UNITED NATIONS INDUSTRIAL DEVELOPMENT ORGANIZATION

Implementing matching procedure and a call centre for
Annual industrial registry updating to improve the production
of industrial statistics

PROJECT: TF/SRL/04/001

Final report to Donor

Project funded by
Norwegian Agency for Development Cooperation (Norad)

Final report

Project number and title:	**TF/SRL/04/001:** Implementing matching procedure and a call centre for annual industrial registry updating to improve the production of industrial statistics
Date:	31 May 2007
Last report (date):	30 May 2006
Total UNIDO Budget:	US$ 140 685
Expected completion date:	November 2006
Originally expected completion date:	April 2006
	14 months after the project start. The first tranche from the Norad was received at the end of Feb 2005.
Objective(s) of the project:	To build the capacity within the Department of Census and Statistics (DCS) of Sri Lanka to build reliable and timely industrial statistics through ongoing annual industrial surveys.

1. The activity report

This report covers the entire project period starting from March 2005 to April 2007. The project was expected to end in 14 months i.e. April 2006, but due to the delay in registry updating subsequently in launching the annual survey of industry (ASI), the project was completed on 30 April 2007. Activities report and major achievements are described below in the sequence of expected output envisaged in the project document.

Output 1: An updated registry of establishments for the entire island

This output is related to the main objective of the project - to establish a computerized registry system that could be regularly updated. The well-updated and completed registry significantly improves coverage and reliability of industrial statistics in general and annual survey results in particular. The Department of Census and Statistics of Sri Lanka (DCS) used to maintain its industrial registry is an informal way. It kept a registry of establishments from the previous Industrial Census (1983), with occasional non-systematic additions from other sources particularly, Board of Investment (BOI).

In 2002, UNIDO sponsored a pilot project in the Western Province, which provided DCS an opportunity to develop a more systematic approach to registry maintenance. Activities carried out under the pilot project resulted in a set of techniques and procedures for directory updating, and a comprehensive manual for use by enumerators and headquarters staff. Four agencies, the Ministry of Enterprise Development, Industrial policy and Investment Promotion (MOID), the Board of Investment, the Employees Provident Fund, and the Ceylon Electricity Board, provided their own lists of Industrial establishments on diskette to DCS. DCS Staff then edited these lists to put them into a standard format. DCS staff used computer assisted matching procedures to reduce all of the lists to a single list of establishment that appeared just once in the combined list and did not appear at all in the current registry. This resulted in a list of candidates for addition to the existing registry.

The current project implemented by UNIDO under NORAD funding is the Phase II of the registry development. Its coverage has been extended from the Western Province to the entire country. The old registry was updated by using form ASI-2 (short form) for establishments that did not respond to the ASI-1 (the Questionnaire for the Annual Survey of Industry). Many establishment in the old DCS registry were found to be no longer active, then were noted as inactive in the DCS 02 file, which was parsed and expanded to include about 60 fields. While the system was effective, it involved too many ad-hoc procedures and software solutions to be implemented on yearly basis.

The goal of Phase II for DCS was to develop a smooth and an efficient procedure for annually updating the directory of industry, building on experience gained in the pilot project on Western Province. From the beginning of the project, DCS prepared an integrated directory based on results of the Census of Industry of 2004 and the Western Province pilot for 2002 including 5000 establishments with 20 or more workers. During the months of May-Sep, 2005 DCS staff entered a number of key identification variables into a new registry (called as Census 04) based on the census list and thus produced an enhanced Census List as Census 04. For the Western Province, DCS had a core registry that was valid for late 2002, when we went to the field and checked the status of establishments. In order to merge two data sets of DCS 02 and Census 04, field sizes of all variables in the both data sets have been set equal and add certain information to various records. Matching process of census 04 list against DCS 02 were completed in the computer with using the customized software during the 1st quarter 2006. The purpose of matching was to identify the duplicates in both files and minimize duplication in field checking of registry candidates. The DCS 02 and Census 04 were combined into a single data set as CORE registry. For the duplicates record in both files, one record was selected with the complete data using the cross editing procedure in software system. This process resulted in establishments of a new registry of almost 5000 establishments with 20 or more persons engaged.

Achievement

The main achievement of the project was a comprehensive register of establishments as of May 2006 that combined records from the different data sources. The project experts also delivered the technical reports presenting updating methods and procedures of the register. The reports were submitted to the Government and attached to the progress report to donor.

Output 2: An improved response rate for Annual industrial survey

For many years, the response rate in the annual surveys of industry (ASI) has been extremely low. The highest response rate of 70% was achieved only when 100% field visits were made in the survey of Large and medium industry. This project was aimed to raise the response rate not through the costly and time-consuming field visits but through setting up of a call centre. A pilot study was carried out to raise response rates by means of phone calls. The study showed a call center could help, by clarifying whether the establishment acknowledged receipt of the questionnaires, resending questionnaires as needed, and eliciting respondent commitments to return the questionnaire by a data certain.

Around one-third of respondents did not sound positive in response to follow-up calls made by the call centre. The percentage distribution of response types can be generalized as follows:

Refusal to any dialogue	7%
Emphatic refusal to cooperate	6%
Response with combative question (e.g. why do you need such information?)	20%
Agreed to response with hesitation	30%
Agreed to return the filled forms	37%

The call center placed about 3000 calls to a total of 1500 establishments at an average of 2 calls per establishment. However, it was concluded that an average of 5 calls per establishments would be required to get a good response rate. (See: Appendix - Final report by Mr. Gunawardena, DCS Director). Experience shows a reliable way to increase response rate is to increase the call rates with other supporting activities.

Achievement

Although, overall response rate could not be improved immediately through new approach, the first stage response in this survey was higher than in previous operations (25% in compare to 20% in earlier surveys). The response rate of 41% was achieved merely through telephone calls. Hence, it was not a full success but partial success that promises better response rate if the call centre is operated more efficiently. The operators have calculated that there was higher response rate in the area where more follow up calls were made.

Output 3: Establishment of a call centre

The purpose of establishing call centre was to improve data collection from establishments. DCS allocated premise for the call centre where additional telephone lines were installed. DCS assigned call centre activists, while UNIDO hired a national expert Mr. Palitha Senanayaka to assist data collection staffs in operation of the call centre.

The call centre organized a number of meetings with the DCS staff members on basics of effective communication. A list of almost 5000 establishments from the register was made available to the call centre in soft and hard copies. A workshop was conducted covering different training modules such as introduction and orientation, reasons of non-response, basics of task oriented effective communication and conversational protocol.

Achievement

A call centre is in operation in DCS premise with the working facilities for 14 call centre staffs. The call centre gets information on status of the completions of ASI questionnaire thanks to the ASI management software. The call centre activists inquire about the delivery of the questionnaire to respondent, response status and returns.

Output 4: Manual and forms prepared for use in annual updating

A number of technical reports and manuals were prepared during the project implementation by UNIDO experts. UNIDO expert on Registry system design Mr. William Weeks undertook one mission of two weeks and prepared a technical paper on List processing and Register updating and a mission report which were submitted to DCS.

UNIDO expert on Industrial Statistics Mr. Alex Korns completed four missions with the total length of more than two working months and prepared specifications for software, manual and reports related to updating procedure of registry. Mr. Korns provided advisory as well as supervisory services to software developers as well as DCS staffs through the project implementation. Away from the mission he frequently provided online guidance and closely monitored the project progress.

Achievement

Manuals that describe concepts, methods and procedures of maintaining and updating business register have been prepared and delivered to DCS. These manuals have also been used in software development, in staff training as well as in registry updating during the project implementation (See Appendix- Paper of Mr. Alex Korns).

Output 5: A streamlined, user-friendly software application in support of registry updating, installed on a DCS server, with a separate version for use at DCS field offices in district with substantial industries

A local software (Genesis) firm was sub-contracted to develop the appropriate and user-friendly programmes for registry updating, while another firm (CNCI) was sub-contracted to prepare the data sets for matching from the different sources. The software firm was regularly reporting to UNIDO expert and DCS on progress made.

The sub-contract comprised of: (a) development of a computer system for the creation of a core registry of industrial establishments (b) through a supplemental contract, the development of enhancements to support the Annual Survey of Industry (ASI). The core registry software was completed and deployed at DCS in November 2005. The system was reviewed during the mission of the UNIDO project manager Mr. Shyam Upadhyaya and Industrial Statistician Mr. Alex Korns. Several corrections were made after this review and DCS started using the software for importing, parsing and editing from December 2005. Currently DCS is using the system also to perform the matching of establishments, while GENESIIS is providing support to DCS in relation to data management activities and supplementary training.

As per specifications given by Mr. Korns ASI software were developed over the period from December 2005 to April 2006. The software was demonstrated to DCS and local UNIDO representatives in Mar 2006. Mr. Alex Korns revised some specifications and necessary changes are being made. The project also provided the computer and related equipment for the registry system. The delivery included a server plus a desktop PC with network hardware system, six additional desktop computers, one notebook, one scanner, two printers, five UPS Power tree as well as network hardware accessories and furniture required for the equipment.

Additional training programme was conducted in software in ASI management of software require to update the system using Microsoft Net programming.

Achievements

A fully computerized registry system has been established in DCS that consists of a user-friendly and fully operational software for registry maintaining and updating, networks and trained national staffs capable in operating and updating the system. DCS has got its data processing facilities significantly enhanced. These achievements are sustainable and instrumental in improving the coverage and reliability of industrial statistics.

Output 6: A continued working partnership between DCS and MED in the collection and use of data for the Registry and the Annual Industrial Survey

The project was implemented in close cooperation of the Ministry of Enterprise development, Industrial policy and Investment promotion. The registry updating as well as other statistical operations were discussed with the Ministry and its data needs were addressed. The registry data from the system established in DCS as well as the results of the industrial survey will be shared with the Ministry for its regular use of statistics. Ministry staffs are also being directly involved in collecting data from the field.

Achievements

There is increased awareness and knowledge of industrial statistics to be available from ongoing statistical operation, which the Ministry can use in formulation industrial development programmes and in monitoring of its implementation. The Ministry staffs were involved in survey and other programmes conducted during the project implementation.

2. The financial report

Thanks to rise of Norwegian Krone against US$ there was a gain of around US$ 6000 in the project budget, which allowed to increase the time of expert services. Most of the procurement and expert services have been completed as envisaged in the project document. The final allocation and expenditure are depicted below.

Financial status[1] of project TF/SRL/04/001 as of 1June2007 in US$

Budget Line		Earlier allocation	Revised allocation	Total expenditure	Balance
1150	International consultant	54811	61174.00	61204.17	-30.17
1600	Other personnel	5000	3646.00	3645.90	0.10
1750	National consultant	6000	3151.00	3159.53	-8.53
2100	Sub-contracts	47687	45190.00	45190.00	0.00
4500	Equipment	14000	16937.00	16945.31	-8.31
5100	Sundries	2978	378.00	377.94	0.06
	Total	130476	130476.00	130522.85	-46.85

Note: The above allocation was made after deducting the UNIDO support cost from the total project budget.

Initial budget including support cost	US$ 140 685
Initial budget excluding support cost	US$ 124 500
Revised total with the exchange gains	US$ 130 476

[1] A certified financial statement will be submitted to donor separately. Figures in the table are presented to illustrate the overall delivery.

Appendices

Guidelines for Software for Updating of DCS Industrial Registry

By Alex Korns

I. Introduction

This note provides guidelines for the software developer, Genesis, in preparing a system for use by the Department of Census and Statistics (DCS) annually to update its registry of industrial establishments with 20 or more workers.

