

On the structuring of statistical information

Abstract

*The starting point of the presentation is to examine the character of statistical information, the way in which data and metadata are interconnected in statistical information and how the entity formed by them can be modelled. The basic model to be created for statistical information is an **informative table** within which framework metadata and data are combined conceptually as one complete entity.*

***Structuring** of the combination of statistical metadata and data, the informative table, creates possibilities for the development of IT applications for production and dissemination based on structured statistical information. In this sense included are document definitions that are intended for testing of solutions based on the standard technologies of structured information (SGML/XML).*

1. Introduction

Barring a few exceptions, the process of producing statistics has, as a rule, been depicted during the past couple of decades as a production chain¹ stretching from the collection to the dissemination of data. The “standard” components of the chain are the data collection phase, the production phase and the dissemination phase (see Figure 1).

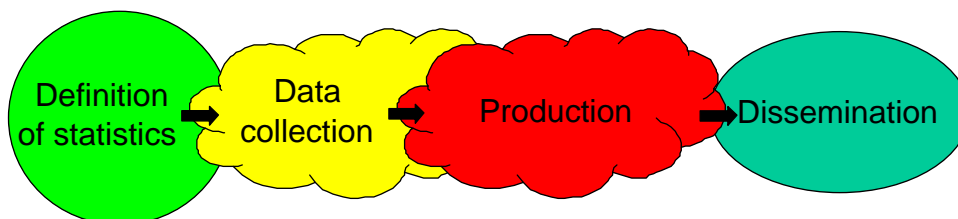


Figure 1. Statistical production process

A part of the statistical production process that is significant from the point of statistical information is left outside the description and, thus also, outside the examination as a whole. The part concerned is the phase during which the topic the statistics depict and the contents of the statistics are defined. This phase is essential because it produces most of the metadata for statistics. The metadata produced during this phase are utilised in a number of contexts during the other stages of production (see Appendix 1 Data and Metadata in Statistics Production). When statistics are published, some of the metadata are attached to them, either as direct quotes or annotations. Through these the user of the statistics receives information on e.g. the

¹ Examples of the few exceptions include Statistics Sweden's metadata system descriptions, see cf. Sundgren "The Swedish Statistical Metadata System, Workshop on Statistical Metadata, Working Paper No. 1.6, Eurostat, Luxembourg 2000.

used international classification standards and their pertinent specific national application rules.

This “trimming” of the statistical production process has had certain aggravating, one could almost say fatal, consequences. The “overlooking” of the first, genuine phase of the production process has meant that no scope has been given for metadata within statistical production systems themselves. Different procedures for processing metadata and actual data have been adopted in these systems and combining them with the available technology has proven problematic and laborious as far as the design, implementation and maintenance of the systems, as well as the required statistical work are concerned. Additional work is required in adjusting the metadata into system-compatible contents and formats and in re-entering the metadata more or less manually into different systems serving dissemination.

The motive for the present study was the desire to eliminate the productional disparity between metadata and actual data, and its side effects. In an effort to do this, statistical information must be examined as a complete entity. Answers must be sought to questions such as what is meant by statistical information, does it have a certain structure or form and, if so, can they be universally defined? In seeking these answers one has to bear in mind that the principles of statistical production stem from the empirical research tradition and from the statistical science, not from data processing and its inherent, urgent questions.

2. Statistical information

Statistical information refers to empirical data, collected from a specific subject for a specific purpose on variables characterised as measurable in one way or other.

The traditional way of presenting the measurement results is the observation matrix (see Figure 2).

		Variable					
		x_1	x_2	...	x_j	...	x_p
Statistical unit	a_1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}

	a_i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}

.	a_n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

Figure 2. An observation matrix

In an observation matrix, the vectors of the attributes of characteristics (list of variables, names of variables $x_1...x_p$) and the observation identification codes ($a_1...a_n$), as well as the name of the matrix constitute its external elements. Information that is external to the matrix but is necessary for its interpretation constitutes the metadata of the matrix. In addition to the above-mentioned external elements, the metadata also comprise e.g. descriptions of data collection methods and times, definitions of used concepts, and descriptions of used measurement methods and operational equivalents of the used concepts. The metadata of a matrix are usually presented as descriptions in

the context of survey or similar reports or, alternatively, as separate documents describing the data and their processing.

