COLLECTION OF RAW DATA

TASK FORCE

MEETING Nº 3

4 OCTOBER 2000

Doc. CoRD 019

A brief outline of existing data collection tools and new developments at Statistics Netherlands

For information

Statistics Netherlands

Division Research and Development

P.O.Box 4000
2270 JM Voorburg
The Netherlands

A brief outline on existing data collection tools and new developments at Statistics Netherlands

K.J.Metz [1]

[1] *Remark*: The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

A BRIEF OUTLINE ON EXISTING DATA COLLECTION TOOLS AND NEW DEVELOPMENTS AT STATISTICS NETHERLANDS

Keywords: Data collection, Blaise, Blaise Internet Services, EDISENT, CBS-IRIS, E-quest, EDISheets

1. Introduction

In the CoRD [2] meeting of 27 March 2000 it was agreed that members of this Task Force will provide a short overview of available data collection tools and new developments in the data collection area that may be of interest to other NSIs (National Statistical Institutes). This paper provides information from Statistics Netherlands in this context.

The data collection tools described in this paper are a mixture of mature, long-standing software products and some more recent pieces of software. The Blaise suite and CBS-IRIS are mature software packages in full operational use. CBS-IRIS is being adapted to the current practice in intra-community trade on a yearly basis. Blaise as a software product is very much alive. It is improved and extended on a continuous basis, also taking into account the ideas and wishes of its user group, especially of the large customers who have a full corporate Blaise license. EDISENT, though already used for statistical production in the Netherlands at a fairly large scale, still needs some further improvement.

In addition to these software products, several other activities in the data collection area are carried out at Statistics Netherlands, some of which still are in an experimental phase.

2. Data collection tools at Statistics Netherlands

2.1 Blaise

2.1.1 Short description

Blaise is an integrated system designed for efficient collection, editing, and analysis of survey data. In its present state it is a powerful and flexible system for computer assisted survey processing. Therefore the scope of Blaise is wider than just data collection. Based on a formal survey definition, the Blaise system automatically generates programs for computer assisted face-to-face interviewing with laptop computers (CAPI: Computer Assisted Personal Interviewing), for Computer Assisted Telephone Interviewing (CATI), for Computer Assisted Self-Interviewing (CASI) and for interactive entering and editing of data collected on paper forms (CADI: Computer Assisted Data Input) [3]. Blaise also provides

[2] CoRD is an acronym for Collection of Raw Data. It is a Eurostat Task Force reporting to the STNE Working Group (Statistics, Telematic Networks and EDI).

[3] A new addition to Blaise's functionality is CAWI: Computer Assisted Web Interviewing. This is provided by Blaise Internet Services, described hereafter.

for various data manipulation facilities and has full survey management capabilities. With Blaise one can perform various activities in an easy and user-friendly way.

For the analysis of data files, Blaise can create descriptions of a Blaise data model for use by other software packages such as SAS, SPSS, Stata, Paradox and Oracle. In addition, users can define additional setups and system files of their own choice. Future releases of Blaise will be more and more "open", using a.o. component-based technology.

Though Blaise was initially designed for interviews among persons and households it nowadays has developed into a tool with a much broader scope that can be used for all kinds of surveys, among enterprises as well.

Blaise is used world wide by many types of survey organisations including government institutions, universities, and private research foundations. These organisations conduct a wide variety of surveys with it, such as labour force surveys, consumer price surveys, multilevel rostering household surveys, panel surveys, business and economic surveys, institutional surveys, health surveys, energy surveys, environment surveys, agricultural surveys, and programs of related surveys.

2.1.2 More detailed information

Since Blaise 4 Windows is the current version of *the* prominent data collection tool at Statistics Netherlands, some additional information giving a better insight in its possibilities is reproduced here from the Blaise web site.

