for National Statistics (UK), Statistics Finland, Instituto Nacional de Estatística de Portugal and Statistics Netherlands. The last institute acted as project coordinator.

The imputation software that has been developed is based on automatic interaction detection (AID) trees (see Sonquist et al., 1971). Because the developed algorithm gives lower weights to outliers while constructing regression trees, the technique is referred to as weighted automatic interaction detection (WAID).

A WAID-tree, or generally a tree-based model, classifies the data in terms of the values of a set of categorical predictor variables. It is a binary tree that is generated by successively splitting a training data set into smaller subsets. These subsets are increasingly more

homogeneous with respect to a selected response variable. This response variable may be either categorical or numerical. The process of recursively splitting the data set into two subsets continues until a stopping criterion is met. The terminal nodes in this tree form homogeneous clusters.

The developed imputation software, also called WAID, can impute for missing values of both categorical and numerical variables. It uses the homogeneous clusters to draw donor records out, or to calculate cluster means.

WAID is a stand-alone program, and can be used under Windows 95/98 and Windows NT. The core of WAID is an algorithm to construct WAID-trees developed by the University of Southampton. Many features have been added to this core in order to make WAID suitable for the imputation of missing data. These features allow WAID to be used in the day-to-day routine at statistical offices.

The remainder of this paper is organised as follows. Section 2 provides an overview of WAID 4.1. How to operate WAID is illustrated in Section 3 by means of a sample session. The imputation methods that are supported by WAID 4.1 are described in Section 4. Section 5 concludes the paper with a brief discussion.

In this paper, we do not provide details of how the WAID-trees are generated. These details can be found in a paper by Tsai and Chambers, which is contained in the compilation report by Chambers et al. (2001b).

## 2. A brief overview of WAID 4.1

To impute for missing values in a data set, the software first determines the missing data patterns (MDPs) in this data set. The user can select which MDPs, or parts of MDPs, he or she wants to impute.

WAID can impute several categorical variables simultaneously using the same donor record. Technically, these categorical variables are combined into a single compound categorical variable. A numerical variable cannot be imputed simultaneously with other variables, either categorical or numerical ones, using the same donor. The only way a numerical variable can be imputed in combination with other variables is by allowing multiple donors. The MDP involving the numerical variable then has to be split into several MDPs, one of which involves only the numerical variable.

For all (parts of) MDPs to be imputed, the user has to select a set of categorical predictor variables. After selection of the predictor variables, WAID-trees are grown for each (part of an) MDP using a complete training data set. This training data set may be (a subset of complete records of) the data set to be imputed, but it may also be a different data set.

The terminal nodes of the generated WAID-trees form clusters of records that are as homogeneous as possible with respect to the variables involved in this (part of an) MDP. The homogeneous clusters themselves, however, are not used by the computer program. Only the classification rules that define these homogenous clusters are used. In this way, we can use a training data set to determine the classification rules, and later use another data set with donor records to actually impute for missing values.

Automatic generation of homogeneous clusters using user-specified predictor variables is the greatest strength of WAID. This functionality can save users a lot of time, which would otherwise be required to analyse the data and determine appropriate homogeneous donor groups themselves. The ability to determine homogeneous clusters automatically sets WAID apart from many other computer programs for imputation of missing data.

After generation of the WAID-trees, and hence generation of the classification rules for constructing homogeneous clusters of records, we supply a data set with donor records to the computer program. This data set may be the same as the data set to be imputed, or it may be a different data set. In the data set with donor records, we apply generated classification rules to construct homogeneous clusters of donor records.

To impute for missing values in a certain record in the data set with missing values, we determine which WAID-trees correspond to the MDP of this record. More than one WAID-tree (and hence more than one homogeneous cluster) may correspond to a particular record, because separate WAID-trees may have been generated for different parts of its MDP. Subsequently, we determine the homogeneous clusters corresponding to this record by using the classification rules to classify the values of the predictor variables. The records in the data set with donor records corresponding to those clusters are used to impute for missing data in the record under consideration. WAID supports several imputation methods, for example nearest neighbour or random donor records may be selected from a homogeneous cluster.

A previous version of WAID, version 4.0, has been tested on realistic data sets. For details concerning these evaluation tests we refer to evaluation reports by Crespo, Laaksonen and Piela, and Tsai and Chambers that can be found in Chambers et al. (2001b).

## 3. A sample session

In this section, we illustrate the use of WAID by describing a small sample session. This sample session is discussed in more detail in the manual to WAID 4.1 (see de Waal et al., 2001). We start by describing the data set used in this sample session. Subsequently, we briefly describe the menu system of WAID. In Subsections 3.3 to 3.6, we describe the steps in the imputation process of WAID: determining missing data patterns, generating WAID-trees, imputing values and examining the output. Subsection 3.7 concludes this section by mentioning a few extra options of WAID that are not illustrated in our sample session.

## 3.1. The data

The database that we use in our sample session consists of 999 records. Each record contains a record identification number called 'Id', plus six categorical variables and one numerical variable. The six categorical variables are called 'Genpuk' (with two categories), 'Stupuk' (two categories), 'Illpuk' (two categories), 'Marital_status_group' (3 categories), 'Ethnic_group' (5 categories) and 'Occupation_group' (10 categories). The numerical variable is 'Age' (values ranging from 0 to 80).

The data set has been synthetically generated in order to demonstrate the use of WAID. It is based on actual data, which have been slightly perturbed for confidentiality reasons. The name of the data set is 'Example'. It will be used to determine the missing data patterns, to generate WAID-trees and to select donor records.

In the sample session, we assume that we want to impute all MDPs, except those involving the variable 'Genpuk' or 'Marital_status_group'. We assume that those MDPs will be imputed by means of other software. As predictors, we use all possible variables except the variables 'Stupuk' and 'Illpuk', which we assume to be bad predictors for the other variables. For all final nodes of all selected MDPs involving 'Age', we use nearest neighbour imputation. For all final nodes of all other selected MDPs we use random donor selection as the imputation method. In our sample session, we are especially interested in comparing the univariate distributions of the variables 'Ethnic_group', 'Occupation_group' and 'Age' on the non-imputed records with the distributions of these variables on the imputed records.

## 3.2. The menu system of WAID

The main menu bar of WAID contains six menus: File, View, Options, Process, Window and Help. The File menu contains options to open, close and save data sets or WAID projects. The View menu contains options to view various data sets, MDPs, WAID-trees, and output such as distributions. The Options menu contains many different options: options to open data sets, to split MDPs, to select MDPs for which WAID-trees will be generated, to select predictors for WAID-trees, to set parameters for WAID-trees, to specify imputation methods, to set parameters for the imputation methods, and to select variables for which distributions will be generated. The Process menu contains options for which WAID has to carry out computations, such as determining MDPs, generating WAID-trees and imputing for missing values. The window is just a standard Windows menu and can be used to cascade or tile windows. Finally, the Help menu provides online help.

## 3.3. Determining missing data patterns

To start a WAID session, we first select a data set with missing data, which will be referred to as the **Data Set with Missing Data**. After selection of our 'Example' data set, we specify that the data set contains a record identifier with the name 'Id'.

The actual imputation process starts with determining which MDPs exist in the **Data Set with Missing Data**. To determine MDPs, we simply choose the appropriate option, 'Missing Data Patterns', from the Process menu. Now two data sets are generated: the **Missing Data Patterns** data set and the **Missing Data Patterns per Record** data set. They can be shown on the computer screen by choosing options from the View menu.

The result of viewing the **Missing Data Patterns** data set is shown in Figure 1. There are 22 MDPs in total in our 'Example' data set. Five of these MDPs, patterns 9, 10, 11, 17 and 18 are coloured in red on the computer screen. In these MDPs, a numerical variable, 'Age', occurs in combination with other variables. As has been explained before, WAID cannot impute such an MDP using a single donor record. To process such an MDP, it has to be split into subpatterns.

MDPs involving only categorical variables may also be split into smaller subpatterns if the user prefers. For each part of these MDPs a separate donor record will then used. Splitting an MDP involving only categorical variables may, for instance be desirable when the MDP involves many variables.

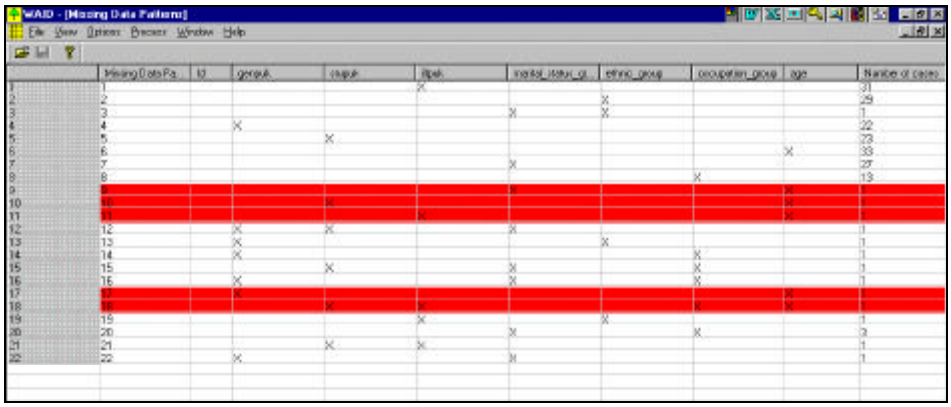


*Figure 1: The missing data patterns*

In our sample session, we split MDP 10 (involving 'Stupuk' and 'Age'), MDP 11 (involving 'Illpuk' and 'Age'), MDP 18 (involving 'Stupuk', 'Illpuk', 'Occupation_group' and 'Age') and MDP 21 (involving 'Stupuk' and 'Illpuk').

Note that it is not necessary to split MDP 21, but we are allowed to if we wish to do so . We do not split MDPs 9 and 17, because in this sample session we assume that we do not want to impute MDPs involving 'Genpuk' or 'Marital_status_group' by means of WAID. We assume that those MDPs will be imputed by means of other imputation software.

To split the abovementioned MDPs, we first have to tell WAID that we want to split MDPs. This can be done by selecting the appropriate item, 'Split Missing Data Patterns', from the Options menu. After this item has been chosen, we see Figure 2 on the computer screen.

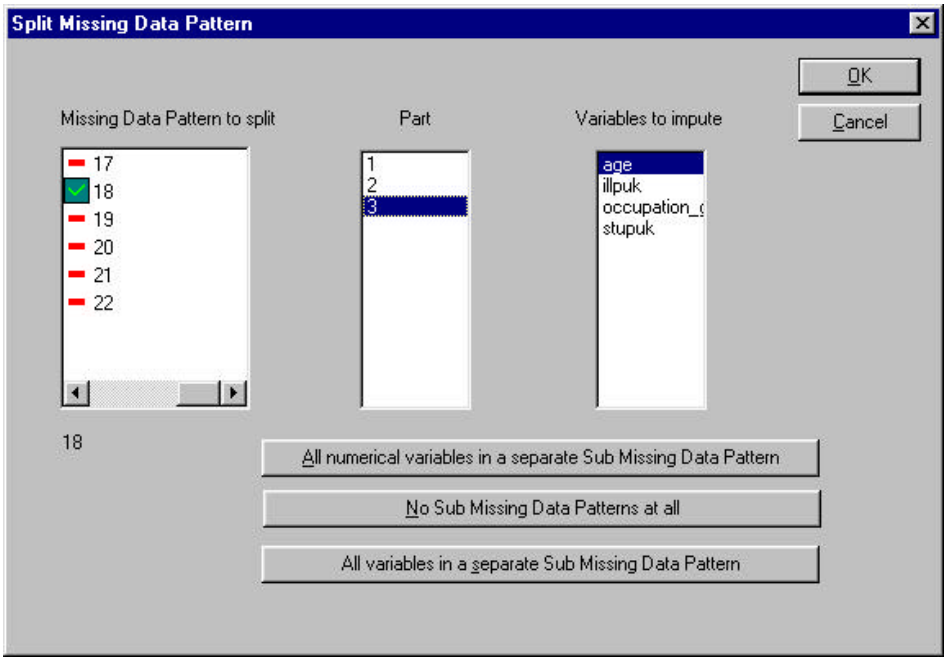



*Figure 2: Splitting an MDP into subpatterns*

To split an MDP into subpatterns, we first click on this MDP to select it. Next, the variables in each part are selected by clicking on these variables and parts. After splitting an MDP, say MDP 18, visual feedback is given in the ' Missing Data Pattern to split' box to indicate that MDP 18 has been split into subpatterns.

In our sample session, we split MDP 18 into three subpatterns: Part 1 involving only 'Stupuk', Part 2 involving both 'Illpuk' and 'Occupation_group', and Part 3 involving only 'Age'. We split MDP 10 into two subpatterns: Part 1 involving only 'Stupuk' and Part 2 involving only 'Age'. We split MDP 11 into two subpatterns: Part 1 involving only

'Illpuk' and Part 2 involving only 'Age'. Also, we split MDP 21 into two subpatterns: Part 1 involving only 'Stupuk' and Part 2 involving only 'Illpuk'.

Determining and splitting MDPs into subpatterns are relatively simple processes in WAID that do not require much time. Once we have split MDPs 10, 11, 18 and 21 in this way, we press **OK**. Splitting the MDPs is now completed.

## 3.4. Generating WAID-trees

The next step in the imputation process is the generation of WAID-trees. To generate WAID-trees, first a data set that will be used for WAID-tree generation has to be chosen. This can be done by selecting the appropriate option from the File menu. This data set will be referred to as the **Data Set used for WAID-Tree Generation**. It may be the same as the **Data Set with Missing Data**, or it may be a different data set.

