UDC 004.82:004.62:314.1

O.R. CHERTOV

National Technical University of Ukraine "Kyiv Polytechnic Institute", Ukraine

MODELS OF STATISTICAL INFORMATION DESCRIPTION METADATA

In the paper, we discuss the problems which arise when modeling metadata used for describing various kinds of statistical information, including demographical one. We also implement advanced semiotic metadata model. Moreover, we enhance a well-known Generic Statistical Business Process Model by introducing new "Depersonalization" and "Providing confidentiality" processes, and also by separating search activity. It's expected that the developed models will be applied to constructing the Integrated statistical data processing system of Ukraine and the computer-aided system "Census-2012" for census data processing.

Key words: *semiotic metadata model, statistical business process model, depersonalization, data anonymity.*

Introduction

The process of developing general structure of a statistical metadata system was initiated in 2004 during the joint meeting of United Nations Economic Commission for Europe (UNECE), Eurostat, and Organization for Economic Co-operation and Development dedicated to the problems of statistical metadata information system (METIS). Today, there have been developed standards and recommendations concerning statistical metadata which already are used in practice by national and international statistics organizations. Some of them are as follows: technical standards and instructions complex called SDMX (Statistical Data and Metadata eXchange) [1], The OECD Data and Metadata Reporting and Presentation Handbook [2], and Generic Statistical Business Process Model [3].

But, all the regulatory or recommendation documents mentioned above are meant either for describing the architecture and generic statistical business processes [1, 3], or for describing the metadata from their interaction with other (not statistical) information systems point of view [2]. At the same time, to practically implement an integrated statistical system driven by metadata, one needs also a detailed examination of a metadata model to be used for describing statistical information and specific way of its processing.

Moreover, it is necessary to explicitly mark out the problems of protecting confidential information because they are of an exclusively big importance in systems of providing or disseminating statistical information. Strasbourg "Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data" [4] implies a peculiar status of statistics, and, thereupon, declares necessity of restricting access rights for statistical information and computer files or databases used in statistics and scientific researches. Importance of providing depersonalization and confidentiality has increased rapidly in the recent decades due to a bunch of reasons [5]:

– increasing of demands for the data of small political subdivisions (to objectively form the state regional policy);

– popularization of selective surveys and dissemination of microdata;

– growing number of cross-reference information sources (statistical, administrative etc.), which increases the risk of unveiling private or confidential information through data comparison;

– extending of information systems capabilities, in particular, rapid development of data mining methods that aid in defining structures and patterns in statistical data.

Therefore, the aim of the current work is to improve existent models of statistical information description metadata. This could make it possible to take into account peculiarities of data processing in statistics, and to explicitly support providing confidentiality of information being stored.

Microdata, Macrodata, Metadata

From the very beginning of their development, information systems in statistics supported storing and processing the data of two kinds, namely, microdata and macrodata.

Microdata mean data about objects of statistical survey. For instance, those could be respondent's sex or date of birth, a number of persons in a household and so on.

Macrodata are estimations of statistical indicators which belong to a group (ensemble) of objects. They aggregate this or that way information about group of homogeneous objects. For example, these might be the mean value of household members in some district, the number of the most numerous ethic groups in some region and so forth.

To describe both microdata and macrodata, one can use metadata. We will further on refer to the following definition. Statistical metadata are structured data which characterize various aspects of statistical data (their contents, accuracy, accessibility etc.).

Statistical Business Processes Model

Statistical community yet a very long time discusses the problems of creating a generic statistical business processes (statistical cycle) model. The aim of such works lies in developing consistent terminology and description of applied business processes for any statistics organization. They are necessary for producing official statistics. It would contribute to standardizing activities of processing and disseminating statistical information, and also to harmonizing statistical calculating infrastructures, along with easing of exchanging parts of corresponding software.

Up to now, we can point out 4 main international models and standards in a way connected with statistical business processes model. They go as follows:

1) handbooks and recommendations concerning the structure of information systems for national and international statistics institutions [6] contain a generic model of stages and activities for statistical monitoring data processing system;

2) Eurostat model called "data life cycle" ("Cycle de vie des Donées") [7] provides a full list of concepts, metadata structures, and information technologies to apply in all statistical branches;

3) combined life cycle model developed for Data Documentation Initiative (DDI) [8], an open international project of developing the standard for technical documentation describing sociological data;

4) generic statistical business processes model (GSBPM) [3] which we will describe in details further on.

