INTEGRATION OF STATISTICAL INFORMATION SYSTEMS

-THEORY AND PRACTICE

Bo Sundgren and Erik Malmborg

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1995:7

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1995:7. Integration of statistical information systems – theory and practice / Bo Sundgren, Erik Malmborg.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1995-X101OP9507

INTEGRATION OF STATISTICAL INFORMATION SYSTEMS

-THEORY AND PRACTICE

Bo Sundgren and Erik Malmborg

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1995:7

Från trycket December 1995
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Lars Lyberg

Förfrågningar Bo Sundgren
tel 08-783 41 48
telefax 08-783 45 99

Erik Malmborg
tel 08-783 40 27
telefax 08-783 45 99

ISSN 0283-8680

Printed December, 1995
Producer Statistics Sweden
Publisher Lars Lyberg

Inquiries Bo Sundgren
telephone +46 8 783 41 48
telefax +46 8 783 45 99

Erik Malmborg
telephone +46 8 783 40 27
telefax +46 8 783 45 99

ISSN 0283-8680

Integration of statistical information systems - theory and practice

Erik Malmborg and Bo Sundgren

Statistics Sweden
S-115 81 STOCKHOLM

Abstract

A theoretical framework for integration of statistical information systems is presented. This framework is compared with practically oriented work to create a European "Distributed Statistical Information system". The development of the "GESMES" EDI-message is part of this work. The structure of GESMES is presented from both semantic and syntactic perspectives.

1 Integration needs

There are growing demands for more advanced integration of statistical information systems, inside statistical organizations (like national statistical offices and international statistical agencies) as well as between such organizations. To a great extent, the demands for integration have their origin in the needs of the users of statistical information. The statistics users have tasks, which imply needs to

- describe and understand a certain "reality" - the so-called **object system of interest**; and/or
- plan, implement, monitor, and evaluate decisions and actions vis-à-vis the object system of interest.

The particular object system, which is of interest to a particular user, or group of users, does not always coincide with the object system of any one particular statistical information system. Instead the statistics user will often have to combine statistical information from several statistical information systems. In such situations it is highly desirable that the different statistical information systems are well integrated with one another.

2 Integration approaches

Integration of statistical information systems can be accomplished in different ways. The traditional way is **physical integration**, whereby the data and processes of several more or less autonomous statistical information systems are brought together into one physically (and logically) integrated system, which is under the **full centralized control** of one management.

Physical integration of statistical information systems is associated with many problems of physical, logical, and organizational nature. Among other things physical integration will easily lead to a lot of duplication and redundancy, which in turn will cause complex and error-prone update procedures. For example, suppose that there are three (groups of) users, and that each one of the users needs statistical information from a subset of a set of 15 different statistical information systems, as illustrated by figure 1. If we create three physically integrated information systems, corresponding to the needs of the three users, respectively, the 15 original systems will be duplicated as shown in table 1.

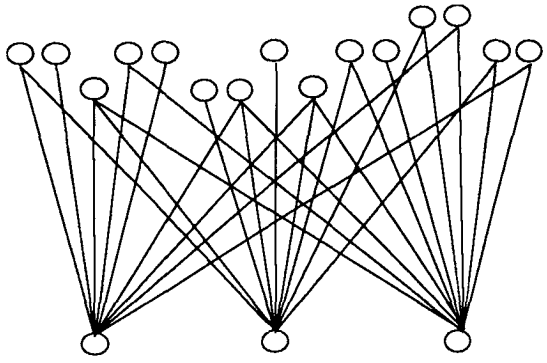

Figure 1. *One way of organizing the interaction between two sets of systems.*

An alternative to physical integration is some kind of **soft integration**, where the original information systems remain as autonomous systems, constrained only by some requirements to be able to exchange information between themselves. The communication requirements in such a system of **loosely coupled systems** may be decided upon globally (as the result of dictates or negotiations) or agreed upon after separate negotiations carried out between each user and the manager of each information system from which the respective user is interested to obtain

statistical information. In its genuine form the latter model will be very complex and resource-consuming. For example, in a real situation corresponding to figure 1 there would have to be 28 different negotiation processes, resulting in 28 individual agreements between a user and the manager of an information system.[2] describes the complexities of "schema integration for heterogenous federated data bases"

System	Number of representations
1	2
2	1
3	3
4	2
5	1
6	1
7	3
8	1
9	3
10	2
11	1
12	2
13	2
14	2
15	2

Table 1. *Quantification of the redundancy in a system of information systems, where the interaction is based upon the model in figure 1.*

Figure 2 illustrates a model for soft integration, which avoids the problems of completely decentralized negotiation processes by introducing a very small amount of centralized control in the form of a **standardized interface for exchange of information**: every user and every information system should be able to communicate with the standardized interface, but every user/producer autonomously determines *how* this requirement should be fulfilled. In the terminology of [14] (and[2]) our approach is somewhere in between "distributed databases" and "federated databases". The semantics of the interchange format gives a certain amount of harmonization, without imposing the constraints of a distributed database architecture.

