Towards a system of official statistics based on a coherent combination of data sources, including surveys and administrative data

Bo Sundgren
2011-07-01

Professor Bo Sundgren
Stockholm University
Department of Computer and Systems Sciences (DSV)
Affiliated with Dalarna University, Department of Informatics
bosund@dsv.su.se, bo.sundgren@gmail.com, bsu@du.se
https://sites.google.com/site/bosundgren/

This paper is inspired by, inter alia, a vision for the European Statistical System (ESS) as outlined in European Commission (2009): "*Communication from the Commission to the European Parliament and the Council on the production method of EU statistics: a vision for the next decade*", document COM(2009)404.

The vision is based on the assumption that the traditional stovepipe approach to official statistics, both in member states and within Eurostat, is going to be replaced by a more holistic, integrated, coherent, and systems-oriented approach, both as regards the statistical contents, and as regards the technical solution. The term "data warehouse" is used to cover both these aspects. The vision will be called ***"Vision 404"***, for short, in this paper.

Background

There are different types of statistical systems (and subsystems). For example, statistical systems may be categorised according to the following paradigms: bookkeeping systems, registers, censuses, sample surveys, archives, etc, and combinations of these. Alternatively, statistical systems may be categorised along geographical/political dimensions: national, international, regional, local, etc.

Statistical systems are often categorised by contents, but it is becoming more and more difficult to find a natural hierarchy of topics, since different sectors of society (and hence different kinds of official statistics) are becoming more and more interrelated and overlapping. Nevertheless most national and international organisations try to maintain some kind of hierarchical categorisation of official statistics by contents, at least for practical and organisational reasons.

Gottfried Achenwall, a German political scientist, used the term "Statistik" for the first time in his work "Staatsverfassung der heutigen vornehmsten europäischen Reiche und Völker im Grundrisse" (1749). By the term "Statistik" he meant a comprehensive description of the social, political, and economic features of a state, a definition which corresponds fairly well to the present understanding of the term "official statistics", as used by national statistical institutes and international organisations. An even shorter definition along the same lines is "data about the state".

Thus originally statistics was a kind of bookkeeping, organised by statesmen and other power centres, such as the Church. People and resources were enumerated, counted, and summarised in simple, but systematic ways. The bookkeeping paradigm is still clearly visible in official statistics, reflected in the use of registers, as well as in the bookkeeping approach to operations such as data editing, where it still prevails to a great extent, although it is now finally (maybe) being replaced by more economical methods, such as significance editing (also called selective editing or macroediting). The bookkeeping approach means that all observation data (primary data) should ideally be complete and correct, regardless of their relative importance for the finally produced, aggregated statistics.

As a matter of fact, the National accounts, may be regarded as an advanced, analytically oriented application of the bookkeeping paradigm of officical statistics, “data about the state”.

During the more than 250 years that have passed, since the seminal work by Achenwall, several paradigms have been applied to European statistics: mathematically based official statistics with its roots in probability theory, probability-based sample surveys, invented and perfected during the 20th century, register-based statistics and reuse of administrative and archival data, initiated in the Nordic countries during the 1960s, step by step replacing traditional censuses from the 1970s and onwards, and being now a major ingridient in official statistics in many European countries.

Maybe now is the time to integrate all these approaches to official statistics into one systems-based theory and practice for official statistics, in Europe and elsewhere. See Sundgren (2010a). “Vision 404” seems to represent a major step in this direction.

Summary of “Vision 404” for the European Statistical System

The vision is based on the assumption that the traditional stovepipe approach to official statistics, both in member states and within Eurostat, is going to be replaced by a more holistic, integrated, coherent, and systems-oriented approach, both as regards the statistical contents, and as regards the technical solution. The term “data warehouse” is used to cover both these aspects.

A visualisation of the vision

Figure 1 visualises a statistical system as a reasonably complete and coherent reflection of important aspects of a society. There is an ongoing, complex exchange of data and actions between the statistical system and its environment. The situation and changes in the environment (a society) is reflected by data and data updates in the statistical system. Decision-makers interpret data from the statistical system, assisted by analysts and researchers, and they make decisions that changes reality (the society), and in the following iterations they may see and evaluate the effects of their decisions as reflected by the statistical system (and other information systems in the environment of the statistical system). Thus the statistical system is a part of a gigantic feedback looop.

