Management & Marketing (2009) Vol. 4, No. 2, pp. 119-124.

SOME CONSIDERATIONS ABOUT THE MANAGEMENT OF METADATA

Viorel Gh. VODĂ
Institute of Mathematical Statistics and Applied Mathematics
„Gheorghe Mihoc – Caius Iacob" of the Romanian Academy

Abstract. *The aim of the present paper is to clarify as much as possible the notion of metadata in connection with the so-called statistical integration for systems designed for agricultural statistics. First, we tried to trace the origin of this concept that unexpectedly goes back to Sir R. A. Fisher's work at the famous Rothamsted Agricultural Experimental Station in Great Britain. Then, we will discuss Bo Sundgren's approach to metadata management. A personal view on metadata from statistical perspective is also presented.*

Keywords: metadata, informational warehouse, statistical integration.

1. Historical preliminaries

In a paper entitled *„Statistical aspects of the design of experiments"* (1955), Samuel S. Wilks (1906 - 1964) quotes some opinions from Sir John Russell's volume *„The Book of Rothamsted Experiments"* (1917) regarding the analysis of field data accumulated at Rothamsted:

– How to prove that these data are accurate?

– What we do want from data is valuable information for agricultural practice.

– New methods for data analysis are obviously necessary.

It is already well known – at least for statisticians – that in 1919, John Russell employed at Rothamsted a Cambridge educated mathematician – the future titan of statistics – Ronald Aylmer Fisher (1890 – 1962). Facing the huge amount of „raw material" – that is the data Russell put in his arms – Fisher realized that he knows nothing about: who collected these data, under which conditions, in which period of time and to what purpose. In this instance he has probably formulated the first definition of what we call now metadata. Fisher considered this information about data in the sense that the experimenter has a total control on the conditions of the experiment itself. Designing and collecting himself the data provided by the experiment, the analyst expects the required accuracy from statistical viewpoint: homogeneity, absence of aberrant values (outliers), obedience by an identifiable law of variation.

Fisher's ideas on experimentation and data control have gained followers in India where in 1929 a „statistical section" has been set up in the Indian Council of Agricultural research (ICAR).

The leading Indian statistician Prasanta Chandra Mahalanobis (1893 – 1972) has presented a report to the Indian Central Jute Committee in 1937 on the *experimental crop census* (performed that year) in which he also stresses the fact that it is not only useful but extremely necessary to have information about the data obtained in agricultural censuses. He refers mainly to the special conditions in India where the vast territory, endless river floods, ethnical and political disorders affect drastically on data accuracy.

2. The modern background of metadata concept

Samuel S. Wilks wrote in an older paper (1939) entitled *„The rise of modern statistical science"* that the conditions and hypotheses on which the statistical methods are based are much better satisfied from probabilistic viewpoint by the populations generated by industry and science than those we meet in economy, agriculture, social affairs.

A larger variability and a larger degree of uncertainty characterize data obtained from these later three domains. These data are organized in general in huge „collections" handled by specialized agencies (namely national statistical agencies) and are obtained from *inquiries, surveys, censuses* or other special activities. A key feature of the quality of those data is the so-called mutual consistency of the data output from these various collections. This mutual consistency is necessary in order to use them jointly and in a meaningful combination that avoids confusions.

Inconsistency is a major nonconformity especially in a census. Various experts – see for instance Colledge (1999) – consider that the background for consistency is the so-called *statistical integration* – a term invented around 1970 by ABS – Australian Bureau of Statistics.

Nowadays, by *statistical integration* one understands *„the state of having statistical data which are mutually consistent and are related to the greatest extent possible. It also refers to the set of activities which aim to produce data in this state"* (Colledge, 1999, p. 80).

We easily see that, in fact, this definition is an adaptation of the well-known concept of the *state of statistical control* introduced by the „father of SQC", Walter Andrew Shewhart (1891 – 1967) (SQC = Statistical Quality Control).

Statistical integration occurs in two large stages, namely:

1) an *intangible one*, which implies the use of the same common notions terms and standards;

2) a *material one*, which refers to the harmonization of data inputs via a process under control.

For instance, we can see the evidence of integration within a statistical agency in the use of a certain register to provide a common frame for a specific survey, in the use of standard classifications and definitions and in a centralized and easy accessible storage of information about the whole collection of data.

To create this kind of integration it is necessary to have not only data but also metadata and an actual link between these two.

By *metadata* we understand *„data about data"* (or outside information about data) and this refers to: definitions; descriptions; procedures; system parameters; operational results, which characterize statistical programs.

G. Priest (1996) classifies metadata into two categories, namely:

(i) *prescriptive metadata* (or active ones), determining the actions of automated survey processes (i. e. they are driven by machines);

(ii)*descriptive metadata* (or passive ones) which are presented in the form of documentation which is used by a specialized agency (i. e. they are driven by humans).

