AN ARCHITECTURE FOR EDI IN BUSINESS SURVEYS BASED ON THE USE OF BLAISE

Hans R. Stol [1]
Netherlands Central Bureau of Statistics, The Netherlands

1. Introduction

There are external as well as internal reasons for the Netherlands Central Bureau of Statistics (NCBS) to stimulate the use of electronic data interchange (EDI) in business surveys as one of the most important issues for the coming years.

The most important external reasons are:

- Businesses use more and more automated information systems to carry out their administrations.[2]

- Often similar or the same questions are asked in different questionnaires. The questions do not always harmonize with the automated administrations of the businesses.

Some internal reasons are:

- Collecting and processing data can be made faster. The elapsed time between the moment of collection and the moment of publication will be considerably reduced.

- The statistical integration will be improved by using the same data as sources for different statistical figures.

This paper consists of four sections. The first section describes the desired situation in the year 2000 concerning the business surveys in the view of the NCBS. Especially the data traffic between the respondents and the NCBS is mentioned.

The second section emphasizes the use of Blaise III in the process of data gathering by EDI.

The third section is about the construction of the database that will be needed to receive the electronic transferred data correctly.

In the last section the experiences of the NCBS with EDI in collecting foreign trade statistics are described in short, as well as some organizational issues concerning the change to the desired EDI concept.

The term EDI used in this paper means: directly transmission of data from the automated administrations of a respondent to the NCBS computer system. The transmission can be done by floppy disk or — more efficient — by data transmission.

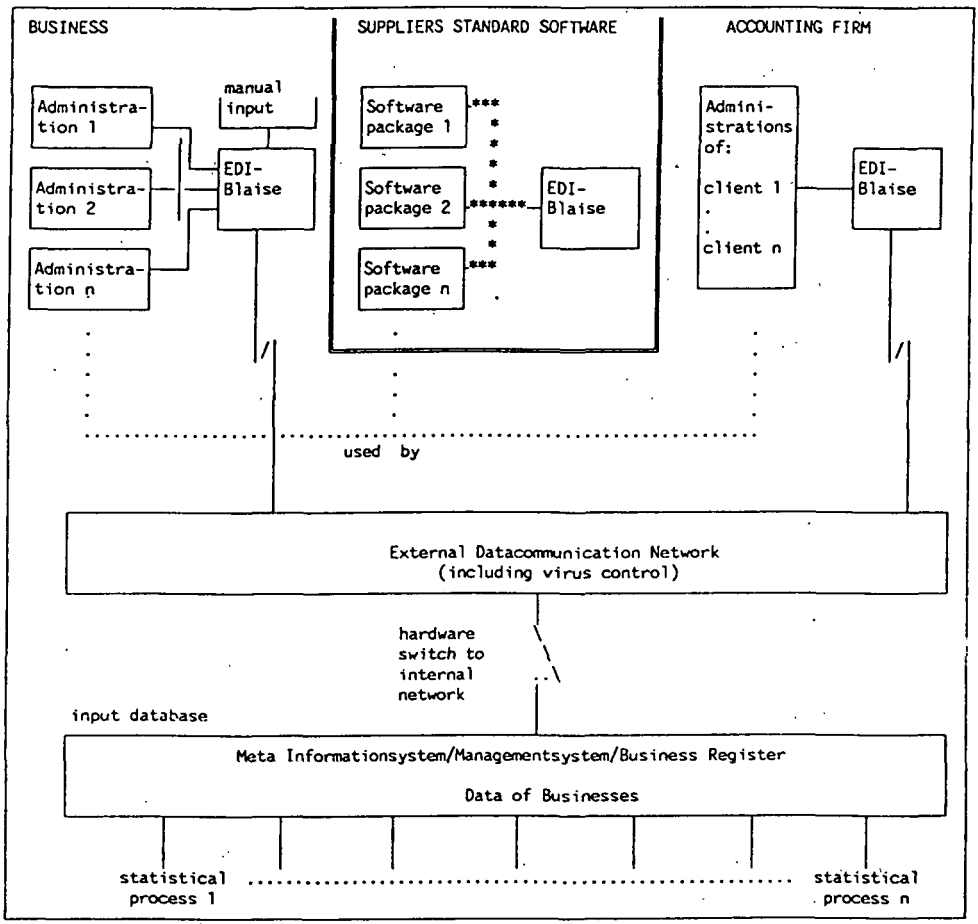
The change to EDI as described in this paper is in a technical, and even more in an organizational sense, a difficult one. It is nevertheless also a challenge and a must to survive in an environment that is heavily automated. The effectiveness and the efficiency of the statistical process as well as the effort for diminishing the administrative burden of information suppliers and the demand of the users for timely statistics will profit from the introduction of EDI.

2. An architecture for EDI in business surveys

Figure 1 shows how data collection in business surveys can be structured using EDI.