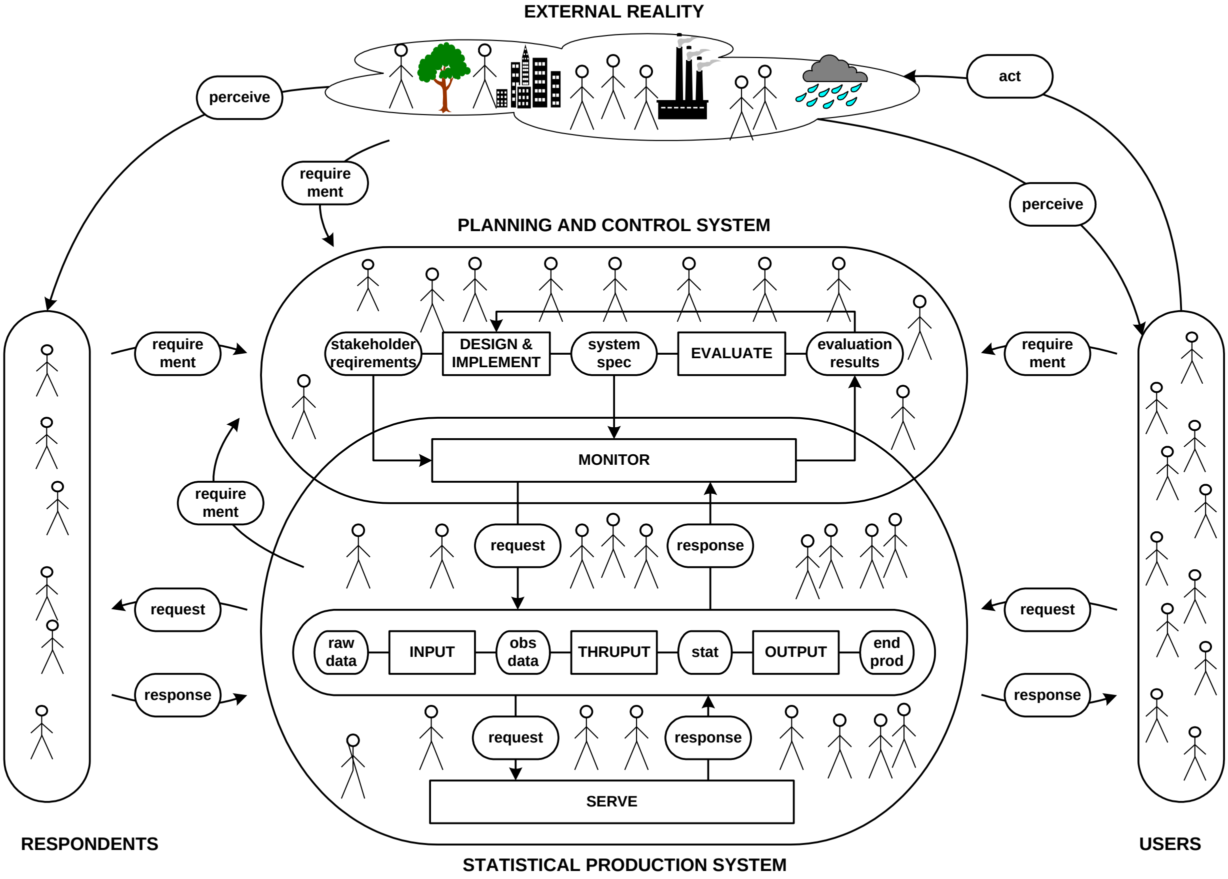


Figure 1. A statistical system in its environment. From Sundgren (2004a).

Desirable features of the system to be designed and implemented

We will discuss desirable (and to some extent even necessary) features of a future European Statistical System based upon a data warehouse approach in accordance with “Vision 404”.

Registers based on standard definitions and unique, stable identifiers

“Vision 404” does not explicitly mention registers, but it mentions administrative systems as important sources of data for official statistics, and registers are obviously important parts of such systems.

Basic definition of a register

What then is a register? A basic, orthodox definition of a register is the following one:

- A register is an authorised, up-to-date list of all objects belonging to a certain population

- The objects listed in the register are uniquely identified by an authorised identifier, such as person number for persons, organisation number for organisations, etc

- In addition to the identifier, a register may contain additional basic and up-to-date information about the objects, such as name (not necessarily unique) and location and other contact information, e.g. address and telephone number

Example: A person register is an authorised list of persons belonging to a certain population, e.g. all persons living in a certain country at a certain point of time.

A **core set of registers for official statistics** of a country (or of the European Union as a whole) may contain (at least)

- a register of persons
- a register of organisations
- a register of real estate objects: houses, dwellings, localities

Registers of this nature and status are sometimes called **base registers**. They define both semantically (by means of inclusion rules) and pragmatically/operationally (by means of enumeration) some basic object types and populations of fundamental importance for the administration of a modern society, as well as for official statistics.

Links between registers

The base registers should also be linked to each other, thus materialising important relations between the basic object types, e.g. “the dwelling were a person lives”, “the organisation where a person works”, “the locality where an organisation is located”.

Unique and informationless identifiers

The identifiers of the objects in a register must be unique and stable over time. The safest way of ensuring stability of identifiers over time is to make them informationless. If an identifier contains information, there is always the risk that this information changes, either because the information turns out to be wrong, or because the status of the object really changes.

For example, person identifiers sometimes contain information about the birth date of the person. This information is believed to be stable, but sometimes it is discovered that the information is wrong and has to be corrected. However, if the information about the birth date is corrected, it will also cause the identify of the person to change, which will inevitably cause a lot of problems, both for the person concerned and for others.

The identifiers of the objects in a register must be unique (over time) within the population covered by the register, e.g. a country or the European Union as a whole. The European Union should have a plan for ensuring that national registers use identifiers and definitions that will make it easy to integrate the national registers into European registers, when this is becoming desirable. For example, a growing number of enterprises are likely to conduct business all over Europe (and beyond), and these enterprises will appear in many national business registers. Even before a European register of organisations is possibly created, it is important that the same enterprise is defined and identified in a compatible and linkable way in all registers where it appears.

Standard definitions of registered objects

The definition of a person is relatively straightforward, but for many other objects it is not. For example, there are many possible defintions of an organisation or an enterprise. Different definitions may be suitable for different purposes. In official statistics, one is usually interested in establishments, which are ideally localised in one place, and engaged in one kind-of-activity only. In practice it may be very difficult to obtain data for such “ideal” establishments, and operational definitions based on compromises have to be used.