Although these models differ in their representation, in fact they are quiet close. GSBPM beats other three models when it comes to performing statistical tasks. For instance, DDI model includes the "Reshaping" process which seems to be unique at first. But, GSBPM contains two processes called "Verifying accessibility of existent data and their usage if possible" (during "Requirements refinement" stage) and "Data integration" (during "Processing" stage) which correspond to the one of DDI model.

Generally speaking, GSBPM includes description of 9 stages, each of them consisting of different number of subprocesses (3 to 8).

Considering experience of conducting statistical monitoring in Ukraine gained throughout the years of its independence, and explicitly marking out processes which provide depersonalization and confidential statistical information preservation, GSBPM can be specified. This could be done by defining 14 processes and regional levels of statistical processing (see Fig. 1). At the local level, there are 498 district and municipal statistics offices. Official statistics bodies of Ukraine at a regional level are the following ones: Senior statistics department in Autonomous Republic of Crimea, 24 Senior statistics departments in every Ukrainian region, and two departments for the cities of Kyiv and Sebastopol. The central level is represented by the State Statistics Committee of Ukraine. We will also relate to it the Senior interregional statistics department in Kyiv (which used to be the Senior Computation Center up to 2000).

It is important to note down that the proposed model does not include general-system processes like strategic planning, human resources, finance, and organizational structure managing etc. Besides, due to the methodological nature of refining requirements, designing, and constructing a statistical process, which are usually done for a specific one, they were left out of the discussed model. On the other hand, systematic processes of statistical data handling have been thoroughly detailed compared to GSBPM.

For the sake of making statistical data closer to their users, it is of a vital importance to implement access to necessary information search:

– with the help of non-regulated queries to an OLTP (on-line transaction processing) statistical system;

– through analytical queries to an OLAP (on-line analytical processing) statistical system;

– by applying data mining to automatic search for certain patterns in them.

That is why, in a model proposed in Fig. 1, we explicitly marked out "Search" process, which in GSBPM is distributed between "Processing", "Analysis", and "Disseminating".

Instead of subprocess "6.4. Information disclosure control" from GSBPM, we introduced two generic processes called "Depersonalization" and "Providing confidentiality".

By personal data depersonalization we understand eliminating information which enables identifying a person (like passport number, full name and so on). The term of "data depersonalization" was brought in a model law "On the personal data" passed on October 16, 1999 at the 14 plenary session of Interparliamentary Assembly of Member Nations of the Commonwealth of Independent States. Principles and regulations of this law this way or the other have to be included into corresponding states' national legislations.

When providing confidentiality of statistical information, we can define three subprocesses as follows:

