Distr.
GENERAL

CES/SEM.46/WP.4
25 June 2001

ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE/EUROSTAT Seminar on Business Registers
(Geneva, 28-29 June 2001)

ON DECISIONS AND QUALITY OF REGISTER DATA[1]

Note by Statistical Office of the Republic of Slovenia[2,3]

SUMMARY

In the paper the authors discuss a rather new field of quality of data. The introductory chapters discusses decision making and importance of data in the process. Further, definitions of quality and data related to the quality of data are considered and proposed. As a general discussion of quality of data is rather difficult, the quality of register data is discussed further, proposing a definition and an approach to be adopted to at least estimate quality of data. Based on that, an example of a method of estimate of the quality of data of the Slovenian Business Register is shown. The final part of the paper contains a discussion and recommendations about a possible future approach to this issue.

.

1. INTRODUCTORY[4]

1.1 On Decision Making

We speak of making decisions when there is a task to be accomplished or a goal to be reached. It is not important whether a task is an easy or a difficult one, or whether a goal is near or remote, or whether ways to solution of the problem are known or not. One of myths of information sciences is that of the correctness of decisions based on appropriate information. Yet there is plenty of examples of wrong decisions even at having all the necessary information at hand. More rare and more surprising are cases where decisions were correct although based on irrelevant or missing information. For this reason, the statement about relevant information as a basis for correct decision is an oversimplification and thus deceptive. Information (or data) is a necessary, but not the sufficient condition for a correct decision. Additional to *data* there must exist an appropriate *model* of a process and a relevant *context*, all of which is also time dependent.

A model is a body of information on a system or a process that enables to simulate the system or a process. The body of information may be formalised and structured, or unformalised and unstructured, or a combination of both. Strictly speaking a model is only present if it is being structured and formalised. In this case a model of a system is a valid functional simplification of a part of a system or of a system as a whole. A model is context-dependent as it depends on a particular part or a property of a system or a process that is under examination. To test a system, various methods are available. In most cases it is more convenient to test a model of a system rather than the system itself, so a model has to be set up to simulate the real system. A model is tested for its validity by entering data into the model, by observing the data that is the result of operations required by the model, and by comparing the results against the required or desired values [2]. This is normally done by iterations whereby during the simulation process the model itself may change too.

When considering decisions we are dealing with real situations in real systems and with real consequences. In professional as well as in everyday life decisions are made daily. Undoubtedly, the concept of decision is common and familiar. It is also commonly agreed upon that

(1) for decisions, information is needed, and
(2) it cannot be certain in advance whether a decision is to be a correct one.

The definition implies that the correctness of a decision can be proved only subsequently, i.e. after the decision has had its effects. Instead of the expression “a correct decision” and “correctness of a decision” it is therefore more appropriate to deal with a *probability of a correct decision*, which may be decided upon only afterwards. Correct (or quality) data is therefore an important prerequisite for correct decisions.

1.2 On Quality

Quality as an important element of competitiveness is a rather new field of concern in development in all industries. By that it is not meant that it is a new concept at all as we all deal with it and understand it as individuals as well as institutions. That which is new is a systematic approach in studying, measurements, and developing means for measurement and improvement of quality. Rather new international and national standards covering quality and related issues, the first of them having been adopted even less than twenty years ago, bear witness of validity of this statement.

The concept of quality has been first applied to commodities, i.e. material products the characteristics of which can be described and measured by physical quantities, and as a consequence of which also the quality can be quantified and measured. Later developments have extended the concept of quality to services where too certain characteristics and properties can be measured and expressed by means of physical quantities. If services are related to production of goods, then they can be described as any other products. It is also possible to describe the way the services are performed, to define it by measurable quantities and eventually to measure, to compare, and to evaluate it. By now both quality of products and quality of services have become recognised as an asset and they are both well regulated by numerous standards. Less has been done to apply the concept of quality and relevant approach to abstract entities an example of which is data.

Fundamental for knowledge workers is use of data of different kinds, from different fields, and from different sources. All that use data at their work have come across the concept of quality of data. Intuitively we are all aware of the fact that, much like goods and services, the data also may be of higher or lesser quality. Rather difficult is it to describe relevant properties of data, to define respective quantities and measures, and to perform suitable measurements, all of which must be done if the quality of data is to be made impartial. The beauty of measurements is just in that they eliminate biased ratings and personal judgements. If for example a data on a length of an object is given as being long, it has implicitly been set in a relation with another object and evaluated, whereby the same object for another observer may be regarded as being short. If however a length is given in terms of a number of units of measure, the data on the length is the same and equal for all. The same approach must be applied to data.