Administrative registers and statistical registers

One may distinguish between administrative registers and statistical registers:

- An **administrative register** is a register used for administrative purposes
- A **statistical register** is a register used for statistical purposes

A statistical register may be created from one or more administrative registers, possibly in combination with data from other sources, such as traditional statistical surveys. See Wallgren&Wallgren (2007) for more details.

Traditionally, registers are used in official statistics as frames for surveys. When conducting a statistical survey, one has to create a frame, or a register, of the objects in the population to be investigated by the survey. When conducting a sample survey, a sample of the objects in the frame is drawn, and then these objects are observed by some kind of measurement method and measurement instrument, e.g. mailed questionnaires or telephone interviews.

In order to be useful for statistical purposes, a register should contain contact information for all objects in the register, that is, name, postal address, physical address, email address, phone number, etc, for the object itself and/or a human respondent representing the object.

It is also useful if a statistical register contains classification variables that makes it easy to create strata and subpopulations (domains) to be used in the design, production, presentation, and analysis of survey data.

Extended registers and satellite registers

A register may also be extended with other data concerning the objects in the register, which can be used in combination with, or instead of, data collected by surveys. Such register data may reduce the response burden and improve the quality and efficiency of official statistics.

Links to other registers and data sources; satellite registers

As has been mentioned already, a register should contain links to other registers, thus representing relations between object types and between populations of objects. Such links between registers effectively multiplies the amount of information contained in a system of registers. For example, by linking persons with their dwellings, we indirectly associate persons with variables and properties of their dwellings, e.g. “persons living in one-family houses”, “persons living in the same dwelling” (possibly forming a household), etc.

Furthermore, a register may be linked to other databases or files containing data about special subdomains (subpopulations) of the population of objects contained in the register. Such data sets are sometimes regarded as **satellite registers** to the registers that they are linked to; they may or may not fulfil the strict definition of a register stated above. For example, a person register may have satellite registers like “student registers” and “patient registers”. Satellite registers are sometimes named after some important (type of) variable in the registers, e.g. “income registers” and “education registers”.

Event registers

Some registers contain event type objects, e.g. road accidents and crimes. Such registers are cumulative in the sense that they contain all events of a certain kind that have occurred since the time when the registration of such events started.

Base registers containing all basic objects of a certain kind, e.g. all persons living in a country, are often closely associated with one or more event registers, containing information about certain types of events that the basic objects are involved in, e.g. birth and death events, migration events, marriages and divorces, etc.

Life history registers
Base registers in combination with event registers may be used for forming life histories of the basic objects in a base register (or objects belonging to a certain subpopulation, registered in a satellite register), e.g. the life histories of patients or criminals. Life history registers, or **life history databases**, may be used for **longitudinal studies**.

Administrative data sources

"Vision 404" emphasises the use of administrative data sources.

In countries like Sweden, up to 99% of the data used for official statistics emanate from administrative data sources, and only 1% from traditional surveys. According to a Dutch study, referenced in Wallgren&Wallgren (2007), it is roughly 100 times more expensive to collect data by a traditional census than by using administrative data.

If we generalise the Dutch findings to all statistics produced by Statistics Sweden, the 1% of the data that emanate from traditional surveys would still account for about 50% of the costs. If Statistics Sweden would, for some reason, no longer have the option of using administrative data, the Swedish government would have to increase the budget of Statistics Sweden by a factor 100 in order to secure the same amount of official statistics with the same quality. This would never happen. The government would more likely leave the budget as it is, demanding Statistics Sweden to decrease the number of topics covered, and to decrease the quality of the estimates in terms of bias and precision, e.g. by decreasing sample sizes and introducing less costly procedures for non-response management.

From another point of view, we may argue that the use of administrative data for statistics production will not only drastically reduce costs. It will also contribute positively to several quality dimensions, e.g. the richness and coherence of the statistical data, as well as reduced bias and increased precision of the estimates. This is true if sound statistical methods are used, and it is particularly true, if administrative data and survey data are systematically used to strengthen each other and reduce each other's weaknesses.

In the future the concept of "administrative data" as a source for official statistics is likely to become generalised to many different kinds of "operational data" generated as side effects of different kinds of processes: public processes, business processes, and private processes (e.g. social media).

Combining surveys and administrative data

While "Vision 404" emphasises the use of administrative data sources, it does not exclude, of course, continued use of traditional statistical surveys as well. We will now briefly discuss how traditional surveys and administrative data sources could best be integrated within a common framework and a common system, and how they could complement each other.

Traditional statistical surveys and statistics production based upon adminstrative data have their relative advantages and disadvantages. The strongest merits of administrative data have to do with costs, response rates, response burden, coherence, timeliness, and flexibility.

The strongest merits of traditional surveys, on the other hand, are within the area of relevance, the possibilities to tailor the design of a survey to specific statistical needs.

Costs

The merits of using existing administrative data sources as regards costs have already been mentioned and examplified above. On the other hand, the costs of collecting data by means of traditional surveys are typically increasing all the time, mainly because of growing difficulties to find and establish contacts with respondents, and to motivate them to participate in surveys.