In the situation illustrated by figure 2, there is a need to design (15 + 3) = 18 communication procedures, each one of which can be autonomously decided upon by a single user/producer without any negotiation and with the only restriction that it should be compatible with the standard interface. This should be compared with the situation in figure 2, where there is a need to design 28 communication procedures, each one of which must be negotiated by a user/producer couple.

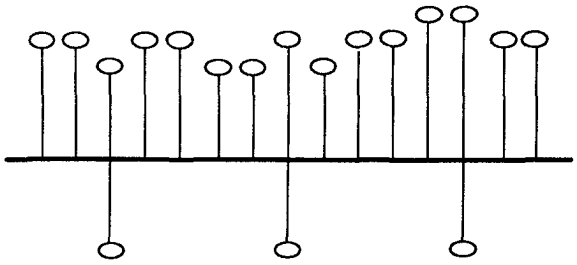

Figure 2. *A model for "soft integration" based upon loosely coupled systems communicating via standarized interfaces.*

More generally, if we assume that there are m users and n producers of information, Communication Model 1 (CM1), corresponding to a generalized version of figure 1, will require in the order of (m x n) communication procedures to be designed and bilaterally agreed upon after the same number of negotiations, whereas Communication Model 2 (CM2), corresponding to a generalized version of figure 2, will require in the order of (m + n) communication procedures to be designed and unilaterally decided upon by each user/producer.

Furthermore, if a new user (producer) is added to the scheme, CM1 will require up to n (m) new communication procedures to be designed and bilaterally negotiated, whereas CM2 will require only 1 (1) new communication procedure to be designed and unilaterally decided upon.

3 Systems of statistical information systems

Integration *between* statistical information systems can be regarded as integration *within* a **system of statistical information systems**, which can again be regarded as another (more complex) statistical information system. Systems of statistical information systems can be (more or less) **open** or (more or less) **closed**.

A statistical office, controlled by one management, could - at least in theory - design its internal systems for production of statistics as a relatively closed system of physically integrated statistical information systems. Most component systems of such a system will be very dependent on the behaviour of other systems. Thus a change in one system may easily trigger of a chain of (necessary) changes in other systems, and the introduction of a new system into the system of systems will often cause complex integration problems.

In a more loosely controlled environment, like an international community of souvereign member states, a closed system of statistical information systems is hardly even theoretically conceivable; an open system of cooperating system is the only practical possibility.

Figure 3 gives an overview of the system of statistical information systems of a statistical office. The individual statistical information systems of such a system of systems belong to some typical categories like

- systems for **primary statistics production** (the traditional type of statistical surveys);
- **instrumental systems** like **registers** consisting of **base registers** and **code registers**;
- systems for **secondary statistics production** like the system of national accounts;
- **retrieval systems** like user-driven **search systems** and **presentation databases**.

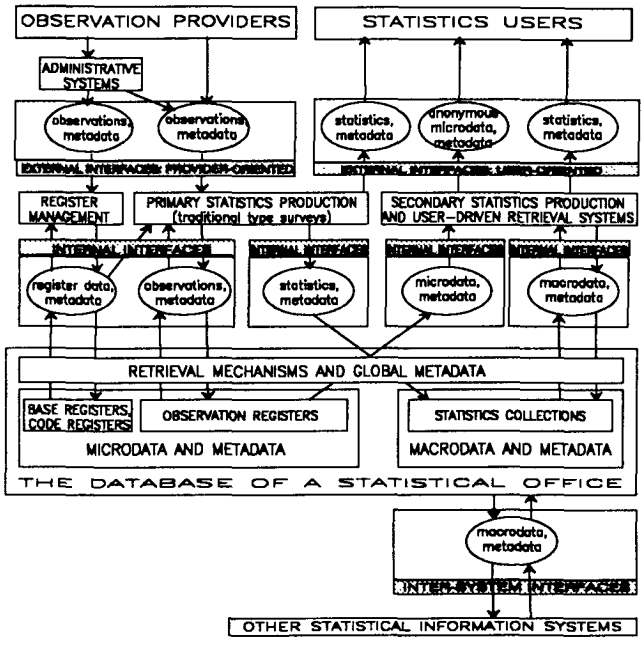

Figure 3. *The statistical information system of a statistical office.*

4 Interface levels

Three major interface levels can be identified (cf fig 3):

Level 1: Interfaces between a statistical information system (or a system of such systems) and external users/producers of statistical information: **external interfaces**.

Level 2: Interfaces between subsystems of statistical information systems: **intra-system interfaces** or **internal interfaces**.

Level 3: Interfaces between statistical information systems (or systems of such systems): **inter-system interfaces** or **interfaces between systems**.