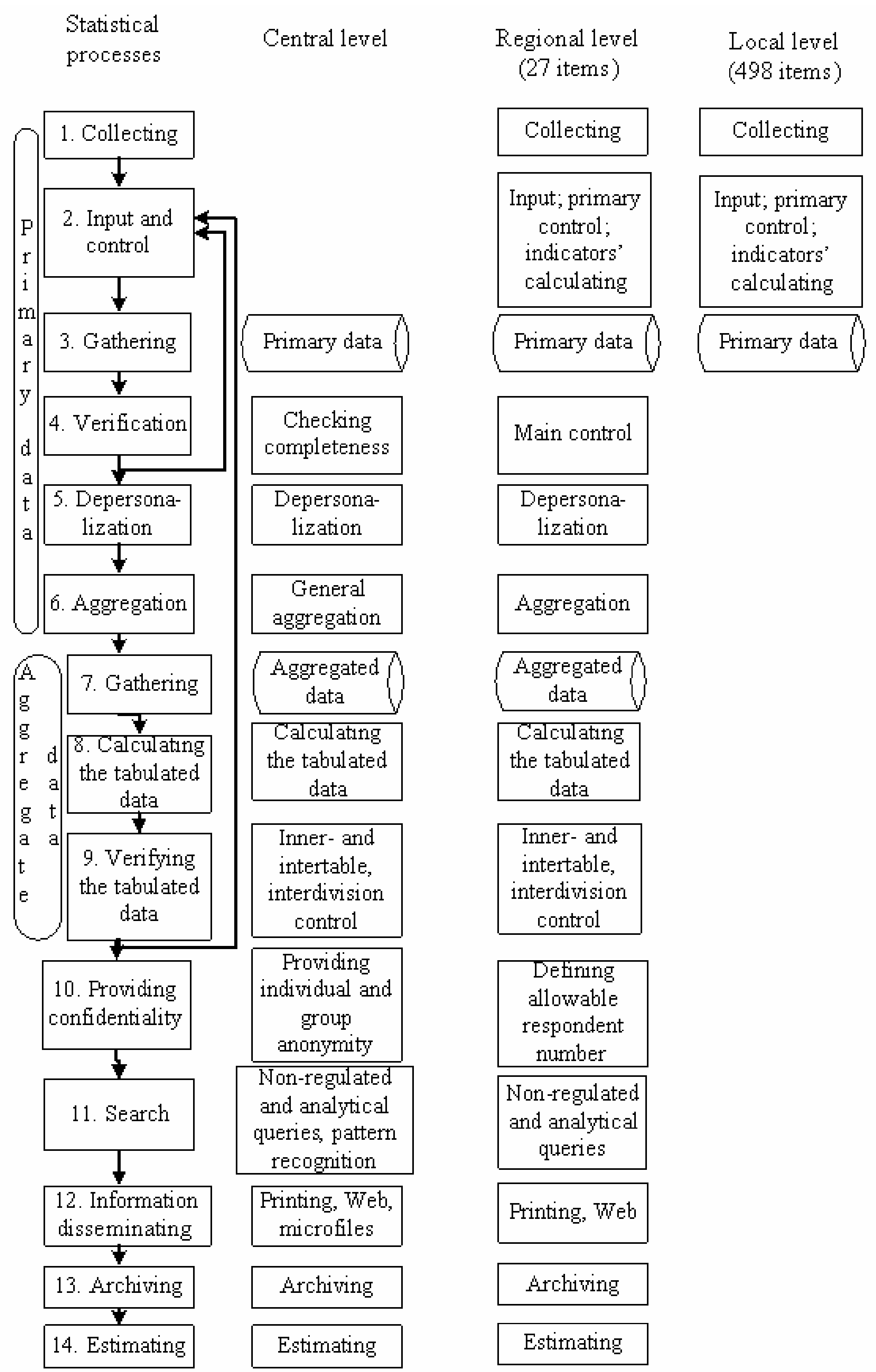
Statistical processes
Central level
Regional level (27 items)
Local level (498 items)
Primary data
Aggregate data
1. Collecting
2. Input and control
3. Gathering
4. Verification
5. Depersona-lization
6. Aggregation
7. Gathering
8. Calculating the tabulated data
9. Verifying the tabulated data
10. Providing confidentiality
11. Search
12. Information disseminating
13. Archiving
14. Estimating
Primary data
Checking completeness
Depersona-lization
General aggregation
Aggregated data
Calculating the tabulated data
Inner- and intertable, interdivision control
Providing individual and group anonymity
Non-regulated and analytical queries, pattern recognition
Printing, Web, microfiles
Archiving
Estimating
Collecting
Input; primary control; indicators’ calculating
Primary data
Main control
Depersona-lization
Aggregation
Aggregated data
Calculating the tabulated data
Inner- and intertable, interdivision control
Defining allowable respondent number
Non-regulated and analytical queries
Printing, Web
Archiving
Estimating
Collecting
Input; primary control; indicators’ calculating
Primary data

Figure 1. Processes and territorial levels of statistical handling in Ukraine

1) determining minimum possible number of respondents to publish information about (in an output table, as a response to an analytical query etc.). For instance, when the data of All-Ukrainian population census (2001) was being processed, the number of 3000 was taken for this matter [9];

2) providing individual data anonymity [10, 11]. As a matter of fact, there exist other methods apart from depersonalization which allow hiding information on a certain respondent (such as lessening specification level for a respondent by, say, referring to him not as "district attorney" but as "law-enforcement officer"). These methods are mainly applied when preparing microfiles, i.e. datasets with information about a sample of respondents;

3) providing group anonymity [12, 13]. This is important because providing user access to primary or aggregated data in a microfile form often leads to enabling of defining distinctive data distribution features. In particular, maximums in the regional distribution of military personnel concentration might point at military bases on a corresponding territory.

Advanced Semiotic Model of Statistical Information Description Metadata

In [14], there was proposed an idea of constructing a model for statistical information description metadata which comes from defining three projections of a three-dimensional metadata representation (see Table 1). Taking into account classical division of semiotics into syntax, semantics, and pragmatics, we propose to call such models *semiotic* ones.