Relevance

It is virtually impossible for designers of official statistics to have any influence over the definitions used by administrative registers and administrative processes generating administrative data. The administrative processes are governed by laws, and these laws may become changed now and then, as the result of political decisions. Those responsible for official statistics will have to adapt to these definitions, and changes of definitions, as intelligently as possible, in order to make the administrative data useful also for statistical purposes, and in order to maintain continuity in official statistics despite the changes in definitions resulting from political decisions.

In contrast, designers of traditional surveys, at least theoretically, have the freedom to tailor the definitions used in those surveys to the needs of official statistics. However, in practice the tailoring may not be as simple as it seems. Using definitions that are different from definitions used in administrative processes may create great difficulties for many respondents, not least in companies, whose information systems have to be adapted to laws and administrative processes – but not necessarily to the needs of official statistics.

Non-response

Non-response has been a steadily growing problem for statistical surveys over several decades. Administrative data sources, on the other hand, have relatively little difficulties with non-response. The persons and enterprises concerned by the administrative systems usually have a strong motivation to provide data to these systems; they may even be forced to do so by laws, or the data may even be generated automatically by actions and events that take place anyhow.

Actually, data from administrative registers may also be one of the best tools for coming to grips with the growing non-response problems in traditional surveys, and to reduce the bias of the estimates in such surveys. Administrative data may be used as supplementary information to intelligently designed estimation processes. A lot of literature on this topic already exists, and it is a very promising research area among statistical methodologists.

Bias

As was just mentioned, administrative data may be used for reducing bias in data resulting from traditional statistical surveys. On the other hand, administrative data have their own problems with bias. It was mentioned above that an advantage of using administrative data is that respondents are often strongly motivated to provide data to administrative processes. Unfortunately, the motivation may not necessarily be to provide correct data. For example, if a citizen or a company has to provide income data to the tax authorities, there is an obvious risk of bias in the data provided.

Coverage
Administrative registers are not free from problems of overcoverage and undercoverage. For example, a population register may contain persons who have immigrated or emigrated without reporting this properly to the administrative authority responsible for population registration. Here traditional statistical surveys may have a role to detect such errors in the administrative systems.

Timeliness
If reasonably relevant data are available from an administrative source, the lead time for producing official statistics from these data will be very short – much shorter than if a traditional statistical survey had to be designed and executed. Moreover, statistics production based on administrative data could be repeated as often as you like, at a very low cost. “We can make a census every day”, is a proud (and almost true) statement from a statistical agency that has replaced traditional population censuses, carried out every 5 or 10 years, with statistics based on administrative registers.

Time series and comparability over time
Provided that events and life histories are maintained in historial versions of registers, registers are excellent, inexpensive, and flexible sources of time series data, both on micro and macro level.

Coherence
Official statistics based on a core of well designed registers are likely to become much more coherent, at a much lower price, than official statistics based upon traditional, stovepipe-organised statistical surveys.

Information potential and flexibility
As emphasised by “Vision 404” the need for official statistics will continue to grow. Because of growing complexity of the decision problems of a modern society, official statistics will be required to be more coherent than is possible in the present stovepipe model. Finally, ad hoc requirements for new statistics will occur more frequently because of the appearance of new and complex problems that require immediate attention by analysts and politicians, e.g. financial crises.

Conceptual standardisation and integration

Conceptual standardisation and integration is a necessary precondition for a holistic and highly integrated European Statistical System, as envisaged by “Vision 404”. We will now briefly discuss what conceptual standardisation and integration could mean in practice.

Standard classifications

National statistical agencies and international organisations have a long and reasonably successful history of creating and maintaining standard classifications on national, regional, and international levels. Standard classifications are extremely important for the coherence and comparability (in time and space) of official statistics, and together with registers and standard concept definitions, standard classifications will be the backbone of the future European Statistical System – as they already are to a considerable extent.

Standardised time and space coordinates

Almost all official statistics are associated with time and space coordinates, and standardisation of these coordinate systems are (like standard classifications) of utmost importance for ensuring the coherence and comparability of official statistics.

Time: frequencies, delays, current data and historical data

Most official statistics have a typical **frequency** – yearly, quarterly, monthly – but there are other frequencies as well, including no frequency at all – **ad hoc surveys**. Sometimes a statistical survey is associated with time coordinates in a more complex way: the survey may be conducted at a certain point in time (or during a relatively short time period), but the data collected by the survey may be associated with other points in time and/or time periods.

A register may be very flexible with regard to time. Primarily, a register should be **up-to-date**, that is, it should reflect the **current status** of the objects in the register with as little delay as possible. However, a statistical register should not be limited to reflecting the current status; ideally it should also contain (or be associated with) a complete history of all objects, including those that have ceased to exist. Such a **historical register** (or system of registers) will make it possible to reconstruct the status of all objects at an arbitrary point of time, and to reconstruct statistics for the population associated with the register, as it was at an arbitrary point in time.

Space: coordinates, maps, regional classifications, and geodata

Thanks to modern technology it is nowadays relatively easy and inexpensive to associate any objects with their current locations in space, and to create high-quality maps, possibly animated, of all locations and objects and data associated with these locations (over time).

Before the technological development made it easy to associated data with “real” space coordinates, **regional classifications** were used as proxies for space coordinates. Regional classifications are still very useful – but now also as aggregation levels for microdata associated with “real” space coordinates, obtained by GPS devices etc.