During 2002 and early 2003, with UNIDO assistance, a prototype system was developed at DCS for registry updating, and was applied for the Western Province. That project is broadly referred to herein as phase I, to distinguish it from the current project, phase II. The new system will involve the following main tasks, on an annual basis except where one-off is mentioned:

A. Migrating data from the old system to the new system on a one-off basis;
B. Importing data from three external sources;
C. Parsing and editing the imported data;
D. Matching the external sources against the DCS registry;
E. On a one-off basis, create a core registry by combining data from Census04 and DCS02
F. Identifying the set of establishments from external sources that are not duplicated in the registry, known as the Results of Matching with External Sources (RMES);
G. Prioritizing RMES establishments by the likelihood that they will qualify for the registry, in this way identifying a subset of candidate establishments for field checks;
H. Entering data from the field check questionnaire (DI-2) into the system, and copying data for qualified candidates to the registry;
I. Managing the registry, including tasks such as editing, sorting, printing and routine tabulations.
J. Regenerating the registry at the end of each survey year.

The remainder of this note will discuss the 10 tasks.

II. History and overview

Before 2002, DCS maintained its industrial registry in an informal way. It kept a registry of establishments from the previous industrial census (1983), with occasional nonsystematic additions from other sources, particularly BoI. The registry included a 100% sample of establishments above a certain size threshold and a less-than-100% sample of smaller establishments. The registry contained about 14 fields, covering name and address, employment class and ISIC code.

UNIDO organized a registry-updating project for the Western Province starting in March of 2002 and lasting a little over a year. The purpose was to develop a methodical approach to registry updating that could be applied annually by DCS. CEMIS wrote software for phase 1. Matching in phase 1 involved the following sources, all for the Western Province:

- Old DCS, which was based on the registry that DCS had kept up through 2001. A subset of this, consisting of establishments that had responded at least once in several years prior to 2002, was called the DCS core.
- 'Old' MoID data, which were imported into the system in mid-2002, parsed and edited.
- 'Old' BoI data, which were imported into the system in mid-2002, parsed and edited.
- 'Old' CEB data, which were imported into the system in mid-2002, parsed and edited.
- 'Old' EPF data, which were imported into the system in mid-2002, parsed and edited.

The old DCS registry was updated during phase 1 by checking on whether establishments in it were still active or not, using form ASI-2 (the null return) for establishments that did not respond to the ASI-1 (the questionnaire for the Annual Survey of Industry). Many establishments in the old DCS registry were found to be no longer active; these were marked closed in the DCS02 file, which was parsed and expanded to include about 60 fields.

The external sources (MoID, BoI, CEB and EPF) were matched against the old DCS registry. A record of matches is preserved in the old system, either in the RMES table or in a separate match table (I am not clear on this). Establishments that were found in an external source but not in the DCS core were used to form the RMES list. High and medium priority segments of the RMES list were field checked using the DI-2 form. Results were entered into the system and stored in the RMES file, I believe.

Among the 2000 or so establishments field checked with the DI-2, about 680 qualified for the registry. The system was supposed to transfer the qualifying establishments to the registry, but this was never done due to bugs.

The old DCS registry is another matter. This time, it will be necessary to migrate 2 versions of the DCS registry into the new system, on a once-off basis.

- Because the CEMIS system was unable to merge the DI-2 results for 2002 with the registry, DCS will do this manually. Furthermore, DCS will add several hundred additional records from the old (pre-census) DCS registry outside the Western Province, and will add some new fields to the file, so that the total will exceed 60. CEMIS will provide the resulting file of over 2000 establishments for migration into the new system as DCS02.

- In 2003, DCS conducted a Census of Industry. The census registry was based mainly on notations made during door-to-door canvassing in mid-2001 for the Census of Population and Housing. Lists of establishments by GN with 10 or more workers (culled from the CPH lists) were sent in mid-2003 to each GN with a request that they be updated for industrial establishments (mostly newer ones) that were lacking in the 2001 list, and for closures of older ones. The updated list became the basis for the census. During 2004, extensive data collection took place for reference period 2003 for establishments in the census list, including some 4900 establishments with 20 or more workers. About 70 percent of the establishments responded on a long form similar to the ASI-1. The form included extensive questions about phone numbers and an exact address for the establishment. During 2005, DCS staff entered this contact data into a new registry based on the census list and thus produced an enhanced census list. Unfortunately, however, the census registry was never linked to the older DCS registry for the Western Province, so we don't know yet whether the census registry includes every active establishment in the older registry. This new registry, with about 60 fields, will be provided to Genesis for migration into the new system as census04.

Once the two versions have been imported, the system can match them against each other; the two sources must then be combined inside or outside the system. In this way, a new DCS registry, to be called DCS05, will come to exist, as will be explained below.

III. Migration will involve the following files from the old system (and perhaps other files in the old system of which I am not yet aware)

- Old MoID data, which will be migrated to an MIIP file, for which original data specs will be exactly the same.
- Old CEB data, which will be migrated to a CEB file, for which original data specs will involve the addition of 5 variables, 3 of which will be empty. DCS will specify.
- Old BoI data, which will be migrated to a BoI file, for which original data specs will involve the addition of many variables. DCS will specify.
- The old RMES data set, which will be migrated to a new RMES data structure that will include a few new variables for the DI-2 specified in the file coremapping3.xls.
- The old match table, which specifies all the matches among records in each of the above data files and in the 02 DCS core registry. This may be included in the RMES file. Note that the old match table matches to EPF, which will not be included in the system.

The files are in Access. Mr. Warnapuspha and Mr. Fernando can identify the file names. The 02 core registry constitutes an exception to the above story, as was explained above. Data structures for the enhanced 02 core, Census04, and the new core registry have been provided shortly to Genesis.

IV. Geo-codes and ISIC codes. Unfortunately, special problems will arise during migration for geo-codes and ISIC codes. The problem arises for all the files in the old system -- the older BoI, MIIP, CEB, RMES and DCS core records -- for which DCS staffs have already assigned geo-codes and/or ISIC codes.

Since phase 1, DCS has totally revised its system of geo-codes, with the result that geo-codes in Census04 are inconsistent with geo-codes for the old core registry or the phase 1 parsed data. Fortunately, the conversion is completely clear and unambiguous – each old code converts to one and only one new code. Conversion upon migration is essential for geo-codes, which provide a basis for matching. Therefore, DCS must prepare a conversion table, which Genesis would then use to program a mass conversion of the codes upon migration. To minimize the risk of mistakes, it is also advisable to preserve the older codes, while writing the newer codes to a newer set of cells. Obviously, this will require that the record structure for each of the data files from phase I be expanded to include new geo-codes, in addition to (but not replacing) the older codes. Furthermore, the system will need to load the full hierarchy of names for both new and old systems.

Conversion for ISIC codes presents a different challenge. Computerized conversion is not feasible, as many codes do not match in a deterministic way, but instead in a branching way. Nor is conversion even required before matching, as ISIC codes are not involved in the matching algorithms. For ISIC codes, therefore, it should be sufficient to rename the old ISIC codes 'ISICv2,' while introducing new ISIC codes under 'ISICv3.' All census data is already coded to ISICv3. After records from the old DCS core are copied to the new registry, coding clerks could assign ISICv3 codes to the records that are not in the census data. Again, the system will need to load lookup tables for both ISICv2 and ISICv3.

V. Importing

Each year the system needs to import data from the 3 external sources. The system will compare record numbers for new and old records. New ones will be imported outright and given a new YRGET. Records for older establishments will be compared with the previous records, and changes will be highlighted for certain fields to be enumerated. Other fields will be simply overwritten.

During importation, it needs to be examined whether a more efficient system can be developed for automatically parsing phone numbers, as it will be tedious to have to parse each of the many numbers being provided by BoI. The most difficult issue may involve area codes. If nothing better is feasible, the following can be done at the very least for new records (i.e., ones not parsed in 2002): Copy the entire number to be parsed to the Tel number in a parsing field, and let the editor decide whether to move part of that number to the preface field.

The field YRGET is filled by the system at the time of importation. Mr. Weeks has suggested that this variable (perhaps under the same name) be converted from a 2 digit year to an 8-digit date, to allow for the more general possibility of importation more than once in a year. A related issue involves how to indicate the year of most recent update. I would prefer that the system not overwrite 'YRGET,' as that information would remain useful. Instead, a second variable needs to be computed at conversion for the date of most recent update whenever an older record is updated. This would be called YRUPDATE

Pursuant to our discussion on the need for matching new and old variables for the same source and identifying where changes have taken place, it seems the problem will be much less onerous that may have first appeared.

- For MIIP, this facility may not even be necessary, as DCS seems inclined to restrict imports to records for newly registered establishments that were not in the 2002 data.
- For CEB, such coding would be very limited, as the number of previous fields was so small – only eight, three of which required parsing – name, address1, address2, and city. It seems unlikely that city could change for a given meter, but perhaps we must allow for the unlikely case that the

city name changes. The 2 new fields with content (kwh and yr of registration) will not require parsing. There are another 3 new fields with no content (tel, fax, and email address). Tel and fax would require faxing when the data becomes available, perhaps next year, but would involve no change from previous data until 2007 at the earliest. So maybe we could forget about an update check for those fields for now.

- For BoI there will be extensive changes in the file format this year, which I need to review. In practice, therefore, there would be need for extensive code to check whether data has changed or not only for BoI and not for the other sources.

The updating of older records has certain implications that need to be addressed systematically in the new system.

- Records that have been updated would in principle need to be re-parsed. However, the older parsed data should not be erased by the system, as it may in most cases remain largely valid. Furthermore, the supervisor may decide that this task is not as urgent as the task of parsing and matching new records. Therefore, the task of re-parsing older records should be separated from that of parsing newer records, with the supervisor having the option to schedule that task, by district, for execution immediately or at a later date.
- For re-parsed records, past matches (and non-matches) should in principle be reviewed, as they may no longer be valid. However, they should not be simply cancelled, as in practice they may rarely need to be reversed. The supervisor needs a choice whether to review the older matches right away or another time. Another useful option would be to allow the operator, during parsing, to check a box indicating no need for re-matching for cases where the changes appear minor.
- In the course of time, more and more registry records will be linked to records from the various external sources, either because they were matched onscreen, or because they became the basis for new establishments that, once field checked, have entered the registry. The match table should of course keep track of all these linkages.
- The only updated data from external sources that can be automatically written to the registry would be the 4 CEB data items for electricity usage, plus the web page address from BOI. In fact, the data item for monthly kWh usage (one month per year) should be tracked by the system from year to year, and for that purpose a series of (now empty) fields has been made available in the core registry. It is believed that this variable may eventually prove very useful for detecting closures.
- Other updated variables – for example, name, address, telephone and fax numbers -- will not be automatically written to the registry, as there is no guarantee that the data from the source meets the needs of DCS and is superior to field data collected by DCS. Instead, the updating should trigger a cross-editing session during which the operator would have the option of copying some of the updated data to the registry. While cut and paste remains a useful editing technique, Genesis may also be able to provide an optional short cut for fields with the same name, such as a button that would copy the contents of a certain cell from the source record into the corresponding registry field with a click. Again the supervisor needs the option to schedule this task, by district, at his

convenience.