Businesses or accounting firms send their data, extracted from their administrations, to the statistical office. Most of the businesses use automated information systems to carry out those administrations. They developed and implemented those information systems dedicated to their own specific needs or purchased them from suppliers of standard software.[3]

Figure 1. The future of EDI in business surveys



Accounting firms carry out the (financial) administration of many of the smaller businesses. Those firms also use dedicated information systems or information systems based on the software purchased from suppliers of standard software.

As the meaning of the data, even inside the businesses, can differ, depending on the use of the data, an interface will be needed to translate the data in a suitable way for the statistical processes.

The basic functions needed for an interface between the administrations of the businesses and the statistical process are provided in Blaise III (see also section 3). It is not always necessary to have a interface like EDI-Blaise. If the data are well structured and have the same meaning as or can be translated to the data needed for the statistics, an interface from the respondent itself or made by the software supplier involved will be sufficient.

When the automated administrations do not contain the data needed for the desired statistics, the respondent must have the opportunity to fill in the data and/or manipulate the data by data entry.

The data arrive at the statistical office at first in a computer installed in an "external network". In the external network the incoming data are checked for executables and viruses. The second function of the intermediate external network is to secure the data stored in the internal network from unauthorized use.

Inside the NCBS an input database takes care of the storage and dissemination of the data of the businesses to the subject matter departments.

The main functions and the data structure of the input database are summarized in section 3.

3. The functions of EDI-Blaise

Figure 2 shows EDI-Blaise related to the business administration. Note that

the business administration can be at the business itself or at an accounting firm.

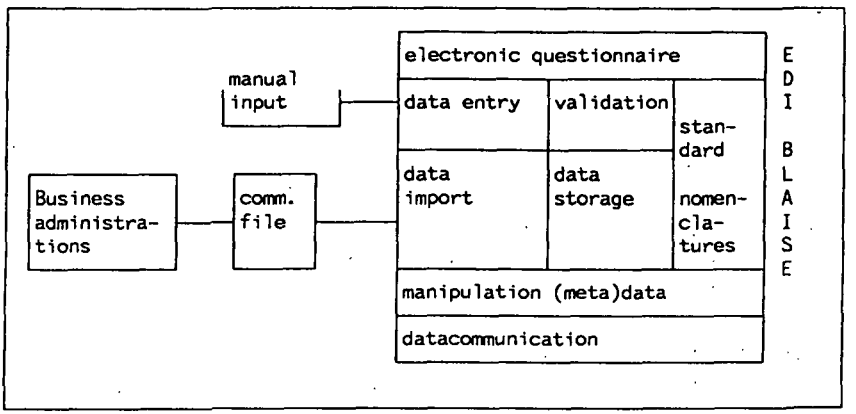
The main functions of EDI-Blaise are:

- Of course, there is the electronic questionnaire. In this case the questionnaire is not dedicated to one survey but programmed to extract the data from the administrations. There will be different questionnaires for different branches as the meaning of the data can vary by branch (e.g. turnover).

- A data entry function is available to add data that are not stored in the administration and can not be derived from it.

- Data are imported from the business administration by an import function that is controlled by the questionnaire. The respondent has to put the data in the given order and format beforehand. The communication file contains data in an appropriate way for EDI-Blaise: flexible but nevertheless standardized.

Figure 2. Some important functions of EDI-Blaise



- There are several ways to manipulate the data. In the first place, the data and the metadata can be manipulated (checking and interactive correction) to export them properly to the NCBS. Also the respondent can make its own figures based on the stored (meta) data.

- The selected data are transferred by an integrated data communication function.

EDI-Blaise offers standard nomenclatures that are used to validate the selected data. The data are stored with the corresponding metadata.

4. The input database

Once the data have arrived at the internal network they have to be stored in an input database. A simplified structure of the input database is shown in figure 3. The input database consists of three parts:

- A registration of the businesses and their mutual relations.