It is important to notice that from „pure statistical" viewpoint metadata are considered in general an extra-mathematical concept. Although metadata are not involved directly in the inferential process, sometimes, data obtained as a result of a statistical inference is considered as metadata on the population investigated. For instance, if we study the length of a given assortment of sunflower seeds, we extract from a bulk of such seeds a sample of size n which is inspected 100% (that is each seed in the sample is measured as regards its length; the value of n may be established using appropriate documents such as ISO2859 for inspection by attributes).

What we obtain after the inspection of that sample are lengths x_1, x_2 … x_n expressed in a certain conventional unit (centimeters or inches). These are data. Computing the average length x we can make a statistical inference on the actual mean-length of the seed population. Let us suppose that we accepted the hypothesis that the mean-value in m = 1.5 c. u. (conventional units) with a given risk, say 5%. For this m we may offer a confidence interval say (1.1 – 1.9) if the variable length obeys a normal law, with a confidence level of – say 95%.

Of course, the example is didactical but the datum m = 1.5 and the corresponding confidence interval are metadata about the underlying population – as a result of an inferential process.

3. Bo Sundgren's approach to metadata

In 1973, in his PhD thesis entitled *„An infological approach to data bases"* (KTH, Stockholm: KTH = Kungl Tekniska Högskolan = Royal Institute of Technology), Bo Sundgren introduced probably for the first time this word – metadata – in a context he called „infological problem" which can be defined as the problem of how to describe information that should be provided by a system in order to satisfy clients/users needs. The author regarded metadata in the frame of bringing user needs to a computing system, the main goals of these metadata being:

(i) To supply information that allows the finding and interpretation of electronic data in a given complex situation;

(ii)To supply machine processing data that facilitates the flow (exchange) of information between systems (and, of course, within a system).

Some years later, Borje Langefors (a follower of Sundgren) discussed in his *„Essays on Infology"* (1993) some of the notions used by Sundgren and tried to define statistical metadata via some additional concepts and stressing on three features – he believes – the Statistics itself has. Among these notions, we mention:

1. *data item* = a characteristic (or an attribute) of a population unit or of the population itself;
2. *data management* = management of data and related metadata throughout the life cycle of a statistical collection;
3. *statistical collection* = the activity of gathering data primarily for the goal of extracting a minimal information about the subject and to perform a scientific inference;
4. *data set* = the metadata describing a table of data regarding a given specific population and a specific set of data items (in some instances together with actual data/measurements).

The reader will easily observe that these notions are quite different from those used in classical statistical inference. As regards the peculiar aspects of statistical metadata, they derive from the following features of statistical science:

a) statistical procedures are in fact a special kind of mathematics applied to numerical data (usually measurements of physical entities) that are used to describe actual phenomena; this implies that real world phenomena are classified into numerical codes;

b) statistical methods are transferable that is they can be applied to a large variety of subject matter domains: therefore, it is necessary to explain the core notions of the specific domain, since SSI (Statistical Information Systems) do not contain these concepts within their processes (the idea of transferability belongs to the famous American statistician and Guru of Quality, William Edwards Deming (1900 – 1993)).

c) Statistical inference is usual applied by subject matter specialists who are not always experts in statistical science; therefore, there is necessary to explain the main statistical concepts and conditions in which the methods could be used in order to perform a correct interpretation of data.

In Langefors' opinion – reiterated in 2001 by Joanne Lamb- these characteristics of Statistics as a discipline *„led to the early development of primitive metadata"*.

4. The database approach

In accordance with UN (United Nations) document entitled *„Guidelines for Statistical Metadata on the Internet"* (2000), the introduction of a successful database approach for the statistical production system requires well organized metadata which have to fulfill the following purposes or functions:

1) metadata destined to *help users* to find the needed information as for example:

- descriptions of the investigated population;
- which data are available;
- searches by key words;
- searches by logical search menus;
- interfaces between the various parts of the information system.

2) metadata which *interpret raw data*:
- description of the statistical units;
- measurements and classification used;
- periodicity of data collection;
- time-lag between collection and releasing information.

3) metadata which describe the *quality of statistical data*:
- information source;
- the level at which data were collected;
- the nature of data (survey data, accounting data)

As regards the quality of a statistical product we can say that quality in general is equated directly with the satisfaction of user needs. In particular, from the viewpoint of an end user, the quality of a statistical product is characterized by the following seven elements:

1) Relevance – whether the product meets user needs;
2) Accuracy – the difference between an estimate and the true value of the parameter of interest;
3) Timeliness – if timely decisions need to be taken;
4) Accessibility – whether the users can easily use the data;
5) Comparability – reliable comparisons possible;
6) Coherence – various sources based on common terminology, methods;
7) Completeness – available statistics should reflect all user needs and priorities.