A special issue involves what to do about geo-codes when new data is imported for older records. In general, I suspect that the true location of a factory will change rarely, so that older geo-codes would rarely become obsolete. This is especially unlikely for CEB meters, but is also not likely to be very common for factory locations in the BoI data. Therefore, I would not agree with the suggestion that the codes be discarded whenever there is a change in the address. At most, if feasible, I would recommend that a flag be inserted to show that the operator should review the old geo-codes for that establishment. Similarly, the ISIC code should be highlighted whenever there is a change in the product description. The operator could then either update the code or simply erase the flag if convinced there was no change. The supervisor needs the option to schedule the task of reviewing such codes, by district, at his convenience.

VI. Parsing

Parsing and editing will be combined in a single module.

An MIS should still keep track of how many records in each imported data set have been imported and what percent have been parsed, for monitoring by supervisors. This could be calculated, for example, simply on the basis of whether the data field for parsed establishment name has or has not been filled in for each record.

A special issue arises for re-parsing of data from a previous year for which there have been changes in data fields. Such records should be shown on separate lines in the MIS table. For CEB, there will be at least 2 new data fields – electricity usage, and year of registration. No need to parse those items, just use them as is. There are also 3 fields that CEB intends to fill out in future – telephone, fax and email. Telephone and fax would need to be parsed in future, but not this year.

Parsing for MIIP and CEB will be the same as in phase 1. Parsing for BoI will involve much more data, and DCS will provide guidelines for the required parsing.

VII. Editing.

All parsed fields can be edited. Most non-parsed fields cannot be edited – for example YRGET, YRUPDATE, and ID numbers. Other non-editable fields include EMP_CAP, DES_1(description of product), DISTRICT_NAME, GROUPCODE and STATUSCODE for BoI, TARIFF and PEAK_LOAD for CEB, PMNUM (or whatever is the name of the employment bracket field), PMNAME and PMSERIAL from MIIP. For each external source, Genesis needs to add three new fields for editing:

- A remarks field.
- EMPLOYEES (4 digit numeric, for use in case this info is obtained over the phone),
- PHONE_CHECK (a 2-digit code for recording findings over the phone, with lookup table to be prepared by DCS).

No parsing is needed for Census04 or old DCS. Editing will however be needed for both files to record findings of matching between the two during matching with the system in the field for PHONE_CHECK05 and to edit other fields that may need to be updated on the basis of phone checks. Editing will be largely the same for BoI, MIIP and CEB except for the addition of remarks field.

The phase 1 system included an 'error code' as a free-text notation to be used by editors while parsing when they noticed problems that required editing. This feature would still be useful, as an operator may wish to finish parsing before editing, even though editing and parsing will take place on the same screen. Such notes could be cleared by operators once the required edit has been carried out. Therefore, we would prefer that the MIS keep track of how many records had text in the error code field. It would also be useful if a browsing screen would offer the option of a filter that would only look at records with a note in the error code cell.

A possible issue for CEB involves multiple meters for one and the same establishment. This could in principle be handled within the system because the system is well designed for detecting doubles. However, as the number of such establishments is small (only about 60 in phase 1), it can also easily be handled outside the system, prior to importing the data. For certain practical reasons, especially the need to be consistent with how this was done in 2002, I recommend that this be taken care of outside the system, before importation.

Another issue for CEB involves the year of registration, which can be gleaned from the account number for 2000 forward. It remains to be seen how the year can be made to appear as a separate variable – whether this can be programmed upon import or must be done manually before importation.

Editing mainly involves correcting parsed data – for example geo-codes that were assigned based on a head office and now need to be reassigned based on factory location. Two exceptions to this rule involve an activity code and a remarks field. During the matching process, staff may learn that the establishment is closed or duplicate. As was done in phase 1, this information will be entered into a field for activity status – a field that will in most cases be left blank, but will be filled with codes for closed and duplicate establishments if need be.

VIII. Matchable files. Here is a summary of what files the system needs to match:

A. Matchable files include:
1. Old DCS
2. Old DI-2 that have any 'situation code' except 1. Keep in mind that these are a subset of the RMES file.
3. Parsed BoI (including migrated older records)
4. Parsed CEB (including migrated older records)
5. Parsed MIID (including migrated older records)
6. Census04 (before the new core is formed but not afterwards)
7. New DCS 05 core (after it is formed but not before)

B. A file will be matched against itself at all times, to identify doubles.
C. During matching for 2005, the system needs to match each of the first 6

files against the others.

D. During matching for 2006 and subsequent years, the system would match each of 5 files against the others (2, 3, 4, 5 and 7).

E. At the end of RMES selection, the selected RMES records need to be matched against sources 7 and 2 to check for missed matches that would indicate no need for a field check.

F. After DI-2 data has been collected, the DI-2 candidates that qualify for the registry need to be matched against each other and against the core registry, to avoid creating duplicates.

IX. Matching: The sequence of review by operators

In both the prototype version and the new version to be built, the structure of matching involves a source list and several target lists.

A. In the prototype version, the sequence of the matching work was as follows:

1. Matching was done on the fly, matching a single source record against all possible target records, once source record at a time.
2. For a single source, match for DSD 1, then DSD 2, then DSD 3, etc. In practice, staff made a single pass through each DSD. If they wanted to take a second look at a specific case, they needed to start over at the beginning of the alphabetical list for the DSD. Once an item in the source list had been matched it no longer appeared in the list of items to be matched.
3. For pending cases, operators could take a closer look at a potential match using the zoom facility. If they wanted to investigate more carefully, for example with a phone call, they could use the 'print screen' to print out the zoom comparison for subsequent investigation. However, there was no way to identify the pending cases within the system or to systematically review the pending cases.
4. After finishing the first source, proceed to the next source, and repeat until all sources had been exhausted. Except that, at each stage, 'reverse matches' were not examined (that is, if source N had already been matched to target P, then source P would not be matched to target N).
5. Results of matching were summarized in a browsing screen that was intended to give a supervisor an opportunity to review and accept or reject. However, the browsing screen in practice failed to give supervisors an opportunity to change anything.
6. There was no opportunity to edit any of the data during the match process. This was a disappointment for DCS staff, who often saw the need to edit the data during the matching process.
7. No procedure was implemented for identifying 'non-matches' or reviewing these.
8. Some tables were provided to monitor progress of the work, but these were not very effective as management tools.

B. The proposed main mode for reviewing the blocked records:

1. Use several algorithms to calculate potential matches en masse, for all sources and all districts. For each pair-wise comparison (e.g., P to Q), make the reverse match as well (Q to P), take the average of the two

matches under the same algorithm, then select the highest score among the algorithms. Store the high score for each pair above a certain threshold to be set by the supervisor (default value 25%), and when deleting data below the threshold take care to make the tabulations required for the matching reports. (Alternatively, save all the high scores, even down to zero – whatever is easier for the system and the programmers).

2. All potential matches will be classified into four groups: Best likelihood (A), good likelihood (B), fair likelihood (C), and lowest likelihood (D). The supervisor would be allowed to reset the cut-offs for A, B, C and D but the default cut-offs would be 85%, 70%, 60% and 40%. During examination of the B group, the screen would also display any remaining unclassified cases from the A group.
3. Operators will begin by examining all of A group, review these in sequence by district and within district by source. This will allow separate operators to work on separate groups of districts. Within district and source, operators will examine cases in order of highest match score. Those with 100% scores would be examined first, then 99% etc. The screen will show the highest score on top but will show all other scores above the threshold as well.
4. For a specific source record, the screen will show the highest scoring match, and all lower-scoring ones for the same target that have a match likelihood index above the cut-off for the likelihood group being processed. If the number of candidates from the same target with MLI's above the cut-off is fewer than 3, the screen will show candidate matches with scores below the cut-off, down to a minimum cut-off level (such as 50%) that can be set in a master table.
5. The operator will identify confirmed matches, confirmed non-matches, and doubtful cases. Doubtful cases would be printed out for further investigation (using a form under preparation by DCS) and results would be input to the system.
6. A review process would allow an operator to re-examine all pending cases in order to input findings. It would also allow a supervisor to browse all the findings – confirmed matches, confirmed non-matches, pending cases and unclassified, by classification.
7. An MIS function will keep track of how many in the A group, by district and source, have been matched (by target), declared non-matches, declared doubtful, or remain unclassified. Jack will describe this system in more detail.
8. Once the supervisor has approved the results for the A group, these results would be written to the database. Matching could now begin on the B group. Steps C – E (above) will now be repeated. Cases that were already declared as matches or non-matches will not be shown, nor will candidates for dual matching be shown. However, unclassified cases from higher probability levels (in this case, from the A group) will also be shown.
9. Similarly for the C group.
10. The D group is a special case, as the probability of matching will be relatively low. Here are some guidelines for this residual group:
    - The supervisor (or operator) needs to set a lower threshold on the match score, below which it is not worth examining the possibility of a match.

- o If the operator so wishes, the D group should only be examined in respect of source data that was registered with the external source before a certain cut-off date, to be defined by the supervisor. This will focus inquiries on cases that overlap with the core registry, and minimize inquiries for newer registrants that are less likely to appear in the core registry.
- o It is unclear whether marking non-matches for this group will be efficient, as the number may be large. The operator should be given the option to mark them or not.

11. There is a need for a cleanup phase as well, in which someone (perhaps the supervisor) examines all the remaining candidate matches starting with the highest MLI score, regardless of district or source.
12. This sequence will enable operators to identify the largest number of matches early in the process, and thereby simplify the remaining task.

C. Some notes on the user interface for the primary mode

1. No need for several options on the L side of the old screen – matching on which variables, weights, or word/character choice.
2. The top window should show establishment name, source (abbreviated), Prov-dist and DSD codes. It doesn't need to show a lot of other data, as it would be tedious to scroll to the other data.
3. The system needs to recognize when a source record has already been matched to a target record, perhaps to display that with special shading. It also needs to recognize when a target record is already matched to a record from another target file and to display that info.
4. Zooming screen would show all of the matching variables on top. Any variables where there was full or partial agreement would be given special shading. If the operator wishes to view a comparison of all info for a pair of records, there should be a convenient option to toggle to the full screen and then back to the normal screen.
5. A comment box will appear whenever a non-match is declared. Furthermore, whenever a supervisor undoes a match that was previously declared by an operator, this will become a non-match with a mandatory comment.

D. A special mode is also needed for reviewing selected records. This would involve a situation in which the operator was trying to match specific source establishments of priority interest, regardless of whether a high scoring block existed or not. The difference would be like the difference, on a camera, between automatic operation and aperture priority operation. Normally, an operator would use this mode to explore certain high-priority cases only, not to examine all cases.

1. In this mode, the operator will start with a browsing list of records in a chosen source list. The records may be sequenced by ID number, by name alphabetically, by size, or by district and/or DS. The operator would select a specific record for examination. Beside the name, the browsing screen should show the district code, indicator of size (employment or electricity usage), and an indicator of vintage (year of registration or some such).

2. Once the operator has selected the specific record for examination, the system would show the top 3 available matches from each source, taking algorithms 1 through 5 into consideration and showing the highest of the 5.
3. If the operator so wishes, he can ask to see all blocked records including ones that would ordinarily be suppressed. For example – target records previously declared non-matched and target records matched to another record in the source. These 'forbidden' choices would be given special shading.

D. **Editing**. DCS has asked for a convenient way to edit records during the matching process. This basically means that the operator would take a break from matching to carry out an edit, then return to matching where he left off.