In the latter case, the structure (e.g. names and types of variables) should be the same in both data sets. The selected data set is used to generate WAID-trees. This is similar to using a training data set as in the case of a neural network. In our sample session, we again choose our 'Example' data set.

Now we have to specify for which MDPs WAID-trees have to be generated. To do this, we first have to instruct WAID that we want to specify the MDPs for which WAID-trees will be generated. This can be done by selecting the appropriate option, 'Missing Data Patterns for which WAID-Trees will be generated', from the Options menu. After selecting the option, we see Figure 3 on the computer screen.

In our sample session, we select all (parts of) MDPs not involving 'Genpuk' or 'Marital_status_group', i.e. we select patterns 1, 2, 5, 6, 8, 10 SUB 1, 10 SUB 2, 11 SUB 1, 11 SUB 2, 18 SUB 1, 18 SUB 2, 18 SUB 3, 19, 21 SUB 1 and 21 SUB 2. After we have done that we press >>**Add**>> and the **OK** button.

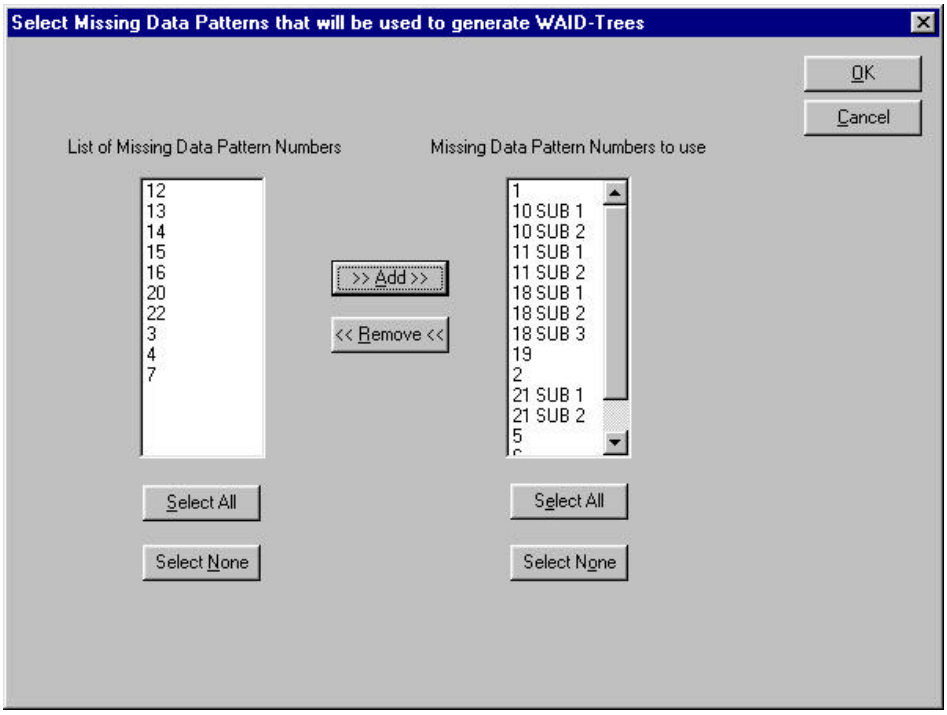



*Figure 3: Selecting MDPs for which WAID-trees will be generated*

To generate the WAID-trees, the WAID-tree algorithm has to know which variables have to be used as predictors for each selected MDP. This can be done by first selecting the 'Predictors' option from the Options menu. We then see the 'Specify predictors' window on the computer screen (see Figure 4). For each selected MDP, at least one variable has to be selected as predictor.

In our sample session, we select all possible variables except 'Stupuk' and 'Illpuk' as predictors for each selected (part of an) MDP. We do this by first clicking on a **Selected Missing Data Pattern Number** to highlight the MDP. For the (sub)pattern highlighted, we click on all shown variables except 'Stupuk' and 'Illpuk' in the 'Predictors to use' box until all these variables are highlighted. We then click on another **Selected Missing Data Pattern Number** and repeat the procedure until predictors have been selected for all (sub)patterns.

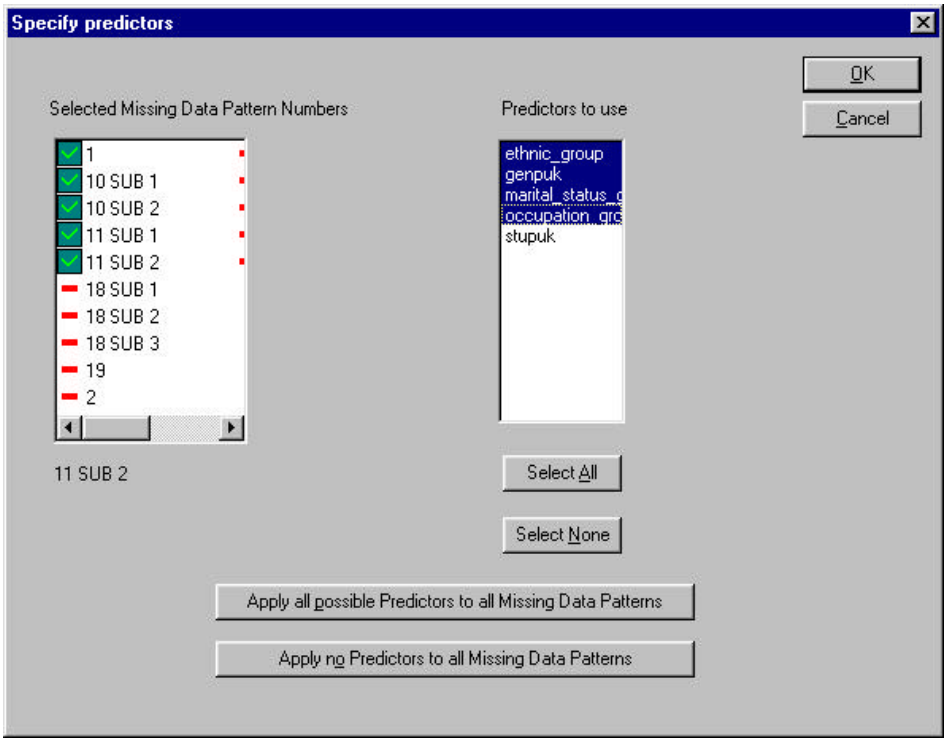



*Figure 4: Selecting predictors for an MDP*

Visual feedback symbols appear in the 'Selected Missing Data Pattern Numbers' box, to indicate for which MDPs predictors have already been specified.

We would like to draw attention to the fact that although both patterns 10 SUB 2 and 11 SUB 2 involve only the variable 'Age', the variables in the 'Predictors to use' box are shown as different. The reason is that these subpatterns have been derived from different MDPs. The variables that may be used as predictor variables for a particular subpatter n are all categorical variables that do not occur in the original MDP.

All that remains to be done in order to generate the WAID -trees is to select the appropriate option, 'WAID-Trees', from the Process menu. Specifying MDPs for which WAID -trees will be generated, specifying predictors for the selected MDPs and generating WAID -trees are simple processes in WAID that do not require much time.

The actual WAID -tree algorithm supports only one dependent variable per WAID -tree. So, if (a part of) an MDP involv es several categorical variables and no numerical ones, a compound variable is generated by constructing the full combination of values that are found for the variables in this (part of the) MDP in the **Data Set used for WAID-Tree Generation**. This compound variable acts as the dependent categorical variable in the

WAID-tree algorithm. Both the compound dependent variable and the WAID -trees can be viewed by selecting the corresponding options from the View menu.

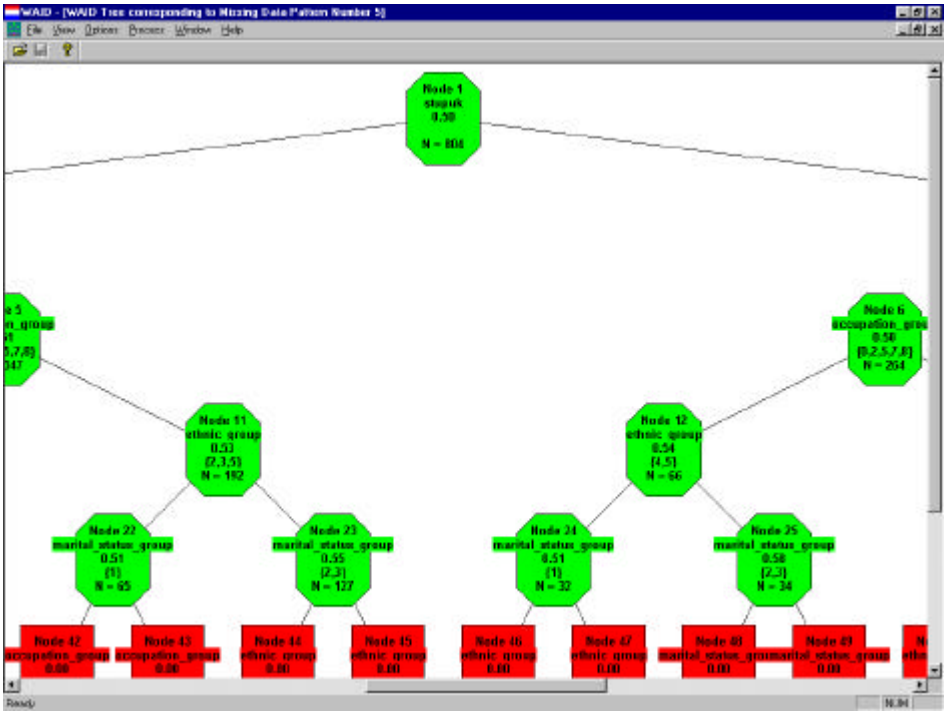



*Figure 5: Viewing (part of) a WAID-tree*

In Figure 5, we see part of the WAID-tree for MDP 5. We can see, for example, that the top node is called Node 1. The variable mentioned below the phrase Node 1 is the variable to be imputed, in this case the variable 'Stupuk'. If a compound dependent variab le is to be imputed, '**CDV**' instead of the variables constituting this compound dependent variable is mentioned.

In the other nodes, the variable below the node number is the predictor variable that has been used to split the parent node. The set mentioned on the fourth line in the node defines the node in terms of the predictor variable. In particular, this set describes the categories of the predictor variable that define the node. For instance, Node 22 was the result of splitting its parent node, Node 11, into two child nodes (Node 22 and Node 23). In Node 22, we can see that to split Node 11 the predictor variable ' Marital_status_group' has been used. In Node 22, only records for which 'Marital_status_group' equals 1 occur. In Node 23, only records for which 'Marital_status_group' equals 2 or 3 occur. To determine how a particular node is defined in terms of all predictor variables, one should follow the path from the root node to this node.

Below the name of the predictor variable that has been used to split the parent node of a certain node, a measure for the homogeneity of the node is shown. If the dependent variable is categorical, a homogeneity measure based on the so-called Gini index for the node under consideration is shown. If the dependent variable is numerical, a (weighted) mean is shown. For instance, in Figure 5, the homogeneity measure based on the Gini index of Node 11 is 0.53.

We have already explained that the fourth line in a node consists of the categories of the predictor variable that define this node. The last line in a node describes how many records correspond to that node. In Figure 5, for instance, 192 records correspond to Node 11.

## 3.5. Imputing values

The last step in the imputation process is actually imputing for missing data. For this step, a data set that will be used for imputation has to be loaded. This can be done by selecting an option from the File menu. This data set will be referred to as the **Data Set used to Impute**. From this data set, donor records will be selected during imputation. In our sample session, we once again choose the 'Example' data set.

To specify the imputation methods that have to be used, the appropriate option, 'Imputation Method', from the Options menu should be chosen. We then see the 'Specify Imputation Method' window on the computer screen (see Figure 6).

For each final node of each WAID-tree, an imputation method can be specified. This can be done by clicking on an imputation method and then clicking on the nodes for which it should be applied. In our sample session, we decide to use random selection for all final nodes of all selected (sub)patterns, except for the (sub)patterns consisting of the variable 'Age' only, i.e. patterns 6, 10 SUB 2, 11 SUB 2 and 18 SUB 3. For those MDPs, we decide to use nearest neighbour donor imputation for all final nodes.

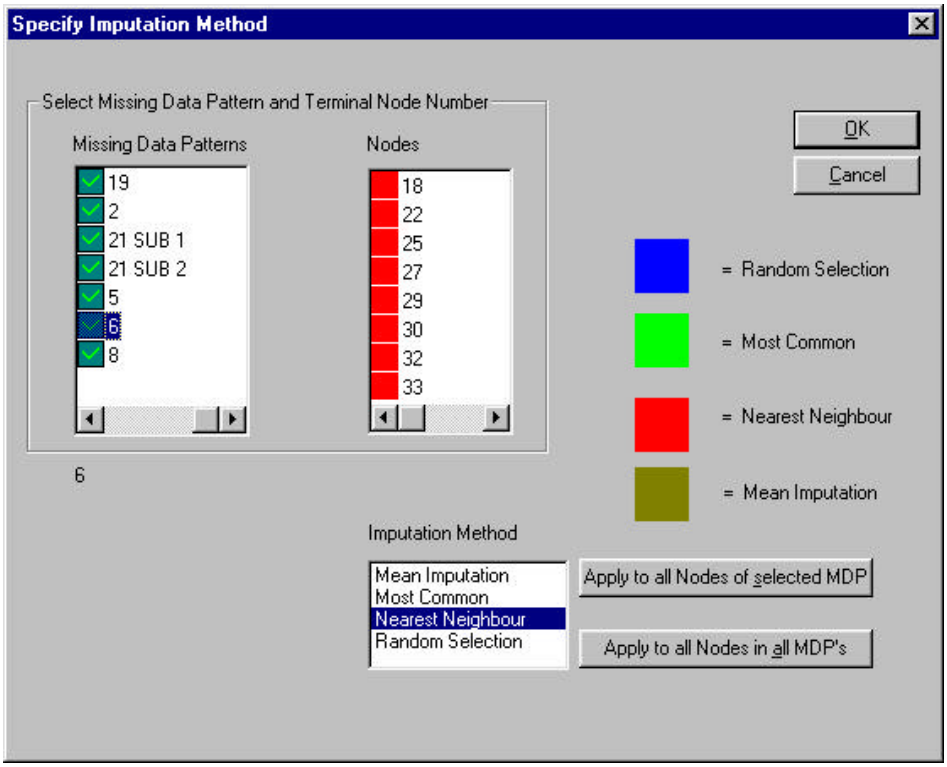



*Figure 6: Specifying imputation methods*

Having selected the data set that will be used for imputation and the imputation methods, we start the actual imputation process by choosing the appropriate option, 'Impute', from the Process menu.