Blaise 4 Windows is a powerful and flexible system for computer assisted survey processing. It provides a multitude of options and features for the survey developer:

1. Virtually unlimited capacity for extremely large numbers of questions, edits, and hierarchies.
2. Constant enforcement of all appropriate routes and edits without slowing down during long interviews.
3. Concurrent interviewing of two or more respondents.
4. Hierarchical, alphabetical, and trigram coding schemes which can be used together.
5. Lookups of information held in external files.
6. Metadata management and manipulation.
7. Data manipulation, re-coding, exporting, and importing.
8. Language switching during interviewing programmed to happen automatically, through menus, or with a keystroke.
9. Question-by-question interview aids within Blaise itself or through WinHelp.
10. Multimedia capability for graphics, video, and audio.
11. Mouse, pen, and touch-screen support.
12. Survey management and reports.
13. Sophisticated CATI call scheduling, including time zone adjustments, interviewer assignments, and time slices to target respondents.
14. Full control of fonts and font sizes for questions, response text, and entry cells.
15. Customisable user interfaces which can be modified for the user's organisation.
16. An audit trail that can be customised for the user's own needs.

As a development system, Blaise is suitable for both the individual and the large survey organisation. A *Control Centre* integrates many tools that help the developer produce and test instruments. Blaise can be used for surveys in multiple modes such as CATI and CAPI combined collections.

For interviewers, data entry personnel, and data editors, the interfaces are powerful and elegant and have proven to be very efficient and popular. For methodologists, Blaise allows data to be gathered correctly by using edits during the interview.

For high-level managers, Blaise can increase productivity. During instrument development, one system specification handles many tasks. During survey production, data collection, coding, entry, and editing are all combined into one or a few steps.

For systems managers, Blaise is a powerful, generalised system that can be customised to the organisation's needs, avoiding the need to develop expensive in-house systems.

Subject matter specialists, statisticians, and programmers can become adept at authoring Blaise instruments. The modular and reusable structure of the language allows many surveys to use the same blocks of code with little or no modification. This results in faster, surer development and better comparability between surveys. The multimode nature of Blaise encourages (and can *enforce*!) consistent specifications and conventions between multiple modes of use.

2.1.3 Blaise related products

Abacus

Abacus is a fast tabulation program. Simple and complex tables can be specified in a user-friendly way, and the generated tables are camera ready. Abacus is completely menu-driven, which makes it easy to use, even for people with little computer experience. Abacus is included in the basic set of Blaise.

Bascula

Bascula is a package that implements various techniques for computing adjustment weights. It can carry out post-stratification. If insufficient population information is available, Bascula can carry out either linear weighting or multiplicative weighting. It can be used both in combination with Blaise and as a stand alone package. Bascula is *not* included in the basic set of Blaise; a separate license has to be contracted.

Manipula

Manipula has been designed for manipulating ASCII and Blaise files. It is a versatile general-purpose data manipulation tool. It can split and merge files, sort and aggregate data, create and delete variables, convert files and create reports. Manipula is included in the basic set of Blaise. It can be used both in combination with Blaise and as a stand alone package.

Maniplus

Maniplus has been designed for the development of *survey control systems*. Some examples are: a CAPI laptop management system (single- or multi-survey), an in-office data flow system, a data editing organising system or a top-down data review with access to micro data. Maniplus can also be used to enhance CATI management. It is an addition to Blaise and can only be used in combination with it. Maniplus is *not* included in the basic set of Blaise; a separate license has to be contracted.

2.1.4 Platforms

At present Blaise exists for the MS/DOS platform (Blaise III) and the Windows platform (Blaise 4 Windows). Current Blaise III users can continue their license until 2001, but no new Blaise III licenses are issued anymore. Blaise 4 Windows is a *full 32-bit* Windows program, suitable for Windows 95, 98, and NT 4.0 [4]. Up to version 4.2 the metadata and data file structures are compatible with its predecessor, Blaise III. This allows users of previous versions of Blaise to easily change to Blaise 4 Windows. From the current version 4.3 onward there is no such compatibility, because of the introduction of new internal storage structures.

2.1.5 Availability to other NSIs

At the moment there are 76 registered institutional users of Blaise (excluding the old MS/DOS version 2); among these 76 are many NSIs. Other NSIs wanting to use Blaise must obtain a license. The same holds for the auxiliary products Maniplus and Bascula. There are different types of licenses, depending on the intended use and the required support. Details on these types of licenses can be found at the Blaise web site: http://neon.vb.cbs.nl

Additional information can also be acquired from the Blaise support group at Statistics Netherlands via e-mail: blaise@cbs.nl

2.2 Blaise Internet Services (Blaise IS)

2.2.1 Short description

Blaise IS is the first step of Blaise on the Internet. It is a new data collection mode in Blaise (CAWI: Computer Assisted Web Interviewing) and as such it is an addition to the more traditional data collection modes CAPI, CATI, CADI and CASI. Blaise IS runs the Blaise questionnaire on the web (WWW). It supports two modes: on-line and off-line; the results of these modes can be combined.