F. Reports. Mr. Weeks has provided a note on matching reports.

X. A new core registry (DCS05) needs to be formed on a one-off basis within the new system, after matching has been completed and before additions to the RMES file can be defined. It is especially important that the matching of the DCS03 data to Census04 be finished before this is done. The core will comprise:

1. All establishments in the census04 list, with entry path recorded as 3 and Activity Status as active.
2. Establishments in the DCS02 core that do not match to census04 will be handled in the following ways:
   - Those confirmed still active (based on PHONE_CHECK_05 code) would be added to the DCS05 registry at this time using data from DCS02, with activity code as active.
   - Those identified as closed in the phase 1 data (DCS02) should also be copied to the new core, with activity status other than active. Similarly for establishments that are confirmed closed by phone (PHONE_CHECK_05).
   - Those that are neither confirmed active nor closed will be given an activity status UNKNOWN and copied to RMES.
3. For active establishments in census 04 that match to DCS02, care must be taken so that DCS can select the best of the available data. Especially for employment, take census data if available. For fields available from only one or the other source, marked 'A' in the core-mapping file, data will be taken from that source. Whereas for fields for which data may be available from either source, marked 'SA' in the core-mapping file, data will be taken from census04; however such data must be highlighted as subject to further edit.
4. Formation of the new core will involve the following additional steps:
   - When copying takes place, the following entry path codes should be assigned: '1' for closed establishments from DCS02 or confirmed active ones not matched to census, '3' for records that have been matched from DCS02 (not from DI-2) to Census04, '4' for records that have been matched from DCS02 (originally from DI-2) to Census04, '5' for records that are copied from unmatched Census04.
   - New EIN. A 6-digit sequence number is sufficient. It is a pure sequence number and can be assigned in order of accession to the

system.

5. After the core registry has been formed, a special editing step will be needed on a one-off basis to review all records with entry codes 3 and 4. This will require a split screen showing the data selected for the new core on the left (with the 'SA' fields highlighted), and the equivalent DCS02 data on the right. Editors will have the option to copy data from DCS02 to the equivalent field in the new core. This kind of editing facility, which I sometimes call 'cross-editing,' is similar to one that will be needed elsewhere in the software. It will be convenient if the editing of those with entry paths 3 and 4 is done separately, as the value of the DCS02 information will be different in the two cases. The editing screen should also show prominently the value for RESPONDED_03, as the value of census04 information depends on that.

There is no problem if this is done outside the system. In subsequent years, the core registry will already exist as a product of updating from the previous year, so there will be no need for the system to create it.

XI. RMES prioritization rules

A. RMES formation and initial reporting

Once the new core (DCS05) has been created, the RMES will be well defined. Establishments qualify for the RMES list if they are found in a source list (including the remnants of the DCS02 old core) and not in either the current core registry or in the old DI-2 list of establishments that were checked and found not to qualify for the registry (situation code other than 1 or 3).

A report should show the number of RMES establishments for each source, including the following 5, crossed in five alternative ways: province (but district within WP), yrget, vintage, size, and results of any phone checks for BoI, MIIP and CEB but not for DI-2 or old core. CEB could not of course be tabulated by employment size, however when we get the data we will be able to calculate a size scale for CEB as well. For size, classify by employment (except for CEB), with employment set as unknown if an employment figure or bracket is not available, as it will not be for the DI-2 cases.

1. First would be DI2 cases that were checked last year and identified for rechecking the following year – provided they do not match to the core registry.
2. Second would be BoI candidates.
3. Third would be MIIP candidates.
4. Fourth would be CEB candidates
5. Fifth would be establishments from the old core (DCS02) in group 1c (but only for 2005).

Establishments that appear in more than one source will be counted for each source. However, column totals will count each establishment only once, so that the totals will as a rule be less than the sum of the numbers for each source.

	A	B	C	Unclass	Total
Leftover DI-2 cases				19	19
BoI	75	75	75		225

MIIP	15	25	25		65
CEB	100	500	1400		2000
DCS02	15	50	75		140
	150	450	1200	19	1819

B. RMES Prioritization

For each source, operators will need to divide RMES establishments into 3 groups: Those (A) to be checked with certainty, those (B) to be partially checked (for example, to be checked by phone but perhaps not necessarily to be checked in a follow up visit) and those (C) not to be checked at all. In practice, it may be easiest for the supervisor to first decide on cutoffs for group C, then for group A. Group B would be a residual. A status report should show, for each source, the number of establishments in the A, B, and C groups once the supervisor has selected the cutoffs, beginning with cutoffs for the top priority source. All these decisions would be revocable and status reports would enable the supervisor to see the numbers involved in the A, B, and C groups before committing himself to a final configuration.

1. The employment variable for BoI is EMP_CAP, for MIIP is PMNUM (unless DCS revises this name), for DCS02 it will be EMPLOYEES_DI2 if available and otherwise EMPLOYEES_CORE02. For the DI-2 cases being considered here, there will not be an employment variable.
2. The vintage variable, a year, for BoI is YEAR_DECL or a new variable for date of registration. For MIIP it is CDATEREG and for CEB it will be a year of registration for which DCS will provide a field name. For DI2 and old core, the vintage year would be 2002. When tabulating vintage, distinguish 05, 04, 03, 02, 01, and everything before 2001.
3. For breaking the establishments from the various source lists into the A, B, and C groups, the operative variables would be as follows:
    i. For BoI establishments: EMP_CAP, YRGET, YEAR_DECL or a new variable for date of registration, and the result of any phone check during matching.
    ii. For MIIP establishments: PMNUM (or the new variable name for employment bracket), CDATEREG and YRGET, and the result of any phone check during matching.
    iii. For CEB establishments: Two size measures (PEAK_LOAD and a measure of kWh usage), YRGET, and the year of registration, which is indicated by the registration number for years beginning with 2000. There may be some result from phone check during matching, although this should be minimal as we gave no phone numbers for CEB establishments.
4. For establishments listed in more than one source, the highest classification will predominate in deciding whether it belongs to the A, B, or C group. Thus, if an establishment is already included in the A group for BoI, it will be counted for the MIIP classification, but if it is in the C group for BoI, it will be counted under MIIP if selected among the A group for MIIP. For establishments at the same classification level, the program will prioritize sources as shown in paragraph 3. Thus, an establishment in BoI, if selected for the C group, would be dropped from MIIP if selected among the C group there as well.

5. Records that are matched to any record which have been showed closed by a phone check will be considered closed and put automatacally in the C group. For example, if a CEB record is matched to a BoI record that has been found closed, the CEB record cannot be selected for the A or B groups.
6. The definition of the A group will as a rule be in more than one piece – for example, establishments with 20 and more workers that registered after December 2003 and establishments with 50 and more workers that registered after June 2001, etc. If the second piece duplicates part of the first piece, only the nonduplicative part of the second piece will be added to the A group. The sub-groups that have been added to the A group, etc., need to be documented on screen and in the report showing the number of unduplicated establishments in the A, B and C groups, so that the supervisor can review the cutoffs and reconsider them, before deciding on a final classification.

C. RMES data structure

The RMES data structure has been expanded from its phase 1 version to include additional source information. This includes old and new fields specific to BoI (EMP_CAP, extra telephones and faxes, but not telex) and to MIIP (employment bracket), as well as four size-related variables from CEB -- TARIFF, PEAK_LOAD, NUM_LIST, and kWh. Furthermore, some variables will be copied to fields with different names, for example, the fields for product description, date of commencement, contact name and contact telephone from BoI or MIIP. The latter fields will of course be overwritten by information from the DI-2 if and when available.

When creating RMES records, data will always be taken from BoI if available, except for the 4 variables from CEB. After BoI, the preferred data sources in order of priority would be (2) DCS02 (one-off for 2005), (3) MIIP, (4) those DI-2 marked to be rechecked, and (5) CEB. This is true irrespective of whether the establishment was identified as part of the A group for BoI or for another source.

Another variable needed in the RMES data structure is a source code showing whether the establishment is comprised in each of the possible sources: BoI, MIIP, CEB, DI-2 (those designated for rechecking next year) and old core (one-off for 2005). DCS can prepare a set of codes for all the various source combinations. The variable should also indicate which source is the main source, defined as the first available source among the 5 prioritized sources mentioned in section D.

D. RMES Seq No.

In the old system, the RMES seq No. was in 7 digits based on 2 for the year, 2 for the district code and 3 for a sequence number. Beginning this year, I would propose a new formula based on 2 digits for the year and 5 for the sequence number. Obviously the older district codes are now obsolete. Nevertheless, I still don't see any need to change or replace the older numbers, unless DCS says otherwise.

E. Browsing

After prioritization, the operator needs to browse in the records for the A, B and C groups, by source. For the A and C groups, he could select individual establishments during the browse, and designate them for transfer to another group (for example, B). For the B group, he could select individual establishments during the browse and designate them either for transfer to another group (that is, A or C), or for designation as a unit within the B group to be field checked or not.

F. Editing

This is a special case of 'Cross-editing.' The operator needs a way to compare the records selected for field checking (i.e., as defined in C above) with the records from the other (non-selected) source or sources that are matched to the selected record. Then the operator needs to designate a specific matched establishment for a more detailed comparison (as on the zoom screen for matching), during which address and telephone information could be copied from the matched record into the designated record as the operator chooses. Near the top of the list of matched variables should appear the code for phone check, so we will know if there was a phone check or not for the record.

G. Reports

The reports need to shows the numbers of establishments in the A, B, C and duplicate groups for each source. The duplicate group is the group of records that is duplicate with a record that has already been designated in another for another source, following the priority rules mentioned earlier. For the B group, the report needs to distinguish those designated for the sample, those designated for exclusion from the sample, and those not yet designated.

Specifically for the A group of each source, it would be useful somehow to show each subgroup that was selected for the A group (for example, establishments with 20 or more workers and vintage after 2003, etc), together with the number of establishments in the subgroup (measured both as total establishments and as total unduplicated with the previously selected subgroups).

H. RMES Labels

The system needs to print information for each establishment selected for the DI-2, in three ways:

1. A 'browsing' list of the establishments, sorted by DSD. This would show name, address, RMES #, main source, and 1-2 telephones, whatever fits on a line.

2. An information sheet that would show complete info for the establishment, as available in the RMES file, to be designed by DCS.

3. A concise label for pasting on the DI-2 form. The label would show the name, address, and phones for the establishment, RMES # and main source. DCS wo;; provide the layout.

XII. DI-2 Module

A. Mr. Gunawardena will provide the new DI-2 questionnaire, which will be slightly expanded from the old one. It will include text for the owner name and may include other items. The data entry screen needs to make room for the new item or items.

B. During data entry, a browsing screen is needed for all DI-2 establishments, as defined in the RMES module (i.e., those records for which labels were printed). This screen should then provide information on how many have already been entered and how many have not yet been entered, in the form of both lists and reports. Both totals should be broken down by district.

C. When the supervisor declares that data entry has ended, the system would match the qualifying DI-2 establishments against the core registry. Alternatively (ask DCS about this), this could be done at the time each qualifying record is entered. Operators would evaluate the blocked records, one by one, and declare them as either matched or not until a clear result was obtained for all blocked records. For any that match, the relevant codes in block III would be changed from qualify or the registry (code 1) to either codes 2 or 3, at the choice of the operator.

D. The Genesis flow chart can just show browsing DI-2 results as a single box, but the user would then be offered the option to browse those that qualify or that don't.

E. At the end of matching (point C above), the supervisor will decide when to add the qualifying candidates to the registry. After that is done, the system should issue a brief report showing the number of active and non-active establishments already in the registry and the number added in this step, broken down by district.

F. Another required facility here is to match the nonqualified cases to the core registry, and then to 'cross-edit' useful info from the nonqualified cases to the core.

G. Tabular reports on the DI-2 survey (basically summarizing the results for qualifying and non-qualifying establishments) would be as in the 2003 version, except that additional columns will be needed for two additional situation/status codes.