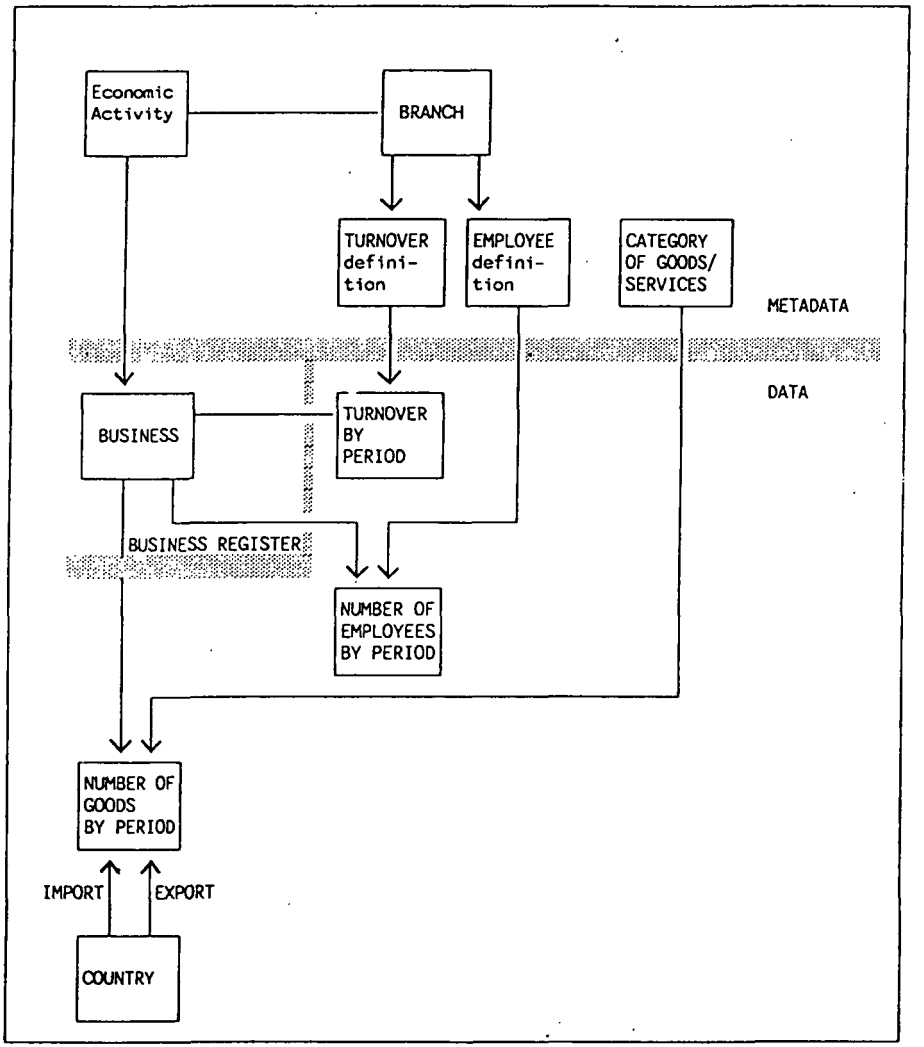
- A registration of the metadata. The meaning of the metadata (including standard nomenclatures), their format, period of validity, etc. are stored. The metadata are related to the corresponding branches.

- The registration of the transferred data itself, related to the corresponding metadata.

A rectangle in the figure represents a data type.[4] An example of a data type concerning metadata is "TURNOVER". Attributes of the data type are: branch id., period, definition. An arrow represents a 1:n relationship between the data types. For example, for one BUSINESS there are more "TURNOVERS" (one for each period).

The structure of the input database will be far more complex than is shown in figure 3. The database has to be structured in a way that the "users"

Figure 3 A simplified data model as a base for the " input database"



(statisticians) can select the desired data for their surveys. Therefore it is not only necessary to know the meaning of the data but also the number of businesses involved related to the population in the survey.

Note that in the simplified data model only one data type "BUSINESS" is mentioned. The current business register of the NCBS (ABR) consists of 8 data types and 12 relationships to represent the "players" in the economy.

As stated earlier, the input database will be very complex. Some people will say that the design and implementation is an ideal instead of a realizable target. The advantages (in terms of statistical coordination and integration) of getting a database as shown are nevertheless unmistakable.

A careful strategy has to be used for the design and implementation of the input database and the functions to fill and use it. Starting with a relative simple structure, prototyping and little by little enlarging it will be keywords in the mentioned strategy. Relational databases like Oracle offer the possibility to realize the concept as described in this section.

It is interesting to notice that the foreign trade can be pictured by four related data types: COUNTRY, NUMBER OF GOODS (by period), BUSINESS and CATEGORY OF GOODS. There are of course two relationships from COUNTRY to NUMBER OF GOODS: one for the import and one for the export of the goods.

5. Experiences with the foreign trade statistics and organizational problems to solve

The first EDI-application of the NCBS — based on the concept as described in section 2 — was realized in the field of foreign trade statistics.

In the end of 1992 the NCBS started with the distribution of a software package called IRIS (Interactive Registration for International trade Statistics) for the data collection at the suppliers of the data. IRIS can be seen

as a program for computer assisted self interviewing. It is built as a Blaise application and consists of the most important functions as described in section 3.

At present, foreign trade data of almost 14000 businesses are transferred to the NCBS every month by floppy disk. A data communication function is built in and has been activated at 100 businesses for testing purposes.

One of the problems to solve was the diversity of computers in the field (IRIS had to work on all IBM-compatible PCs). There were not only technical problems to solve by introducing IRIS. The help desk capacity needed to support the users was underestimated at first. Also the statistical data processing was re-designed because of the automated checking and correction of the data in IRIS.