2. QUALITY OF DATA[5]

2.1 General

Related to data we are all familiar with an idea of correct (or incorrect) data. In general data that is in agreement with objectively provable fact is considered as being correct. Data that do not agree with facts is regarded as incorrect or false. In considering the quality of data we are dealing with a generalisation of the concept of correctness of data that is used for a specific purpose. Whereas correctness of data can only have two values (true or false), the quality can be expressed similarly as probability, i.e. with a value between, and including, 0 and 1.

When dealing with a new field it is essential not to introduce new categories unnecessarily as well as not to apply the existing ones improperly. As data and quality already have a recognised definitions it is advisable to examine if they are acceptable for the purpose, use them, and deduce whatever conclusions follow based thereupon. Only if the existing definitions do not apply is it reasonable propose or introduce new ones. Basic definitions on quality are given in standards that are dealing with quality. Slovenian standard SLS ISO 8402 defines the vocabulary, that is definitions and use of expressions on quality of products and services. In the standard, the quality is defined as follows:

> *Quality is the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs.*

Additionally to that it also defines products and services, which are concepts that appear in the upper definition:

> *Product or service can be a result of activities or processes (material product, non-material product such as a service, a computer program, a blueprint, instructions for use) or an activity or a process (such as performing of a service or carrying out a production process).*

It is obvious that the definition of quality covers material categories (products) as well as non-material categories (services, computer programs), *but not the data.* This is remarkable in that computer programs are dealing with data, which may be a part of a program or an input to it, and upon the correctness (but more generally on the quality) of data results of the program depends vitally. Whereas the above definition of quality is rather undisputed, there are much more aspects and definitions related to data. The common feature of all of them is that they put into relation *data*, *information*, and *knowledge*. Let as therefore take the following definitions as a platform for further consideration[6]:

> *Data is representation of objective facts by means of physical processes.*

> *Data is carrier of information.*

> *Information is generated during an experiment; it is a quantity that extends knowledge about a certain phenomenon.*

The above definitions illustrate a rather clear concept that data provide for information and that the information increases knowledge. It is therefore important to operate with quality data in order to increase information and consequently the knowledge. But there is much more to the knowledge than just pursuit for as it is rarely the purpose in itself. The real importance of knowledge becomes important in making decisions.

2.2 Quality of Register Data

It is difficult to offer valid findings, conclusions, and recommendations on quality of data in general. It is advisable to discuss the domain of quality of data that is related to something more specific such as a service or a product. For further discussion let us take the field of register data which is general enough, but rather more defined in terms of inputs, processes, and outputs. It is still about data, the definition, collecting, sources, definitions, semantics, and use of which are however rather well regulated by formal acts of various levels such as acts, decrees, and methods for keeping of registers.

A register is defined as[7] follows:

> *A register is an organised collection of data on a specific phenomenon.*

A phenomenon is something that exists in reality[8]. The phenomenon is comprised of units[9] that in turn are described by data[10] which are descriptions of properties or states of a unit. In general data is information on units. By the word "organised" it is to be understood that all cases of a phenomenon are described by the same set of data and that there is defined a source of every single data and a rule on how the data will be physically represented (e.g. put down, coded, etc.). Phenomena and units can be described by various data the multitude and number of which are unlimited. How many and which of them will be selected depends primarily on the purpose of describing and on various additional considerations such as of processes to be deployed, on relations between data, and similar. It is important to remark that for every particular data there must also exist data on time of collection the data. Only so it is possible to perform a correct comparison of data from different sources and to determine past situations ("archive data", "historical data").