4.1 External interfaces

A statistical information system exchanges information with

(a) **statistics users**;
(b) **input providers** (respondents and/or intermediaries like interviewers or administrative systems).

4.1.1 Input-oriented external interfaces

The input providers provide information about **observations and measurements** of a number of object characteristics (states and events) for a number of individual objects in the object system. An **object characteristic** can be formally represented as an ordered pair

(4.1) $C_O = \langle O, V \rangle$ or, with dot notation, $C_O = O.V$

where O is an object type and V is a variable. Sometimes O will rather be a vector of object types, in which case V will be a relation or a variable that is based upon a relation, e g "quantity (of a commodity) exported (from one country to another country)".

The basic building-blocks of information about observations and measurements of object characteristics are so-called (micro)object level **elementary messages** (e-messages) with the semantical structure

(4.2) $m_O = \langle o_i, p, t \rangle$ or, with an alternative notation, $m_O = [o_i.V(t) = v_j]$

where o_i is an object instance belonging to the object type O, p is a property, typically expressed as a value a_j of a variable V, and t is an instance of time (point or interval) at/during which the object is supposed to have (had) the property p. Alternatively o_i could be a vector of objects, p being a relation (like "married") or a $\langle V, v_j \rangle$ pair, where V is based upon a relation (like in the "export" example above).

In a typical interaction between a statistical information system and an input provider, the latter receives a set of questions, often **hierarchically structured by respondent**. The respondent is sometimes identical with (one of) the object(s) observed. The questions are accompanied by some **metainformation** (explanations, instructions, etc). In some systems additional metainformation may be requested interactively by the respondent, if and when it is needed. When observation messages are returned to the statistical information system, they may be accompanied by other types of metainformation, informing about, say, some exceptional circumstances noted in connection with the observation process.

When the hierarchically structured sets of observation messages enter the statistical information system, they are often - sooner or later - transformed into **flat files** or **relational tables** in accordance with relatively well standardized procedures, supported by many commercial software products (cf form handling tools of relational database management systems). The accompanying metainformation should ideally be systematically taken care of by a parallel process, but this part of the external interface on the input side has not yet reached any degree of standardization.

4.1.2 Output-oriented external interfaces

On the output side the external interface traditionally consists of statistical tables accompanied by metainformation in the form of headings, column and row labels, footnotes, comments, etc. Today electronical equivalents of statistical tables are at least equally important, and such outputs are often the result of interactions, which are initiated by a user, and which involve the processing of metadata provided alternately by the user (search questions etc) and by the statistical information system.

The basic building-blocks of the aggregated statistical information contained in statistical tables are **statistical e-messages** with the semantical structure

(4.3) $m_s = [e(O(t_1).V(t_2).f) = a']$

where

(i) $O(t_1)$ is a **population of objects** existing at/during time t_1;
(ii) $V(t_2)$ is the status of a (vector of) variable(s) at/during time(s) t_2;
(iii) f is an **aggregation function** like count, sum, average, correlation, etc;
(iv) e is an **estimation function**, providing estimates of the true values of the **statistical characteristic** or **statistical concept**

(4.4) $C_s = O.V.f$

Statistical macroinformation is often organized in certain typical structures. For example, statistics users are often interested to obtain estimated values of "the same" statistical characteristic for

- a series of time periods (rather than a signle one) - "**time series data**"; and/or
- a sturctured set of object populations (rather than a single one) - "**cross-sectional data**".

The following format is general enough to cover most structures of statistical metainformation that are demanded by statistics users:

(4.5) $O(t_a)(\underline{with}\ p_a)(\underline{by}\ V_g(t_g)).V_b(t_b).f$

where

(i) $O(t_a)$ is a (series of) **population(s) of objects** existing at/during the time t_a, which, in the case of time series information, is a parameter varying over a certain range of times;

(ii) p_a is a property, the **alfa property**, selecting a subset of $O(t_a)$;

(iii) $V_g(t_g)$ is the status at/during the time t_g of a vector of variables, called the **gamma variables**, which are crossclassifying the population(s) $O(t_a)$; in the most general case t_g will be a vector of time parameters (corresponding to the vector of variables), and each time parameter will vary over a range of values that is matching the range of values of the time parameter t_a;

(iv) $V(t_b)$ is the status at/during the time t_b of a (vector of) variable(s), the so-called **beta variables**; in the most general case t_b will be a vector of time parameters (corresponding to the vector of variables), and each time parameter will vary over a range of values that is matching the range of values of the time parameter t_a;

(v) f is an **aggregation function** like count, sum, average, correlation, etc.

The GESMES format is a proposed standard for representation of statistical macroinformation and accompanying metainformation. "GESMES" stands for "Generic Statistical Message", and the standard proposal is developed by the UN/EDIFACT Message Development Group 6.1. GESMES is described in sections 6-8 of this paper.