Although we know the solution created by IRIS and the mentioned production system is only serving the foreign trade statistic, we have identified a lot of technical and organizational problems that we will have to overcome before we have "edify-ed" our economical statistics.

Some organizational issues to discuss in connection with EDI are the following:

- Until now we design paper and pen questionnaires to collect data for one or more surveys. The perception of reality of the statisticians, who are responsible for the survey(s), determines the content of the questionnaire. The respondent has to translate the questions to his figures, often registered in automated information systems. In the EDI-era the statisticians have to accept that the administrative data are the starting-point for the statistics. This means an important "change of mind" with enormous organizational impact.

- New organizational structures have to be designed and implemented to cope with the new way of collecting the data. Examples are help desks and account management on behalf of the respondents (businesses and accounting firms) and suppliers of standard software.

- Design of appropriate statistical methods to make statistics out of the administrative data.

- Define and fulfil the necessary change management: on the one hand continuity in the statistics, on the other hand changing the organization in the above mentioned way. A complication is the n:m relationship between the respondents (and the suppliers of the standard software) and the individual statistics. There will inevitably be a mixture of data collected by conventional questionnaires and data collected by EDI in the transition period.

Apart from the organizational issues, a number of technical problems have to be solved to realize the aimed situation. The right concepts to solve these technical problems are given in sections 2, 3 and 4 of this paper. Summarized, the following actions are needed in this field:

- Harmonize Blaise with the requirements of EDI-Blaise.

- Implement the (concept of) EDI-Blaise in the software packages of the suppliers of standard software.

- Implement the (concept of) EDI-Blaise in the software that accounting firms use for the financial administration of their clients.

- Design, develop and implement the appropriate meta information system (to recognize and interpret the data) in accordance with a management system (to control the logistics of the data) and related to the Business Register.

Notes

[1] The views in this paper are those of the author and do not necessarily reflect the policies of the Netherlands Central Bureau of Statistics.

2. In 1991 the following percentage of businesses in the Netherlands were automated (a organization is defined as "automated" when it had a computer of at least Dfl 2000 and/or specialized automation personal).

Number of employees	5 - 10	10 - 20	20 - 50	50 - 100	100 - 200	200 - 500	500 and more
automation degree	67%	76%	88%	96%	96%	98%	99%

Source: NCBS, *Automation Statistics private sector, 1991*

3. In 1991 the following percentage of businesses in the Netherlands used standard software packages for different applications.

applications / number of employees	financial	sale	production	other
50 - 100	96%	51%	42%	25%
100 - 200	94%	45%	47%	28%
200 and more	89%	44%	47%	45%

Source: NBC, *Automation Statistics, software 1990-1992*

4. This representation of data structures was first published by: Bachman, C.W., 'Data structure diagrams' in *Data Base*, Summer 1969.

Trigram coding in the Family Expenditure Survey of the CBS
Martje Roessingh and Jelke Bethlehem

Errata

Figure 3.2 (page 123)

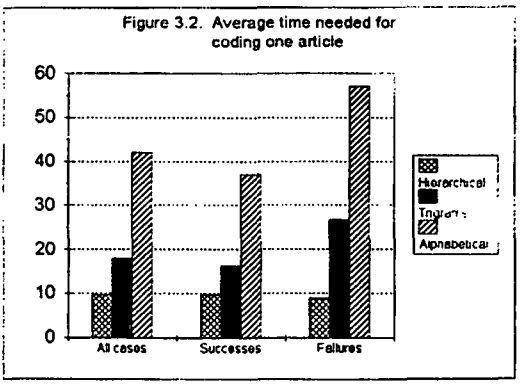


Table 5.1 Use of aphabetical coding (page 127)

Success cases	78 %
Direct alphabetical coding	38 %
First hierarchical, then alphabetical	7 %
Hierarchical, alphabetical and again hierarchical	8 %
First hierarchical, then alphabetical without text	6 %
First with trigrams, then alphabetical	15 %
Other	4 %
Failure cases	22 %
Alphabetical, no final code	8 %
First trigram, then alphabetical, then text edited	8 %
First trigram, then alphabetical, no final code	4 %
Other	2 %

Table 5.2 Mean time required for alphabetical coding (in seconds) (page 128)

Success cases	37
Direct alphabetical coding	22
First hierarchical, then alphabetical	49
Hierarchical, alphabetical and again hierarchical	51
First hierarchical, then alphabetical without text	70
First with trigrams, then alphabetical	43
Failure cases	57
Alphabetical, no final code	72
First trigram, then alphabetical, then text edited	42
First trigram, then alphabetical, no final code	31

Paragraph 6, last sentence has to be: (page 131)
This does not include typing in the text.