Regarding the above considerations on quality, data, and registers, it is possible to offer the following definition of quality of register data:

> *The quality of register data is a totality of features and characteristics of data that bear on their ability to satisfy stated or implied needs that were the reason of establishing the register.*

If the upper definition is used considering the definition of a register, quality register data can be described in more detail, so that we set quantities to measure quality as well as methods to measure it. Quality register data must satisfy three requirements: at any given time data must offer an adequate picture of a complete phenomenon in that they contain all the units of a register (*the completeness of coverage*); all data must exist for every unit (the completeness of description); and data must be coherent in terms of time (*synchronicity*). Based on these requirements the quantities as well as units of measures of quality can de decided such as the percentage of the coverage; the percentage of correct data on a unit; the percentage of data on a unit at a given time; and similar. Ideally this can be carried out so that at any time all the register data on all the units of a phenomenon is compared with the situation in nature. Obviously, such a method is impossible to carry out, so other approaches must be adopted that will still give satisfactory results, based on which it will be possible to at least estimate the quality of data and to propose measures to improve it accordingly. One such possibility is use of statistical methods for data quality estimates, an example of which is shown below on the data of the Business Register of Slovenia.

3. QUALITY OF DATA OF THE BUSINESS REGISTER OF SLOVENIA

One of the ways of measuring the quality of business registers is to compare some register data with the same kind of data in other sources, such as statistical surveys, registers kept by other institutions and special quality evaluation surveys (control surveys of business registers).

Since 1995, statistical offices of eleven Central European countries have been carrying out the DoSME project (Demography of Small and Medium Sized Enterprises) under Eurostat's guidance. The aim of the project, which used to be named PECO, is to monitor demography of the business population of co-operating countries on the basis of eight panel surveys, to measure quality of business registers and to obtain knowledge and experience in the field of panel surveys. Panel surveys are used in data collection and processing. They are necessary for implementing comparative analyses between input

data, intermediate data and final data, and for assessing certain characteristics of the observed phenomenon on the basis of common methodologies in different periods.

In all co-operating countries the project is carried out according to the same method and on the basis of the same questionnaire, which is, if necessary, adjusted to the type and extent of the survey. In Slovenia the panel survey sample for newly created business subjects covered 1,400 units of the Business Register of Slovenia (BRS) in 1995, 1996 and 1997. According to the methodology prescribed by Eurostat, the project does not cover enterprises the main activities of which belong to the fields of agriculture, hunting, forestry, fishing, public administration, defence, compulsory social security, activities of associations, private households and extra-territorial organisations. In surveys, business subjects created in individual years are checked if their data in business registers correspond to actual data in the surveys.

For business subjects created in 1995, 1996 and 1997 we compared the following data stated by them in surveys with data in the BRS: main activity at the 3-digit level of NACE[11], legal organisational form and size class. Results of comparisons for eleven co-operating countries, prepared by Infostat[12], are shown below.

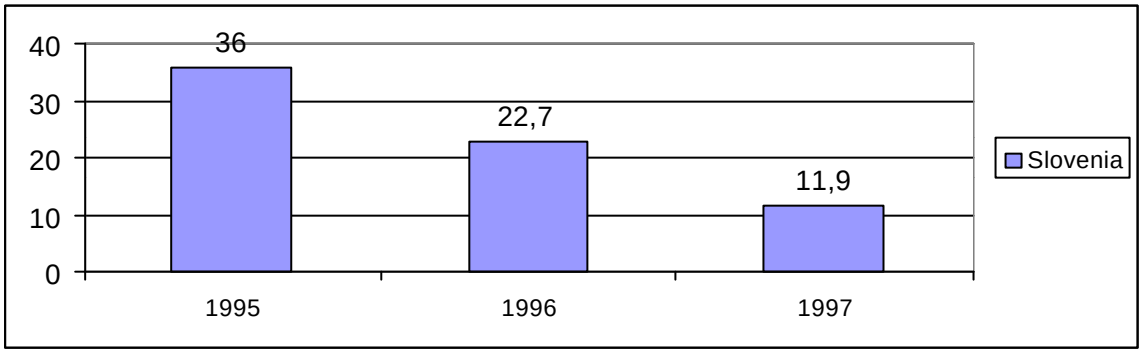

Chart 1: Shares of incorrect main activities at the 3-digit level of NACE in the BRS[13]

Data for Slovenia show that the share of incorrect activities at the 3-digit level of NACE is decreasing. In 1995, 36% of units were misclassified, in 1996 the share was 22.7% and in 1997 only 11.9%. If we were to make assumptions based only on these numbers, we could claim that the 12% share of incorrect activities is rather big. But if we take into account that our survey only dealt with newly created business subjects (which may not know at registration which activity they will be predominantly engaged in) and if we compare our data with those collected in other countries (where incorrect activities mostly present over 50% of activities), we can see the rather high quality of BRS data. In addition to Slovenia, in all three years under review the share of misclassified activities at the 3-digit level of NACE was under 50% only in Albania and Poland.