Observation registers, containing observed and/or derived microdata, are - beside collections of statistics/macrodata - the other important type of information output from statistical surveys. More and more competent users of statistics demand access to microdata, for their own analyses, in their own computer environments. Statistical offices are responding to such demands by preparing files of **anonymized microdata**, for example so-called **public files**.

An external user who is about to (re)use the microdata in an observation register may not be in a position where he or she has access to the staff in the statistical office, who once (maybe years earlier) produced the data. Thus

the observation register will have to be accompanied by an appropriate documentation, that is, a set of metadata.

4.2 Internal interfaces

Figure 4 illustrates the typical structure of one single statistical information system, corresponding to one statistical survey, in a statistical office. A database-oriented architecture is assumed with three major types of subsystems exhanging data and metadata with a common statistical database. The subsystem types are labeled

- input acquisition;
- aggregation;
- output delivery.

In a database-oriented statistical information system most exchange of data takes place via one or more databases. Thus the most important internal interfaces in such systems are the interfaces between the data base management software and the various software products, which are used for performing the basic functions in a statistical information system (cf figure 5). Today the most widely accepted, relevant standard for this type of interface is the Structured Query Language (SQL).

It should be noted that SQL does not contain a standard for the exchange of semantically oriented metadata accompanying the data. In the future there may be more complete general (and commercially supported) standards for the exchange of data/metadata between databases and application functions, possibly based upon object-oriented data models rather than relational ones.

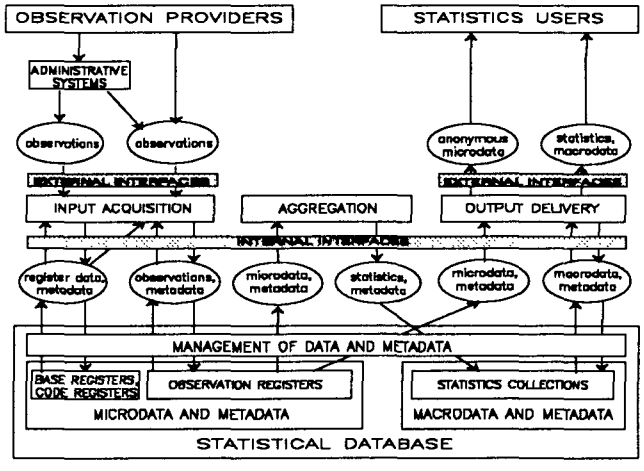

Figure 4. *A model of a self-describing database-oriented statistical information system.*

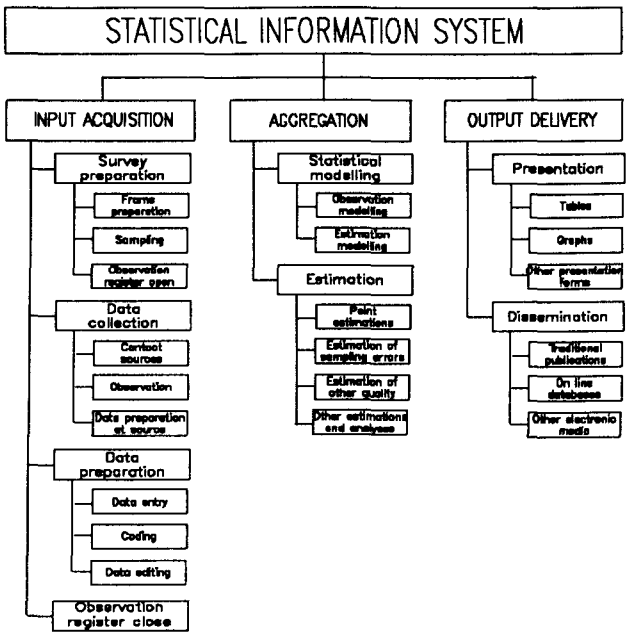

Figure 5. *A functionally oriented model of a statistical information system.*

4.3 Inter-system interfaces

As long as we are within a system of statistical information systems, which is - at least in principle - controlled by one single management, this management has certain possibilities to impose standards for internal properties of the interdependent systems and subsystems as well as for interfaces between them. When we consider systems (like the statistical system of the European countries belonging to the European Communities), which are not controlled by a single management, imposing standards for internal properties will be virtually impossible, and the necessary negotiation processes for reaching agreements on standards for information exchange between the different member systems. Very open system design principles become a necessity in such situations.

As an example we may consider the model of a proposed Distributed Statistical Information System (DSIS) developed in a study initiated by the Commission of the European Communities (EC). A key element in this model is the setting up of a European reference environment, which is illustrated on a conceptual level in figure 6. Each set of three boxes represents a DSIS organization and the production (P), reference (R), and dissemination (D) environments within those organizations. The shaded portions of the reference environments are those accessible across the DSIS network, and these comprise the European reference environment.