Other standardised concepts and variables

A wide range of concepts needed to be standardised in a future European Statistical System have been covered above: objects as defined by registers, qualitative variables as defined by classifications, time and space coordinates. However, there are even more concepts and variables that need to be properly defined and harmonised, e.g. important quantitative concepts and variables like “income”.

A coherent conceptual model and ontology

Traditionally, the official statistics of a country consists of a hundred or more statistical surveys. According to the stovepipe approach to official statistics, each one of these surveys is designed without very much coordination with other surveys. Each survey has its own main purposes, its own main users, and its own design. As is pointed out in “Vision 404”, this approach easily leads to higher costs, and higher response burdens, than necessary. It also makes the data from one survey less comparable and less coherent with data from other surveys than desirable, especially when socio-economic developments become more complex, and when more analysts and politicians and other main users of official statistics become interested in effects of decisions and actions on society as a whole. A more holistic approach is needed – also in official statistics.

How can official statistics become more holistic? The main problems and opportunities are in the so-called conceptual models that are applied – explicitly or implicitly – in the design of the individual surveys that together form the system of official statistics. Today, at best, the designers and main users of a statistical survey discuss and decide upon a conceptual model for the particular survey under consideration. They analyse which objects and variables need to be observed, according to which ideal and operational definitions, in order to be suitable for producing the statistical estimates required and asked for. They may even visualise the conceptual model graphically, with boxes representing objects and populations, lines between the boxes, representing relations between the object types, and other symbols representing variables observed and parameters estimated, as we shall discuss more in detail below.

These individual conceptual models for individual surveys may be the starting-point for an integration process of the surveys that together comprise the statistical system of a country (or the European Union). Here are some methods for making the individual conceptual models of stovepipe-designed surveys become more integrated, compatible, comparable and coherent with each other:

- Use object types, identifiers, populations, subpopulations, links to other object types, etc, as defined in base registers
- Use standard definitions of variables and standard classifications
- Use standard time scales, space coordinates, and regional classifications

A conceptual model of a statistical survey or a domain of statistics

Figure 2 gives a graphical illustration, a so-called **object graph**, of a concrete domain of statistics. More exactly it illustrates the conceptual model of UNESCO's education statistics. See also Bruneforth&Sundgren (2007).

A concrete conceptual model of a piece of reality, looked at through the glasses of the designers of a statistical survey or a statistical information system, may be based on a **generic conceptual model**, or an **ontology**, described as follows; see also Sundgren (1973, 2004a, 2004b, 2005, 2006):

- There are **objects** (e.g. persons or enterprises) belonging to **object types** (e.g. PERSON or ENTERPRISE) and **populations** (e.g. persons living in a certain country at a certain time), visualised by rectangular boxes in the graphical version of the conceptual model.

- The individual objects (also called object instances) are uniquely identified by means of an identifying variable (also called **identifier**). Ideally an identifier should have no other function than identifying objects uniquely within a certain domain; thus it should be something like a random number or some other informationless variable, which is guaranteed to be stable over the lifetime of the object. Unfortunately, in practice designers often choose identifiers, which are not informationless, and which cannot be guaranteed to be stable over time.

- Objects have **properties**, which may be qualitative or quantitative. Properties used in statistics are often formalised as **<variable, value>** pairs. For example, the property "female" of a person may be formalised as "sex=female".

- Quantitative variable may be transformed into classifications by means of **grouping** into intervals, e.g. age groups, income groups, etc.

- The properties which correspond to the set of all possible values of a qualitative variable, the **value set** of the variable, for example the values “female” and “male” of the variable “sex”, are often chosen so as to classify the objects which are associated with the variable, e.g. “persons” associated with the variable “sex”. However, sometimes this is not the case. For example, one and the same company may be associated with more than one “kind of activity”.

- A **hierarchical variable** is a variable with a value set that is structured into a hierarchy, consisting of several levels. Examples: “region”, “kind of activity”. The values on each level of such a hierarchy usually constitute a classification of the objects with which the variable is associated, and the set of objects associated with a value on level i in the hierarchy is the union of the objects associated with the values on level i+1 that are subordinate to the value on level i.

- Like in figure 2, the variables associated with an object type or a population may be represented by a list of variable names in the box representing the object type or population in the object graph.

- Objects may be related to each other by means of **object relations** (of degree 2, 3, or sometimes even higher). For example, a trade relation may relate three objects: a seller, a buyer, and a commodity. Both the seller and the buyer may be a person or an enterprise.

- Object relations may themselves be “objectified” into **relation objects**, which have their own properties and variables. For example, the trade relation just mentioned may be objectified into a trade transaction object with the variables “quantity in tons” and “amount in Euro”.

A higher level ontology or generic conceptual model
We have just describe a generic conceptual model, or ontology, on a certain level of abstraction. In order to facilitate the integration of specific conceptual models corresponding to individual surveys or domains of statistics, into more general and holistic conceptual models of major sectors of society, or even society as a whole, we need a coneptual model, or an ontology, on a higher level of abstraction.

For example, we may categorise all object types that occur in specific conceptual models of official statistics into the following three categories:

- **actors** (also called agents), e.g. persons and organisations

- **utilities** (also called “things” or “items”), e.g. buildings, vehicles, commodities, real or financial assets

- **complex objects** (based on relations between actors and/or utilities), e.g. events, transactions, relationships (e.g. employments, marriages, ownerships)

Figure 3 illustrates this high-level, generic conceptual model.