Specifying imputation methods is quite a simple process in WAID that can be done quickly. Actually imputing for missing data usually does not require much time, except when one wants to impute very large data sets using nearest neighbour imputation.

## 3.6. Examining the output

After we have imputed the **Data Set with Missing Data**, we can examine the output. Four data sets have been generated: the **Imputed Data Set**, the **Flag Data Set**, the **Records that cannot be Imputed**, and the **Imputed Records**. We can view these data sets by selecting the appropriate options from the View menu. In the **Imputed Data Set**, where some fields are coloured blue, those fields have been imputed using random selection. Some other fields are coloured red; those fields have been imputed using nearest neighbour donor imputation. The other fields are coloured black in our sample session. Those fields have not been imputed. In the **Imputed Records** data set, we can only see the records that have

been imputed. In the data set called **Records that cannot be Imputed**, we can see all records that could not be imputed. The reason why a record could not be imputed is mentioned in the last column. Finally, in the **Flag Data Set**, we can see which fields have been imputed (indicated by a one) and which have not been imputed (indicated by a zero).

In our sample session, we are also interested in examining the univariate distributions of 'Ethnic_group', 'Occupation_group' and 'Age'. To generate those distributions, we specify these variables using the 'Distributions' option from the Options menu. To view the distribution of the numerical variable 'Age', we decide to divide its range into 20 intervals of equal size. Now, we can view distributions of 'Ethnic_group', 'Occupation_group' and 'Age' by selecting the 'Distributions' option from the View menu.

For all three variables 'Ethnic_group', 'Occupation_group' and 'Age', a graph is made for the distribution of the non-imputed values and another graph for distribution of the imputed values. These graphs help us determine whether or not we can consider the imputation process to be a success (see Figure 7).

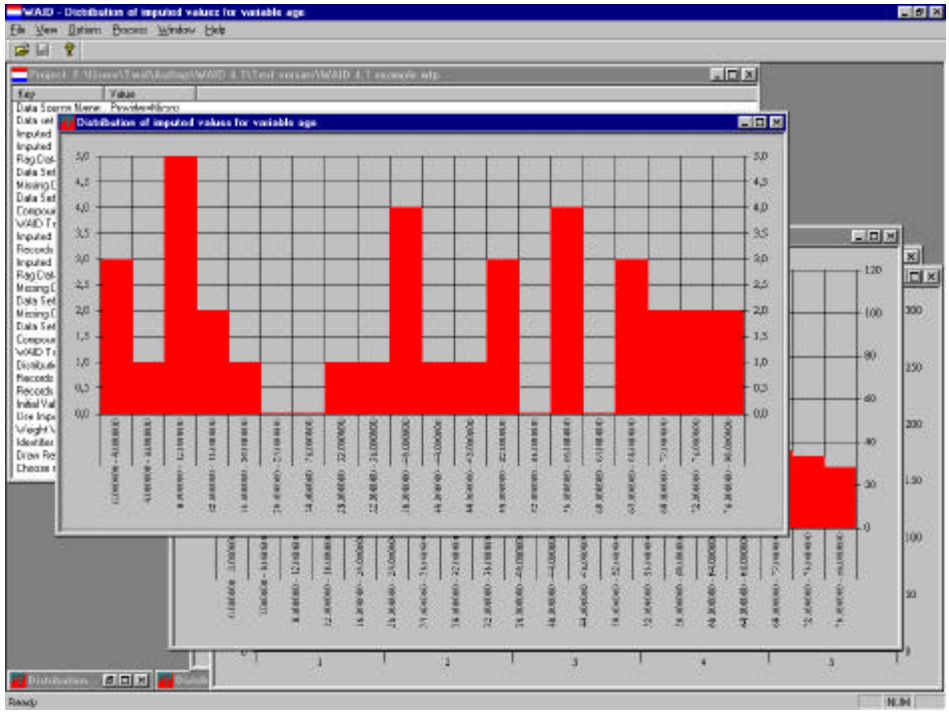



*Figure 7: Viewing distributions*

## 3.7. Further options

WAID 4.1 offers more functionality than described in our brief sample session. Below, we mention a few more options of WAID.

- The user can save all current settings (data sets, changed options, etc.) in a WAID-tree project (.wtp) file. He or she can then continue working on this WAID project another time.

- The user can specify whether a categorical variable is nominal or ordinal. Nominal predictor variables are handled differently to ordinal ones (see below).

- The user can specify the homogeneity measure for the nodes of each WAID-tree. If a categorical variable (or a combination of several categorical variables) is to be imputed, homogeneity of the nodes is measured by the Gini index. For an ordinal variable, the Gini index is computed for every binary split that preserves the ordering. For a nominal variable, WAID supports two options. One option computes the Gini index for every possible binary split. The other option creates a pseudo-ordering to transform the nominal variable into a pseudo-ordinal variable. Subsequently, the Gini index is computed for every possible binary split that preserves the pseudo-ordering.

  If a numerical variable is to be imputed, homogeneity can be measured by an ordinary least squares criterion, by Tukey's biweight function, by Huber's min/max, or by Andrews' sine. The last three criteria are robust against the presence of outliers in the data. For more details, see the paper by Tsai and Chambers in Chambers et al. (2001b).

- The user can specify which records are used for WAID-tree generation. The same is possible with respect to the donor records that are used to impute. The user can either use the complete records only or, for each MDP, he or she can use the records for which both the predictors and the response variables are complete.

- The user can specify a weight variable for the **Data Set used to Impute**. During imputation, a weight is then used for each record in the **Data Set used to Impute**. For instance, if the user applies the random selection imputation method and specifies a weight variable, weighted random selection will, in fact, be the imputation method used (see also Section 4).

- The user can specify whether or not donor records may be selected many times to impute for the variables in an MDP.

- The user can specify whether imputed records can serve as potential donor records. This option, **Use Imputed Records for Imputation of Others**, can only be set to **Yes** when the **Data Set with Missing Data** is the same as the **Data Set used to Impute**.

- The user can specify in which order the records in the **Data Set with Missing Data** are imputed. This can be important when the **Data Set with Missing Data** is the same as the **Data Set used to Impute** (i.e. when hot-deck donor imputation is used) and when the option **Use Imputed Records for Imputation of Others** is set to **Yes**. In that case, a certain imputation order could make some records imputable, whereas these records would not be imputable when another imputation order would be used.

- The user can choose what to do when certain predictors for MDPs have values in the **Data Set with Missing Data**, but these do not occur in the **Data Set used for WAID-Tree Generation**. There are two options: **None** and **Random**. If **None** is selected, no value is imputed for this MDP in the record under consideration whenever a predictor value that does not occur in the **Data Set used for WAID-Tree Generation** is encountered in the **Data Set with Missing Data**. If **Random** is selected, WAID chooses one of the two child nodes according to a probability distribution whenever a predictor value that does not occur in the **Data Set used for WAID-Tree Generation** is encountered in the **Data Set with Missing Data**. The probability for WAID to choose a child node is proportional to the numbers of records in this node.

## 4. Supported imputation methods

For each final node of each WAID-tree, i.e. for each (part of an) MDP for which a WAID-tree has been generated, the user can specify the imputation method that has to be used. WAID 4.1 supports four imputation methods. For all four methods, WAID first determines the final nodes in the WAID-trees that correspond to a given record in the **Data Set with Missing Data**. To impute the record under consideration in the **Data Set with Missing Data**, WAID only uses those records in the **Data Set used to Impute** that are classified in these final nodes. The set of records that are used by WAID to impute a certain (part of an) MDP in a certain record are referred to as the donor cluster for this (part of the) MDP and that record. The imputation methods supported by WAID 4.1 are: most common category; nearest neighbour; random selection; and mean imputation (only if a numerical variable is to be imputed).

When most common category imputation is used, the most common category of the (compound) variable — corresponding to the (part of the) MDP that is being imputed — in the donor cluster is imputed for this (part of the) MDP.

When nearest neighbour imputation is used, the record in the donor cluster that is closest to the record to be imputed is used to impute for the (part of the) MDP under consideration.

When random selection imputation is used, a record from the donor cluster is randomly drawn and used as the donor record to impute for the (part of the) MDP under consideration.

When mean imputation is used, the mean value of the variable to be imputed over all records within the donor cluster is computed. This mean value is imputed in records corresponding to that donor cluster for which the value of the variable to be imputed is missing. Mean imputation can only be applied to (univariate) numerical variables. The default imputation method is random selection.

In the case where nearest neighbour imputation is used, a distance function should be specified to measure the distance between two records. The distance functions that can be used are quite simple, because only categorical predictor variables are allowed in WAID. To measure the distance between a potential donor record and the recipient record in a final node of a WAID-tree, distance functions of the following type are allowed:

$$\sum_{i \in S} a(R_i, D_i), \tag{1}$$

where $S$ denotes the set of predictor variables, $R_i$ the value of the $i$th predictor variable in the recipient record, and $D_i$ the value of the $i$th predictor variable in the potential donor record. The values $a(R_i, D_i)$ depend only on the values of $R_i$ and $D_i$.

Together, the $a(R_i, D_i)$ values form a matrix. This matrix may be specified by the user of WAID. The requirements for the matrix are that it is symmetric, that the entries on the main diagonal are equal to zero, and that the other entries are positive. The potential donor record that is the closest to the recipient record is chosen as donor record.

The user can use a weight variable during the imputation process in combination with either of the four supported imputation methods.

(i) In the case where the user uses a weight variable and in some final node of a WAID-tree applies the random selection method, the probability to select record $i$ from the donor cluster is equal to

$$p_i = \frac{w_i}{\sum_{j \in DC} w_j}, \tag{2}$$

where $w_j$ is the weight of record $j$, and $DC$ denotes the donor cluster.

(ii) In the case where the user uses a weight variable and in some final node of a WAID-tree applies the most common category method, a category $c$ is chosen for imputation for which

$$\sum_{i \in DC} w_i \, d(D_i = c)$$

(3)

is maximal.

Here $d(D_i = c) = 1$, if the value of the $i$th record in the donor cluster $DC$ equals $c$; otherwise $d(D_i = c) = 0$. In formula (3), if several categories have a maximal value, one of these categories is randomly selected.

(iii) In the case where the user uses a weight variable and in some final node of a WAID-tree corresponding to a missing numerical variable applies the mean imputation method, the value

$$\frac{\sum_{i \in DC} w_i y_i}{\sum_{j \in DC} w_j}$$

(4)

is imputed in records corresponding to the final node in this WAID-tree for which the value of the numerical variable under consideration is missing. Here $y_i$ denotes the value of the numerical variable to be imputed in record $i$ of the donor cluster.

(iv) Finally, in the case where the user uses nearest neighbour imputation, specification of a weight variable has no effect.

## 5. Discussion

WAID 4.1 is an easy-to-use computer program for the imputation of missing data. Hopefully, the sample session as discussed in this paper illustrates this. Applying the steps in our sample session requires only a few minutes. WAID offers several imputation methods to impute for missing data: four methods for missing numerical data, and three for missing categorical data. It allows the user substantial control of how the imputations are carried out. The main strength of WAID is its ability to automatically generate homogeneous donor clusters using predictors specified by the user. This ability drastically reduces the time required to analyse the data before imputing for missing data.

The quality of the imputations carried out by WAID seems to be good in comparison with other imputation software. For detailed evaluation reports of WAID, we refer to Chambers et al. (2001b).

One important question remains to be answered: Is there a future for WAID? The answer seems to be confirmative. At several institutes, WAID is being used or further developed. At the University of Southampton, the methodology of WAID is currently being further developed. This work is being carried out under the Euredit project, a fifth framework programme project partly funded by the European Commission. At Statistics Finland, more evaluation experiments with WAID 4.1 are currently being performed. This work is also being carried out under the Euredit project. Finally, at Statistics Netherlands, WAID 4.1 has been applied to produce actual data sets that are used to publish statistical figures. Hence, support for using, developing and maintaining WAID at Statistics Netherlands is growing.

The precise form in which WAID will be further developed in the future, for example as an open-source code or as commercial software, has not yet been decided upon. This decision does not only depend on the original developers of WAID, but also on the rest of the world, for example on producers of commercial software who may be interested in incorporating WAID into their software products.

# 6. References

[1] Chambers, R. L., Hoogland, J., Laaksonen, S., Mesa, D. M., Pannekoek, J., Piela, P., Tsai P. and de Waal, T. (2001a), *The Autimp-project: evaluation of imputation software*, report, Statistics Netherlands, Voorburg.