XIII. Registry management

A. The system should allow an operator at any time to see all records from other files that match to the registry. This would include the external sources (BoI, MIIP, CEB) and RMES (including the DI-2 info). The editor would then have the option to 'cross-edit' from one of those records to the registry. This facility should also enable the operator to go through all registry establishments that match with newly matched BoI records, for example, or with newly matched and/or updated BoI records, to look systematically for info that needs to be copied to the registry.

B. During the matching phase, especially the investigation of doubtful cases, phone calls will be made to establishments. These calls may result in updated info for a registry establishment. The system needs to enable an operator to enter such info into the registry. Date stamps will be automatically saved for any updating of the activity status or EMPLOYEES.

C. Specifically during the receipt of questionnaires for the ASI (to be discussed in another note), when employment from the questionnaire is entered at the time of receipt, that number will be entered in the field for employment for the relevant year. This number will also be copied to the EMPLOYEES field, as the most recent employment number for the establishment. Subsequently, whenever the operator obtains new employment info, he/she can enter that info under EMPLOYEES.

D. The operator needs to be able to enter a new establishment directly into the registry. Such a case would be given a special ENTRY_PATH code for an ad hoc entry.

E. When viewing establishments, the operator needs to have a choice of filters by district or sub-district and by activity code. The default filter would be for all active establishments, but the operator could also view all establishments including non-active ones, or all closed establishments, etc. Another filter would be for changed status during the survey year. Other filters that would sometimes be needed would be by industrial estate, by employment size and by ISICv3. The latter grouping should be able as well to show only those establishments for which ISICv3 is still empty.

F. A facility is also needed for exporting a given list of records to Excel. The user would merely define the filters to be applied in selecting the records. DCS can say whether they would always need all the variables, or only a subset, or perhaps a menu for deciding which variables would be exported.

G. Printing would usually be for active establishments, sorted down to the lowest geo-code, as this would be most convenient for dividing the work among enumerators. The operator would need a way to decide which variables to print so as to fit the data into the available space.

H. Reports. I would recommend that standard tabular reports on the registry would be as in Indonesia. The row tabs would show the names of districts, while the column tabs would show size classes for employment (20-49, 50-99, 100-199, 200-499, 500+). Another option to save space would be to show provincial data in the rows, except that the three districts of the Western Province would be distinguished. The cells would show either the number of active establishments or the number of employees at those establishments, at the choice of the operator. The user could also filter for a specific ISIC (often at the 2-digit level). Another filter could limit the report to new discoveries (newly active). 0ther reports could summarize the number of nonactive establishments by Activity status.

XIV. Regeneration

A. The main function of the regeneration command is the conversion of status codes, as will be explained. Another function has to do with resetting the ASI year.

B. As is mentioned in Coromapping7, on the lookup tables tab, there are 10 activity status codes. These would convert upon regeneration as follows:

		During first year upon acquiring the status*	Upon regeneration, the previous codes would convert to:
1	Active (A)	Newly active (NA)	Active (A)
2	Temporarily closed (T)	Newly temporarily closed (NT)	Temporarily closed (T)
3	Permanently closed (P)	Newly permanently closed (NP)	Permanently closed (P)
4	Double (D)	Newly double (ND)	Double (D)
5	Merged (MG)	Newly merged (NMG)	Merged (MG)
6	Out of Scope (O)	Newly out of Scope (NO)	Out of Scope (O)
7	Moved (MV)	Newly moved (NMV)	Moved (MV)
8	Small active (SA)	Newly small active (NSA)	Small active (SA)
9	small closed (SC)	Newly small closed (NSC)	small closed (SC)
0	Unknown (U)	Newly unknown (NU)	Unknown (U)

C. The advantage of this system is that it enables managers to keep track of establishments where changes are taking place. For example, when DI-2 records that qualify for the registry are added they will receive a status of 'newly active.'

D. The only exception to the above is for the one-off formation of the core registry during 2005. At that time, no codes will be designated 'newly,' only active and closed will be recognize (please check with Mr. Warnapuspha if the old registry includes other codes).

E. A related issue is whether certain conversions will be forbidden. DCS has agreed that duplicate, merged and permanently closed will not be allowed to reactivate, whereas temporarily closed and small will be allowed.

F. When reports are prepared, the operator should have the option whether to combine newly active with active, etc. or to show them separately.

G. The other change that takes place during regeneration is that a new ASI year begins. That means that receipts can no longer be entered for the previous ASI, and activity for the new ASI can begin.

H. I propose that the current year be designated as 2005, even though the reference year for the ASI in 2005 is 2004. Most likely, DCS would wish to undertake regeneration sometime around April, when it is too late to receive any more questionnaires for the previous year and when it is time to begin updating the registry again.

Annex 1: The matching algorithm

I. Name algorithm

A. Analyze the words in the name of the source establishment. If any of the 130 words in the lookup table are present, separate the words out. Call these generic words; call the remaining words specific text. (For this purpose, spelling variations are treated as a single word – thus apparel, aparel, aperel, etc. are all treated as the same word).

B. Compare all of the specific text with the names of target establishments, using the bigram method. As in the prototype version, prepare for the comparison by converting all letters to capital letters and discarding all spaces and punctuation. Calculate a blocking score.

C. Compare each of the generic words with the names of target establishments, on an all or nothing basis.

D. For generic text, calculate a frequency index, f, by taking the cube root of the number of times the word appears and rounding that off, as shown in sheet 2 of the file 'frequency1.'

E. Combine the results of steps B, C and D as follows. Let the number of bigrams of specific text be s, while the number of bigrams of the first generic word is g_1, the number of letters of the second generic word is g_2, etc. Let the weight for s be one, while the weight for g_1 is

$w(g_1) = 1/f_1$, where f_1 is found in the lookup table.

Then the formula for the combined weights is:

$$W = S + g_1/f_1 + g_2/f_2 \text{ etc.}$$

Where s is the number of identical bigrams of specific text, g_1 is the number of bigrams for the first generic word, etc.

In other words, if the specific text contains 12 letters with 11 bigrams, while the generic text contains one word with 8 letters with 7 bigrams and with a frequency index of 4, the 7 bigrams of generic text would be given an importance in the blocking formula equivalent to only 1.75 bigrams of specific text. W would equal 12.75, calculated by summing the specific identical bigrams and a 1.75 bigram equivalent value for the 7 bigrams of generic text.

F. The matching likelihood index (MLI) M would then be as follows:

$$M = (m_s + m_{g1}/f_{g1})/W$$

$$M = (m_s + m_{g1}/4)/12.75$$

Where m_s is the number of bigrams that agree for specific text, and m_{g1} is either 7 or 0, depending on whether the generic word agrees or not.

The maximum score for perfect agreement between the text in the source and target list would be 100 %. It would be the result if some text was generic, while both specific and generic text agreed perfectly. It would also be the result if all text was specific and agreed perfectly. If specific text agrees less than perfectly, the formula serves to combine the results of comparison for the various components in such a way as to discount the importance of generic text. If no text agreed for either specific or generic text, M would equal 0.

II. Combined algorithm – method 1

A. Assign the name comparison a weight of 80 %

B. In source and target records, compare the fields for

- DSD code (always within a specific Province-district code, observing the hierarchy of geocodes)
- City name
- Telephone number (if available for both source and target)
- Assessment number (if available for both source and target)
- Street name (if available for both source and target)

C. Comparisons for the street name and assessment number, when involving data files that provide both location and head office addresses, should be to either address, whichever scores highest.

D. Comparisons for the DSD code and assessment number should probably be on a simple all or nothing basis. Comparisons for the street name could admit of partial agreement. Comparisons for the city name should be defined in such a way as to distinguish Colombo 9 from Colombo 10, but to accommodate spelling variations (in other words, if the numbers do not agree, the match value for the field would be zero).

E. Comparison for the telephone number should accommodate the recent addition of a digit to many telephone numbers. Also, due the multiplicity of phone numbers in the registry, it will be necessary to compare every available telephone number in the source record with every available one in the target record, including as well numbers for both location and head office if both are available. If a single number agrees, even partially – 6 out of 7 or 5 out of 6 or 4 out of 5 – consider it a match with a high partial score.

F. If one of the five fields in B agrees perfectly, assign it a contribution of 10%, if two or more agree, assign it a contribution of 20% (the maximum for this group), except that if the city and DSD codes agree assign it a lesser contribution. If no blocks are perfect but one is partial, assign it a proportionately lower contribution.

G. Combine the results of A and E. The maximum combined score should be 100%. Calculate each of the above scores going both forward and backwards (source to target and target to source), and take the average of the two as the MLI for the pair.

III. Combined algorithm – method 2

A. Assign the name comparison a weight of 50%

B. In source and target records, compare the fields for

- DSD code
- City name
- Telephone number (if available for both source and target)

- o Assessment number (if available for both source and target)
- o Street name (if available for both source and target)

C. Assign the telephone number a weight of 35%. Partial agreement is allowed for the telephone, as mentioned above.

D. If any of the remaining four fields agrees, assign it a weight of 15%.

E. Combine the results of A, C and D, and take the average of the forward and backward scores. The maximum combined score should be 100%.

III. Combined algorithm – method 3

A. Assign the name comparison a weight of 50%

B. In source and target records, compare the fields for
- o DSD code
- o City name
- o Telephone number (if available for both source and target)
- o Assessment number (if available for both source and target)
- o Street name (if available for both source and target)

C. Assign the combination assessment number & street name a weight of 35%. If both agree perfectly and if they are in the same district, assign 35 points (if the districts differ assign only 25 points). If only the assessment number agrees, and the street name is lacking for one or both fields, assign it a full 35%. If the assessment number matches, there are two possibilities: if the street name does not agree at all, assign the combination a value of 20 points, while if the street name agrees partially, assign the combination a blocking value between 20 and 35 points. For all these combinations, subtract 10 points if the 2 records are in a different district. If any of the remaining three fields agrees, assign it a weight of 15%

D. Combine the results of A, C and D, and take the average of the forward and backward scores. The maximum combined score should be 100%.

IV. Special search

A special search will occasionally be required, as discussed in Matching Note 2. This would require 2 new algorithms, ones that involve a modification of algorithms 2 and 3. *Algorithm 4* would be based on algorithm 2, except that the value of a match for the phone number would be raised to 65 percent, and the value of the name alone would drop to 25 %, with the other matching variables receiving 15%. *Algorithm 5* would be based on algorithm 3, except that the value of a match for the street name alone, in the same district, would be raised to 50 %, while the value of a match for the street name and number would reach 65%. The name would be worth only 20 percent and the remaining variables would be worth 15%. Again, take the average of the forward and backward scores. The maximum combined score should be 100%. The results of these algorithms would be used in a separate mode only, to be described in note2, and would not be used during the normal mode of searching for the most likely matches.

V. Blocking process

The matching process would be carried out at the operator's command. The system would compare entries in each source with all the target entries, using all 3 algorithms. The system would then record the highest score for each pairwise match among the 3 scores thus calculated, together with a code for the algorithm with the highest score to facilitate review of the usefulness of the algorithm.

Blocking results would be calculated in batch and stored in the system. They should be updated while the data is still being parsed and edited, but would not need to be edited if no data changed. It remains to be seen whether it would be more efficient to have an automatic daily update of the blocking scores, or just to update them after some data had changed.

However, there is also a need to update the matching scores partially (say for matches involving the records for which data had changed). The operator would be given the option to request a recalculation for changed records at any time, and would avail him or herself of the option after up-dating data related to the match under consideration.