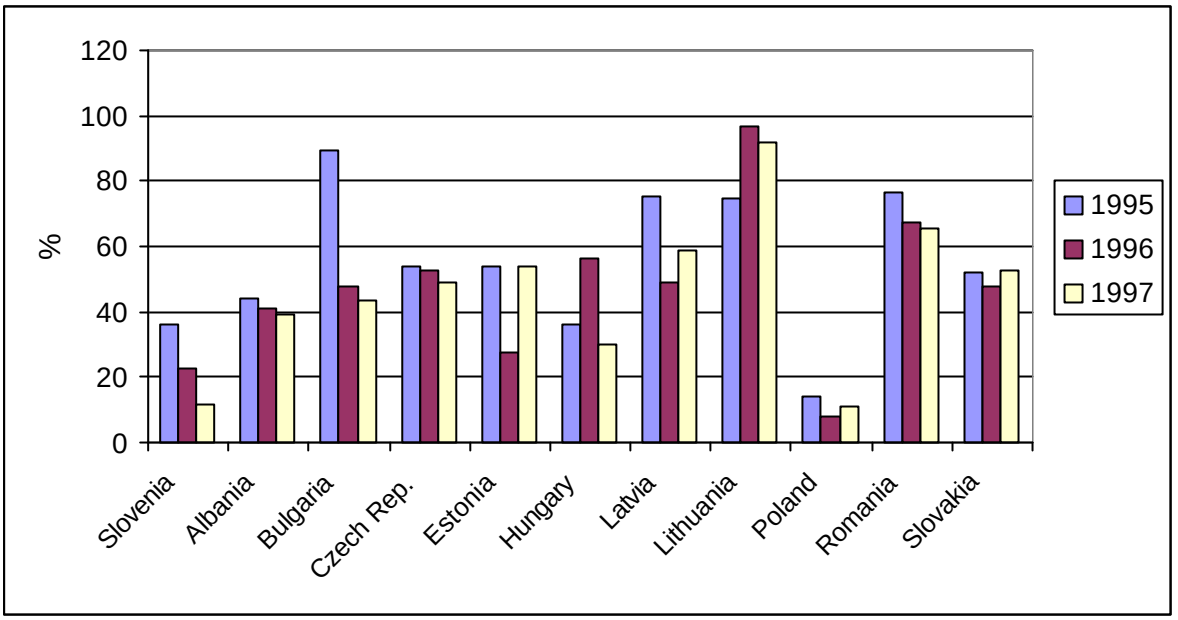

Chart 2: Shares of incorrect main activities at the 3-digit level of NACE, by countries

The next criterion in comparing the data was the number of employees on the basis of 7 classes: *0* - 0 employees, *1* - 1 employee, *2* - 2 employees, *3* - 3 to 4 employees, *4* - 5 to 9 employees, *5* - 10 to 19 employees, *6* - 20 to 49 employees, *7* - 50 and more employees.