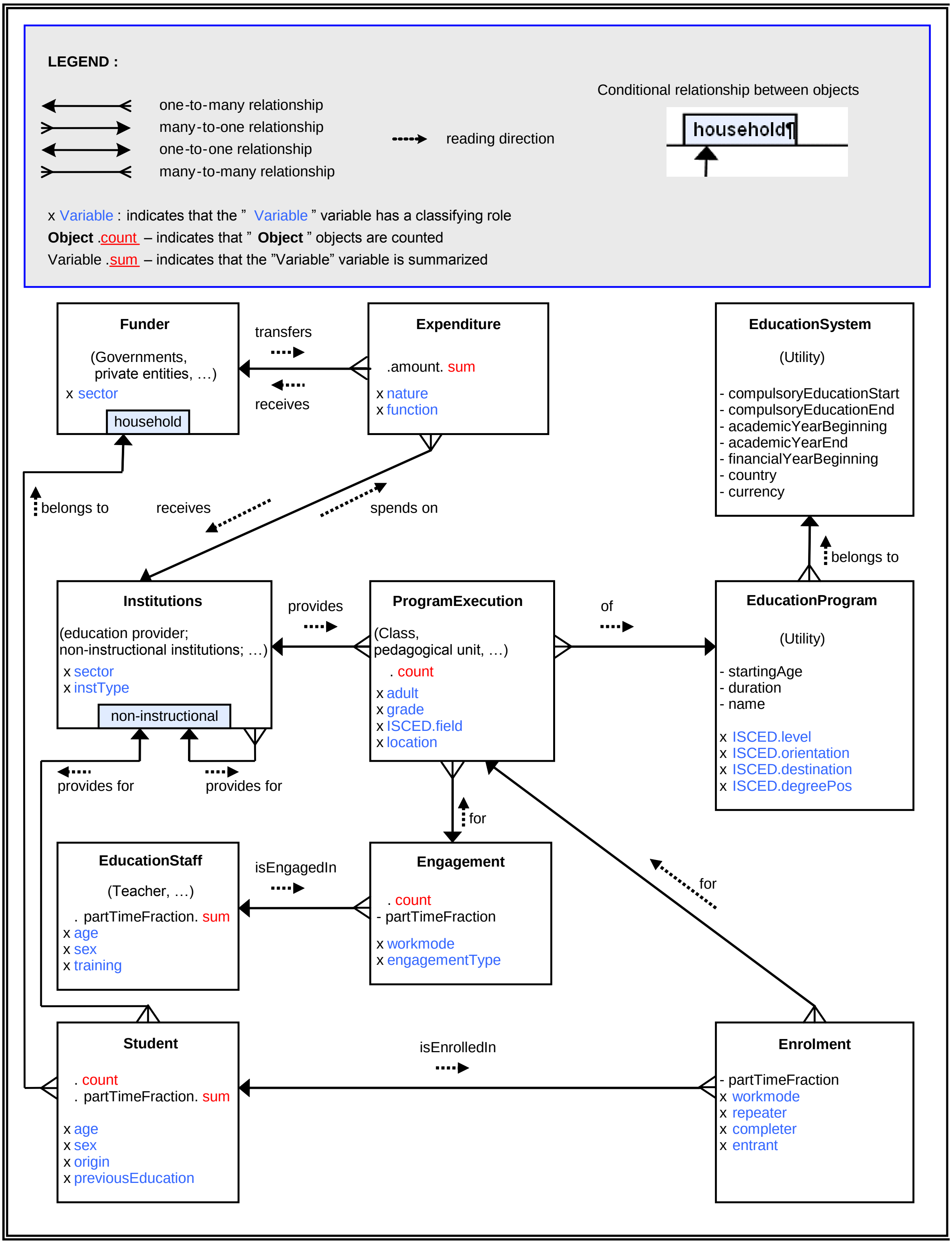


Figure 2. A conceptual model of education statistics collected and produced by UNESCO.