[2] Chambers, R. L., Crespo, T., Laaksonen, S., Piela, P., Tsai P. and de Waal, T. (2001b), *The Autimp-project: evaluation of WAID*, report, Statistics Netherlands, Voorburg.

[3] de Waal, T., Plomp, R. and de Waard, J. (2001), *Manual WAID (4.1)*, report, Statistics Netherlands, Voorburg.

[4] Sonquist, J. N., Baker, E. L. and Morgan, J. A. (1971), *Searching for structure*, Institute for Social Research, University of Michigan.

# FORUM

This section of the ROS Journal contains contributions which are mostly for information purposes. Such contributions should present reports on:

- specific statistical research projects and programmes;
- statistical research activities in offi cial statistical institutes;
- experience on practical application of new techniques and technologies for statistics;
- experience on transfer of technologies and know -how both from the perspectives of those making the transfer and those to whom the transfer is being made;
- book reviews, etc.;
- other information of general interest.

Imperatively, papers published in the section have not been put through the usual full review process. Their review has been light and has been dictated by the nature of the p aper.

# Dissemination of business data: MASQ — a software program for single-axis microaggregation of quantitative variables

Daniela Pagliuca and Giovanni Seri

*ISTAT*
*Via C. Balbo, 16, I-00184 Rome*
*E-mail: pagliuca@istat.it, seri@istat.it*

**Keywords:** statistical disclosure control, confidentiality, microaggregation

**Abstract**

*Microaggregation consists of a class of disclosure control methods for microdata bas ed on the perturbation/substitution of the observed values. The idea behind microaggregation is to create pseudo-enterprises by means of a synthesis of the observed values of each variable in a small size - group of k units. This paper outlines the main feat ures of MASQ, a program implemented through the SAS language, developed for the microaggregation of quantitative variables using single -axis methods. MASQ is a software program designed to provide the Italian National Statistical Institute (ISTAT) with an easy tool to create public use files based on microaggregation techniques and to check the safety and quality of the microaggregated data.*

## 1. Introduction

Microdata stemming from official statistics before reaching the public have necessarily been treated for disclosure limitation concerning respondents. Disclosure limitation is a very recent research area of the national statistical institutes (NSIs). The development of new, fast, flexible and user-friendly technologies for the release of data has increased the demand for more detailed information and raised the awareness of respondents and NSIs regarding the problem of confidentiality — see Willenborg and de Waal (1996). NSIs usually assume that a user has an archive containing names of individuals and some publicly available variables that are also present in the released microdata file. Disclosure can occur if a match between a record in the released file and an individual is possible by means of the overlapping information.

The concept of risk of disclosure is mainly based on the idea of rareness with respect to a set of key variables. A protection scheme based on this idea is not often suitable for business microdata because of the characteristics of the population under investigation and the nature of the data collected. In many cases, populations of enterprises are sparse and firms are easily identifiable simply by their economic activity and geographical position (firms tend to group themselves in the same region and, therefore, identification outside these regions is easier). Moreover, a lot of the information collected is referred to quantitative characteristics with asymmetric distribution. Quantitative variables are often

representative of the firm size and this can lead to identification of enterprises. Many of the specific business data-protection techniques proposed are thought to be perturbative of the original data (see for examples McGuckin and Nguyen, 1988) in a way that disclosure of information is doubtful. It is obvious that, at the same time, the information content of perturbed data should be as similar as possible to the original data, in order to preserve the quality of statistical results. The hard question is to prove that data modified in some way can avoid breaches of confidentiality. To our knowledge, effective criteria for evaluating risk of disclosure in perturbed economic microdata have not been defined in practice. Therefore, the extent of disclosure protection provided by any type of perturbed data is mostly uncertain.

Recently, ISTAT has focused its attention on microaggregation methods, which are a set of techniques that proved to be useful in disseminating business data (Defays and Nanopoulos, 1992; Defays and Anwar, 1998) particularly for their peculiarities of conforming to ISTAT's confidentiality rule. The aim of this paper is to present MASQ, a software program for single-axis microaggregation of quantitative variables. In Section 2, we produce a brief overview of microaggregation methods. In Section 3, we describe the MASQ software developed at ISTAT.

## 2. Microaggregation methods

The idea behind microaggregation is to transform data in a way that the confidentiality rule of the 'threshold-*k*' is respected. On the basis of this rule, data can be disseminated without breaching confidentiality only if they are referred to at least $k$ individuals (in practice $k = 3$ or $k = 4$ are used). Microaggregation thus consists of the substitution of the values associated to an enterprise with a synthesis of the observed values in a small size-group of $k$ units (the size of a group is not necessarily a constant). For quantitative variables, the synthesis is operated by the average. This means that the more the enterprises are similar in the group, the less information is lost due to the microaggregation process. From this point of view, microaggregation can be thought of as a problem of cluster analysis, with the additional constraint that the number of enterprises in a cluster has to be equal to or greater than $k$. Given that the enterprises in the initial file are N, then the problem is to find a partition of N units in size-groups greater than or equal to $k$, while reducing the information loss as far as possible.

Information loss is usually measured by the distance between the original data and the transformed data. The smaller the distance, the lower the information loss. Given that we consider only continuous variables, we get the decomposition of the **total sum of square** (T) by adding the **within groups sum of square** (W) to the **between groups sum of square** (B). The partition-minimising W (or equivalently, the partition-maximising B, since T = B + W) is then an optimal solution for the problem.

Domingo Ferrer and Mateo Sanz (1998) outlined a way to reduce the number of partitions that have to be investigated in order to find an optimal solution via exhaustive search,

although they show that computing time grows exponentially with N. However, this is certainly the best solution if N is small enough to make it affordable.

In order to group enterprises, a variable, usually selected from those collected or computed as relevant, is used as a similarity criterion. This variable is called an **axis (or sorting) variable**. Data are sorted according to the axis values (in descending or ascending order) and then groups of *k* consecutive enterprises are formed. We refer to these methods as **single-axis methods**. If N is not a multiple of *k*, then the last (or the first) group has more than *k* units. Each method is characterised by the similarity criterion used to create groups of enterprises.

Different methods have been proposed according to different proximity measures among enterprises. Following Anwar (1993), we consider single-axis ranking methods as well as other microaggregation methods. In general, concerning single-axis methods, it is suitable to choose an axis which is used to sort and aggregate data, highly correlated with all the variables in the data set; a single variable or the first principal component method or the sum of standardised scores are possible axes. In effect, size variables such as turnover are suggested as a criterion for sorting individuals, but, in principle, any variable associated with the individual enterprises can be used (single-variable method). Therefore, it seems quite natural to choose the first principal component as the axis variable — the advantage is that all (or a subset of the most influencing) variables are taken into account in computing the axis (first principal component method). With the same reasoning, a method based on the sum of standardised scores of the observation has been proposed as an alternative way to compute the axis variable (the sum of z- scores method).

Other methods have been developed that relate to more than one axis. This set of techniques proved to be less invasive than single-axis methods by supplying lower indices of information loss (Defays and Nanopoulos, 1992).

Because of concerns about possible breaches of confidentiality, we concentrate our attention on single-axis microaggregation methods (Corsini et al., 1998; Franconi et al., 1998).

As shown in the example above, the application of single-axis methods produces identical pseudo-enterprises *k* by *k*, which implies that each cell in all the possible cross-tabulations between variables counts at least *k* units. This is a very strong issue for statistical disclosure control because it abides to the confidentiality rule used to disseminate tabular data in Italy (the abovementioned threshold-*k* rule). This is the most important advantage of this kind of method of protecting business microdata. Another is that it is easy to implement routines for single-axis methods into software applications with negligible computing costs.

In general, data microaggregated by using individual ranking does not conform to the threshold-*k* rule.

# 3. MASQ: a single-axis microaggregation tool for quantitative variables

## 3.1. About the MASQ project: a brief history

An extensive study on the application of single-axis methods was carried out on Italian business data, in order to evaluate the performance of different methods as far as the maintenance of the characteristics of the original data was concerned (Franconi et al., 1998).

During the study, we implemented single-axis microaggregation methods using ad hoc routines, with SAS language v 6.12 (SAS Institute Inc., 1990). Afterwards, we implemented a software prototype that creates microaggregates using the sum of z-scores method (Pagliuca and Seri, 1998). Our purpose was to supply a flexible tool, useful to disseminate business data. The software was subsequently developed by introducing the possibility to use other single-axis methods. We called the software MASQ, which is the Italian acronym for single-axis microaggregation of quantitative variables (we want the name MASQ to recall the English verb 'to mask', as masking data are often used in statistical disclosure control literature).

We used the **SAS/AF** software (SAS Institute Inc., 1989) that allows the creation of interactive-window applications and the **screen control language SCL** (SAS Institute Inc., 1994), a programming language for driving **SAS/AF** applications. SCL provides functions, routines and statements for manipulating data values and data sets, managing applications and controlling the application environments. The idea was to use a general-purpose statistics package, such as SAS, that incorporates extensive descriptive statistics, many statistical procedures, including factor and principal components analysis, cluster and discriminant analysis, etc., to build software. Moreover, SAS is the standard statistical package at ISTAT. We used SAS for the Microsoft Windows environment (Win 95), but it should be quite easy to take advantage of SAS portability to other platforms.

## 3.2. MASQ: an interactive application

MASQ provides a user-friendly interface to access a data set, to select variables and input parameters to create a microaggregated SAS data set. Users communicate with the application through a main window and subsequent menus and selection lists. Figures 1 to 4 show the main window and menus in MASQ.

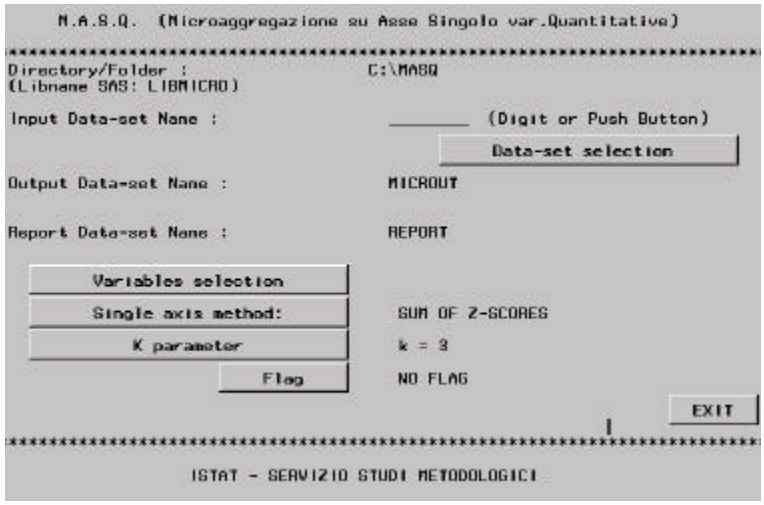



*Figure 1: The main window in MASQ*

*The main window: input data*

First of all, the user has to type the path of the file system where the software will find the input data set. A default path is suggested (C:\MICRO). The user can move data into this directory or change the path by typing a new one. Input data have to be in SAS data-set format. An SAS library (named Libmicro) is assigned to the directory chosen.

Using the **Data-set selection** button, all the data sets in the library are shown in a selection list window and a single choice is allowed. As an alternative, the user can type the name of the data set above the black line. The default name of the output data set (Microut) can be changed by typing another name. This will be a permanent data set under the SAS library Libmicro. In the same way, the default name of the report data set (Microut) can be changed. A fuller description of the output produced by the program is given at the end of this section.



The steps described above have to be completed before starting the program, otherwise only the **Exit** button can be clicked.

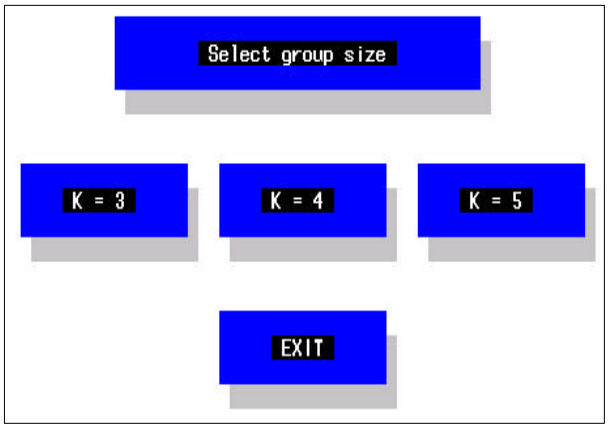
*The main window: the Flag button*

The **Flag** button allows the user to switch on or off a flag to mark a percentage of microaggregates according to a measure of distortion with respect to the original data. The mean of the Euclidean distances between the enterprises in a group and the pseudo-enterprises which substitute them is computed, so as to measure distortion in each group. Variables involved in the distance are standardised beforehand, in order to have 0 mean values and unit variances. Pseudo-enterprises are

sorted according to this measure of distortion. The program marks with '1' the pseudo-enterprises with low distortion and with '2' the pseudo-enterprises with high distortion. We set the percentage of pseudo-enterprises marked 2 at 10 %. If the number of enterprises in the population (or subpopulation) is less than 300 (consequently, 100 is the minimum number of groups in a subpopulation to activate the routine), then the routine does not run and the flag variable is set to 0. This flag, for our purpose, could be useful to avoid the effects of outliers (users of business data could, eventually, drop records marked 2) without increasing risk of disclosure.

*The* k *parameter menu*

Using the **K parameter** button, a block menu is opened (see Figure 2). This menu allows the user to change the value *k* to either 3, 4 or 5 (default value is *k = 3*).