In an observation matrix, the identification of an observation or data can be based on the order or on the location of the data. Using this kind of identification method, the data in an observation matrix could in the past be transferred to a punchcard row by row or, later on, entered as observation-specific records into a file or database.

To simplify the identification of observations or data, the information contained in a matrix can be presented in the form of a table (see Table 1).

Table 1. The observation matrix as a table.

Statistical unit identifier	Variable space			
	Characteristic 1	Characteristic 2	...	Characteristic n
Xxx	X_{11}	X_{12}	...	X_{1p}
yyy	X_{21}
.....
.....
.....
.....
.....
.....
.....
zzz	X_{n1}	X_{np}

The table contains the same microdata as the observation matrix does. The additional metadata of an observation matrix that are included in its table format presentation are the heading, the variable vector data of the matrix as table column headings, and the unit identification data of the matrix as table row headings.

However, basing on their historical times of origin, the relationship between a matrix and its table format presentation could also be characterised as reversed: the matrix is the data part of a table, from which the parts containing metadata have been removed in order to simplify the numerical processing of the data.

The logical similarities and dissimilarities between a matrix and its table format presentation are, however, more important than the likely historical relationships between them:

- Both may contain microdata and this property does not constitute a distinguishing factor between a table and an observation matrix.
- By virtue of its logical structure, a table may be perceived as a matrix in which the identification data of a variable form the first row and the identification data of the observation form the first column. The distinguishing factor then would be that the variable names have been entered into the original observation matrix as the first element group and the observation identifier data have been attached as the first column, whereby the row and column headings can be used for the locating and identifying of data.
- If we leave the heading of a table outside the examination, then a table represents a special case of a matrix also in the sense that its first row and column contain the essential matrix metadata with which the data content of the rest of the table are interpreted. In this sense, a table is a structured matrix in which data and metadata appear together.

The foregoing represents an effort to describe succinctly the structure of a table, the main parts of this structure, and their meaning and feasible uses. The described idea about the parts that belong to a table (heading, column and row headings, data) is also

commonly applied in the production of statistics². However, the intention of tables produced for statistical purposes has not been to describe statistical data as such, but to present the ratio of one characteristic to one or more other characteristics the data contain as a result of a numerical analysis. The result, mostly obtained by cross-tabulation, is presented so that the characteristic whose impact on the frequency of other characteristics is to be described is placed into the table as the row dimension in place of the characteristic identifier data, and the remaining other characteristics are, correspondingly, presented as columns of the table. The data part of the table is the parameter describing the frequency of the characteristics.

Once a table contains a heading and, as its column and row headings, the metadata that are necessary for the interpretation of the empirical data it describes, it does not yet mean that it contains all the metadata of its empirical data to the extent that would be necessary for its interpretation. Considerable amounts of metadata that are necessary from the empirical research frame perspective are left out of the table. The metadata not included in the table comprise e.g. description of the research subject, definitions of the used concepts, and descriptions of their operational equivalents and the used measures.

3. The informative table

So that all the metadata that are necessary for data interpretation could be presented within the framework of a table, the structure of the table must be enhanced. The expanded table format is here called the informative table.

An informative table is defined as a table that contains comprehensively all the metadata that are necessary for the interpretation of the data it contains. The reference data system is used as a means of including metadata in a table. How the metadata that are necessary for the interpretation of a table are attached to a table, thereby expanding it into an informative one, is illustrated by the adjacent table (Table 2).

Table 2. Informative table structure.					
Explanation of row headings	Set of column headings which may comprise a number of levels				
	Division heading A		Division heading B		Heading C
	Column 1	Column 2	Column 1	Column 2	Column
Row heading 1					
Row heading 2					
Row heading 3					
Subtotal row					
Row heading n					
Total row					
¹⁾ Footnote containing annotation data ^{a)} Metadata reference containing more incisive metadata Used symbols and their explanations:					

In the reference data system the reference data included in a table fall into three categories (see Table 2):

- 1) Footnote,
- 2) Metadata reference, and
- 3) Explanations of used symbols.