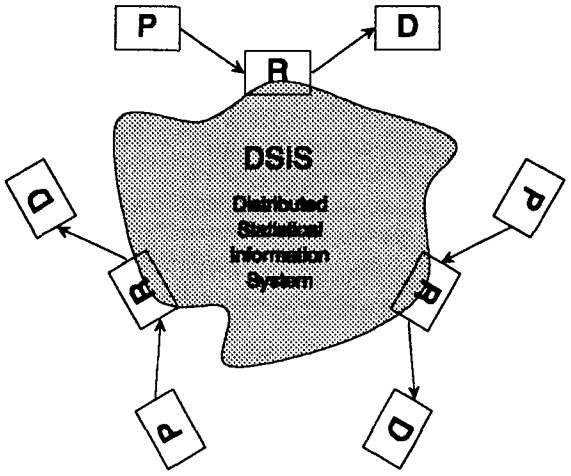

Figure 6. *A loosely coupled system of statistical information system: the European reference environment.*

4.4 Concluding remarks concerning interfaces

For each one of the interface levels discussed in the previous sections standardization processes are underway, aiming at *de facto* and/or formal standards for the exchange of data and metadata within and between statistical information systems and between such systems and external users/producers of statistical data. Some partial standards have already become widely accepted, others exist but are not very well known, and finally, some standardization issues, notably those dealing with metadata, are still in a relatively early stage of development.

5 Requirements on interchange formats

This section will discuss three requirements on interchange formats. The requirements are from the perspective of this paper, i.e. Integration of Statistical Information Systems.

5.1 Interchange formats should be standardized

A more theoretical discussion on this topic can be found in section 2 of the paper. Here we want to stress the practical and economic consequences of neglecting this requirement. If routines are to be developed for electronic data interchange between two parties, there always exist some tailored solution. If one of the parties works towards several other parties, he will have to develop and maintain several solutions over the time. This can be resource-consuming.

In order to handle the problems there is an international standardization process to create standardized interchange formats for statistical data. The chosen basis for this standardization work is the UN EDIFACT framework for Electronic Data Interchange (EDI). This is an international standard for EDI, which is to replace national standards such as the ANSI X.12 standard.

An EDIFACT message for interchange of statistical data has been developed. The name of the message is GESMES, which is an acronym for GEneric Statistical MESsage. The work has been sponsored by Eurostat, i.e. the EC Statistical Office in Luxembourg. More information on this work can be found in [5] and [3]. The specification of GESMES is published by Eurostat in [1]. The published version is for trial use, and is named GESMES-93. A more official "UN EDIFACT Status-1" message can be expected in 1995.

5.2 Interchange formats should be semantically rich

As many readers have experiences from wordprocessing on PCs, we will use this as a starting point. Almost all word processors on PCs have their own "native" formats for storing text. When necessary text can be "exported" into or "imported" from another format. If we are to move a text from wordprocessor A to wordprocessor B we may be lucky so B can import the native format of A, or A can export into the native format of B. If this is not feasible, we have to find a common format for A and B. This might be "DCA/RFT" (an older IBM "standard") or "ODIF" (an ISO standard). In any of these cases we can move the text including **bold** text, *italicized* text, page breakings etc. If we are unlucky we may have to export from A as "ASCII" and then import the text into B. In this case we loose the editing information (bold, italicized ..). We say that DCA/RFT and ODIF are "richer" or "semantically richer" than pure ASCII-text. There is more editing information included in the format.

A similar argument is relevant for statistical information. In this case the ASCII-file corresponds to simple "spread-sheet" formats such as "comma-delimited", "PRN, "WK1" etc. Richer formats exist for time-series (AREMOS TSD ..) and for matrices/tables (e.g. PC-AXIS format, c.f. [5]). GESMES is a very rich format for time series, matrices and statistical tables.

As a general rule it is always possible to transform from a richer into a simpler format. If you want to go in the other direction you must add information. This may be done by filling in forms in a translation system. An example might be loading data into PC-AXIS or LOTUS IMPROV from an ordinary spreadsheet. You must interactively add information about the spreadsheet data ("meta-data"). The benefits gained are new possibilities for analyzing and presenting the data.

5.3 Interchange formats should be developed to facilitate the building of cooperative databases

A typical competent user of statistical data will use data from several sources, and sometimes his own data as well. If the user is putting several different sets of data together to compile statistics, we can consider him to be building a database. Interchange formats can be seen as a link between the data sources. This can be seen as a simple form of cooperating data bases.

A much more sophisticated situation with cooperating data bases is if the user has e.g. a PC, which whenever necessary will automatically phone different on-line data bases to fetch data. If this is to function, the PC must be aware of which information is available at the different sources. We don't have to assume this knowledge to be complete initially. It can be built gradually in dialog with the different data bases. This situation puts new demands on interchange formats:

- It should be possible to separate data and metadata (e.g. information about available data).
- We should be able to send "queries" based on the metadata locally available.