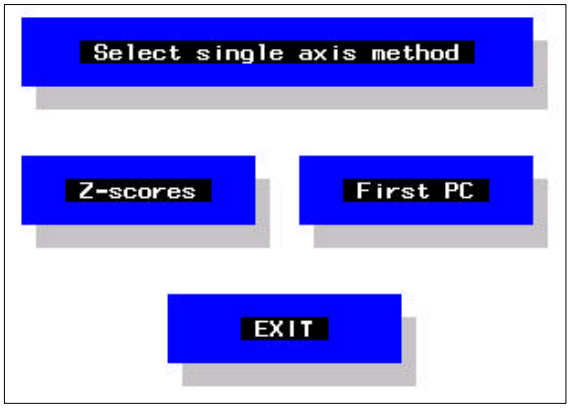

The larger the value of *k*, the safer the data released, but too large a *k* is not compatible with the idea of microaggregation.

In practice, only k = 3 or k = 4 is used.

*Figure 2: Group size selection menu*

*The methods selection menu*

Using the **Single-axis method** button, a block menu (see Figure 3) with two options is opened. At the same time, the sum of z-scores method and the first principal component method are implemented.




New methods can be introduced as a new option in this block menu.

Notice that single-variable selection can easily be applied in the particular case of both sum of z-scores and first principal component methods being used.

*Figure 3: Methods selection menu*

*Stratification:* Variables defining the subpopulations where methods are applied (usually subpopulations are defined by the geographical detail and the level of the hierarchical classification of the economic activities).

*Axis:* Variables that contribute to computing the axis (in order to apply the sum of z-scores method, it is necessary to compute the sum of standardised values over a set of variables that is thought relevant to define similarity between enterprises). The user selects variables, selecting only one variable; thus a single-variable method is applied.

*Microaggregation:* Variables that have to be microaggregated in the output data set.

*Individual ranking:* We are currently implementing this routine in the program. However, we introduce the individual ranking option here because it is not a single-axis method. As stated above, we think that individual ranking microaggregation can be used in a profitable way for some kinds of variables, such as percentages or indices.

*Unmodified:* Variables that have to remain unmodified in the output data set.

*Weight:* In the case of sample surveys. This appears separately because it could be used in averaging the observed value and/or modified by randomisation or microaggregation.

*OK:* To start the application.

*Exit:* Return to the main window and delete previous selections.

Each of the first five blocks corresponds to a selection list window showing the input data-set variables. We decided to give the user a way to select variables from among those collected in the survey and to choose from possible variable combinations.


*Figure 4: Variables selection menu*

*Output data sets*
MASQ produces three data sets as the output of the program. All of them are SAS data sets.

The first is the data set containing microaggregated data and an identification code that allows a link with the original data set. Such a link could be useful to do some checks. Other variables are included in the output data set if the user has selected them by using the 'Stratification', 'Unmodified' or 'Weights' lists. The flag variable will also appear in the data set if the corresponding button in the main window is selected.

The second data set is a report containing information on user choices (list of selected variables, names of the data sets, path of the files, value of *k*, method used).

The last is a data set containing indices about the microaggregation process. For each variable, the ratio between the variance of the microaggregated data and the variance of the original data is computed. The more this ratio is nearer to 1, the more the quality of the representation of the original data is high and the microaggregation process provides good results. If stratification variables are selected, such indices are computed for each subpopulation. As with the other two output data sets, it will be placed under the same library (Libmicro).

## 4. Conclusions

We implemented MASQ because it supplies the releaser of business microdata with a user-friendly interface to access a data set, to select variables and input parameters to create a microaggregated SAS data set. We considered the software on a wider project basis and, in the near future, some other methods based on cluster analysis are to be included in MASQ, that will allow an option to take into account the **dominance rule** (see Willenborg and de Waal, 1996).

## 5. Acknowledgements

The authors would like to thank Luisa Franconi for helpful comments.

The views expressed are those of the authors and do not necessarily reflect the policies of ISTAT.

## 6. References

[1] Anwar, M. N. (1993), 'Microaggregation: the small aggregates method', Eurostat internal report.

[2] Corsini, V., Franconi, L., Pagliuca, D. and Seri, G. (1998), 'An application of microaggregation methods to the Italian business survey', *Proceedings of the Statistical Data Protection Conference '98,* Lisbon.

[3] Defays, D. and Anwar, M. N. (1998), 'Masking microdata using microaggregation', *Journal of Official Statistics* , Volume 14, No 4, 1998.

[4] Defays, D. and Nanopoulos, P. (1992), 'Panels of enterprises and confidentiality: the small aggregates method', *Proceedings of Statistics Canada Symposium '92* , *Design and analysis of longitudinal surveys* .

[5] Domingo Ferrer, J. and Mateo Sanz, J. M. (1998), 'Practical data-oriented microaggregation for statistical disclosure control', *Report de Recerca*, DEI-RR-98-005, Departament d'Enginyeria Informàtica, Universitat 'Rovira i Virgili', Spain.

[6] Franconi, L., Pagliuca, D., Piersimoni, F. and Seri, G. (1998), 'SDC techniques for microdata: study of some perturbative techniques', Deliverable MI-3/D1 — SDC project, Esprit No 20462.

[7] McGuckin, R. H. and Nguyen, S. V. (1988). 'Use of ''surrogate file'' to conduct economic studies with longitudinal microdata', *Proceedings of the Fourth Annual Research Conference, US Census Bureau* , 20, pp. 193–211.

[8] Pagliuca, D. and Seri, G. (1998), 'The release of business microdata: a software prototype for microaggregation', *Proceedings of the NTTS '98 International Seminar on New Techniques and Technologies for Statistics* , Sorrento, Italy.

[9] SAS Institute Inc. (1989), *SAS/AF software: usage and reference — Version 6,* First edition, SAS Institute Inc., Cary, NC.

[10] SAS Institute Inc. (1990), *SAS language: reference — Version 6,* First edition, SAS Institute Inc., Cary, NC.

[11] SAS Institute Inc. (1994), *SAS screen control language: reference — Version 6,* Second edition, SAS Institute Inc., Cary, NC.

[12] Willenborg, L. C. R. J. and de Waal, T. (1996), *Statistical disclosure control in practice* , Springer-Verlag, New York.

# Process quality control to prevent non-sampling errors in the Italian multipurpose system of social surveys

Rina Camporese, Saverio Gazzelloni and Paolo Piergentili

*ISTAT*
*Via Ravà, 150, I-00100 Rome*
*E-mail: ca.rina@katamail.com, gazzello@istat.it, piergent@istat.it*

**Keywords:** survey process, quality, monitoring, non-sampling errors

**Abstract**

*ISTAT has been conducting a multipurpose integrated system of social surveys since 1993. Seven surveys share methodological approaches, analysis and dissemination strategies. Different survey phases are part of a unitary process: they are integrated and governed by other transversal phases which last all survey long (e.g. documentation). This work aims to limit and measure non-sampling errors. The whole survey process is accompanied by instruments and indicators to evaluate quality, monitor methodological rules and check the presence of unwanted events. Quality verification is continuous; it is not only* ex post. *Implementation of those instruments is always ongoing and needs the acquisition of competence in powerful and innovative information technology instruments.*

## 1. Introduction: Multipurpose system of social surveys

ISTAT has been conducting a multipurpose integrated system of social surveys since 1993. Seven sample surveys on resident households, one annual on living conditions and another six on specific themes, rotate in a time slot of five years the covering most important topics of social statistics: health and access to related services, citizens' safety, family and social subjects, time use, travel and tourism, leisure time and culture. On average, every survey involves 24 000 households (about 70 000 individuals) and 800 variables are collected. The seven surveys constitute a system because they share methodological approaches, data-collection and data-treatment techniques, analysis and dissemination strategies.

The multipurpose system is planned to collect structural information on the main aspects of citizens' daily life and also to be open to contingent informative needs that can emerge year after year. Therefore, even though basic structures of questionnaires and surveys are defined, innovations and modifications are frequent [1].

(1) This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

(2) The number of visits to a web page is not necessarily equal to a comparable number of unique visitors (D. Coon, e-mail, 12 April 2001).

Moreover, every survey is not only an informative instrument but also constitutes a methodological laboratory to test new data-collection techniques, wording, training and respondent sensitisation strategies, new instruments to check and correct data, and other methodological approaches. Experiences accumulated for one survey are transferred as soon as possible to the following ones [2].

## 2. Non-sampling error prevention and measurement

Attention to quality and possible sources of errors covers all survey phases, starting from project definition and ending with data dissemination. Non-sampling errors are less formally defined and measurable as compared to sampling ones, but this does not allow researchers to consider them less important. Actually, they can have an even stronger impact on estimation quality when a sample survey is concerned.

In order to prevent and keep under control non-sampling errors, survey steps are considered as a unitary process towards the final estimates. For instance, validation of results cannot be unaware of the main fieldwork problems (e.g. respondents' selection bias). Weight calculation and post-stratification are strictly linked, not only to sample design but also to the results of fieldwork (first- and second-stage non-response, checks on interviewers' misbehaviour, etc.).

Therefore, for every multipurpose survey, many separate phases are identified, but these phases are strongly integrated and governed by other transversal phases which last all survey long (e.g. documentation, monitoring, etc.). Every phase is accompanied by a set of indicators planned to check the correctness of operations performed.

For example, for both computer-aided telephone interview (CATI) and paper and pencil interview (PAPI) surveys, a monitoring system of fieldwork has been designed, based on the calculation of performance indicators. The system of indicators keeps under strict control interviewers, interviewees, refusals and other non-response, by cause, territory, time of day, and other important variables. Indicators are calculated and checked every day (for CATI) or at least twice a week (for PAPI) in order to let anomalies and undesired events emerge as soon as they occur. Web technologies are used to transfer data from local data-collection units and ISTAT central offices, where surveys are governed [3], [4], [5].

When the project phase is performed, modes and techniques to check and integrate the subsequent steps are already defined. Not only are separate phases planned, but also the ways to pass from one to the other, using all the documented information coming from previous steps for the subsequent ones. For instance, questionnaire definition is documented through a relational database that stores questions, filters, possible answers, associated codes, and so on; this database is also projected to be the starting point for data checks, correction and tabulation (see Section 3).

One of the main purposes of this work is to limit and measure non-sampling errors, which mostly derive from fieldwork, but are also related to data entry, checks and correction, weight calculation, and so on. Therefore, the whole survey process is accompanied by instruments and indicators that enable researchers to evaluate quality, monitor the respect of methodological rules given to other operators (e.g. interviewers, coders) and check the presence of unwanted events.

Quality verification is not only *ex post* (after steps are concluded), but is continuous, as far as possible. By keeping indicators and 'alarm bell' events under continuous observation, it is sometimes possible to intervene and resolve problems during the conducting of the survey. Where non-sampling errors or unforeseen events occur, they should emerge as soon as possible and, possibly, be corrected or evaluated when they occur.

The implementation of monitoring instruments is always ongoing. Usually, a new instrument is tested and refined for one survey and then is transferred and adapted to all the others. To carry on this process towards quality, it has been necessary to acquire profound competence in powerful computers and innovative information technology instruments.

## 3. New software under construction

At the beginning of 1999, a process of re-engineering of the existing software started. Every step of various multipurpose surveys used to be, and still is, managed and monitored by some software program; but even though conceptual frames and theoretical approaches were more or less already shared among surveys, sometimes software and programs developed for different surveys were scarcely integrated. Therefore, after recognising best practice performed for every survey and after identifying improvements required, a phase of standardisation and production of common new software started [6], [7].

According to a modular approach, all old software is progressively going to be substituted by new and improved programs — generally written in SAS language — that take into account all the experiences accumulated in past surveys. The idea is that all survey phases will be governed and monitored by specific software programs and that all these programs will be integrated.

Due to the shortness of this paper, only two software modules are presented. They are probably the more complicated and useful ones and they are strongly connected with many phases of a survey, from questionnaire definition to data dissemination. The modules are Survey Definition DataBase (SDDB) and check-correction tabulation software (SASS: Supporto all'Analisi dei Segmenti Sezione — Support to Segment-section Analysis).

The first, SDDB, is built to store and manage all definition parts of survey questionnaires, describing sections, questions, jumps, filters, possible answers, variables, etc. It also links and integrates such metadata with other, more technical and process-connected information: record formats, fields, identification keys, codes, classifications, domains, etc.

It is an Oracle relational database, resident on UNIX, populated and accessed for query by Access applications using ODBC protocol.

The second, SASS, is an application that produces, stores and handles programs that users create to check, correct and tabulate data. It also allows users to recover easily and reuse old programs, eventually introducing some modifications. SDDB and SASS are, of course, strongly integrated and the first module (SDDB) is ancillary to the second (SASS), which continuously draws information from the first one. The SASS application is written in SAS language, using SCL for the graphic user interface and SQL to access information stored in SDDB.

For multipurpose surveys, check and correction steps are conceptually designed as modular: according to the questionnaire structure, data manipulation is structured in various phases — each phase is often linked to one or more sections of the questionnaire (e.g. structural variables, family composition variables, health variables, dwelling section variables, etc.). All the different phases consist of check, correction and tabulation parts.

Phases are organised in a hierarchical structure: some leading variables have to be corrected first (usually structural and family variables), while other section priorities are defined for each survey on the basis of logical and causal relations, dissemination strategies or other criteria. The hierarchical structure of correction phases often determines the data-treatment schedule over time: some actions cannot be performed before others are concluded, but parallel processes are also possible, for example when two phases treat completely independent variables.