It will not be possible to fine-tune these algorithms until a full data set has been loaded for both core registry and external sources, including the 6 matching variables. The final judgment of what works best will be an empirical one, based on actual data. Based on previous experience, however, it appears likely that these algorithms will catch nearly all the probable matches, while minimizing clutter from extraneous cases.

Annex II. Summary of required functions for ASI management

ASI Management		
Sample flag by year	Designates which establishments are in the sample for which year.	Stage 2
Receipt of questionnaires	Allows operators to record receipt of questionnaire – whether ASI-1 or ASI-2, as in Indonesia. Functions like in Indonesia system. Simultaneously updates registry data as necessary, especially for employment and activity status (e.g., closures or small or doubles).	Stage 2
Follow-up for nonresponse	Monitors actions by DCS to follow-up on non-response – including phone calls, reminder letters and faxes, and site visits. Enables operators to ensure that all non-respondents have received due attention. With details of phone calls in diary form.	Stage 2
Organize and schedule the work flow	Organize the task of follow-up for non-response by district, to facilitate assignment of tasks among DCS headquarters staff. Schedule review of each case based on the results of a previous conversation or the timing of a previous fax. Organize the daily work for each district by prioritizing the tasks, giving top priority to the largest establishments and to overdue cases. Provide special treatment for large, 'difficult,' establishments.	
Reminder letters and faxes	Prepares reminder letters for establishments that have not yet submitted questionnaires. Outputs the faxes to a program that can send multiple faxes from a batch file. Prepares address labels for letters.	Stage 2
Data entry in ASI-2	For establishments that do not respond to ASI-1, enter all ASI-2 data, especially regarding response status and employment.	Stage 2
Lists of establishments	Prints lists of estab's for which no questionnaire (ASI-1 or ASI-2) yet received or for which follow-up actions are indicated, sorted by DS (or lower level of geo-code) & name.	Stage 2
Reports on ASI completion	Similar to those in Indonesia. One page per district, detail by DS. Shows receipts by size class and 5 response codes, with percents shown for receipt of ASI-1, ASI-2, and no document.	Stage 2

To : Dr. Shyam Upahyaya
Copies : Dr. Alex Korns, CNCI

From : D.C.A. Gunawardena, Director, Dept. of Census and Statistics

Subject : Phase II : Final report on updating the registry
Sri lanka Integrated Programme to Update Industrial Establishment Registry

ASI - 2 - Response rate

As in phase 1, enumerators for the Annual Survey of Industry (ASI) were asked to use short form ASI - 2 to document cases for which respondents did not complete a questionnaire. The ASI - 2 was introduced to clarify the reasons for non response such as merge cannot be located, closed, totally refused, non industry and actual non-response due to various problems. Previously, blank questionnaires (Nil returns) were sent for non-responses and no details were furnished by enumerators. This led to difficulties to blow-up the sample to population using the weights. Also this process (ASI - 2) led in the current ASI to the discovery of about 40 establishments that are no longer active or in scope.

The problem of non-response for ASI-1

For many years, ASI response rates have been extremely low. The response rate (measured as the number of completed questionnaires divided by the presumed number of active establishments) was about 77 percent in the Census of Industry for large and medium establishments. This higher response rate achieved with the 100 percent field visits by the data collectors. However, only 41 percent in the ASI for 2005. A response rate of only 41 percent is insufficient for the preparations of reliable data.

The method of data collection for ASI was adopted earlier by DCS was postal inquires and followed by field visits by DCS data collectors. However, the most of instances the data collectors compel to visit the establishment to get the questionnaire completed. The collection of accurate data from industrial establishments is very tedious and time consuming. In most of the cases, the data collectors had to visit them several times in order to get a good response.

During phase 2 a pilot study was carried out to raise response rates by means of phone calls. The study showed a call center could help, by clarifying whether the establishment acknowledged receipt of the questionnaires, resending questionnaires as needed, and eliciting respondent commitments to return the questionnaire by a data certain.

The call center placed about 3000 calls to a total of 1500 establishments at an average of 2 calls per establishment. However, it was concluded that an average of 5 calls per establishments would be required to get a good response rate. (See the annex : detail report submitted by the call center consultant)

Experience shows a reliable way to increase response rate is to spend here more money on piece rates, transportation, telephone calls and other supporting activities. So this becomes a budgetary issues.

Analyse results of DI-2 Survey

A survey of candidates for addition to the directory to be conducted each year, to assure that directory discovers most new large/medium industry.

This survey was carried out in the field during the February/March, 2007 using Questionnaire DI-2 with 1588 candidates. DI-2 data was entered, prepared the error free data file and obtained the necessary tabulation for analyse results.

DI-2 results. Some overall findings are summarized in Table 1 below which shows the number of candidates, by sources, divided into two main groups: successful and unsuccessful.

Table 1 - DI-2 Summary Results, by Source Group

Source	All Checked		Successful Candidates by employment					Un successful candidates, by reason								
	Total	Percent Success	All	100+	50-99	30-49	20-29	All	Commercial Production		Re check next year	Closed	Moved etc.	Head Office	Non-Industry	Not found
									(>19)	(<20)						
BOI	716	63.5	455	199	121	80	55	261	108	24	9	58	3	.	33	26
CEB	630	30.2	190	74	42	38	36	440	233	52	.	23	14	3	53	62
MIIP	242	22.3	54	9	14	9	22	188	17	20	3	32	7	5	9	95
Total	**1588**	**44.0**	**699**	**282**	**177**	**127**	**113**	**889**	**358**	**96**	**12**	**113**	**24**	**8**	**95**	**183**

Note : Unsuccessful candidates in commercial production with 20 or more workers duplicates with the Core registry

Of the 1588 candidates checked in the field, 699 (44 percent), qualified for addition to the new DCS registry. Among the 889 that did not qualify, 113 were closed, 95 had shifted into non-industry, 96 were become small (fewer than 20 workers), 24 were moved to some other unknown district, 183 could note found and 358 in commercial operation with 20 or more workers were duplicated with the Core registry.

Of the 1588 candidates checked in the field, 716 were from the Board of Investment (BoI), 630 from the Ceylon Electricity Board (CEB), 242 from the Ministry of Industries and Industrial Policy (MIIP).

In terms of success rates (percent qualifying for additions to the registry), BoI was the most productive source, 63.5 percent, followed by CEB and MIIP, with success rates of 30.2 and 22.3 percent respectively. For MIIP among 188 that did not qualify for the registry, 09 (5 percent) had shifted to non industry or out of scope, 32 (17 percent) were closed, 20 (11 percent) were small (fewer than 20 workers), 95 (51 percent) could not be found. The high value for could not be found the establishment in the field was due to the that industrialist had given address which was not the location address of the establishment.

Most of the 440 unsuccessful CEB candidates were either duplicates (53 percent), or not found (14 percent), or non-industry (12 percent) or small (12 percent) for various reasons. The fact that so many duplicates cases were found in the CEB data was due to either different name of different address of the Industry establishments provided during the registration of the meter connection.

Among the 699 successful candidates, that qualify for the new registry, nearly 40 percent had 100 or more workers as shown in the Table 2 (See Annex 2). Other tabulation showed that the 699 together had about approximately 134,000 employees. When the discoveries are sorted by year of commercial production, it appears that an average of 57 establishments (20 and more workers) started production in each of the year 2000-2005. This finding indicates that annual updating capture 50-60 new establishments per year.

The DI-2 Survey also asked establishment whether the had registered with MIIP, BoI, EPF, CEB, IDB, EDB etc; Such data may be useful in guiding DCS in the choice of data sources for future updating Process. The results reveal that 40 percent of establishment said they were registered with CEB, while 30 percent said they were registered with BoI. Further more 20 percent were registered with EPF.

Results by district, When the discoveries are broken down by the district in the all island, Gampaha had the largest number, 186, while Colombo had the second largest number 184. The success rate was higher in Gampaha (54 percent) than in Colombo (51 percent).

The DI-2 survey asked to identify the Grama Niladari Division (GND, smaller unit of administrative area in the district) which was located by the establishment. This GND level information is needed to classify the industrial establishment at GND level. In some of the cases, respondents do not know the name of the GNDs. In such cases, address list have to be sent to the district industrial office for further investigation.

MIIP. The 1440 RMES candidates available for selection from MIIP source, only 242 candidates were checked in the field through DI-2 survey. For budgetary reasons DCS had to select 242 MIIP candidates out of 1440 available for RMES selection.

In terms of success rate, MIIP had the low scores, 22 percent.

Of the 242 candidates from MIIP source checked in the field for all island , 54 are active establishments with 20 and more workers. Among the 188 that did not qualify, 32 were closed, 9 had shifted into non-industry, 95 could not be found, 20 were small (fewer than 20 workers) 7 were located in a district other than the one where they were sought, and 3 were temporary closed and to be re checked next year and 5 were head office.

BoI. The 791 RMES candidates available for selection from BoI source, 716 candidates were checked in the field though DI-2 survey.

Highest success rate (percent) qualifying for addition to the registry were turned for BoI, 64 percent when compared with other sources. Of the 716 candidates from BoI sources checked in the field for all island. 455 are active establishment with 20 or more workers. Among the 261 that did not qualify for new registry (workers with 20 or more) 58 were closed, 33 had shifted in to non industry, 26 could not be found, 24 were small (fewer than 20 workers), 3 were located in a district other than the one where they were sough, and 9 were temporary closed and to be re-checked by next year.

Size of the new registry, 2006

A. New Core registry 2005 (DCS05) for all island	5235	estabs.
B. New discoveries (DI-2)	699	estabs.
Size of the new registry 2006 (A+B)	5936	estabs.

A new Core registry (DCS05) has been formed after matching has been completed. The core registry DCS05 consisted of :

1. All establishments in the Census04 list and
2. Establishments in the DCS02 Core that do not match to census04

This Core registry DCS05 for all island included 5235 establishments. Of the 1588 candidates from the external sources that do not match with Core registry DCS05, checked in the field, 699 qualified for DCS registry.

Therefore, the size of the new registry (DCS06) after DI-2 successful candidates were added to the Core registry (DCS05) was 5936 establishments.

Lessons learned from call Center

The total of 3448 calls were taken over four months period by 14 officers during this office hours. 2649 Establishments were effectively contacted and of these 400 were found to be dormant. Final response rate (desired results) was around 200. It should be noted that on the average an establishment was contacted only 1.4 times by our operators.

For the call center to be more meaningful the number of times as establishment is contacted has to be increased to at least 5 times and also the staff engaged in follow-up it should be paid an incentive based on the results obtained. A public relations exercise launched simultaneously could improve the results. For a more comprehensive picture please refer the call center final report attached (see annex 1).

Plans to print the new registry

It was proposed to print the new registry by informing the industrialist in advance whether they are agreed to include their names and address, type of industry and employment category etc. in the directory of industrial establishments. DCS is not in a position to divulge the individual information of establishment to the public according to our statistics ordinance. A sample page of industrial directory is attached (Annex 3)

Annex I

Call Center for Annual Survey on Industry

Final Report

1. Terms of Reference :-

- To strengthen the Industrial statistical operation in Sri Lanka to facilitate the activities of the Ministry of Industries, the Board of Investment, the Central Bank of Sri Lanka and the Department of Census and Statistics, by assisting the Department of Census and Statistics to update its Industrial registry. This is to be achieved by assisting the DCS to develop a modern system of enhanced survey operations at Industrial establishments by making better use of the telephone and the fax communications.

- More specifically UNIDO will provide DCS with the required advice , software and hardware in the exercise of registry updating through the twin tasks of obtaining and processing information

  * Available with external agencies such as the EPF and the CEB and

  * Obtained through a questionnaire addressed to each establishment under survey for specific information needs.