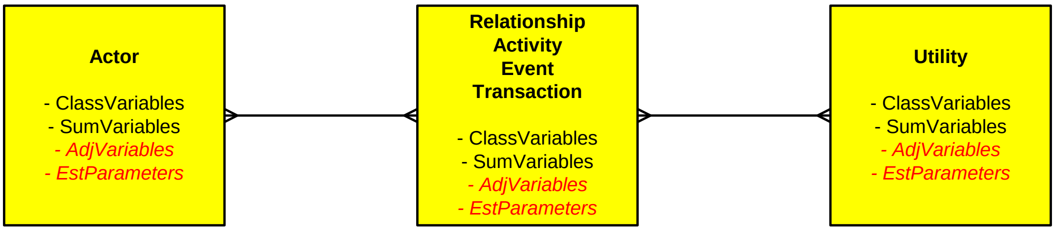


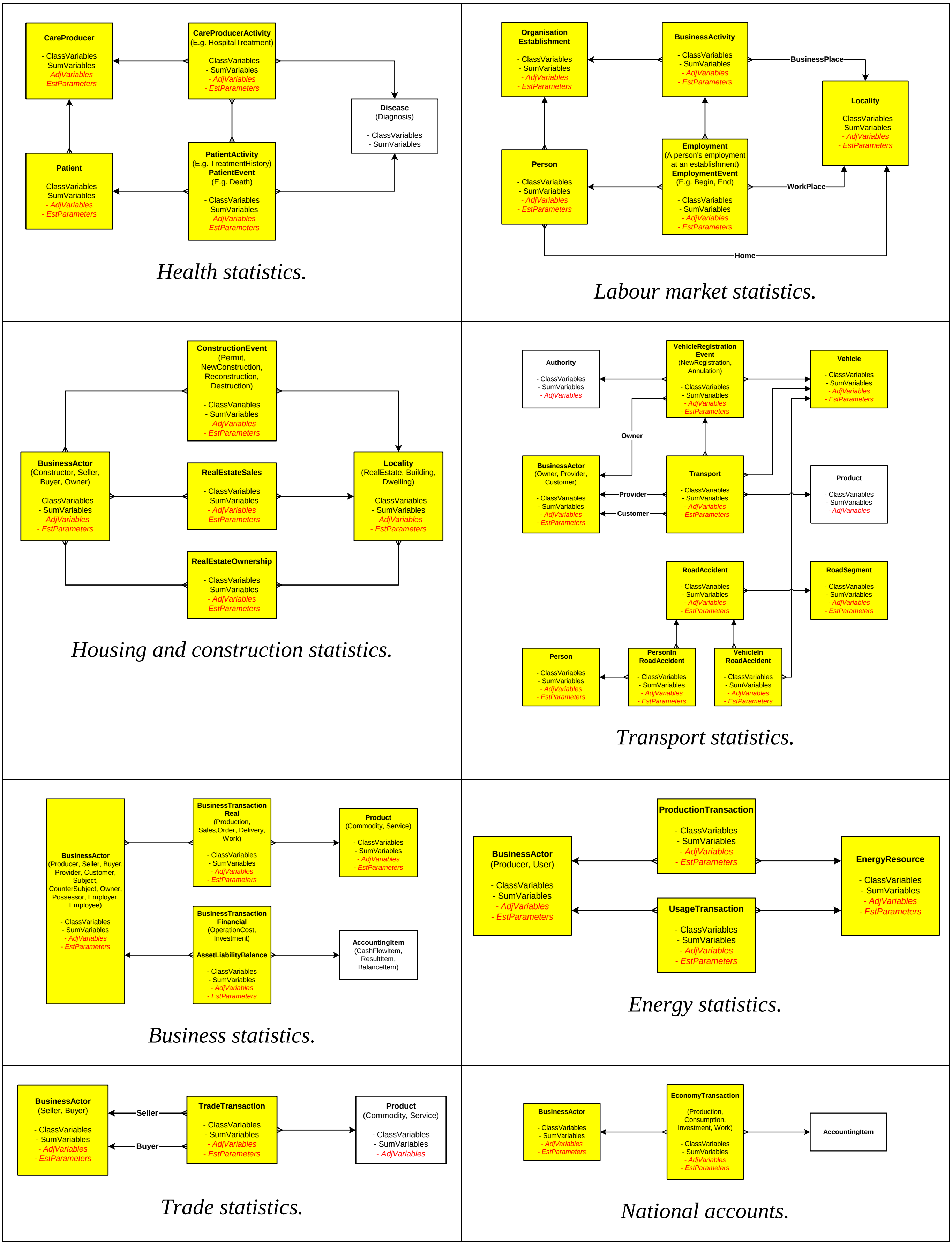
Figure 3. Generic model of the contents of a system of official statistics.

Actually the contents of all branches of official statistics can be expressed as specialisations of this generic model. This thesis has been verified in a large number of practical examples, and no counter-examples have been found.

Figure 4 provides more examples of this kind of schematic object graphs for a number of domains that are typical for official statistics on both national and international level. Actors occur to the left in the figures, utilities to the right, and complex objects in the middle. Further examples can be found in Sundgren (2005), Sundgren (2006), and Sundgren (2007b).

National accounts

The last object graph in figure 4 illustrates a schematic conceptual model for the Swedish National accounts. Figure 5 provides a more detailed model, still relatively simple, though.



Health statistics.

Labour market statistics.

Housing and construction statistics.

Transport statistics.