² A similar analysis of table parts also appears as the specification of a table in recent projects relating to reviews of statistical data dissemination, see e.g. the Statistics Open Source (SOS) Project, Erik van Bracht, Statistics Open Source (SOS), 2nd Technical Meeting – Cube Object Concepts, Statistics Netherlands 2000 – <http://neon.vb.nl/sos> cubes.

The nature and contents of all reference data vary depending on the part of the table they are attached to. The basic structure of the table guides the locating and presenting of the reference data as follows:

- 1) Footnote data can only be attached to a table cell located in the numerical part of the table (shaded area in Table 2). Footnote data are annotation type data about e.g. the used calculation method where it differs from the one used for the rest of the data presented in the table.

Footnote data are most commonly needed in diverse compilation tables in which data from different sources are combined. In normal statistics production, the need for annotations concerning individual figures rarely arises.

- 2) The metadata of a table are attached into different parts of the table as metadata references. Metadata can only be attached to the heading data of a table (coloured areas in Table 2).

The metadata attached to a table heading can be, for example, the following:

- Description concerning the specification of the topic the statistics depict,
- Quality description of the statistics,
- Descriptions of used statistical (estimation, etc.) methods.

The metadata attached to column and row headings can be, for example, the following:

- Definition of a variable concept,
- Description of the operational equivalent of a concept, containing e.g. forming or calculating rules,
- Used classification,
- Or other similar variable-specific metadata.

Therefore, the metadata in a table represent cell-specific reference data of heading data and they cannot be attached to cells in the numerical part of the table.

- 3) Explanations of the used symbols represent a standard part of a table and need not necessarily be connected with the use of an actual specific symbol, although they can also be made into symbol-specific reference data. The number of symbols used in today's statistical tables is small and the symbols are fairly standardised.

The (so far) logical structure of the defined informative table has a place for all the statistical metadata that are produced during the processing of statistical data today. Data concerning the guidelines and monitoring of statistics production do not represent statistical metadata and need not, therefore, be presented together with statistical data. On the whole one can assume that the guidance and monitoring data of the production process are, and will also remain in the near future, insignificant as far as interpreting statistical data is concerned.

3.1. Pointing single data items in a structured table

The problems in using single data items of a table have been associated both with locating them in a table - or in table format data - and with describing their meaning.

In a structured table, data are located on the basis of plain language column and row headings as sections of the columns and rows they define (see Table 3 in which the sought data have, by way of an example, been defined as the section of Column 2 of Division heading A and Row heading 3). In a larger group of tables, the identification of data for the right table is done on the basis of the table's heading data. The co-

ordinators of the data are the table's heading data, the column heading and the row heading.

Table 3. Pointing single data items in a structured table.					
Explanation of row headings	Set of column headings which may comprise a number of levels				
	Division heading A		Division heading B		Heading C
	Column 1	Column 2	Column 1	Column 2	Column
Row heading 1					
Row heading 2					
Row heading 3					
Subtotal row					
Row heading n					
Total row					
¹⁾ Footnote containing annotation data ^{a)} Metadata reference containing more incisive metadata Used symbols and their explanations:					

A data item identified in the described manner can be moved from an informative table to other contexts without loss of the information that are necessary for its interpretation. At the same time as the table's heading and column and row headings act as data co-ordinators, they also indicate the metadata that are necessary for the interpretation of the data. The metadata thus indicated can be picked out simultaneously with numerical data for secondary use.

The "data point" picked out from an informative table is by its logical structure also a table that contains all the metadata required for the interpretation of the numerical data it contains (see Table 4). It can be called informative element data, or an informative data point, and it can be processed with methods following similar logic as those used for processing structured tables.

Table 4. Data point picked out from an informative table.

Explanation of row headings	Set of column headings which may comprise a number of levels	
	Division heading A	
	Column 2	
Row heading 3		
¹⁾ Footnote containing annotation data ^{a)} Metadata reference containing more incisive metadata Used symbols and their explanations:		

3.2 Multidimensional data

By their very nature, statistical data are multidimensional as they are. This is because of the repetitiveness of their collection. The statistical reference period forms a third dimensions in all statistical data. Relative to the statistical reference period, statistical data repeat themselves in more or less the same form. In simplified terms, one could say that relative to *time*, statistical data form a three-dimensional observation matrix (see Figure 3).