An example of an ambitious project with cooperating databases is the above-mentioned (section 4) DSIS. This EC-framework is to interconnect databases at European Central Statistical Offices (CSOs). The planned interchange format is GESMES.

6 Semantics of interchange formats

Semantic modelling is recognized as an important area of data base research. Also for interchange formats there is a need to use semantic models. In many cases the same semantically oriented modelling languages can be used both for data bases and for interchange formats. [4] and [5] elaborates on semantic modelling for statistical data and for statistical interchange formats. In this paper we will use a simple Entity Relationship (ER) type model to explain the semantics behind the GESMES format. In [1] a full ER-model used in the design of GESMES is presented. That model is considerably more complex than the model used in this paper, but the terminology between the models is consistent. Hopefully the model presented here can be used as a tool for understanding the more complex model. The presentation of the ER-model in this section and of GESMES in the next section will use a simple statistical cross-tabulation as an example (Figure 8).

This example is identical with an example used in [5]. One reason for this choice is that the GESMES-example presented in [5] has been outdated by later developments of GESMES and should be replaced.

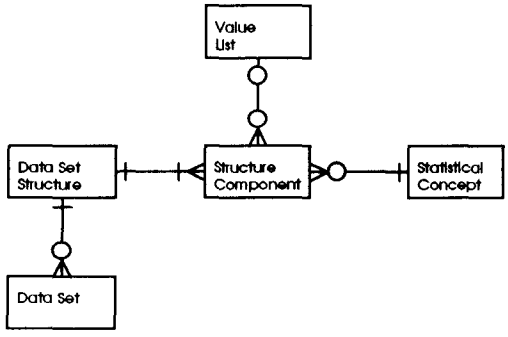

Figure 7. *An ER-type model of GESMES*

The different entities in the model will be explained using the example:

- **Data Set** represents a statistical table, matrix or set of time series to be transmitted using GESMES. Several data sets can be transmitted in one message. Our table in the example above is considered to be the one data set to be transmitted.

- **Data Set Structure** represents a description of the structure for a data set. The data set structure can be described once and applied to several data sets in a transmission.

- **Structure Component** represents the atomic parts of the data set structure. In our example we have four structure components. The most natural representation is to consider the table as having three "dimensions" County, marital status, and sex. The 4th structure component is population, which is "cell data" in GESMES terminology.

  If we compare this terminology with the framework of section 4 we find that "dimension" corresponds to "gamma variable". "Cell data" corresponds to "beta variable" (or to be more exact to, $V_b(t_b).f$ in (4.5)). The GESMES concept of "scope" corresponds to "alfa properties".

  Structure Component corresponds to "attribute" in

Population 1985 by county, marital status, and sex (1000s)

	unmarried		married		widow/widower		divorced	
	male	female	male	female	male	female	male	female
Stockholm	393	363	291	292	17	75	62	84
Uppsala	65	58	49	49	3	11	7	9

Figure 8. *Example of a statistical table*

[7]. In that framework we would consider our table to have three "category attributes" and one "summary attribute".

- **Statistical Concept** is a rather loose concept in GESMES. It is basically a placeholder for definitions. If we have a table with trade between different countries country would be one Statistical Concept, but there would be two Structure Components representing import country and export country.

- **Value List** represents lists of values. In our example each dimension has a Value List. County has the values (Stockholm, Uppsala), marital status has the values (unmarried, married, widow/widower, divorced), and sex the values (male, female). Several Structure Components can use the same Value List

7 Functionality of GESMES

GESMES has been designed to allow for a broad spectrum of different use patterns. The EDIFACT inheritance gives several practical benefits. The X.400 standard for communication (specifically X.435) and EDIFACT combine to give a practical infrastructure where it is possible for large organisations to internally send messages to the correct receiver (whether human or software system). The handling of value lists is a good example to illustrate the flexibility of GESMES:

- One extreme is not use a value list for a specific dimension. In this case the values are sent explicitly with the data. The values are "associated" with the data. If all dimensions are associated each cell value is accompanied by explicit values for all dimensions. In our example each of the 16 cell-values should be accompanied by the corresponding value for all 3 dimensions as illustrated in the tabular structure below:

Stockholm	unmarried	male	393
Stockholm	unmarried	female	362
Stockholm	married	male	291
.....			

- Code lists are supported. When sending a value list both code-values and textual descriptions can be transmitted.

- Typically the dimensional data is factored out and sent as value lists separate from the cell data. This "non-associated" way to handle dimensional data corresponds to matrix storage of statistical data. The GESMES example in the next section handles all 3 dimensions as non-associated.

- If there is a regular exchange of similar data the same value list might be needed several times. This can then be sent once and stored in a local (meta-) database with the reciever. It is then given an identifier according to EDIFACT-standard. When using the value list in a later transmission only this identifier is sent. In this way large classifications (as e.g. the HS classification for goods) can be sent once only.