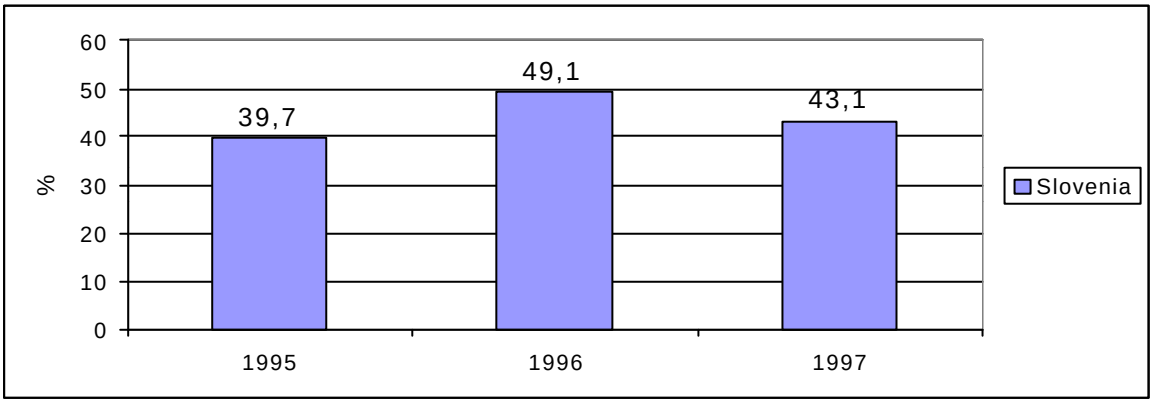

Chart 3: Shares of incorrect sizes on the basis of size classes in the BRS

Data for Slovenia show that the share of incorrect data on the size in the BRS on the basis of the mentioned seven classes was 40% in 1995, 49% in 1996 and 43% in 1997. The shares in the BRS deviate considerably from actual values and show lack of updating. This could be expected since most units observed in the DoSME project have the status of natural persons and the data on their size are at the moment not updated in the BRS. All newly created natural persons are at registration classified into the first class, since they must have at least 1 employee. Our next step in the project will be to take over the data on the number of employees for natural persons from the appropriate administrative source

(e.g. from the Pension and Disability Insurance Institute), which should improve the reliability of these data, which are important sampling stratification variables in many surveys.

Shares of incorrect data on the size on the basis of seven classes in other countries are as shown in the chart 4. Comparison of data for Slovenia with other countries shows that in spite of fact that data on natural persons are not updated their quality is quite good. In addition to Slovenia, in all three observed years the share of incorrect values was under 50% only in Albania and Poland. All other countries have problems with updating these data, irrespective of the status of the business subject (legal or natural person).

The last criterion of data comparison was legal organisational form of the business subject. In Slovenia the distribution of incorrect data on legal organisational forms varies between 0% and 4.6%. In 1997 the share of incorrect legal organisational forms was slightly bigger because of other natural persons (lawyers, doctors, sports persons, freelance artists, etc.), who were in 1997 registered in the BRS for the first time. According to Eurostat's methodology, these legal organisational forms are classified among individual private entrepreneurs. In our

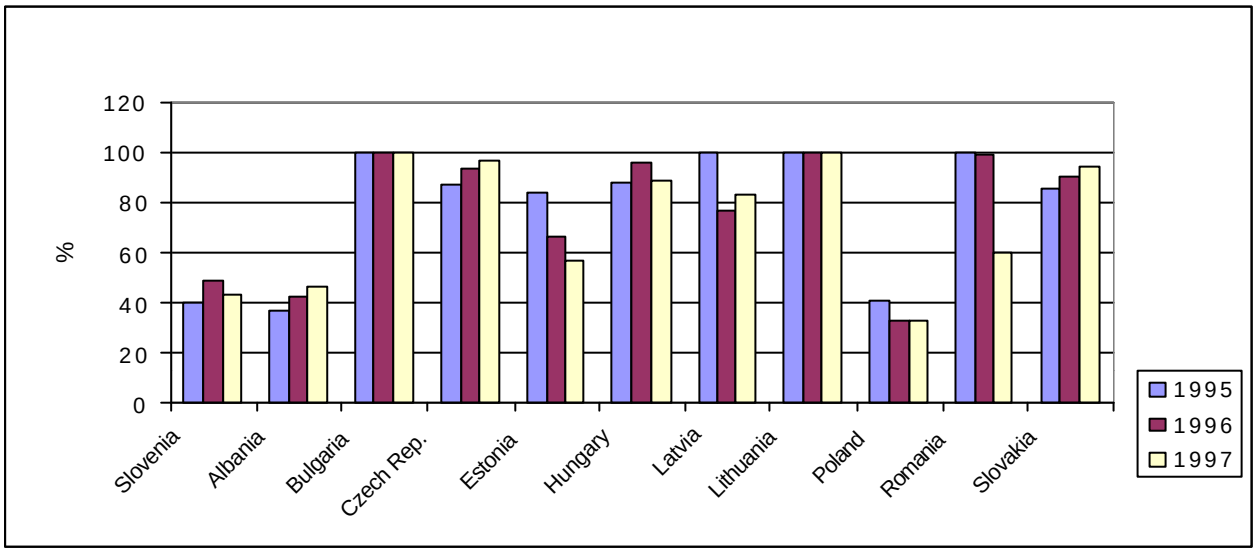

Chart 4: Shares of incorrect sizes on the basis of size classes, by countries

survey the mentioned natural persons were not classified into this legal organisational form, therefore the 1997 share of incorrect legal organisational forms was slightly bigger compared to the previous two years. The quality of these data was good in all countries in all three years, because in most of them the share of incorrect data never exceed 6%.