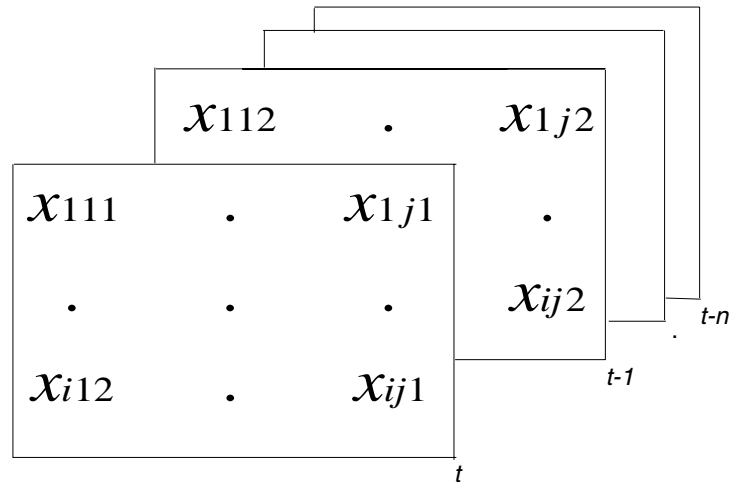


Figure 3. Multidimensional observation matrix

The processing of statistical data from different reference periods as observation matrices would require absolute analogy from the statistical data. In terms of their data contents the statistics should be unchanged to the extent that no variation whatsoever would be allowed even between the statistical units in different sets of data. Contrastingly, when the observation matrices of different statistical reference periods are converted to structured informative tables, i.e. when structured informative tables are produced from them, the requirement of absolute analogy no longer applies. Then the time dimension becomes just one part of the metadata pertaining to the data concerned.

The right table temporally is picket out from the informative structured tables produced from statistics covering different reference periods on the basis of the information in its heading concerning time. After this, the data are located in the previously described manner on the basis of the column and row data.

The metadata of an informative structured table not only help the locating of data but also the interpretation of their time dimension. When data from different time periods are selected, or when they are otherwise used, the metadata make it possible to control their temporal comparability. Correct interpretation of data from each time period can be confirmed with the help of the metadata. The data are interpreted on the basis of the column and row data and the reference data attached to them. If changes are observed in the content of the data, they can be allowed for in comparison conclusions. Another alternative would be to perform a content adaptation at the data set level, with which the changed data content could be rendered temporally as comparative as possible.

In addition to time, the other usual reason for the multidimensionality of statistical data is their **regionality**. Statistical data can be geographically tied in many ways.

At its simplest, the geographic information contained in statistical data may represent just one type of character data among others. In this kind of a case the geographic characteristic does not appear as a kind dimension that would render the statistical data multidimensional and the treatment of the geographic information does not differ from the treatment of other information in the data relating to characteristic.

On the other hand, statistical data can be rearranged on the basis of their geographic characteristic information so that an observation matrix is formed for each geographic unit. Rearrangement of the data is conditional to the inclusion of data on the geographic units in the geographic characteristic information. The result is a multidimensional

matrix like the one in Figure 3, in which geographic codes replace the time dimension. The treatment of data that are multidimensional with regard to the geographic dimension is identical to that of data that are multidimensional with regard to the time dimension, as the time dimension is replaced by the dimension formed by the geographic units. Geographic multidimensionality thus implemented in the informative structured form requires no processing logic that deviates from the one described above. Its processing logic follows the one for multidimensionality presented in the context of Figure 3 above.

Statistical data can also be arranged differently from above relative to their geographic information. An alternative way of arranging geographic information would be to change the statistical data containing the information on the geographic characteristic into geographic unit-specific data. The characteristics data of statistical data can be compacted on the basis of the geographic characteristic information into geographic unit-specific aggregates or into other geographic unit-specific parameters. The adaptation can only be done provided the geographic characteristic information contains data on the geographic unit.

In reality the adaptation produces a new statistical data set in which the statistical units are the geographic units that were used as geographic information in the adaptation. In fact, the situation is similar to when the data are collected direct from the geographic units. Geographic unit-based statistical data can be presented in the table form as follows (Table 5):

Table 5. Geographic data in the structured table form.