- A Call Center is established, equipped, with a specialist consultant to act as a catalyst in improving the response rate of these information needs with the view of improving the overall efficiency of the quality and quantity of data collected.

2. Period of Operation :- From 1st February to 31st July 2006

3. Monthly Progress :-

3.1. <u>February 2006.</u>

Assumed duties on the 1st February 2006

Familiarization meeting with the director DCS and the Staff.

Finalizing the arrangements with regard to the staff allocated to the Call Center.

Drafting the 1st reminder to the non respondents of the ASI questionnaire (already posted).

Familiarization with the ASI survey, the questionnaire and the instruction leaflet that accompanies the ASI questionnaire

Attended meetings at Genesis on soft ware development and it was expected that the software would be delivered for application early March 2006.

Initiation of discussions with each member of the Call Center group to ascertain their views on the practicalities, the issues and the problems associated with improving the response rate for the current ASI questionnaire.

First training Session Module 1- Induction and Orientation.

Conducting study sessions with the Call Center group on the ASI questionnaire and the instruction manual

3.2. March 2006

Recommending amendments to the ASI questionnaire and the instructions manual to the Director DCS with the view of overcoming certain ambiguities in the accompanying instructions to improve the efficiency level of communication.

Installing Phoned and designing the Call Center for two telephones.

Second training Session Module 11 - Possible reasons for not responding to the ASI questionnaire

Incorporating amendments to the approved ASI questionnaire and the instructions manual

Attended the Genesis meeting on software and since there were certain modifications and improvements still pending there was no commitment to finalize the software package by the end of the month..

Initiating discussions with the staff and the Director on the possibility of obtaining a hard copy of the existing Industry frame to commence the Call Center due to the delay in developing customized software.

Third Training Session Module 111 – The Basics of task oriented effective communication.

Preparation of a report addressed to the Director General DCS through Director (Industry and Services) making recommendations on the steps needed to take in order to improve the response rate for questionnaires on the basis of studies carried out to date.

3.3. April 2006

Preparation of a Press release to be made to the public news papers(English and Sinhala) in order to gain some publicity for the new system of collecting information i.e. through the call center. Also proposed was a TV interview with the Director DCS.

Translating the Press release in to Sinhala.

Working with the staff in obtaining a hard copy of the Industrial establishment frame that would be used as a base, in place of the software, to enable the Call center activist to commence calling the non responsive establishments (net of responded to questionnaire and 1st reminder) to solicit their contribution to the survey. A coding system was used to identify the establishments in their different stages of activity such as return from post etc..

Training Session Module 1V – Conversational Protocol.

Updating the hard copy of the Industrial frame with regard to the survey activities that have already preceded, such as the ‘return from post’ and ‘response received’ with the view of obtaining a updated hard copy for the use ion the Call Center.

Adopting the ASI questionnaire and instruction manual from the page maker to the office XL and word formats for mailing.

Fragmenting the existing updated Industrial frame among the Call Center activists in accordance with the areas the activists are involved in ASI management. .

Commencement of Call Center operations.24th April 2006 with the hard copy of the Industrial frame as the base.

Allocating congenial user times for 8 operators to use the 2 phones on two sessions of the day with each operator having two sessions every week..

3.4. May 2006

Introducing and maintaining a diary for each phone giving then operator in each session and also the number of calls taken in each session.

Preparing motivational theory papers drawing parrarels between the objectives of the call center operation and the needs of the society vis a vis individual needs of the call center activist as a member of that society.

Monitoring the calls taken and the response received from various establishments.

Periodic meeting with the center activists to review the situation and take appropriate action. For instance since 90 % of the respondents called within the first three weeks maintained that the questionnaire was not received by them, action was taken to re send the questionnaire to every one of them.

Attending a workshop with the call Center activists at Genesis to familiarize the staff with the custom made soft ware that would be introduced to the system shortly.

Monitoring the progress of the call center and particularly the use of the call center facility and and follow up action on those falling short of the targeted use.

Meeting the activists individually to discuss the problems they have encountered in getting the establishments to submit the information as per our questionnaire.
Monitored the response rate for call center activity and proposed to Director DCS and UNIDO Consultant an incentive payment to call center staff based on the response elicited.

3.5. June 2006

Monitoring the call center results to end of May, appraisal of the results and taking corrective action with regard to areas where there had not been sufficient activity

Discussing with the DCS staff the possibility of increasing the Call Center activists in view of the under utilization of the call center facility.

Inducting six more members of the DCS staff as call Center activists and familiarizing them with the 5 training modules prepared to train the call center activists.

Re allocating the non responded establishments among the new activists in keeping with their individual areas of ASI management so that monitoring of the response also becomes possible.

Reallocating the call center time among the 14 members now supposed to operate the call center depending on times convenient to the 14 operators..

Redesigning the progress charts to accommodate the progress of the now 14 call center activists

3.6. July 2006

A workshop was held in the DCS premises for call center activists by Genesis on the installation and the operation of the finalized ASI software.

A trial runs were performed by DCS staff on the workings of the ASI management software. A few bugs were detected, they were brought to the attention of Genesis and corrected.

Progress of the response was monitored as perfected questionnaires continued to come in trickles.

Review meeting was held with the view of making recommendations for the future course of activity.

4. The Final Results

4.1 ASI Survey results

Total No. of establishments	Response to the Questionnaire	Response to the 1st Reminder	Already responded	Pending
6050	681	679	178	1462
	11.3 %	11.2 %	2.9 %	24.2%

4.2 Call Centre Operation

Total No. of Establishments with employees 20 or more

Districts	No. of Est.	No. of Est. contacted	No. of times contacted	Dormant No.	Response	Pending
Colombo	1122	782	1	168	37	289
Gampaha	977	395	1	86	22	141
Kalutara	302	220	1	32	8	54
Kandy	262	190	1	42	22	36
Matale	61	47	1	6	3	16
Nuwaraeliya	162	61	1	12	1	45
Galle	231	153	2	28	19	78
Matara	157	78	1	21	8	23
Hambantota	44	37	1	5	1	7
Kurunegala	266	90	1	8	2	27
Puttalam	195	92	1	12	1	41
Anuradhapura	54	41	1	6	3	2
Polonnaruwa	52	40	1	5	2	11
Badulla	108	108	1	8	3	15
Moneragala	18	9	1	5	_	4
Rathnapura	192	137	2	21	28	52
Kegalle	129	93	1	8	13	49
Jaffna	33	14	1	3	0	11
Mannar	3	1	1	0	0	1
Vavuniya	11	0	0	0	0	3
Mullativu	14	0	0	0	0	2
Kilinochchi	12	1	1	1	0	1
Batticaloa	21	16	2	4	2	10
Ampara	45	27	1	6	2	11
Trincomalee	13	5	2	0	1	5
Total	**4484**	**2637**	**27**	**487**	**178**	**934**

5.Observations and Comments :-

5.1. As can be observed the final results of the call center is far from being completed as the majority of the results are in ‘pending’ form. It should also be noted that in areas such as Colombo ,Gampaha and Kurunegala, due to reasons mentioned in my previous reports all the establishments were not contacted.

5.2. The survey has received around 25 % response initially (questionnaire + reminder) and that is high compared to the customary 20 % initial response for surveys at DCS. Pushing the response rate beyond 25%, the task the call center attempted, is proving to be difficult unless there is intensive interaction.

5.3. Even in the case of a call center it is a matter of pursuance, as those districts where the establishments were called up twice, e.g Ratnapura and Galle, the response rate is high. Judging by the type of response the call center was able to evoke, it was mentioned in my third progress report, that it would be necessary to give at least 5 calls to every establishment before tangible results could be felt.

5.4. Six months is, too short a period for an operation of this nature since when the preparation takes one months and finalization another, there would be hardly four months of actual work. Specially in this case since we were waiting for the software there was a contingency period and effectively the Call Center operated for only three months. It should also be noted that this particular period experienced the new year vacation, Vesak and Poson holidays, meaning time lost due to annual holidays. In crucial areas like Gampaha, Kurunegala and puttlam the operators were not able to cover all the establishments even once through the Call Center.

5.4. Different establishments react in different ways when they are approached for information. Of the establishments called the response to call center could be generalized as follows,

Refusal to have a dialogue	7 %
Emphatic refusal to cooperate	6 %
Response with combative questions (Such as why you need the information? or what benefit would come to us? etc.)	20 %
Those who agree to send but sounds negative	30 %
Those who agree to send	37%

5.5..Without discounting the need to Motivate, train and guide the staff in an exercise of this nature, it should be recognized that the real limiting factor in the response rate is the ability to motivate the respondent. Hence other strategies have to employed in order to motivate the industrialist and these should mean the deployment of strategies to either impel or compel the industrialist to respond to the survey returns.

5.6.When you consider the impelling strategy organizations such as the Central Bank invoke a prompt and better response rate among the industrial fraternity since the CB has established itself as a principal Government controlling body whose reports are often quoted at national and international forums. As far as the DCS is concerned the public perception of the department is limited to the census performed once in about ten years and now as the department preparing the COL index.

The stature of the DCS as the principal Government body responsible for data collection and processing has to be augmented through a public relations exercise. .

5.7. The compelling strategy is successfully employed by departments such as Inland Revenue where legal action is taken when the returns are not submitted in time. The DCS too has to assert its authority in terms of its Statistical ordinance and such assertion will help improve the response rate without having to resort to field activity which costs time and money. Section 10 of the statistical ordinance 1956, provide for such action and if the provision available at present is insufficient, suitable amendments should be made in view of the importance of collecting this information..

5.8.Another important factor that needs to be taken in to account is the fact that many of these establishments complain that there are number of Government and local Government institutions that keep requesting them for various types of information at various intervals. Certain organizations complain that there are many forms to be filled up that they may need to appoint a separate officer to deal with such ‘form fillings’. This certainly is a factor that dilutes the establishments focus of the need to respond.

6. Recommendations.

6.1. A public relations drive should be launched in line with my previous recommendations to the Director General DCS to augment the Departments stature.

6.2. The Department should take action to prosecute recalcitrant members of public who are openly flaunting their obligation to a Government statutory organization. Just as management information is required to direct a company data collection and processing at the national level is also important to the Government to give direction to the nation. Hence the information requirements of DCS should be treated in that spirit.

6.3. Since there are quite a number of returns that needs editing and completion, the call center should be essentially used as a facility to interact with establishments for such needs.