The on-line mode displays *one question at a time* in the browser. Range checks are done at the browser. After submitting an answer, routing and consistency checking is done at the server. The off-line mode shows *the entire questionnaire* in one HTML page. Range checks are done at the browser. After submitting the complete form, consistency checking is done at the server.

Blaise IS stores the data directly in a Blaise database at the server. This allows for instance to mix the results of CAWI and CATI.

In future releases Blaise IS will gradually be extended to cover more functionality of the Blaise data entry program, such as lookups, classification dialogs and more questions on a page.

2.2.2 Platforms

Blaise IS can only be used in combination with Blaise 4 Windows *version 4.3* and up.

[4] Currently the functioning of Blaise 4 Windows under Windows2000 is under test.

Blaise IS supports Microsoft and Netscape browsers version 3.0 and higher. It requires Windows NT version 4.0 or higher and Microsoft Internet Information Server (IIS) version 3.0 or higher at the server.

2.2.3 Availability to other NSIs

Blaise IS is *not* included in the basic set of Blaise; a separate license has to be contracted. A license policy is currently being defined. More information on this will be supplied on the Blaise web site: see under 2.1.4.

2.3 EDISENT

2.3.1 Short description

EDISENT is an acronym for Electronic Data Interchange between Statistics and ENTerprises. It is a software program that has been further developed within the TELER project, partially funded through the European Union in the 4th Framework Programme. EDISENT was designed to reduce the administrative burden caused by conducting surveys at enterprises and institutes. After completion of a once-only adjustment, EDISENT allows for the *automatic* filling of electronic questionnaires and for sending the results to the data collector(s). It is essential to understand that the use of EDISENT implies more than just entering data into an electronic questionnaire by keying in. EDISENT *can* be used as a data entry engine for keying in data, but the added value of EDISENT is the possibility to fill the questionnaire automatically (“auto fill”) and to also translate the data automatically from one coding system (i.e. that of the enterprise) to another (viz. that of the NSI).

EDISENT has been discussed at several STNE and SERT meetings, where it also has been demonstrated. More detailed information can be found in the TELER deliverables (see below) and in the following STNE documents: STNE/98/8/013B; STNE/97/7/011; STNE97/6/019. In STNE10 there was an unnumbered document (TELER.PDF) under the title: General presentation of TELER.

2.3.2 Platforms

There is an EDISENT version for both the MS/DOS and the Windows platforms. The MS/DOS version is only used by Statistics Netherlands for its own purposes; it currently has been distributed among some 1 200 enterprises in the Netherlands. For the pilots in the TELER project a 16 bits Windows (3.1 or up) prototype version has been made with different, much enhanced functionality. After the end of TELER, EDISENT for Windows has been adapted: some bugs were solved, some functional improvements were made based on the results of the TELER trials in 8 countries and the sending module was rewritten completely and adapted to the Dutch situation. Further improvements on EDISENT are being considered.

2.3.3 Availability to other NSIs

The MS/DOS version will be phased out and will thus not be made available to other parties. Interested parties may try and test the EDISENT for Windows prototype that has been used for the TELER project and use this *as is*, but without any support.

All TELER deliverables, including a lot of information on EDISENT, are available at the web. The first 5 (out of 8) deliverables can still be found on the web site that has been maintained by Datacare (one of the participants) during the TELER project: http://www.datacare.ie/teler/download.htm Since TELER has ended now, this TELER information may be removed from the Datacare site in the near future. A more reliable URL to be used for the coming months (or years?) therefore is http://europa.eu.int/en/comm/eurostat/research/dosis/teler/index.htm where all the (8) TELER deliverables are available in PDF format.