Geographic unit identifier	Variable space				
	Variable group A		Variable group B		Group C
	Column 1	Column 2	Column 3	Column 4	Column 5
Xxx					
Yyy					
...					
...					
...					
Zzz					
¹⁾ Footnote containing annotation data ^{a)} Metadata reference containing more incisive metadata					

The metadata of geographic unit-specific data mostly stem from the original statistical data. The metadata of the data on characteristics are transferred from the original data together with the data on characteristics. Apart from the metadata relating to the collection and processing of the original data, the general data-specific metadata also comprise the metadata relating to the forming of area-specific data. The metadata can also be supplemented with metadata concerning the geographic units.

By nature the metadata related to geographic units are data depicting the interrelationships between the units, e.g. information concerning whether the units form together larger geographic entities and, if so, what these entities are. If such relationships between geographic units can be presented in the form of a classification, the data in question can also be placed into the material as separate characteristic data of their own. In Table 5, Column 1 of Variable group A depicts a situation where a geographic unit belongs to a larger geographic unit. A special feature of data thus supplemented is that they are aggregable relative to the characteristic in question.

The aggregability of data is an additional characteristic that does not influence their processing as informative structured table data. Conversely, the aggregability of data on a specific characteristic is an additional feature that should be taken into account as a factor that automates the calculation and processing routines of applications using the data.

Complex, *multidimensional tabulation arrangements* represent the third basis on which statistical data may be multidimensional. When complex tabulation arrangements are used, multidimensional data arrays are produced as interim results (as an example, see Figure 4).

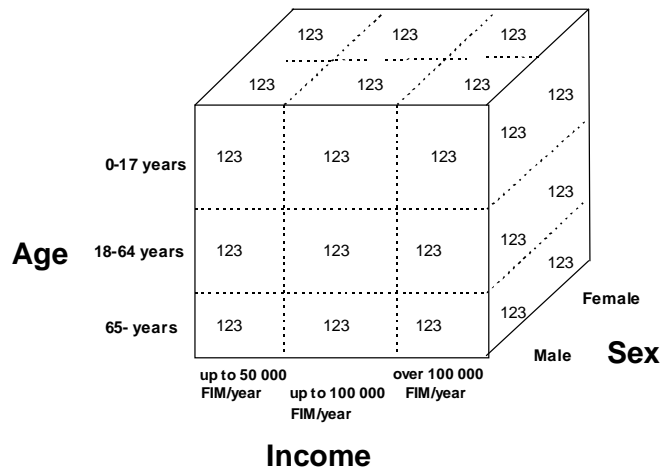


Figure 4. Three-dimensional data array³

In an array, multidimensional data are understood to refer to a numerical value obtained by examining a subject simultaneously in respect of more than two characteristics. In fact, an array represents a two-dimensional table description and not true multidimensionality in the same sense as the time and geographic dimensions were discussed above. The table description on which an array is based originates in a situation where several of the variables classifying the data are used simultaneously in forming a table. In the table form the results are simply presented with the help of nested variables.

4. Benefits of an informative table structure

I have examined above the applicability of the informative structured table concept to statistical data in certain special cases appertaining to statistical information. The question now arises of whether other practical benefits could also be gained from the use of the informative table structure in the production and dissemination of statistical data. The main benefit that can be obtained with it is the integration of data and metadata. The benefits from data and metadata integration discussed below can, unfortunately, only be fully enjoyed in electronic production and dissemination of statistics.

The advantages achieved at the use phase of statistics relate to changes in the management and usages of data. These can be summarised as follows:

- When statistical data are disseminated and used in the electronic form, the informative table creates a foundation for a new use situation in which the metadata associated with the interpretation of the data are readily available. The data available at the use situation also cover the concept definitions and descriptions of the used measures.

³ The visualisation method has been used by Mikko Kurki-Suonio: Modernized user interface for Statistics Finland's online database. Helsinki University of Technology, Department of Computer Science and Engineering, Laboratory of Information Processing Science. Helsinki 1998.