It is worth to notice that although the metadata on quality of data are important they are rather insignificant for the majority of the users. Therefore, they must be concentrated on the most important issues such as:

- source of information;
- overall description;
- assessment of their accuracy;
- the collection method;
- the reference time (e.g. the year of reference).

In a very general sense, the so-called naive (or inexpert) user has some requirements from metadata in order to help him – without knowing very much about statistics or IT – to:

- understand the data he locates;
- provide unified documentation of various meta-information sources for information retrieval;
- combine different data from different sources;

- assess these data in terms of content, importance and so on;
- detect errors or omissions or other non-conformities he is aware of.

As we have seen, this metadata subject matter is quite a vast one. We still do not have yet a sound explanatory dictionary on metadata terminology. The existing ISO or IEC (International Electrotechnical Commission) materials are useful but not complete.

A large reference guide (a specialized bibliography) is also needed as some other domains like SQC or Reliability Theory do possess.

References

Charlton, J., Bailey, S. (2001), Sharing best methods and know-how for improving quality of data, *New Techniques and Technologies for Statistics*, Pre-proceedings, **1**, Hersonissos, Crete

Colledge, M. J. (1999), Statistical integration through metadata management, *International Statistical Review*, **67**, no. 1, pp. 79-98

Ghosh, J. K. et al. (1999), Evolution of statistics in India, *International Statistical Review*, **67**, no. 1, pp. 13-34

Jug, M. et al. (2002), *Metadata quality. The quality of official statistics?* retrieved from http://europe.eu.int/comm/eurostat/research/

Kokolakis, G. et al. (2001), The role of metadata in the intelligent use of data: the user approach, New *Techniques and Technologies for Statistics*, Pre-proceedings, **1**, Hersonissos, Crete, pp. 257-262

Lamb, J. (2001), Sharing best methods and know-how for improving generation and use of metadata, New *Techniques and Technologies for Statistics*, Pre-proceedings, **1**, Hersonissos, Crete, pp. 175-194

Petrescu, E., Vodă V. Gh. (2008), *Managementul Fiabilităţii,*Editura ASAB, Bucureşti, Colecţia „Sisteme de Management"

Priest, G. (1996), *A corporate metadata information system,* Internal Working Paper, Statistics, Canada, Ottawa

Wilks, S.S. (1939), The rise of modern statistical science *MIT Industrial Statistics Conference*, Proceedings, New York: Pitman Publ. Corp, pp. 283-310

Management & Marketing (2009) Vol. 4, No. 2, pp. 125-150.

CREATIVE ECONOMY'S GAUGES. ROMANIA'S CASE

Mina IVANOVICI, **Anca MÂNDRULEANU**
Academy of Economic Studies, Bucharest

Abstract. *The aim of this paper is, on the one hand, to outline the main theoretical directions pertaining to the concept of creative economy and, on the other hand, to provide an overview of the creative economy in Romania. Thus, this paper analyses Romania's position as concerns the creative economy and presents several differences between our country and other European countries. In addition, a study on the Romanian youngsters' attitudes is provided in the last part of the paper. Its main purpose is to weight Romania's degree of preparedness for implementing and promoting the creative economy and the creative industries.*

Keywords: creative economy, innovation, creativity

1. Introduction

The new technological conditions consisting in more powerful and cheaper technology enable learning, consumption and creation processes of virtually anything. Under these conditions, the public has access to much more and more complete information and becomes more demanding, but also more creative. Whence the change in the consumer's behavior is reflected by the fact that consumers experiment creative products; increased consumer diversity needs to meet various requirements (mainly referring to ethnicity, gender, age, religion, cognition) thus generating competition; demand for the creative products is highly uncertain because products are experimental and there is no available information about them a priori, while the obtained utility is subjective and intangible; consumers have become richer and better educated and have developed a taste for individualized, customized products.

The creative economy represents one of the greatest contributors to the gross domestic product in many countries, such as Great Britain, France, Germany or the United States of America. The creative economy does not represent, however, a mere sum of the creative industries. Its meaning is far larger and it can only be understood within the context of the relation between information, knowledge and creativity. This context was evoked within the *Lisbon Agenda* as well. Knowledge and creativity together play an essential role in the economy.

2. Theoretical landmarks for the creative economy

The creative industries are, according to the British model: advertising, architecture, arts and antiques markets, crafts, design, designer fashion, film, interactive leisure software, music, television and radio, performing arts, publishing and software. From another standpoint, the one of intellectual property, these industries can be classified within the following categories: *process businesses* (architecture, advertising and PR, marketing services), *product businesses* (film,