The data-treatment structure is usually defined during a survey project phase, but it may vary from survey to survey and it also varies for the same survey performed on different occasions (e.g. the annual survey questionnaire on living conditions varies a little every year). It can also vary for the same data set because of urgent or unforeseen needs for information (e.g. changes in dissemination plan). Therefore, to handle and manage multipurpose data sets, very flexible software programs are needed, not only integrated ones.

Moreover, data access and elaboration have to become free from rigid ties, such as record formats, file and variables names, and other specific and technical aspects that may vary from survey to survey. Every time a survey starts (usually twice or three times a year, not to mention quarterly surveys on tourism), improvements and changes appear. Some are small (only a few changes to questions or possible answers); some are greater (entire sections may be cancelled or introduced, classifications may vary, etc.). All changes have, of course, to be documented, but, above all, they have to produce the minimum (possibly, no) effect in terms of rewriting of programs.

Many people are working contemporaneously on the same data set, checking, making corrections and tabulations, and, conversely, the same person can work on two or more surveys at the same time. These people are usually experts in data analysis and validation, but not necessarily experts in programming (SAS, COBOL, HTML, etc.). They need

instruments which enable them to work on data and which guarantee online access to all documentation but which do not require continuous assistance from programmers.

If the described needs can be satisfied, continuous methodological and content modifications could be faced more easily. Furthermore, changes could have a minor impact on the risk of non-sampling errors due to manipulation of wrong data.

As mentioned before, SDDB is a relational database that documents and provides easy access to all information about questionnaire and data files. Much work was required, but it is not as innovative as the second component (SASS), which will be finished in the next few months (its last component on tabulation is currently under B-test). That is the reason why this paper concentrates on SASS description.

SASS software allows the possibility to open *n* parallel analysis processes: each process can be constituted by *m* segment-sections. For each segment-section, one or two catalogues can be associated. Every catalogue contains information on a set of variables considered, possible new variables calculated ad hoc (macrovar, array, recoded variables), check programs, correction programs and tabulation programs. Every process provides the possibility to write a complete output data file containing modifications and corrections performed during the whole process.

This is hierarchically structured, which may sound peculiar, except that it mirrors the logical approach that is used to check multipurpose data:

- main paths **(processes)** for data checks are conceptually defined, and there is one for every set of variables that can be logically linked to each other (e.g. structural variables, information on sport, 'small' criminality). Ideally, every process contains variables that are independent of the variables included in other processes. Only the first, or leading process, defined for structural information (e.g. sex, age, education) contains variables that appear in all the other processes, because they are fundamental to interpreting and evaluating all other sections on social and daily life phenomena;

- a process can involve one or more sections of the questionnaire **(segments-section)**. For instance, data on health are spread throughout several sections of the annual living conditions survey (e.g. chronic diseases, hospitalisation, medicines' assumption, healthy habits). Each segment-section can be included in a process to be corrected;

- a segment-section can contain one or two **catalogues**, because variables contained in one section can sometimes be considered and treated in different ways (e.g. qualitative and quantitative variables have different approaches to correction) or may be analysed sequentially (e.g. a first set of variables has to be definitively corrected before approaching the second set).

Only at the final level (catalogue) is specific information on variables and programs defined, while the whole hierarchical structure is more apt to describe and document the logical path followed to analyse data.

A special remark needs to be made about the word 'hierarchical' used to describe the SASS structure. In this context, it has a general meaning, because data checking sometimes is a recursive process, and not always predefined at the beginning. Moreover, the same sections and variables can appear in various processes playing different roles (for comparison and validation, for cross-tabulation, for correction, etc.).

Logical links between daily life and living condition variables are infinite: behaviour and lifestyle variables influence one another through an intricate net of ties and connections. Therefore, a too-rigid structure would not have been appropriate to perform the required job. In social surveys, it is not possible to subdivide variables into mutual exclusive subsets, because the same variable can appear in many different check phases and can have links to different phenomena. It is not possible to classify variables in a priori categories (e.g. descriptive, interpretative), because the same variable can assume different meanings; for instance, type of household can be analysed in itself or can be used as a cross-tabulation variable to understand and interpret leisure time habits. It is not always desirable to define a fixed priority schedule in data correction, because only a few rigid causal links are present among social variables.

The following is a description of the SASS instrument. All SASS application modules are integrated with SDDB software, so as to allow the user to navigate among metadata to choose elements on which to work. A segment-section is connected to a questionnaire section by questions, possible answers and variables.

In concrete terms, a segment-section is a micro data set (one record for each sample unit, usually one individual) containing all the variables included in the selected questionnaire section. However, any other single variable present in the record can be included and some ad hoc variables can be created, thus modifying the existing ones (macro variables, arrays, identification keys, or others). This technique allows the dimension of the working data set to be minimised, so as to consider only the variables that are actually needed for elaboration. By means of reducing information redundancy, data processing time and the risk of performing wrong corrections are also reduced.

Permanent links to SDDB meta-information allow users to attribute meaning to the selected micro data set. Not only single variables, questions and classifications are accessible, but also the entire structure of the questionnaire, in terms of jumps and filters, is considered. Actually, the application automatically adds one dichotomous (true/false) variable for every filter present in the selected section and for the questions included. For instance, if a section is allocated to people aged 14 or over, individuals will be classified as aged 14+ or not, depending on their age; if question X has to be answered only if the answer to the previous question is affirmative, a new variable will be created to identify respondents who answered affirmatively to question (X – 1).

Once the segment-section is defined, it can be accessed and used on every occasion for analysis and tabulation. Main check tabulations are already automatically predefined, mostly those linked to the questionnaire structure in term of filters and jumps (conditioned and unconditioned). However, for data analysis, users can perform every cross-tabulation needed, simply by clicking on the variables and accessories (e.g. filters, formats, labels) appearing on the screen. Elaborations are performed online and when a tentative tabulation is recognised as useful and definitive it can be saved and stored in a catalogue so as to be documented and eventually used in the future. This approach derives from the fact that data checks are sometimes born of attempts and ad hoc tabulations that become necessary and clear to researchers, usually after some basic data elaboration has already been performed.

If a value inconsistency appears among variables, then a modification can be made to data. Every modification is identified by a rule that can be stored in the catalogue linked to the analysed section. Every time the segment-section is accessed by some user, all stored rules (corrections) are performed before the user can have access to data. In this manner, data treatment can be interrupted and restarted whenever necessary, without losing the work previously done.

Alternatively, tentative corrections that do not lead to satisfying results can be eliminated.

Once all checks and corrections are defined, work on a section can be considered as concluded and a new, corrected data file can be produced. Two possible choices are available: a complete data file can be written, or a partial data file can be produced containing only identification, structural and segment-section variables. The second choice is used, for example, when some additional corrections to the analysed variables (e.g. probabilistic imputation) are necessary.

In case observed errors are randomly distributed (non-systematic) and no clear priority or causal link can be identified among inconsistent variables, random or donor imputation offers a better approach to correction. At present, only deterministic interventions can be made by SASS (if A, then B), while probabilistic or donor imputation is performed via external software (SCIA, RIDA, ad hoc programs).

To limit the possibility of wrong and incongruent interventions on data, only a recognised manager of the segment can store data on the catalogue by use of the check and correction programs and/or write a new data file, whereas other users can perform only extemporaneous analysis.

After data are corrected and validated, tabulations can be made to perform analysis or disseminate information. Also, the SASS tabulation module can draw all the information available from SDDB and help users in writing SAS Tabulate programs which can be permanently stored in the catalogue. Crossings, subsettings, formats, labels and titles can be made user-friendly on the screen, so that each click is automatically translated into SAS or SQL instructions. A data step can be defined and performed before tabulation, in order to create ad hoc variables or to recode existing ones.

In the end, each catalogue, associated to a segment-section, stores and documents all checks, corrections and tabulations performed on a set of chosen variables (usually all variables included in one section). The whole set of operations can be easily recovered and performed again on a new data set coming from another survey, or, more frequently, from another wave of the same survey (e.g. the health survey conducted in 1999–2000 was split into four quarterly data-collection parts, to take into account the seasonal nature of many surveyed phenomena).

In the near future, the tabulation application will be linked and integrated with the 'aggregated data application', which is in its project phase. For recurrent questions and variables (that are present in several surveys), many tabulations have already been defined, which continue to be the same every time the survey is carried out. In these cases, an aggregated data set would be very useful to avoid online data processing of micro files, as well as for quick access to already elaborated data. At the same time, all data coming from the same survey and disseminated in different publications can be available in a unique aggregated data set, ready to be consulted without having to go from one book to another.

The existing SASS tabulation application is already storing some information that will be useful to create aggregated data files according to the shape defined by a SAS tabulation program stored in a SASS catalogue. Information necessary to perform this future step mainly consists of used variables, their classifications, their position on the table (page, row, column), statistics calculated (e.g. sum, percentage, mean), weighting system, etc. Given this information, it will be possible to widen navigation to the multipurpose surveys database, by asking, for instance, which surveys produced tables for one particular phenomenon, or to gain access to the meta information.

## 4. Conclusions and future perspectives

All actions performed on survey data (fieldwork indicators, micro, macro and meta files) are continuously revised and improved. Therefore, every software program produced has to be as flexible as possible. Its functions have to be 'soft', in the sense that they have to be open to possible changes or innovations. More emphasis is put on documentation and easy access to existing information than on rigid and prefabricated elaboration modules.

For the future, implementation of other modules of the survey information system are planned, so as to cover all survey phases (e.g. dissemination of aggregated data, as previously described).

A possible evolution of all software used could be towards the web environment, translating, for example, SCL graphic interfaces into Java. Transferring SDDB and SASS to a web environment (intranet or Internet), for instance, could open a new perspective for data dissemination.

## 5. References

[1] Camporese, R., 'La qualità nel sistema di indagini sociali Multiscopo', Atti IV Conferenza Nazionale di Statistica, Rome, 1998.

[2] Sabbadini, L. L., 'La strategia di qualità nelle indagini sociali telefoniche dell'ISTAT', Atti IV Conferenza Nazionale di Statistica, Rome, 1998.

[3] Iannucci, L., Quattrociocchi, L. and Vitaletti, S., 'A quality control approach to CATI operations in safety on citizen survey: the non-response and substitution rates monitoring', *International Seminar on New Techniques and Technologies for Statistics,* Volume II, Sorrento, Italy, 1998.

[4] Quattrociocchi, L. and Sabbadini, L. L., 'Telephone survey, new problems and new solutions in monitoring data quality: the Italian experience', *International Seminar on New Techniques and Technologies for Statistics,* Volume I, Sorrento, Italy, 1998.

[5] Muratore, M. G., Quattrociocchi, L. and Sabbadini, L. L. (a cura di), 'Indagini sociali telefoniche: metodologie ed esperienze della statistica ufficiale', ISTAT, Metodi e Norme (in the course of publication).

[6] Bagatta G. L., Giusti, M. V. and Perez, M., 'The transition process for the implementation of the statistical information system on tourism demand in Italy', IVth International Forum on Tourism Statistics, Copenhagen, 17 to 19 June 1998.

[7] Camporese, R., De Francisci, S. and Piergentili, P., 'Integrated data system for integrated system of social surveys in Italy', *Proceedings of the NTTS International Seminar on New Techniques and Technologies for Statistics* , Sorrento, 1998.

# Official macroeconomic statistics in the European Union: an interesting topic for political economists

Klaus Reeh ([1])
*European Commission, Eurostat*
*5, rue Alphonse Weicker, L – 2721 Luxembourg*
*klaus.reeh@cec.eu.int*

**Abstract**

*In this paper I describe the institutional nature of official macroeconomic statistics. I hope to encourage political economists to look into this subject, since macroeconomic statistics are not being discussed as much as they ought to be and this might hamper the understanding and proper use of these statistics. I start with a look at the political role of macroeconomic statistics in the EU. Then I sketch the consequences of their political use and concomitant politicisation, as this has prompted important changes in the content and structure of official statistics. Finally after having looked more closely at Excessive Deficit Procedure statistics I turn to the worries that are occupying EU statisticians and the challenges ahead of them.*

## 1. The role of official macroeconomic statistics

National accounts are at the centre of official macroeconomic statistics. They are supplemented by balance of payments and other satellite accounts. Aggregate economic statistics have also to be added: production, prices, labour force and unemployment, credit and external trade. All these statistics provide a perception framework for charting economic trends and sketching a given economic situation. Growth, inflation, productivity, unemployment, debt, trade - indeed, the changing face of the economy as a whole - are viewed against the backdrop of categories underlying these statistics and charted numerically by them. Policy measures are assessed, and successes or failures captured in terms of these categories. The statistics are therefore more than a too l for measuring the economy - they embody a convention on how we should see the economy, what we should pay attention to and aim for.

These categories are not an invention of statisticians, but the result of theories devised by economists. Statisticians are responsible for squaring statistical records with theoretical concepts. They make categories empirically ascertainable, attribute a numerical significance to them, and bestow on them a kind of natural objectivity that makes them

([1]) Klaus Reeh is head of the business cycle statistics unit at Eurostat, the Statistical Office of the European Communities and also lectures in European economic statistics at the University of Trier. The views expressed in this paper, which is based on an original presentation made in May 2000 at the meeting devoted to institutional economics of the *Political Economy Working Group* in Heppenheim, Germany, are his own.

appear less notional. These categories and their numerical manifestations play by now a central role in the political debate.