The situation differs considerably from the present status quo where the metadata influencing the interpretation of statistics are published as separate reports, explanatory notes or handbooks. Separation of data and metadata and poor accessibility of metadata are, in fact, the main factors limiting increased professionalism in the use of statistical figures and general statistical literacy.

- When the metadata guiding the interpretation of statistical data are freely available at the stage when statistical data are being searched, the modes of searching for data become more diverse and specific. Basing on the metadata, and with natural language descriptions, the user can, on the one hand, describe more accurately the required data and, on the other, control the match of the search results against the required purposes of use.
- Several, different scenarios changing according to interest can be formed into structured statistical data for the user. These can include, for example, the metadata perspective, i.e. an examination and comparison of statistical figures relating to the same topic in which data search is preceded by examination of concepts, definitions and classification, or the methodological perspective, in which an approach is made via examination of methodology to establish e.g. the comparability of statistics. Defining the tabulated data constitutes a perspective of its own. The purpose for the defining could be, for example, the wish to select a certain part of the table.
- The use of data archives becomes simpler and easier. This also applies to statistical data accumulated over an extended period of time when informative tables are introduced into "the nation's memory".
- When the amount of information attached to a basic table increases and its intelligibility grows, interpreting the table becomes easier. The ease or effortlessness of interpretation in turn eliminates the need of an interpreter between the customer and the statistics.

Full exploitation of the usability of informative structured tables of course also requires development of the other elements of usability. These other elements of usability of statistical data include, among other things, improvement of the readability of tables by means of content design (e.g. by simplifying table compositions), development of statistical literacy through communication and training and enhancement of the illustrativeness and instructiveness of the user interface applications for statistical data.

The introduction of informative and structured table logic into statistics production means combining of the documentation and processing of statistical data. In production, each action performed to statistical data will be documented into their context at later production stages and for eventual use in their dissemination. Thus, metadata are not treated in production as a separate "commodity" that must be saved, adapted and treated several times over during the production process.

The benefits obtainable from the informative structured table logic in production are connected with production process changes that simplify and rationalise statistics production:

- Use of the informative structured table format facilitates simpler data structures than now. Simplification of data structures is a prerequisite for data system rationalisation.
- On the other hand, data structures and concept interpretations must be rendered transparent, to enable checks to be made at every stage as to what has been meant with each concept and why the performed actions, data modifications and the like have been carried out. This improves the controllability of the production process. The visibility of metadata at all stages, for its part, creates new possibilities for controlling the quality of the production process. Well-functioning quality control, in turn, is a prerequisite

for appropriate planning and implementation of quality improvement measures.

- Integration of data and metadata reduces the need for separate data systems, and the amount of work needed to maintain metadata systems in production and archiving alike.
- It is easier to meet the users' continuously growing and diversifying data needs within the framework of the informative structured table logic than it is by increasing the numbers of interim aggregate databases. Aggregation always relates to data needs that have been specified in advance and is, therefore, an inflexible and heavy implement for meeting new data needs. The forming and maintenance of aggregated interim data stores always requires extra work, which is then lost from other activities. On the other hand, aggregation into interim stores prevents the use of statistical analysing and further processing methods in both production and in post-completion information services.

The yield from better controllability of statistical data is that their secondary use becomes easier and modifications to meet the diversity of data needs can be produced automatically or semi-automatically direct from them with only minor production process rearrangements. At the same time, the need for the specification and application work pertaining to predetermined data needs is reduced.

Despite the benefits presented above one should also not forget other, important measures for quality improvement in statistics production. As a rule, statistical production processes should be improved so that the factors affecting their quality can be taken into account better than before. In this context one factor that has received less attention than it should have done should be brought to the fore, that is the contents and standard of metadata.

If the quality of metadata improves so that, e.g. productional know-how, i.e. all the production skills and knowledge of the producers of statistics, is gathered into them it will increase the statistical organisation's available, well-organised and easily exploitable knowledge capital. Better exploitation of knowledge capital increases overall professionalism in the organisation's activities.

The benefit from the informative structured table logic to the international data exchange between statistical organisations can be seen as reduced need and amount of upkeep of harmonised data stores, as harmonisation can be defined with the metadata attached to the statistics. As far as concepts are concerned, the harmonisation functions performed to metadata for a specific purpose change slowly; therefore they are automatically revisable as the amount of data relating to the purpose increases.