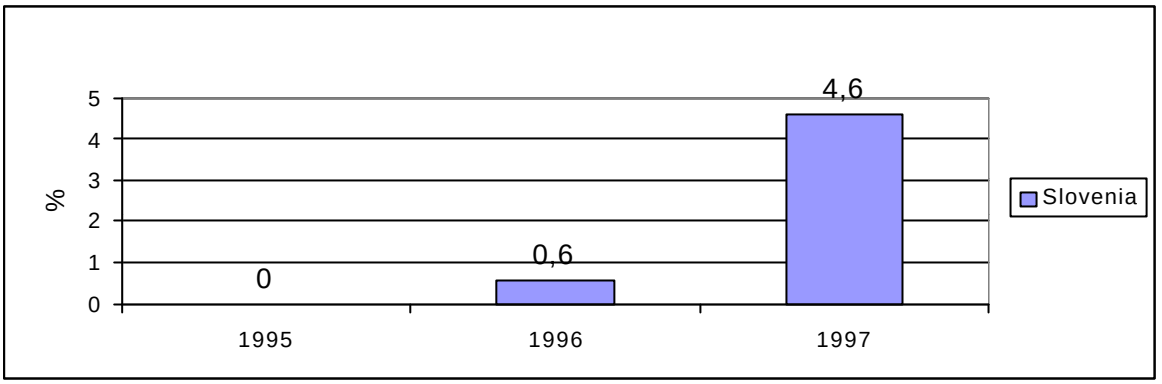

Chart 5: Shares of incorrect legal organisational forms in the BRS

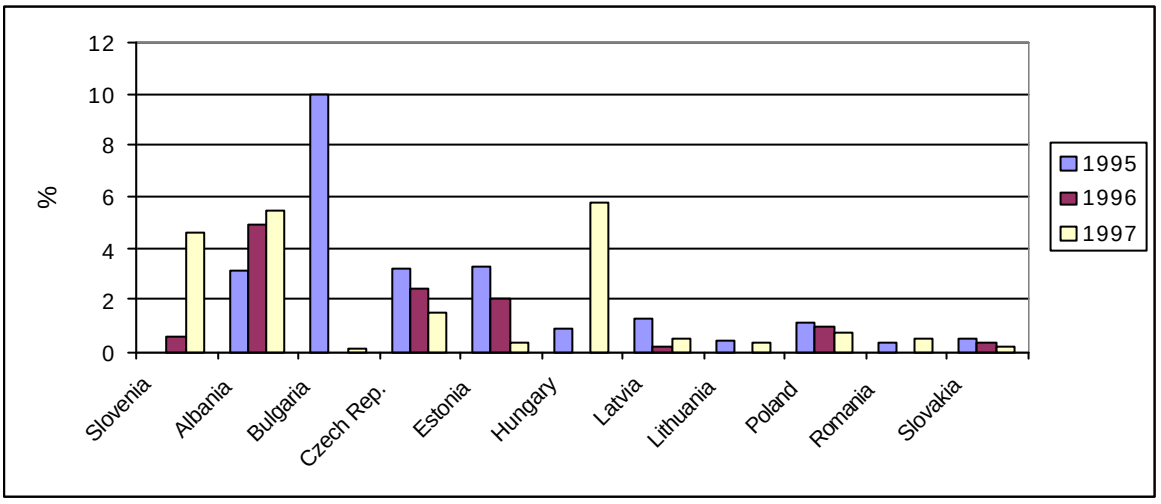

Chart 6: Shares of incorrect legal organisational forms, by countries

From the examples we can see that comparison of data in business registers with data from surveys often reveals the quality of registers, since they are an important source of information on the quality of data in the BRS and tell us which data should be devoted more attention to in updating. We can monitor the quality of all data in the register or we can limit ourselves only to the most important data. The criterion is further applicability of data for users of the business register. If it proves that the data are not reliable and we cannot find the reason for deviation from the actual situation, more attention has to be focused on the source and/or the method of updating. In certain cases it can turn out that it is necessary to use other data sources or to change the method of updating.