- The value list for a non-associated dimension can be a subset of a large classification. In this case a value list with only the codes for the subset need to be sent with the data. The codes and texts of the full classification can be sent in the same or in another message.

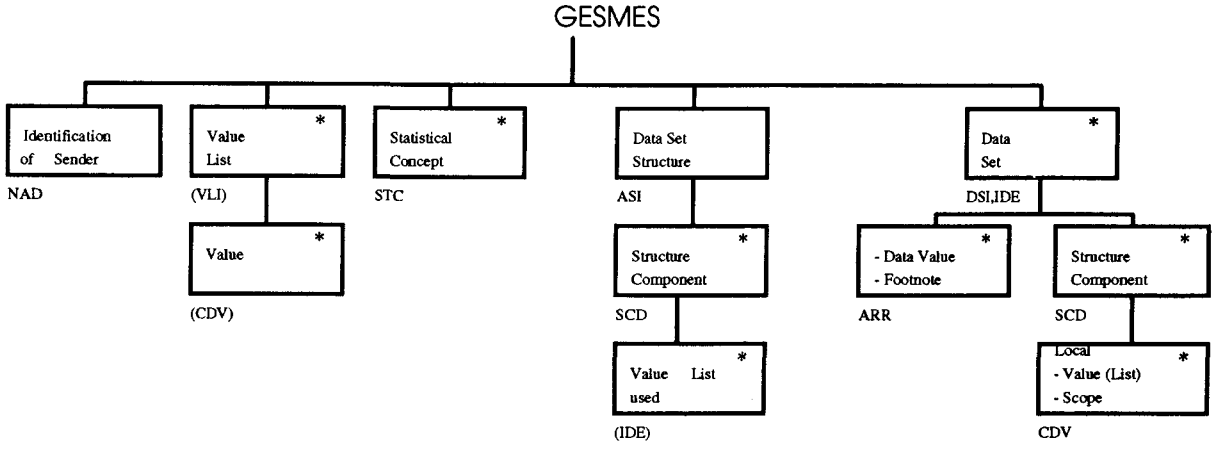

Figure 9. *Structure of GESMES*

- The value list can describe the time dimension for a set of time series using special facilities. The need for such facilities was discussed in [5].

Fig 9 is a tree-structured diagram showing the different parts of a GESMES-message. Note that value lists can be sent at 2 different places in the structure. The first place is used for the transfer of e.g. a code list to be stored in the recievers metadata base. If a value list is used for several data sets in the same message it is also sent here. In this case the value list is given an identifier that is local to the message. The second place is used for non-associated value lists local to a data set. The * in the graph indicates multiplicity. Under each box the corresponding segment identifier(s) are given as a help for the study of the example in section 8.

The graph only gives an overview of the structure. In addition there is e.g. an elaborate mechanism to support footnotes for different components. Such facilities for giving "quality information" are very important in practical use

8 An example of GESMES

An EDIFACT-message is composed of segments. Each segment is identified by a three-letter identifier. Some of the segments are standard segments used in many different EDIFACT messages, some are specific for GESMES. All EDIFACT segments, data elements and codelists are parts of the globally maintained EDIFACT directories.

The syntax rules for EDIFACT messages are given by the ISO-standard ISO9735. We don't argue for GESMES to be especially readable! The corresponding PC-AXIS file can be found in [5], and certainly is more readable. In practice this is not a problem. The EDIFACT message is either interpreted by the receiving software system or translated by a table-driven EDIFACT translator.

```
UNH+001+GESMES:0:27:M6'
BGM+:::SSDBM EXAMPLE'
NAD+MS+SWEDSTAT'
STC+COUNTY'
FTX+Z08+++Standardized regional
  areas'
STC+MARITAL STATUS'
FTX+Z08+++Marital status'
STC+SEX'
STC+POPULATION'
ASI+SIMPLE.DEMOTABLE'
SCD+Z01+COUNTY++1'
SCD+Z01+MARITAL STATUS++2'
SCD+Z01+SEX++3'
SCD+Z02+POPULATION++4'
DSI+POPULATION EXAMPLE'
FTX+Z10+++Population 1985 by country,
  marital status, and sex (1000s)'
ARR++393:363:291:292:17:75:62:84:65:5
  8:49:49:3:11:7:9'
IDE+Z08+SIMPLE.DEMOTABLE'
SCD+Z01+++1+Z02'
CDV+Stockholm'
CDV+Uppsala'
SCD+Z01+++2+Z02'
CDV+unmarried'
CDV+married'
CDV+widow/widower'
CDV+divorced'
SCD+Z01+++3+Z02'
CDV+male'
CDV+female'
UNT+31+001'
```