In harmonisation, the "value" of the original measurement does not change into some other "value" but, in exact terms, harmonisation means producing an adapted interpretation of the original value of a variable for a specific purpose. The harmonised value cannot, as such, replace the original measurement definition or the corresponding result. In international comparisons, too, it must be possible to verify the original data and the metadata that influence their interpretation.

Use of the informative structured table logic does not make the implementation of productional revisions more cumbersome. Rather, it makes the planning and implementation of the transition phase more flexible.

Partial revisions bring immediate benefits. For example, metadata can be attached to tables produced with conventional tabulation systems and this does not require massive, simultaneous revisions of other parts of the production system. Neither does doing it call for changes to the contents of the tables or customary groupings. The reporting set-ups that have proven necessary in practice can be retained as they are, and the table specifications need not be revised simultaneously if no other pressing needs for it exist. At the first phase of introducing informative structured tables, the metadata that illustrate the use, and dictate the interpretation of, the tables are added to them. As

the tables and table groupings remain unchanged, the introduction of electronic services can be built on the users' existing routines for searching and using statistical data.

5. Implementation technologies

Structural data have conventionally been handled using diverse database models. When it comes to structured data in the table form, upon which greater demands of informativeness are imposed than on the conventional table, the traditional database models offer no easy and economic solutions.

As a concept, the informative structured table form is actually closer to an independent document than to a database or its element. If the above-described structure of an informative table can be formally defined universally and unambiguously, structured document technology can be applied to table format data. Because a statistical table is made up of data and the attached pertinent metadata, the aforementioned claim means that the metadata attached to a table should also be structured before structured document technology can be fully applied.

The formal defining (marking up) of structured documents has been standardised with the ISO-ratified SGML (Standard Generalized Markup Language) standard (ISO 8879:1986). Based on this, two well-known implementations of different type have been produced. The first one of these was HTML. HTML is a structural definition of a document, i.e. its DTD (Document Type Definition) complying with the SGML standard. The second implementation is XML, which could be characterised as a part group of SGML definitions.

Before SGML technology can be used on documents, they must be given structural definitions, i.e. DTDs, based on their structural analyses. The existence of a separate DTD is not absolutely necessary in XML-based solutions, but in practice the existence of a DTD produces a better and more controllably documented solution.

Existence of a separate DTD is also preferable because the document data can then be structured piece by piece, thereby producing modular DTD definitions based on pieces of the document. This mode of working is well suited for producing definitions for informative structured table data. The starting module can be the conventional table, and this can then be gradually expanded with modules containing the structural definitions of diverse reference data. A preliminary structural definition of a table document produced in this manner is presented in *Table.dtd*.

The produced table document definition follows the presented informative structured table logic as far as possible, but the DTD does not describe the complete structure as yet. Reference data have been included in the DTD as general reference data but no separation into footnotes and metadata references has been made in the DTD. As a further remark it should also be emphasised that the DTD does not contain any structural definitions of the metadata. The intention is to implement these as their own DTD modules.

The attached initial version of a definition of a structured table document is meant to be used for testing the concept in practice and is only to be used at the testing phase. The intention is to produce a more precise and exhaustive DTD on the basis of gained experience.