Annex II

DI-2 Table 1 : No.of Qualifying Establishments by Year Commercial Production

	No.of Establishments that began commercial production								
Source	**pre-2000**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**Total**
BOI	200	33	25	46	39	70	42	.	**455**
CEB	135	6	6	4	8	18	11	2	**190**
MIIP	17	5	2	16	6	4	3	1	**54**
Total	**352**	**44**	**33**	**66**	**53**	**92**	**56**	**3**	**699**

DI-2 Table 2 : No.of Qualifying Establishments by Employment Size Class

	Employment Size Class						
Source	**20 - 24**	**25 - 29**	**30 - 34**	**35 - 49**	**50 - 99**	**100 +**	**Total**
BOI	31	24	24	56	121	199	**455**
CEB	13	23	8	30	42	74	**190**
MIIP	8	14	5	4	14	9	**54**
Total	**52**	**61**	**37**	**90**	**177**	**282**	**699**

DI-2 Table 3 : No.of Workers at Qualifying Establishments by Employment Size Class

Source	Employment Size Class						Total
	20 - 24	25 - 29	30 - 34	35 - 49	50 - 99	100 +	
BOI	675	638	779	2,310	8,254	75,456	**88,112**
CEB	279	609	244	1,245	2,677	36,933	**41,987**
MIIP	161	350	150	168	983	2,011	**3,823**
Total	**1,115**	**1,597**	**1,173**	**3,723**	**11,914**	**114,400**	**133,922**

DI-2 Table 4 : Success rate by Source

Source	Qualifying	Not Qualifying	Total	Percent Successful
BOI	455	261	716	(63.55%)
CEB	190	440	630	(30.16%)
MIIP	54	188	242	(22.31%)
Total	**699**	**889**	**1,588**	**(44.02%)**

DI-2 Table 5 : Success rate by Source and District (Number of qualifying candidates)

District	BOI	CEB	MIIP	Total
Colombo	146	13	25	**184**
Gampaha	156	14	16	**186**
Kalutara	32	3	2	**37**
Kandy	20	27	4	**51**
Matale	3	3	.	**6**
Nuwara-Eliya	8	15	.	**23**
Galle	6	16	.	**22**
Matara	.	3	.	**3**
Hambantota	1	5	.	**6**
Vavuniya	1	4	.	**5**
Ampara	1	2	.	**3**
Trincomalee	1	2	.	**3**
Kurunegala	21	3	2	**26**
Puttalam	33	2	2	**37**
Anuradhapura	2	9	.	**11**
Polonnaruwa	2	7	1	**10**
Badulla	9	11	.	**20**
Moneragala	1	3	.	**4**
Ratnapura	8	41	1	**50**
Kegalle	4	7	1	**12**
Total	**455**	**190**	**54**	**699**

DI-2 Table 5(contd.) : Success rate by Source and District
(Number of non qualifying candidates)

District	BOI	CEB	MIIP	Total
Colombo	62	39	76	**177**
Gampaha	75	29	53	**157**
Kalutara	22	6	17	**45**
Kandy	7	51	11	**69**
Matale	6	22	1	**29**
Nuwara-Eliya	8	69	1	**78**
Galle	16	53	3	**72**
Matara	5	31	2	**38**
Hambantota	3	8	4	**15**
Jaffna	1	.	.	**1**
Mannar	.	2	.	**2**
Vavuniya	.	4	.	**4**
Batticaloa	2	.	.	**2**
Ampara	.	9	.	**9**
Trincomalee	1	4	1	**6**
Kurunegala	20	5	6	**31**
Puttalam	9	2	5	**16**
Anuradhapura	5	18	1	**24**
Polonnaruwa	3	10	1	**14**
Badulla	2	26	1	**29**
Moneragala	5	4	.	**9**
Ratnapura	6	37	4	**47**
Kegalle	3	11	1	**15**
Total	**261**	**440**	**188**	**889**

DI-2 Table 5 (contd.) : Success rate by Source and District
(Total Number of candidates)

District	BOI	CEB	MIIP	Total
Colombo	208	52	101	**361**
Gampaha	231	43	69	**343**
Kalutara	54	9	19	**82**
Kandy	27	78	15	**120**
Matale	9	25	1	**35**
Nuwara-Eliya	16	84	1	**101**
Galle	22	69	3	**94**
Matara	5	34	2	**41**
Hambantota	4	13	4	**21**
Jaffna	1	.	.	**1**
Mannar	.	2	.	**2**
Vavuniya	1	8	.	**9**
Batticaloa	2	.	.	**2**
Ampara	1	11	.	**12**
Trincomalee	2	6	1	**9**
Kurunegala	41	8	8	**57**
Puttalam	42	4	7	**53**
Anuradhapura	7	27	1	**35**
Polonnaruwa	5	17	2	**24**
Badulla	11	37	1	**49**
Moneragala	6	7	.	**13**
Ratnapura	14	78	5	**97**
Kegalle	7	18	2	**27**
Total	**716**	**630**	**242**	**1,588**

DI-2 Table 5 : Success rates by Source and District
(Number of qualifying candidates)

District	BOI	CEB	MIIP	Total
Colombo	70.19	25.00	24.75	**50.97**
Gampaha	67.53	32.56	23.19	**54.23**
Kalutara	59.26	33.33	10.53	**45.12**
Kandy	74.07	34.62	26.67	**42.50**
Matale	33.33	12.00	0.00	**17.14**
Nuwara-Eliya	50.00	17.86	0.00	**22.77**
Galle	27.27	23.19	0.00	**23.40**
Matara	0.00	8.82	0.00	**7.32**
Hambantota	25.00	38.46	0.00	**28.57**
Jaffna	0.00	_	_	**0.00**
Mannar	_	0.00	_	**0.00**
Vavuniya	100.00	50.00	_	**55.56**
Batticaloa	0.00	_	_	**0.00**
Ampara	100.00	18.18	_	**25.00**
Trincomalee	50.00	33.33	0.00	**33.33**
Kurunegala	51.22	37.50	25.00	**45.61**
Puttalam	78.57	50.00	28.57	**69.81**
Anuradhapura	28.57	33.33	0.00	**31.43**
Polonnaruwa	40.00	41.18	50.00	**41.67**
Badulla	81.82	29.73	0.00	**40.82**
Moneragala	16.67	42.86	_	**30.77**
Ratnapura	57.14	52.56	20.00	**51.55**
Kegalle	57.14	38.89	50.00	**44.44**
Total	**63.55**	**30.16**	**22.31**	**44.02**

DI-2 Table 6 : No.of Qualifying Establishments by Registered Agencies

Source	Registered Agencies								Total
	MIIP	BOI	EPF	CEB	IDB	EDB	TQB	Other	
BOI	2	455	134	446	1	7	1	3	**1049**
CEB	33	44	149	190	19	6	13	32	**486**
MIIP	54	2	47	49	6	4	2	0	**164**
Total	**89**	**501**	**330**	**685**	**26**	**17**	**16**	**35**	**1699**

DI-2 Table 7 : No. of Establishments by Source and Employment size class
No. of qualifying candidates

Source	Qualifying Candidates by Employment Size Class				Total
	20 – 29	30 - 49	50 - 99	100 +	
BOI	55	80	121	199	**455**
CEB	36	38	42	74	**190**
MIIP	22	9	14	9	**54**
Total	**113**	**127**	**177**	**282**	**699**

DI-2 Table 7(contd.) : No. of Establishments by Source and Employment size classes
(Number of non qualifying candidates)

	Non Qualifying Candidates by Reason									
Source	Com.Prod. 20+	Com.Prod. <20	Recheck next year	Closed	Moved to known address	Moved to unknown address	Head office	Non-Industry	Not found	Total
BOI	108	24	9	58	.	3	.	33	26	**261**
CEB	233	52	.	23	9	5	3	53	62	**440**
MIIP	17	20	3	32	2	5	5	9	95	**188**
Total	**358**	**96**	**12**	**113**	**11**	**13**	**8**	**95**	**183**	**889**

DI-2 Table 8 A : No.of Establishments by Source Group

		Qualifying Candidates by Employment Size Class				
Source Group		20 - 29	30 - 49	50 - 99	100 +	Total
High Priority A	BOI	24	39	58	76	197
	CEB	35	37	40	45	157
	MIIP	7	9	8	5	29
	Total	**66**	**85**	**106**	**126**	**383**
Medium Priority B	BOI	31	41	63	123	258
	CEB	1	1	2	29	33
	MIIP	15	.	6	4	25
	Total	**47**	**42**	**71**	**156**	**316**
Grand Total		**113**	**127**	**177**	**282**	**699**

DI-2 Table 8B : No.of Establishments by Source Group

		Non Qualifying Candidates by Reason									
Source Group		Com.Prod. 20+	Com.Prod. <20	Recheck next year	Colsed	Moved to known address	Moved to unknown address	Head office	Non-Industry	Not found	Total
High Priority A	BOI	31	6	2	18	.	1	.	9	4	71
	CEB	209	49	.	20	9	5	2	49	26	369
	MIIP	11	16	1	13	2	3	3	2	9	60
	Total	**251**	**71**	**3**	**51**	**11**	**9**	**5**	**60**	**39**	**500**
Medium Priority B	BOI	77	18	7	40	.	2	.	24	22	190
	CEB	24	3	.	3	.	.	1	4	36	71
	MIIP	6	4	2	19	.	2	2	7	86	128
	Total	**107**	**25**	**9**	**62**	**.**	**4**	**3**	**35**	**144**	**389**
Grand Total		**358**	**96**	**12**	**113**	**11**	**13**	**8**	**95**	**183**	**889**

Sample page for proposed Industrial Directory - 2007

District : Colombo
DSD : Colombo

ESTABLISHMENT NAME	ADDRESS	ISIC	DISCRIPTION	EMP. CATEGARY
Amuddra Traders	62/26, Sri Kalyani Rd, Mattakkuliya	2520	Manu.of plastic products	A
Ceylon Leather Coporation	115, Kelani Ganga Mola Rd, Mattakkuliya, Colombo 15	1911	Dressing and tanning of leather	D
Mackie Garment	241, Leyards Brdway, Colombo 14	1810	Manu.of wearing apparel, except fur apparel	C
Little Lion Bakers	105, Vivekananda Hill, Colombo 13	1541	Manufacture of bakery products	D
Mihiri Bakery	38, New Chetti St, Colombo 13	1541	Manufacture of bakery products	B
W N Canyas Bakery	820, Maradana Rd, Colombo 10	1541	Manufacture of bakery products	C
Hasthigiri Hotel	Dam St, Colombo 12	1541	Manufacture of bakery products	B
Little Lion Associates	46, Vivekananda Hill, Colombo 13	1541	Manufacture of bakery products	C
Trans Asia Hotel	151, Sri Chittampalam A Gardinar Mw, Colombo 02	1541	Manufacture of bakery products	B
Hilton Hotel	P O Box 1000, Lotus Rd, Colombo 01	1541	Manufacture of bakery products	A
Galleface Hotel	2, Galle Rd, Colombo 03	1541	Manufacture of bakery products	C
Taj Samudra Hotel	25, Galle Rd, Colombo 03	1541	Manufacture of bakery products	D
K Thilleye Nardhan	148, Justin Akbar Mw, Colombo 02	1541	Manufacture of bakery products	B
Deloranpen &	15, Rock House La, Colombo 15	1539	Manufacture of animal foods not classified elsewhere	D
Gold Coin Feed Mills Lanka	205, Vystwyke Rd, Mattakkuliya, COLOMBO 15	1539	Manufacture of animal foods not classified elsewhere	C
Cargills Quality Foods	35/1, Malwatta La, Mattakkuliya, COLOMBO 15	1511	Slaughtering preperatn & preserving of meat & meat pro	D
Readywear Garment	45, Mogun Rd, Colombo 02	1810	Manu.of wearing apparel, except fur apparel	D
Belgium Tape Led	97, Maligawatta Pl, Colombo 10	1810	Manu.of wearing apparel, except fur apparel	A
Waruna Pearl Industires (Pvt.) Co.	50/6, Sir James Peris Mw, Colombo 02	1810	Manu.of wearing apparel, except fur apparel	C
Next Asia Garment	Millenium House, 46/58, Colombo 02	1810	Manu.of wearing apparel, except fur apparel	B
Lies Coleybon	110, Sir James Peris Mw, Colombo 02	1810	Manu.of wearing apparel, except fur apparel	C

NOTE :

EMP. CATEGARY

A = 20 - 29

B = 30 - 49

C = 50 - 99

D = 100 - 999

E = 1000 - 1999

F = 2000 & Above