Parties who are interested in the EDISENT prototype may contact Statistics Netherlands bilaterally in order to acquire the improved demo version of EDISENT (plus the User Manual in English) and discuss possibilities for further use. The contact e-mail address is nkmr@cbs.nl. Note that *no support* whatsoever will be given on this demo version by Statistics Netherlands. As of now, there is no licensing policy for the use of EDISENT as an official product (including the necessary support and maintenance).

2.4 CBS-IRIS

2.4.1 Short description

IRIS is an acronym for Interactive Registration for International trade Statistics. It is Statistics Netherlands' own alternative to the IDEP/CN8 software that is supplied by Eurostat. CBS-IRIS was initially developed in 1992 in order to solve the problem of collecting data on intra-community trade as of 1 January 1993 when the customs formalities regarding foreign trade between EU Member States were cancelled. During the project co-operation with the Dutch Tax Department ("Inland Revenue") has been established because some of the respondents' data were of interest to both parties. CBS-IRIS provides for the filling of an electronic questionnaire by keying in the appropriate data. Also, some import facilities are available for the automatic inclusion of data from the respondent's computerised accounts. Some 50 software suppliers in the Netherlands have made provisions in their software packages to allow for a connection to CBS-IRIS, but the user of CBS-IRIS can make his own provisions for establishing such a connection as well. The final results of data entry can be sent to the data collector by using diskette(s) or via data communication.

Note: CBS-IRIS has been discussed and demonstrated at length within EDICOM.

2.4.2 Platforms

There is a CBS-IRIS version for both the MS/DOS and the Windows platforms (Windows 95 or up). CBS-IRIS is used by Statistics Netherlands and Statistisches Bundesamt (Germany). In the Netherlands CBS-IRIS is being used by some 8 500 enterprises [5], 3 500 of which use the Windows version. Some 3 500 users also use the ICT module (Intra Community Transactions) for reporting VAT data to the Dutch Tax Department. In Germany currently only the MS/DOS version is used, at some 6 000 enterprises.

[5] This numbered was higher in the past (some 12 500). At that time the threshold value for the obligatory supplying of intra-community trade data was lower than it is now.

2.4.3 Availability to other NSIs

No official license policy has been defined for CBS-IRIS. Statistisches Bundesamt and Statistics Netherlands have signed an ad hoc bilateral contract for the use of CBS-IRIS in Germany. Interested other parties may contact Statistics Netherlands (jtlt@cbs.nl) for discussing this topic or bring this up in EDICOM meetings.

3. Some other data collection activities at Statistics Netherlands

In the previous chapter the main data collection tools have been described and some insight was given in their use for statistical production by Statistics Netherlands. In addition to this, several additional data collection activities have been going on in past and present. A few of the current activities are described in this chapter. Paragraph 3.1 contains an example of large-scale, fast implementation for production, in paragraph 3.2 an example of recent research is given and paragraph 3.3 contains an example of co-operation on a national scale. Paragraph 3.4 contains a short remark on future developments.

3.1 Data collection via the Internet

The first experiments with data collection via the Internet were based on the use of e-mail and were directed towards enterprises. Very recently also experiments have been started to collect data via e-mail with persons and households for the Budget Inquiry survey. As for the first category (e-mail with enterprises) several options have been tried in practice. For the monthly survey on Short-term Economic Indicators ("Conjunctuurtest") an ASCII questionnaire has been tried, sent as the simple message body of an e-mail message, to be filled in by the respondent using his/her mail software and to be sent back via e-mail as well. Valuable information has been derived from this experiment, both on the use of this approach proper, but also on the complications of using e-mail exchanges at a larger scale (correctness of e-mail addresses, reception, collection and processing of the results, sending reminders, authentication of respondents, etc.). A more advanced approach was the sending of an HTML questionnaire via e-mail, to be filled in by the respondent using his browser and to be returned via e-mail. This has been tested for the monthly Construction Survey. The acceptance of this approach seems to be higher, both at the respondents and at Statistics Netherlands itself. This approach led to the development of Blaise Internet Services, described in chapter 2.