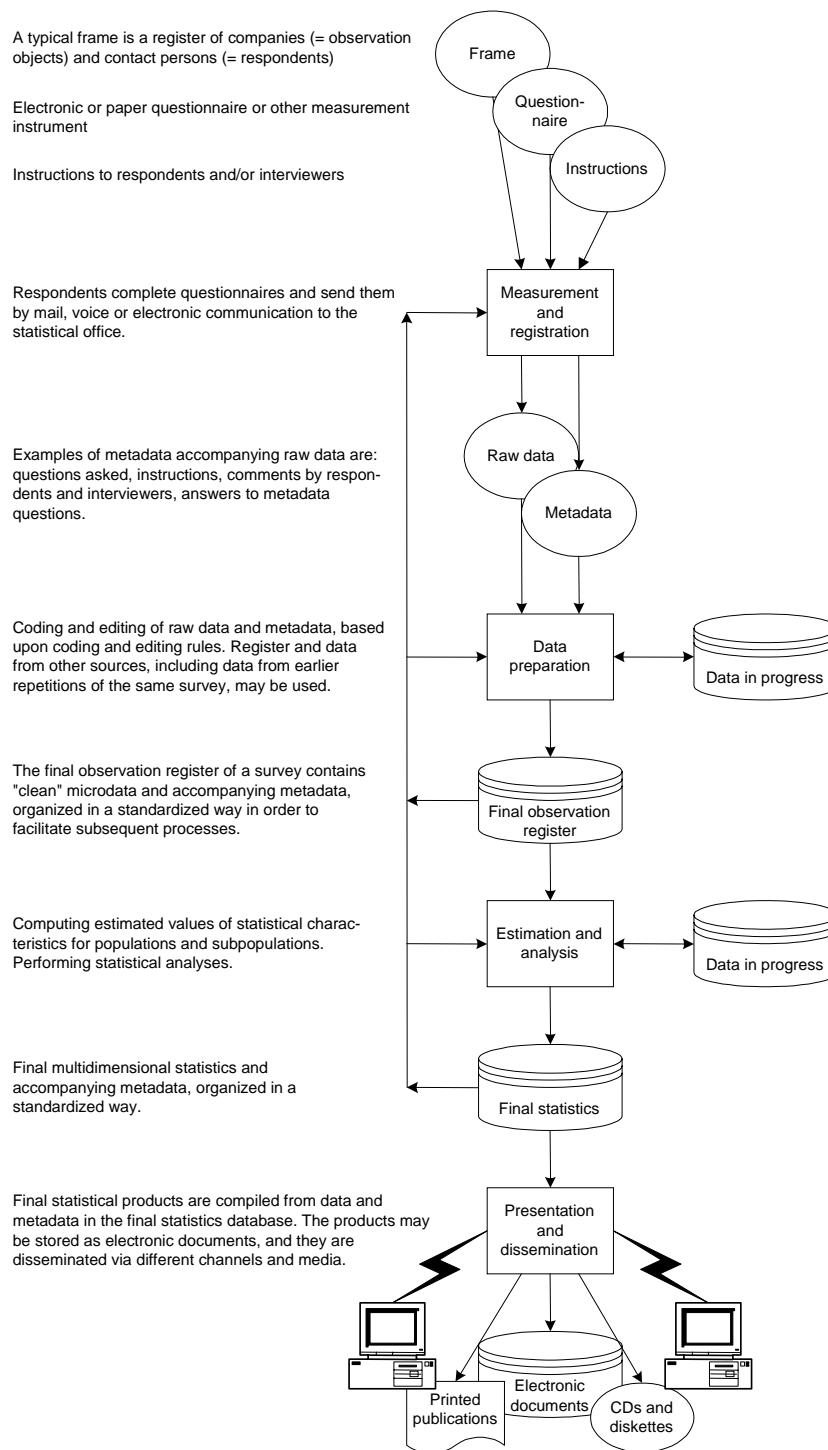
The solution is based on a DTD developed by Statistics Norway which, in turn, is based on a document defining system known as CALS DTD. The CALS DTD adapted by Statistics Norway has, on the one hand, been enhanced here to meet Statistics Finland's special needs by increasing the number of available languages, for example, and, on the other hand, by adding to it the reference data definitions needed in the processing of informative table reference data.

Statistics Finland uses a uniform system to describe the identification data of individual page documents in its HTML page production. The system contains the metadata concerning each document. These document description data have been defined in a separate DTD module, which is presented in *docmeta.dtd*. The definitions of metadata are compatible with the Dublin Core definitions.

Appendices

Appendix 1. Data and metadata in statistics production

Appendix 1. Data and metadata in statistics production



Typical flow of data and metadata through the processes of a survey of a national statistical office (An Information Systems Architecture For National And International Statistical Organizations (submitted by Statistics Sweden, prepared by Bo Sundgren.). Statistical Commission And Economic Commission For Europe - Conference Of European Statisticians Meeting on the Management of Statistical Information Technology , Geneva, Switzerland, 15-17 February 1999. Economic and Social Council CES/AC.71/1999/4, 23 November 19.