Improved comparability has established Community statistics - an amalgam of national statistics derived from nationally collected data - as the reference framework for the political process in the European Union. They help perceptions and opinions to converge. They were important in forging political consensus allowing the Community to act. Recently, however, their role has moved beyond that of merely providing a common perception framework enabling policy co-ordination. They are incorporated into decision-making processes. GNP determines national contributions to the EU budget. Government debt and deficit indicators play a central role in assessing convergence and thus in determining monetary union membership. The eligibility of regions for assistance is determined by per capita GDP (expressed in purchasing power standards). Price stability, the ECB's guiding principle for monetary policy, is rendered operational through a harmonised consumer price index. Under the Stability and Growth Pact, budgetary discipline may be relaxed during a downswing, defined as a not growing real GDP in two successive quarters.

More or less automatic procedures driven by statistics have eased the political process. Statistics have provided categories that are more comparable than most alternatives. By making views converge they have fostered consensus and facilitated integration, dampened conflicts of interest, and provided a benchmark for measuring and assuring compliance with Community procedures.

## 2. The changing face of official macroeconomic statistics

The adoption of this new role has gone hand in hand with changes to the structure and content of macroeconomic Community statistics. They had to possess certain attributes before they could play this role. The key attribute of comparability has always been at the centre of EU statisticians' work. Their endeavours have borne fruit, as the achieved comparability is by now sufficient for the political use of statistics. Such use makes sense only if the statistics provide ample justification for the decisions in question, so they must also be applicable to the issue at stake. The use of purchasing power parities, for example, was a way of justifying and restricting redistributive measures at the same time. Similarly for reasons of tax equity, statisticians had to look for mutually acceptable and convincing ways of including the shadow economy into GNP estimates.

Consistency is another crucial feature. EU statisticians try to overcome the lack of spatial consistency. Major problems arise with the recording of intra-Community flows where, e.g. imports do not necessarily tally with exports. Being able to compare such figures is not synonymous with being able to consolidate them without any problem. The diversity of methodological approaches has a great impact. Statisticians have to reconcile data from a wide variety of sources whose composition varies among Member States. It is not easy for this to be understood and any resultant discrepancies accepted. The lack of consistency is

often taken to mean that Community statistics are of inferior quality, but this lack could also be an indication of the inferior quality of national statistics. Member States, however, are still sovereign, also in statistics, and yet not prepared to make their statistics subject to a Community consistency proviso.

Adequate concepts have had to be developed and fine-tuned for Community statistics to be sufficiently comparable, relevant and consistent.. However, organisational changes have been at least as crucial to the successful use of statistics. An organisational milestone has been the placing of official statistics on a firmer legal footing. The Amsterdam Treaty enshrines a set of basic principles for official European statistics. The Maastricht Treaty contains provisions on official statistics needed for monetary policy. High level committees also for comitology decision making ([2]) have been set up to allow for faster and more purposeful headway. Quite a few important legal acts were adopted such as the European System of National Accounts (ESA) regulation, the Council Directive on the harmonisation of GNP compilation, the Council Regulation on the Excessive Deficit Procedure, or the legal acts on the harmonisation of consumer price indices. The compilation of the main macroeconomic indicators is now codified by legislation, at least for their use in European policy-making.

Planning at EU level has been improved, as has the dovetailing of national statistical programmes. Official statisticians have been working also towards greater transparency. They achieved greater material and methodological independence and strove for mutual recognition. The technical independence ([3]) of statistical agencies has been enhanced, although Eurostat is still part of the European Commission and National Statistical Institutes part of national administrative structures. A closer co-operation between statisticians from NSIs and central banks has been established, notably for ensuring the adequacy of convergence statistics. As of recently the focus of their co-operation is on short-term statistics indispensable for framing monetary policy. There is more emphasis on questions such as EU/EMU aggregation and consolidation, whilst the operational objective is to produce statistics faster and more reliably, because EU/EMU statistics are still incomplete, and far more seriously, too sluggish when compared with their national counterparts.

Eurostat's orientation has changed too. Hitherto geared fairly closely to meeting Community requirements, it is increasingly providing information that goes beyond the requirements of European institutions. Entrepreneurs, financial market operators, all sorts of associations, academia and the media have become its customers. Its presswork has improved, and so has its dissemination of statistics. It also seeks advice from the scientific and professional community and is prepared to expose itself to the criticism of users and academic statisticians.

---

([2]) Council Decision of 13 July 1987 Council Decision of 13 July 1987 laying down the procedures for the exercise of implementing powers conferred on the Commission (87/373/EEC)

([3]) Commission Decision of 21 April 1997 on the role of Eurostat as regards the production of Community statistics (97/281/EC), and Council Regulation (EC) No 322/97 of 17 February 1997 on Community statistics.

Community macroeconomic statistics are much more transparent than before, their legitimacy is enhanced and they command much more authority. It's no wonder that they have become significant and the object of so much attention and lively debate. Although the face of the European Statistical System (ESS) has changed considerably, the ESS is still a long way from being a truly federal statistical body as it is wrongly assumed to be already. Community statistics are still few and far between, and turn out on closer inspection to have a national slant, as administrative structures leave their imprint also on macroeconomic statistics. Furthermore, official statistics are a public institution and as such part of the public infrastructure marked by national traditions and subject to national law. Nevertheless, in spite of the limitations that the subsidiarity principle places on the development of a more uniform infrastructure for official statistics, the ESS is now sufficiently robust and efficient, also when it comes to official macroeconomic statistics. Official macroeconomic statistics have become the mainstay of an institution that has found in Eurostat an organisation that is able to sustain and nurture them.

## 3. Problems and worries

In spite of their undoubted success, perhaps even because of the increased use being made of their macroeconomic statistics, the official statisticians are still confronted with some difficulties. Even worse, how to overcome these difficulties is far from evident. Some are content-related, whilst others are of a technical, organisational and political nature. A closer look at the statistics for the Excessive Deficit Procedure (EDP) may serve as an example.

### 3.1. Macroeconomic statistics and the Excessive Deficit Procedure

Macroeconomic statistics on government deficit and debt play an important role in this procedure, since comparable fiscal indicators are not available. This EDP procedure proved its worth in the decision-making process around the introduction of the Euro, because the statistics that shaped the political debate had met with broad acceptance. That said, they were not completely infallible, their content was not as relevant and encompassing as conceived and they were not beyond reproach from a procedural point of view. Thus they merit closer scrutiny, not least because this procedure plays an important role in the Stability and Growth Pact and will be used for future enlargements of the monetary union.

Council Regulation (EC) No 3605/93 laid down how the key statistical indicators 'government deficit/GDP' and 'government debt/GDP' for the Excessive Deficit Procedure have to be compiled. The regulation stipulates which ESA headings Member States should use to calculate government deficit and debt. It also defines GDP required for the compilation of the statistical indicator, and the extent of the Member States' reporting

obligation. This enables the Commission to check whether a Member State is in danger of incurring an excessive deficit.

The purpose of this regulation was to lay the foundations for a credible start into monetary union. Looking back, it seems to have been quite a success. However, there are some doubts as to whether this will equally allow managing the Stability and Growth Pact, in which statistical indicators play an even more important role. Firstly, the indicators appear to be less robust than they ought to. Secondly, their relevance may also be called into question. Thirdly and lastly, the procedure for setting up a statistical computing system for political ends is somewhat problematic as far as procedural principles of democratic decision-making are concerned. Each point will now be discussed.

### 3.1.1. Indicator robustness

The lack of robustness is attributable to t he fact that the dividing line between the public and private sector, between financial and non-financial transactions, but also between previous, current and forthcoming years in the ESA cannot be drawn as clearly or comprehensively as one would wish. New institutional units are coming into existence (e.g. special enterprises to finance and maintain public infrastructure, pension funds), and so are new types of transactions (e.g. interest and currency swaps, discounted and index-linked securities, financial leasing, or various types of State guarantees for debts). Existing institutional units change their legal forms (privatisation). Furthermore, Member States carry out special transactions to reduce their deficit or debt (e.g. purchase of gold or revaluation of reserves by central banks; changes to the maturity date for taxes, social contributions and insurance; shouldering of pension benefits by government; securitisation).

National accountants have to take classification decisions that have an impact on results. Their choices are politically significant. Whether these decisions ([4]) are right or wrong, appropriate or inappropriate, wise or unwise is open to debate. One thing is certain: the ESA regulation does not always provide them with a clear yardstic k. It is not easy, in view of differing institutional and legal structures, to measure all Member States with one and the same yardstick, and one that remains unchanged over time ([5]) . Such a common yardstick does not exist in fiscal accounting and in national accounting the measuring rod is not a true yardstick.

([4]) Cf. Eurostat press releases No 10/97 of 3 February 1997, No 16/97 of 21 February 1997, No 24/97 of 26 March 1997, No 33/97 of 30 April 1997, No 88/97 of 17 December 1997 and No 05/98 of 27 January 1998 or more recently No 15/2002 of 31 January 2002.

([5]) For further details, cf. Jean-Pierre Dupuis, The Reliability of the National Accounts in the Context of the Excessive Deficit Procedure, Seventh National Accounts Seminar, National Accounts Association, Par is, January 1998.

### 3.1.2. Indicator relevance

Doubts about the relevance of the EDP indicators are related to national accounting concepts. Generally speaking, and notwithstanding the aspect of autonomous decision - making, national accountants allocate an institutional unit to the public or private sector according to the origin of its income. If its income is derived mainly from taxes or levies, it is allocated to the public sector, whereas income from commercial revenu e places it in the private sector. However, many institutional units in mixed economies, whist showing high commercial revenues, nevertheless belong to the State; their commercial - and thus debt - policies are determined by the State, which is ultimately responsible for their debts. These units are assigned to one of the private sectors, and with them not just their debts, but also part of their deficits (provided the deficits are not offset wholly and immediately by State subsidies but are otherwise f inanced, at least temporarily).

National accounts do not take full account of who owns a unit, who is responsible for its deficits and who ultimately shoulders its debt. However, it is just these aspects that should be considered when assessing if the fiscal situation and behaviour of a Member State is consistent with what is required for monetary union. Inflation always comes about when government commitments that can no longer be honoured (or that can only be honoured at too high a social cost) have to be m onetised. If the objective is to avoid inflation at any cost, then countries may participate in a monetary union only if they can honour their commitments, and are not threatened with bankruptcy or crisis. As Member States cannot be "threatened with bankru ptcy" (although bankruptcy is at least thinkable in monetary union), they must scale down their commitments to a point where they can honour them out of taxes or levies. To assess this, it would have been more appropriate to distinguish between sectors tha t are able or unable to go bankrupt. However, the public sector as shown in national accounts approximates only to an "unbankruptable" sector, which goes way beyond the public sector.

Moreover, some government commitments do not even appear in the nation al accounts. Debt that is not bonded or covered by a contract is not recorded. Such commitments are either honoured according to the pay as you go principle by contributions or their contents can be changed anyway at virtually any time by legislation. Howe ver, these too are concrete commitments, at least for the citizens to whom these promises are made and who pay contributions and taxes for this very reason. It would thus be important for them to know whether their country and their partner countries are i n a position to honour these commitments. However, this information requirement cannot be adequately met by the statistical system set up under the Maastricht Treaty.

Finally, as mentioned before efforts are being made when calculating GDP at market prices to take account of the shadow economy, even though this contributes little to a country's ability to honour its commitments. Including these shadowy parts in the national accounts for EDP purposes does not make sense, as it lowers the EPD indicators for no good reason. Furthermore, the comparability of indicators suffers, as the Member States

differ not just with respect to the extent of their shadow economy, but also in the way it is integrated into the accounts.

In short, the logical approach national accountants must follow, the standards and definitions they apply and the methods they use to make their calculations fall somewhat short of the mark and this unavoidably so. The indicators that national accountants calculate, though perhaps better than many others, are simply less pertinent than they appear.

### 3.1.3. Procedural weaknesses

This regulation seems to have failed to provide the Commission with a reliable mechanism that would allow Eurostat, as a subordinate authority and in its capacity as a mere "measuring and calculating agency", to unambiguously measure the government deficit and debt of Member States. This shortcoming is particularly unfortunate, given that Protocol No 5 of the Maastricht Treaty lays down numerical threshold values for the deficit and debt indicators (3% and 60% respectively) that trigger off decisions.

The provisions serve as a basis for expert discussions under the aegis of Eurostat. This has led to the emergence of a benchmark of sorts in this matter, though this is anything but definitive. Specialists are thus availing themselves of the leeway for assessment, and must continue to do so in future. In granting such leeway, the Council dispensed with regulating this matter on the basis of the comitology procedure. The adoption of the rules governing the compilation of these indicators was not seen as a political and thus politically to be controlled process leaving leeway for assessment, but as a purely technical matter in which the Member States had merely to bring their expertise to bear. The fact that Eurostat has been granted "technical autonomy" by the European Commission clearly emphasised the non-political nature of the process.

This procedure makes the decisions taken by statisticians objectively, politically and legally unassailable. The consultation with national experts, no doubt, confers greater authority on Eurostat's decisions. However, if the political bodies (Council, Parliament, or Commission) pull back, the requisite legitimacy does not automatically devolve upon the administrative bodies upon which decision-making is then incumbent, in our case upon Eurostat. For that to happen, Member States must be possible to challenge Eurostat's decisions and subject them to legal scrutiny. However, these decisions are assumed not to relate to the exercise of administrative powers, the relevance and legality of which would have to stand up to such scrutiny. Rather, the macroeconomic statistics based on these decisions are seen as scientifically verifiable findings about the economy by an institution that is considered to be independent and whose findings are beyond reproach and apolitical.