Recently the **E-quest** project has been started to further develop instruments and processes and to implement these for the data collection of the larger part of the simple monthly and quarterly surveys for the Manufacturing Industry, Construction, Short-term Economic Indicators, Domestic Trade & Transport Enterprises and Commercial Services by using simple electronic forms. The idea here is that important efficiency gains can be realised in the rather short term. Logistics are part of the problems to be solved for this project. Overall, about 30 000 enterprises may be involved for a total of 270 000 questionnaires (maximum) on a yearly basis. Keywords for this project are: *large scale, simple approach, fast implementation, efficiency gains*. The approach taken will most probably be based on Blaise IS in off-line mode (see above), because this is the most promising approach at this point in time. A very important consideration here is also that the problem of encrypting the data that are sent by the respondent can best be solved with this approach (the solution will be based on Secure MIME).

In the meantime, development of Blaise IS will proceed, improving both the off-line and the on-line mode.

3.2 The spreadsheet approach

Another development is called **EDIsheets**. Here the use of spreadsheets for filling in the data at the respondent side (enterprises) is tried and tested. Using a mixture of Blaise technology and the possibilities of MS Excel a prototype has been developed in which the questionnaire is generated by using Blaise in the format of an Excel spreadsheet that can be filled in by the respondent. The prototype also provides for the possibility of filling in the answers by using spreadsheet formulas that derive the data automatically from the respondents own Excel spreadsheets that are part of his book-keeping system ("self-filling questionnaires in Excel"). There definitely is an EDISENT-like flavour to these exercises. However, this approach is still in the research phase. Use of this technique in production is still under consideration, but for the time being the attention will be focussed on the E-quest approach mentioned above, where the respondent uses his/her own browser, as opposed to the necessity of using software that is provided by Statistics Netherlands.

3.3 Co-operation with other government institutions

Like Statistics Netherlands as a statistical office, also other institutions within and outside Dutch government are in the process of improving their data collection process. Lowering the administrative burden for the enterprises is a hot political issue for quite some time in the Netherlands and it has been defined as a government priority. In this context a project called "Electronische Heerendiensten" [6] has been started in which Statistics Netherlands participates together with LISV (National Institute for Social Security) and the Tax Department.

The project capitalises on the trend that communication between enterprises and public bodies is done more and more in an electronic fashion. Both the business community and these public bodies can take advantage of a shift from information exchanges on paper towards electronic exchanges. Until now, the software that is used for these exchanges is developed on a case-by-case basis. This leads to an abundance of software programs that are provided to the enterprises and/or of program specifications that are handed over to commercial software suppliers. Each of these programs or specs serves their own specific purpose and uses different ways of linking to the enterprises accounts, different methods for structuring the data, different measures for safeguarding the integrity/authenticity of data and communication partners and different data transport facilities. These differences lead to an increase of costs, they raise irritation and are becoming a barrier for practical use, especially for small and medium-sized enterprises (SMEs) who cannot see the wood for the trees.

In an attempt to stop this proliferation the above mentioned partners have joined their efforts in the EHD project. The main aim is to provide generalised specifications for the software to be used, in such a way that the same software can be used for various purposes. The approach taken here is that of a layered model. Statistics Netherlands is responsible for the application layer and is building a prototype starting from its EDISENT experience. LISV takes care of the message layer and the Tax

[6] It is very difficult to translate this wording from ancient Dutch. Heerendiensten are the labour that servants in the past had to carry out for free for their masters (Heeren). The English word for this is statute labour, I think. Here we talk about comparable services, but these must be rendered in an electronic fashion.

Department is responsible for the transport layer. Pilot tests are foreseen for the second half of 2000, where Statistics Netherlands will focus on the financial accounts, LISV on the administration of wages and the Tax Department on secure transport of data.

3.4 Consolidation

Applying new techniques for data collection is very much in a revolutionary development phase. Many different approaches are being studied and applied in practice for statistical production, but at some point in time reflection is needed in order to consolidate the different approaches taken. For the future, Statistics Netherlands is reconsidering its current line of data collection tools and studying whether the positive features of Blaise, Blaise IS, EDISENT, EDIsheets and even CBS-IRIS may be merged into one universal tool that is able to take care all the needed functions. In the end the choice of how to supply data should be left to the enterprises (and the households and persons). And for quite some time to come, the use of paper questionnaires will have to be one of the options also…