It is important that the quality of data in business registers is monitored regularly, since the BRS is universally used and its data have indirect influence on the quality of information in statistical surveys, analytical, commercial, study and research works derived from them and on the quality of other data collections based on BRS data.

4. DISCUSSION

It is beyond any doubt that data is an important element of decision making. It is also equally obvious that the state statistics is probably the most important single provider of data for government decisions. Consequently, to arrive at quality decisions, quality data must be available. The contribution wants to show that the quality of data in general and the quality of register data in particular is a quantity that can be defined and measured. An example of a possibility of how to measure the quality of the data is shown on the case of the Business Register of Slovenia. There cannot be any discussion about a necessity of measuring quality of data in general as the discussion of importance of correct data for correct decisions has been ended long ago. In making decisions there is more than just data as also models and contexts must be considered, but such an approach would vastly exceed the size and purpose of this paper. It is important that the concept of the quality of data be accepted and consequently applied in practice. In particular this is important in case of the state statistics, the data of which is expected to be reliable as based there upon important economical and political decisions are being made daily.

It has been proved that the quality of data in general as well as the quality of register data in particular can be defined and measured. The example given in this paper can be applied as starting point for measurement of quality of data of other registers and, at a suitable definition, of data of other statistical research in general, provided that there exist definitions of data and methods of keeping of registers and data. The method applied using the business registers data and related to the international project DoSME is but one possibility; there are several others that can be used equally satisfactory. It is important to emphasise that the main purpose of the project DoSME has not been study of the quality of data, but rather the study of demography of business subjects. The results related to the quality of data were, however important and useful in this context, mere side effects. However, they are a convincing proof that a systematic approach could provide even more precise and reliable numbers on the quality of register and other data. Another interesting possibility to improve the quality of register data is to regard the keeping of a register as a service the performing of which can be organised and carried out according to respective standards.

It must be understood that the field of the quality in general, and the quality of data in particular, must be regarded in complete. It is not enough to improve the quality of a certain class of data, but an effort must be undertaken to improve the quality of all the data that are kept, which means the register data as well as all the data of all statistical research. To achieve this there must exist a *data quality policy*. There are organisations that have recognised the need to introduce quality policy and to make formal steps to state it in their business and operative plans. There are also some state statistic offices that have done so although they are more an exception than a rule. *The quality policy*[14] is a generally recognised and established approach. It is defined as *general directions and goals of an organisation that are formally defined by the highest management*. The definition can be extended so as to cover the quality of data. The quality policy on data can be described as *general directions and goals of an organisation regarding the quality* ***of data*** *that are formally defined by the highest management.* The definition indicates that if an organisation decided to systematically handle the quality of data, the responsibility must be that a of the highest management as is the quality policy in general. The necessary activities and processes must be defined, organised, and carried out accordingly as is normally the case with the quality policy.

If we do not want to lag behind as a state and as a state statistics professionally, and considering the role that we have in the state, there is a new quantum leap waiting for us to be made in the near future. This is the effort regarding the data quality. It has been said that the domain is relatively new. The majority of organisations that have begun to explore the field of quality policy in statistics have limited their efforts to quality of statistical processes. The Statistics of Canada for example has been one of a few state statistics that has defined its quality guidelines[15] that regard policy of the quality of statistical data as early as 1987. Rather late, in 1997, they have been extended by guidelines on data quality policy[16], which is rare even more. In a way and taking into account our situation this is a relief as the first steps which are always the most difficult ones have been made by others, so that there are those from whom we can learn. On the other hand the situation is also a challenge. If we decide at this point we may be still early enough for others to learn from us, which has been the case in other professional areas. If we begin now we shall be among the few organisations that have recognised the importance of the quality of data and that have understood the quality policy as inseparable from data quality policy which is -- last but not least -- also a prospect of interesting research work and professional development.

NOTES

[1] Legal disclaimer: The views, ideas, and proposals contained in this article are not necessarily identical with, and supported by, those of the Statistical Office of the Republic of Slovenia.

[2] Meta Blejec, Statistical Office of the Republic of Slovenia, *meta* blejec@gov.si

[3] Niko Schlamberger, Statistical Office of the Republic of Slovenia, niko.schlamberger@gov.si

[4] see more in [6].

[5] see [26].

[6] after [20].

[7] after [21].

[8] in informatised data set the expression *entity* is used.

[9] in relation data base *cases*.