The example is somewhat simplified. In the valuelists each CDV ought to be a CDV, FTX combination. In order to help the reader better understand the example, the names behind the 3-letter acronymes for the different segments is given below. In order to fully "decode" the example there is a need for the full structure of all the segments, and all the code-lists for different data elements. These can be found in [1].

```
UNH   Message  header
BGM   Beginning of message
NAD   Name and address
STC   Statistical concept
FTX   Free text
ASI   Array structure identification
SCD   Structure component definition
DSI   Data set identification
ARR   Array information
IDE   Identity
CDV   Code value definition
UNT   Message trailer
```

9 Concluding remarks

In this paper we have presented a theoretical framework for integration of statistical information systems. The framework is based on Research and Development within Statistics Sweden (c.f. [8]-[13] in the references). This framework is related to the more practically oriented work to create a European "Distributed Statistical Information System". The development of the GESMES EDI-message is part of this effort. It is important to notice that the approach presented could be applied on different scales:

- GESMES is to become an international standard. Most development has been made in Europe, but UN/EDIFACT is an international process for standardization of messages (i.e. interchange formats).
- The use on the European level has been discussed above (DSIS)

- In Sweden the production of official statistics will become more decentralized. This is a political decision to introduce "market economy" in statistical production. Statistics Sweden will have the responsibility to compile central data bases based on the different production systems. Some of these production systems will be inside Statistics Sweden, some of them outside of the organisation. The framework presented in this paper will be used and GESMES will probably be used as an interchange format.

Finally we want to give some remarks on the usability of GESMES:

- GESMES supports makro-data, especially cross-sectional and time series data. There is no explicit support for microdata. It is possible to use GESMES also for this case, but is probably not the best solution.
- GESMES has no special support for Geographical Information Systems (GIS) and coordinate data. This might be good idea for an extension to GESMES.
- GESMES is suitable for all forms of multi-dimensional data. Measurement data from scientific experiments is often in this category (c.f. [7]).

The theoretical framework presented has not been the formal base for the development of GESMES. The working group behind GESMES (EDIFACT Message Design group 6.1) has created its own semantic concepts as illustrated in section 6. Of course it had been benificial if the work had been based on one chosen established framework. The backgrounds (theoretical and practical) of the participants in the group were too different for this to happen. One of the aims of our paper is to relate the rather "pragmatic" conceptual framework of GESMES to a more theoretical framework.

References

[1] Eurostat
GESMES 93 Guidance to Users & Reference Guide (separate volumes)
Eurostat, Luxembourg, 1993

[2] Johannesson, P.
Schema Integration, Schema Translation, and Interoperability in Federated Information Systems
University of Stockholm, 1993 (Doctoral Thesis)

[3] Lebaube, P.
EDI and Statistics - A Challenge for statisticians.
Proc 48th Session of the International Statistical Institute, Cairo, September 9-17, 1991.

[4] Malmborg, E.
On the Semantics of Aggregated Data
Proc. Third Int. Workshop on Statistical and Scientific Database Management, Luxembourg 1986 (Published by Eurostat)

[5] Malmborg, E.
Matrix-based Interchange of Aggregated Statistical Data
Proc. Sixth International Working Conference on Scientific and Statistical Database Management, Ascona, Switzerland 1992 (published by ETH, Zurich)

[6] Malmborg,E., Lisagor, L.
Implementing a Statistical Meta-Information System.
In Eurostat Conference on Statistical Meta Information, Luxembourg, 2-4 Feb 93. Also in Statistical Journal of the United Nations UN/ECE 2/1993.

[7] Shoshani, A.
Statistical Databases : Characteristics, Problems and some Solutions
Proc 8th Int Conf on Very Large Data Bases, 1982

[8] Sundgren, B.
An Infological Approach to Data Bases
Statistics Sweden, Urval Nr 7, 1973 (Doctoral Thesis)

[9] Sundgren, B.
Statistical Metainformation and Metainformation Systems.
Statistics Sweden R&D Report 1991:11. Also in Statistical Journal of the UN/ECE 2/1992..

[10] Sundgren, B.
What metainformation should accompany statistical macrodata?
Statistics Sweden R&D Report 1991:9

[11] Sundgren, B.
Organizing the Metainformation Systems of a Statistical Office.
Statistics Sweden R&D Report 1992:10, also as wp at ECE Work session on Statistical Metadata 1992 (METIS)

[12] Sundgren, B.
Statistical Metainformation Systems - pragmatics, semantics, syntactics.
In Eurostat Conference on Statistical Meta Information Systems, Luxembourg, 2-4 Feb 1993. Also in Statistical Journal of the UN/ECE 2/1993.

[13] Sundgren, B.
Guidelines on the Design and Implementation of Statistical Metainformation Systems.
Statistics Sweden R&D Report 1993:4. ECE Work session on Statistical Metadata nov 1993

[14] Özsu, M.T., Valduriez, P.
Principles of Distributed Database Systems
Prentice-Hall, 1990