However, leaving key decisions to administrations without politicians assuming responsibility for these decisions or without citizens or their representatives being able to

seek legal redress undermines political accountability and the rule of law. Eurostat, by having taken decisions of EDP relevance may not have forced the political decision about the start of monetary union, but it certainly strongly influenced it. All Member States that wanted to enter monetary union announced, after the relevant national statistics had been published, that they had "qualified" for the Euro and pre-empted more or less any political assessment. Moreover, whilst Eurostat was responsible for the accuracy of the figures in numerical terms, it was only able to submit the data from the national authorities to a methodological rather than a detailed numerical check on contents. It had to take the national authorities' word on this. National accountants certified in a way the Euro suitability of their countries and they will continue to do so for the Euro-compatibility of national budgetary policies.

Finally, amongst the politicians that wanted or at least backed this procedure, many could not predict the decisions that would emerge from it with any particular degree of accuracy. If politicians do not readily know what can and cannot be done under the rules they themselves have devised, then they are violating fundamental principles of constitutional democracy. Indeed, they are failing to fulfil the fundamental task for which they were elected, namely to determine, in the light of voters' wishes, what is permissible and what is not according to the status quo and the interests at stake.

## 3.2. Outlook

There is thus still every reason for European legislators to come to terms with the concern for monetary and fiscal stability as required by the Maastricht Treaty. After all, the European Union will acquire new members, as will monetary union. And the European Union may well have to deal with contentious issues as part of the Stability and Growth Pact. The decision-making processes of official statisticians will likewise come in for legal scrutiny. Political decisions remain political even when they are modified to take account of statistical indicators. Care should thus be taken to ensure that only bodies that have an explicit mandate to do so take such decisions. The trend in European and therefore national politics, of shifting responsibility for decisions onto political and administrative semi-automata, albeit temporarily, should be stopped. Unfortunately the temptation to do so will become ever greater in an ever-larger European Union. However, if we are successful in this venture, we will have made a valuable contribution not only to strengthening democracy and the rule of law but probably also to the economic success of the European Union.

However, official statisticians do not have to worry only about EDP statistics. There are more worries of a more general nature. Content-wise, the greatest danger is conceptual erosion of macroeconomic accounting systems. European integration and global liberalisation are just two of the reasons why dividing lines are becoming blurred (see indicator weakness above). This makes it difficult to reconcile what is economically meant with what can be measured in practice. It is ever more difficult to establish links between fiscal or legal categories and economic categories used in national accounts, making it no

easier to convey the message that national accounting is not just bookkeeping at national level.

Inflation statistics also are affected by conceptual erosion. Whilst the findings of the Boskin commission on inflation statistics in the United States cannot be applied to Europe as they stand, European statisticians will nevertheless find it increasingly difficult to isolate price changes from other changes. Technical progress, new marketing methods, the dematerialisation of products and the tailoring of products and services - none of this will make the official statistician's work any easier. The Statute of the ECB stipulates that its primary objective is to maintain price stability. In view of this, conceptual erosion should be taken very seriously, as it may have wide-ranging consequences.

Technical difficulties are also on the rise. National accountants are witnessing a loss or pollution of their data sources. A dramatic example of this is intra-Community trade statistics. Accessing sources is often quite difficult, because of restrictive legislation on data protection. Furthermore, the territorial status of the European Union and degree of European integration are in a constant state of flux (e.g. MUM to non-MUM membership), hampering aggregation and consolidation work and making time-consuming and problematic back-calculations necessary.

Stability in time, by which analysts put so much store, cannot even be approximated to. If it is achieved, it will be cosmetic at best and, at worst, contradictory, irrelevant or meaningless. The pace of economic and political change is too fast, too dramatic, particularly in and around the European Union. If official statistics are to keep abreast of developments conceptually, diachronic stability will increasingly have to be sacrificed. This is nothing new, but the lack of diachronic stability is so great that it is causing economists a great deal of trouble, increasingly forcing them to query their findings. However, instead of questioning their own methods, the empiricists amongst them want to see even more diachronic stability, little realising that they are asking the impossible or the irrelevant.

Organisational difficulties will also increase. Hitherto, technology has always come to the help of statistics, enhancing speed, accuracy and reliability in the capture of data and in the calculation and dissemination of statistics. Obviously these possibilities have not been exhausted, but progress can now be made only if NSIs manage to keep in step technologically - which will allow economies of scale. However, this is no easy thing to achieve. Furthermore, with the shift to infra-annual statistics, a lack of discipline here is much more readily felt. It is also becoming increasingly apparent that legal bases are not nearly as developed as they should be in this area.

All these difficulties are the consequence of the political EU structure. The ESS is not a federal system, and the implementation of Community legislation in the Member States is far more difficult than, say, the implementation of federal laws in a Federation. There are also limits to mutual assistance and closer co-operation between Member States, as the Member States are sovereign and some types of work simply cannot be assigned to other

countries. Finally, the Member States require far more methodological leeway than, say, the German Länder, as the administrative infrastructure of the Member States is far less homogeneous than that of the German Länder.

## 4. What is ahead?

For many, the development of Community statistics has not gone far enough. The strategy has indeed been marked by subsidiarity, aiming mainly at output comparability and going for input harmonisation only if absolutely necessary. However, this strategy seems to ha ve reached its limits in areas where speed and accuracy are of the essence, such as short -term statistics. EMU may, at least in the field of economic statistics, need a truly federal statistical system. This could give fresh impetus to input harmonisation and to the imposition of a Community consistency pr oviso on Member States statistics.

Basic confidence in common official statistics is still fragile, too. The great confidence shown by statisticians contrasts with a certain lack of trust on the part of national politicians, the general public and the media. It is all too carelessly assumed that national official statistics are manipulated, to qualify for a common initiative that is in the "national interest" such as the Euro, or ensure high returns from the EU budget. Consequently checks on national statistics had to be introduced for GNP as well as debt and deficit compilations and recently, for the calculation of harmonised consumer price indices.

It seems that the EU has only superficially extricated itself from an awkward situation by depoliticising decision-making with the help of statistics. Making the Community decision-making more objective through the use of statistics means not so much achieving a political breakthrough as modifying the decisio n-making process. This is probably less than what was intended, but it is nevertheless quite beneficial, but cannot be pushed much further.

For some people, things have already gone too far. They believe, by contrast, that greater use should be made of th e subsidiarity principle. The predominance of Community matters means that national statisticians barely have sufficient latitude to meet national requirements. Which is why the smaller and undoubtedly more heavily burdened Member States are always calling for Community financing for Community statistics. Again, this smacks of a certain reluctance towards federalism, as true federalism would mean that constituent states discharge federal duties as if they were their own.

Legislating on the methodology of official macroeconomic statistics also has its drawbacks. No country would think, for instance, of putting its national accounts on a legal footing. Nor for that matter do price statistics need regulating. However, in a European context, regulation is needed, if only to ensure that the Member States toe the line. Matters are getting complicated when improvements are made to statistics that are already used for political purposes, because these improvements might alter in one way or another the initial political deal. For example, the new ESA entered into force as a Council Regulation

in 1996. Use had to be made, however, of the old ESA dating from 1979 to calculate convergence criteria even though it had no explicit legal status, simply because a use of the new ESA would have changed the content of the Maastricht Treaty.

It should also be borne in mind that the political use of official statistics might have a political rebound effect. Political actions are, in full knowledge about the rules on the classification of institutions and transactions, cast in forms that distort the statistical picture in a way that makes it easier for national politicians to pursue their objectives. This not only limits what empirical scientific study can teach us, it also reduces the political appeal of official macroeconomic statistics, as political decisions are less readily justifiable. This is because the picture that such statistics are supposed to provide of a given situation should be relevant, whereas it is at best incomplete. Furthermore, official statistics are supposed to stabilise the expectations of, say, financial operators in respect of politicians' intentions concerning economic policy making, most notably in the realm of fiscal and monetary policy. However, such rebound effects reduce the ability of official statistics to stabilise expectations.

Finally, mention must be made of a tendency towards ever more "aestheticism" particularly in macroeconomic statistics. The quest for comparability, consistency and completeness, but also for conceptual elegance, leads increasingly to bona fide survey data and administrative records being mixed with "artificial" data. Data have to be extrapolated, estimated, supplemented or allocated. This means making highly specific hypotheses that are naturally paradigmatic. Under such circumstances, the users of such statistics, from researchers, analysts and consultants to commentators and politicians, run the risk of becoming lost in a maze of positive findings and normative positions ([6]).

Another problem is that users of statistics are increasingly falling back on business and household opinion surveys to supplement the statistical picture of the economic situation. What is of interest is not just the GDP change, but expectations about its change, not just change in the level of consumer prices, but the expected change, too. Expectations are turned into statistical information, if increasing use is made of such surveys to compile flash estimates. As expectations about economic trends have an impact on the trend itself, the inherent tendency towards self-referentiality in statistics is compounded. On the one hand official statistics help streamline perceptions and stabilise expectations. On the other hand, their self-referentiality increases arbitrariness and therefore the instability of expectations. Whether the stabilising or destabilising trend prevails ultimately depends on the degree of self-referentiality. This is, however, becoming increasingly common, partly as a result of the greater use being made of opinion surveys that relate directly to macroeconomic categories.

([6]) Cf. articles by H. W. Holub, J. Richter and G. Tappeiner, particularly "Ex post, ex ante und dazwischen" [Ex post, ex ante and everything in between] by Holub and Tappeiner in the "Jahrbuch für Nationalökonomie und Statistik" [Yearbook for economics and statistics], 1995.

Finally the biggest danger is that official economic statistics will encourage technocracy and prompt a concomitant de-democratisation of the political process. Figures that appear to be objective will in a sense make political decisions inevitable, and this not just at European level. Political decision-making is no longer about what is right or wrong, about what is permissible or not - it is about efficiency and inefficiency in the light of macroeconomic indicators that must be increased or reduced. However, increased legalisation, embellishment and self-referentiality appear to be pushing relevance and significance of macroeconomic statistics further and further into the background. Official macroeconomic statistics thus run the risk, if not of launching a debate, then at least of fuelling a debate that will be increasingly removed from political, social and economic realities. This danger is clearly much greater at European than at national level. The official statisticians in charge of Community statistics as a political and social institution should at least be conscious of these dangers, alert their political authorities and make the public aware of them.

# Note to Authors

**ROS** welcomes contributions from authors on results of research activities in official statistics. Contributions will normally be accepted in English. Nevertheless, reports in any other official languages of the European Union will be considered for publication, subject to the author submitting a summary of not more than 200 words in English. This summary must be submitted to The Executive Editor (at the address below)at the same time as the paper.

Before submitting their papers, authors are advised to seek assistance in the writing of their papers for the correct use of English.

**Copyright**: In submitting a paper, the author implies that it contains original unpublished work which has not (and is not planned to be submitted) for publication elsewhere. If this is not the case and the paper has been submitted elsewhere for publication, or actually already published, the author must clearly indicate this on the first page.

**Pre-assessment:** A first evaluation of each paper will be done as soon as possible and authors will be informed of this within a few weeks of the submission. Accepted papers will be published within six months of the author approving the final proof.

**Submission format:** The author should submit only one copy of his manuscript on paper. This should be accompanied by a summary of not more than 100 words. Manuscripts should in addition be sent electronically - that is, on diskette or by electronic mail. This will facilitate the editing process.

If a diskette is used, it must be the 3.5 inch disk in MS-DOS format. It must be a new diskette and must bear very clearly the name(s) of the author(s) and the title of the paper. Authors must ensure that the version of the electronic copy is exactly the same as the paper copy that accompanies it. The software tools used must be Word for Windows or WordPerfect. Authors wishing to use any other software tools must first agree this with the Executive Editor. Neither the hard or electronic copies of manuscripts will be returned to the authors.

**Submission fee**: In line with the policy of providing a forum for dissemination of results of statistical research activities, no submission fees are charged for unsolicited contributions received.

**The author:** Each paper must carry the following information on the front page in this order: (1) the title (2) the name(s) of the author(s), (3) their institution(s)/ affiliation(s), (4) a list of four or five keywords and (5) a short abstract of not more than 100 words. A clear indication of whom the proofs should be sent to (including the name, address, phone number, fax number e-mail address) should be given on this same page.

**Format:** Manuscripts should be printed on one side of the paper only. Pages should be numbered. All diagrams and graphs should be referred to in the paper as figures. Tables and figures are to be numbered in consecutive order in the text using Arabic numerals and should be printed on separate sheets.

**References:** References should be arranged in alphabetical order. Multiple references to the same author should be given in chronological order.

**Footnotes:** Footnotes should be kept to a minimum. When used, they should be numbered consecutively using Arabic numerals. Figures, tables and displayed formulae should not be included in footnotes.

**Reproduction**: Authors should note that printed copies will be made directly from photographic reproduction of final proof copies received from them. It is therefore imperative that high quality camera-ready originals are submitted. Illustrations should be of such quality that they are suitable for direct reproduction and ideally require the same degree of reduction. They should be clearly marked and correspond to references to them in the text.

**Proofs**: Two sets of proof copies will be sent to each author for final review. One of these must be signed and sent back to the executive editor within the time limit indicated in the cover letter.

**Free copies**: For each paper, author(s) will be entitled to one free copy of the journal of the issue in which the paper appears. The copy will be mailed directly to the author(s). Additional copies will be available at a special rate to the author.

**Further information:**
Enquiries relating to submission of papers etc. should be directed to:
Executive Editor
ROS, Eurostat, Room A2/162a
BECH Building
L-2920, LUXEMBOURG
Phone: +(352) 4301 34190 Fax: +(352) 4301 34149
e-mail: journal.ROS@eurostat.cec.be