[10] In data base *attributes*, in statistical data sets *signs*.

[11] Standard Classification of Activities, Statistical Office of the Republic of Slovenia, Ljubljana, 1999.

[12] Institute for Informatics and Statistics, Bratislava.

[13] Source of this chart and the subsequent ones is Eurostat, DoSME Project Documentation [19].

[14] SLS ISO 8402, Quality – Dictionary, first edition, 1993.

[15] Quality Guidelines, Statistics Canada, 1987.

[16] Policy on Informing Users of Data Quality and Methodology, Statistics Canada (1997).

REFERENCES

On Decision Making

[1] Margaret Nicholas: The World's Greatest Cranks and Crackpots, Octopus Books Limited, London, 1982.

[2] Meehan, Eugene J.: The Thinking Game (Chatham House Publishers, Inc.), Chatham, New Jersey 07928, 1988, ISBN 0-934540-64-0.

[2] Gordon, Geoffrey: System Simulation (Prentice-Hall, Inc.), Englewood Cliffs, New Jersey, 1969, 13-881805-3, LC 77-87262.

[4] Slanc, Radovan J.: Polkvantitativno modeliranje kot podpora uèenja z raèunalnikom, (Slovensko druš tvo INFORMATIKA), Uporabna informatika 4/1996, ISSN 1318-1882.

[5] Slanc, Radovan J.: Polkvantitativno modeliranje, (Slovensko druš tvo INFORMATIKA), Zbornik, Dnevi slovenske informatike '97, Portorož 1997.

[6] Banovec, Tomaž ; Schlamberger, Niko: Utilization of Data of National Statistics in the Process of Decision Making, Proceedings, Context-Sensitive Decision Support Systems, IFIP WG 8.3 Working Conference, Bled, Slovenia 1998.

[7] Meehan, E. J.: The Thinking Game (A Guide to Effective Study), Chatham House Publishers, Inc. Chatham, New Jersey (1988).

[8] Riley, D.: Learning About Systems by Making Models, Computers and Education 15, pp. 255-262 (1990).

[9] Schlamberger, N.: Raèunalnik in pomoè pri odloèanju, uporabna informatika, 1994 No. 2, (str. 36 - 38), ISSN 1318-1882.

Prototyping tool for development of semi-quantitative computer models EdusPKVN (Ó Korak d.o.o).

Other

[16] Tomaž Banovec, Niko Schlamberger: Informatika v drž avnih organih, Zbornik referatov, Brdo pri Kranju, 1993.
(str. 5 - 18), ISBN 86-81141-32-5.

[17] Niko Schlamberger: Re-Engineering in Administration, Re-Technologies for Information Systems, (Conference Proceedings ReTIS '95, pp. 153 - 160), 4th International Conference on Re-Technologies for Information Systems, Bled, Slovenia, 19 - 20 June, 1995, ISBN 3-7029-0404-2, R. Oldenbourg Wien Munchen 1995.

On Quality

[18] SLS ISO 8402, Kakovost - Slovar, prva izdaja, 1993 (Slovenian Standard).

[19] DoSME: Document number 62: Register Quality - Surveys B1 - B3, Infostat, March 2000.

[20] Dr. Tomaž Mohoriè: O podatkih, informacijah in znanju (On Data, Information, and Knowledge), Uporabna informatika (journal), Slovenian Society INFORMATIKA, 3/1999, ISSN 1318-1882.

[21] Metodološ ko navodilo za vodenje poslovnega registra Slovenije, Statistièni urad Republike Slovenije, avgust 2000 (veè avtorjev, interno gradivo) (Methodology of Keeping of the Business Register of Slovenia, Second draft, August 2000).

[22] PSIST ISO/DIS 9001, Sistem vodenja kakovosti - Zahteve, 2000 (Slovenian Standard).

[23] Quality Guidelines, Statistics Canada, 1987 (second edition).

[24] Policy on Informing Users of Data Quality and Methodology, Statistics Canada (1997).

[25] Quality Requirements, Measurements and Reports (*Document EUROSTAT/D1/BRSU/98-12), Eurostat, Luxembourg, 1998*.
[26] Leš njek, Aleksandra; Schlamberger, Niko: On Quality of Register Data, Statistical Days 2000, Proceedings, ISBN 961-6349-28-7, Radenci, Slovenia (2000).