Table 1

Projections (semiotic levels)
of describing statistical metadata

№	Semiotic level	Direction of metadata description	Short explanation
1	syntax projection	data representation	metadata are a physical representation of metainformation
2	semantic projection	data contents	metadata provide information about data
3	pragmatic projection	aims of usage	metadata are the data necessary for preparing and using statistical data

The metadata representation presented in Table 1 is a typical one for various semiotic models. But, it needs to be extended, because it does not explicitly take into account another (fourth) level, namely, confidentiality level of statistical information being stored or processed. Information systems in statistics always dealt with dialectic opposites of data accuracy and confidentiality. In other words, primary microdata can provide the highest statistical information accuracy, but, at the same time, they are of the lowest depersonalization level.

In [6] and [14], syntax and semantic projections have been specified to describe metadata traditionally presented in statistics in a table form. Bo Sundgren, the author of this approach, marks out 4 description components (dimensions), which go as follows:

– α: the dimension of population and scope;

– β: the dimension of measurement and summation;

– γ: the dimension of classification;

– τ: the dimension of time.

Appropriate examples are proposed in Table 2.

Table 2

Metadata macrodescription components

	macrodata components			
component name	α-components	β-components	γ-components	τ-components
syntax projection	table row or column	table cells	table row or column	table row, column, or header
semantic projection (component meaning)	What object are being calculated?	What quantity is being calculated?	What features (characteristics) objects are being distributed by?	Which time series is being taken, i.e., of which time interval and with what periodicity?
example	Ukrainian residents at a certain time period	number divided by 100	by place of living	yearly, from 1.01.1970 to 1.01.2010

All statistical metadata traditionally [14, p. 16-19] are being gathered into groups. But, in our opinion, classification proposed in [14], once again, does not consider importance and peculiarity of providing confidentiality for statistical information. That is why it needs to be extended. We reflected that in additional rows of Table 3 (in italics).

Table 3

Groups of statistical metadata

metadata purpose	metadata group name	symbolic code (and name, if present) of a group	explanation
for the users	general (global)	U1	annotation, thesaurus, index
	declarative	U2.1 / meaning	to measure data relevance
		U2.2 / accuracy	concerning calculating or estimating
		U2.3 / accessibility	how to gain access to the data
	process-oriented	U3	applied models, sampling procedures, coding, editing, verification etc.
for the data suppliers	general (global)	S1	applied methodology, a statistical problem close by technique
	feedback	S2	whether the users are satisfied with received data and, if so, how much
	process-oriented	S3	identical to U3
	confidentiality-oriented	*S4.1*	*depersonalization and individual anonymity*
		S4.2	*group anonymity*
connected with software	factographic	C1	table names, data representation formats etc.
	algorithmic	C2	calculating algorithms and their description

According to the process modeling theory [3, p. 3], every process has to possess a certain number of clearly visible attributes, including input and output data, purpose (with additional characteristics), possessor, management (technical documentation), licensing elements (people and systems), means of feedback.

Mentioned attributes have to be by all means taken into account by statistical data description model (if it is supposed to be used in practice).

Conclusions

In the paper, we improved the generic statistical business processes model by introducing new processes of "Depersonalization" and "Providing confidentiality". Also, we separated the process of searching, either in an OLTP statistical system (by means of non-regulated queries) or in an OLAP statistical system (through analytical queries), or when applying data mining to automated defining the certain patterns. All this helps to provide more complete and accurate description of statistical information handling processes, especially – demographical one.

Besides, the semiotic model of statistical information description metadata gained its further development. In it, as opposed to the existent models, we presented an additional group of metadata for individual and group anonymity. The latest term lies in protecting important data features, distributions, and collective patterns which cannot be defined by analyzing individual data records only.

This enables explicit separating the problems of providing depersonalization and information confidentiality both when describing statistical (particularly, demographical) data and further implementing corresponding statistical processes.

We expect that the developed models will be applied to constructing the Integrated statistical data processing system of Ukraine and the computer-aided system "Census-2012" for census data processing.

References

1. SDMX Standards Version 2.0 [Electronic resource] – Access mode: http://sdmx.org.

2. The OECD Data and Metadata Reporting and Presentation Handbook / Denis Ward (ed.). – Paris: OECD Publications, 2007. – 161 p.

3. Generic Statistical Business Process Model, Version 4.0 (April 2009) / Steven Vale (ed.). – Paris: METIS, 2009. – 28 p.

4. Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ETS №108, 28.01.1981 (with Amendments approved by the Committee of Ministers on 15 June 1999). – Strasbourg: The Council of Europe, 1999. – 35 p.

5. Григорьев В.В. Обеспечение конфиденциальности статистических данных в странах Европейского Сообщества / В.В. Григорьев, Е.В. Кузнецова // Вопросы статистики. – 2000. – № 12. – С. 8-14.

6. Sundgren B. Information systems architecture for national and international statistical offices. Guidelines and recommendations // Conference of European statisticians. Statistical standards and studies, № 51. – Geneva: United Nations Statistical Commission, 1999. – 56 p.

7. Eurostat. Data Life Cycle Object Model for Eurostat Production Systems, Cronos Technologies, Unit A1, DG ESTAT. – Luxemburg: European Commission, 2002. – 25 p.

8. The Data Documentation Initiative Specification, version 3.1 (October 2009) [Electronic resource] – Access mode: http://www.ddialliance.org.

9. Перший Всеукраїнський перепис населення: історичні, методологічні, соціальні, економічні, етнічні аспекти / Н.С. Власенко, Е.М. Лібанова, О.Г. Осауленко та ін..; під наук. ред. акад.. І.Ф. Кураса, акад. С.І. Пирожкова. – К.: Державний комітет статистики України; Ін-т демографії та соц. досліджень НАН України, 2004. – 558 с.

10. A Terminology for Talking about Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.32 [Electronic resource] / A. Pfitzmann, M. Hansen. – 2009. – Access mode: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml.

11. Privacy-Preserving Data Mining: Models and Algorithms / C.C. Aggarwal, P.S. Yu (eds.). – New York: Springer, 2008. – 513 p.

12. Chertov O. Group Anonymity / O. Chertov, D. Tavrov // Hüllermeier, E., Kruse, R. (eds.) IPMU-2010. – Heidelberg: Springer, 2010. – Communications in Computer and Information Science, vol. 81. – P. 592-601.

13. Чертов О.Р. Застосування недиадних вейвлетів для забезпечення анонімності даних / О.Р. Чертов // Інформаційна безпека. — 2010. — № 2 (4). — С. 96—101.

14. Guidelines for the modeling of statistical data and metadata // Conference of European statisticians. Methodological material. – Geneva: United Nations Statistical Commission, 1995. – 30 p.

Поступила в редакцию 10.09.2010

Рецензент: д-р техн. наук, проф., декан факультета прикладной математики И.А. Дичка, Национальный технический университет Украины «Киевский политехнический институт», Киев.

МОДЕЛИ МЕТАДАННЫХ ОПИСАНИЯ СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ

О.Р. Чертов

В статье рассматривается вопросы моделирования метаданных, используемых для описания разнообразной статистической информации, включая демографическую. Предложена расширенная семиотическая модель метаданных. Усовершенствована известная общая модель статистических бизнес-процессов путем введения новых процессов «обезличенности» и «обеспечения конфиденциальности», а также выделения процесса поиска. Планируется, что разработанные модели будут использованы при построении Интегрированной системы обработки статистических данных Украины и автоматизированной системы «Перепись-2012» по обработке переписных данных.

Ключевые слова: семиотическая модель метаданных, модель статистических бизнес-процессов, обезличенность, анонимность данных.

МОДЕЛІ МЕТАДАНИХ ОПИСУ СТАТИСТИЧНОЇ ІНФОРМАЦІЇ

О.Р. Чертов

У статті розглядаються питання моделювання метаданих, котрі використовуються для опису різноманітної статистичної інформації, включаючи демографічну. Впроваджено розширену семіотичну модель метаданих. Удосконалено відому загальну модель статистичних бізнес-процесів шляхом уведення нових процесів «знеособлення» та «забезпечення конфіденційності», а також виокремлення процесу пошуку. Планується, що розроблені моделі будуть використані під час побудови Інтегрованої системи обробки статистичних даних України і автоматизованої системи «Перепис-2012» з обробки переписних даних.

Ключові слова: семіотична модель метаданих, модель статистичних бізнес-процесів, знеособленість, анонімність даних.

Чертов Олег Романович – канд. техн. наук, доцент, доцент кафедры прикладной математики, Национальный технический университет Украины «Киевский политехнический институт», Киев, Украина, e-mail: chertov@i.ua.