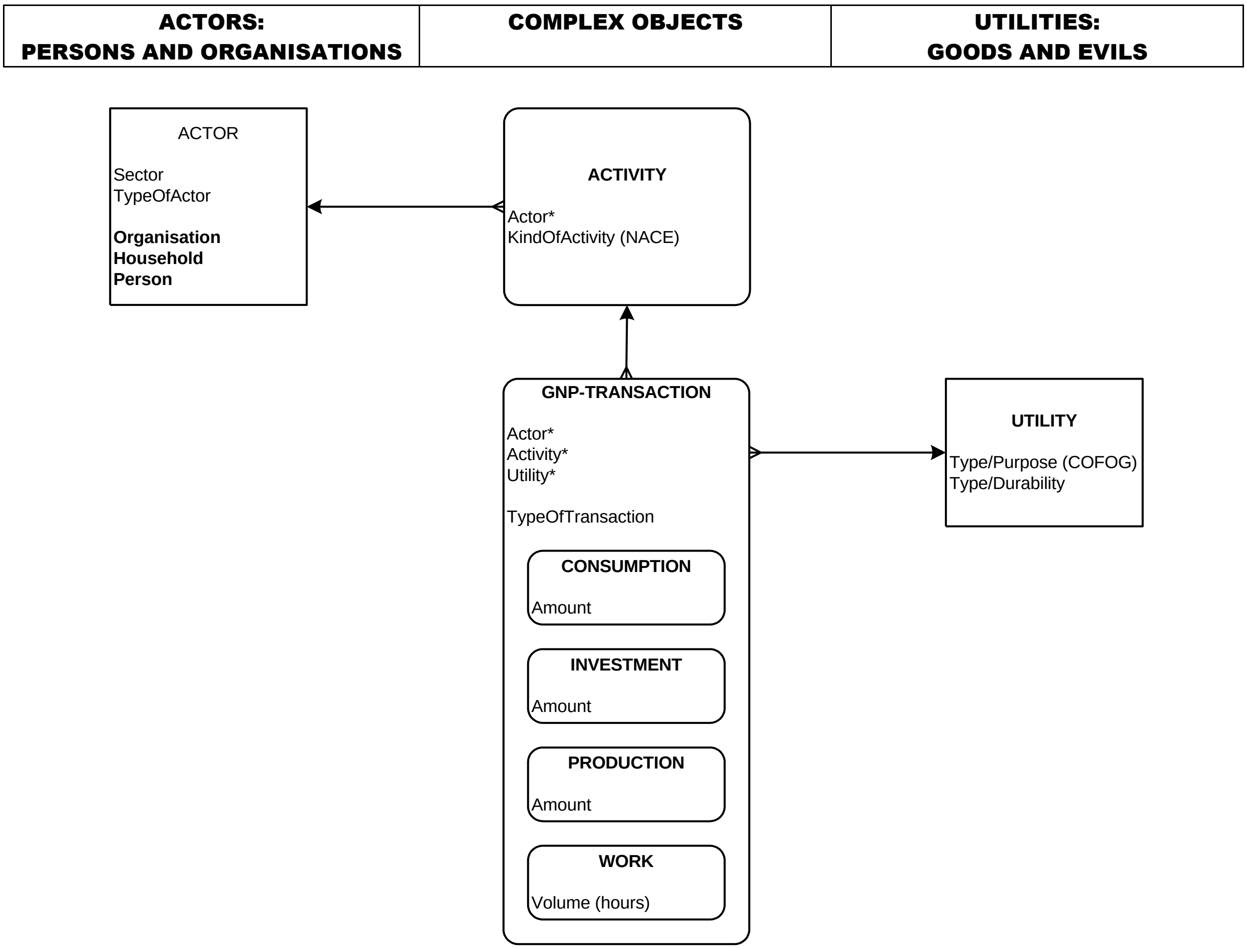
Business statistics.

Energy statistics.

Trade statistics.

National accounts.

Figure 4. Examples of schematic conceptual models for domains of official statistics



National accounts		
OBJECT TYPE/(SUB)POPULATION	**CLASSIFICATION VARIABLES**	**SUMMATION VARIABLES**
GNP TRANSACTION		
GDP transactions in Sweden during a certain time period	Sector Kind of activity (NACE)	
- Consumption transactions	Durability (ESA) Purpose (ESA)	Amount (SEK)
- Investment transactions		Amount (SEK)
- Production transactions		Amount (SEK)
- Work transactions		Amount (hours)
- Import transactions	Type of utility (commodity/service)	Amount (SEK)
- Export transactions	Type of utility (commodity/service)	Amount (SEK)

Figure 5. National accounts.

A data warehouse architecture for a statistical system

One of the main reasons for having a special statistical organisation in a country or within some other kind of organisation (e.g. an international organisation or maybe even a business company) is that both specialised statistical systems and administrative systems in society generate a lot of data over time. These data have a considerable information potential, and they can be used over and over again, for many different purposes, often purposes that are quite different from the purposes for which the data were originally collected and used.

By taking care of all these data and by storing them, well documented, in an organised way, a statistical organisation may accumulate a data capital, the future yields of which may go far beyond the value of the first usages of the data. This advantage of having a specialised organisation for this task arises even if the statistical organisation does not do very much more than we have just described: storing the data together, well documented, and making them available as a collective resource. However, using its statistical competence, the statistical organisation may also add new value to the data, e.g. by integrating the data not only in a physical way, but also from a contents-oriented point of view, by making the data more comparable and coherent, by using standardised concepts and classifications, etc. Some improvements in these directions may be done *a posteriori*, when the data have already been collected, but it is even better, of course, if these aspects are considered already when the data collection processes are planned – data co-ordination *a priori*.

In figure 1 in the beginning of this report we visualised a statistical system as a reasonably complete and coherent reflection of important aspects of a society. Such a system could serve as an excellent basis for advanced analysis, decision-making, and evaluations of decisions already taken and implemented. Now we shall investigate in some more detail what the architecture of such a system could look like. We will start with a brief historical background of the data warehouse approach to statistical data, and then proceed to a modern version of the same concept.

Historical background

The roots of the modern data warehouse approach to official statistics are to be found in the works from the 1960's of Svein Nordbotten on what he called a **statistical file system**, or an **archive-statistical system**. This approach was further developed and put into practice in some statistical agencies, notably Statistics Sweden, where particular emphasis was put on the emerging database technology, development and use of standardised software for the typical processes of statistics production, and – last but not least – the development and use of advanced metadata systems; the availability of high-quality documentation and metadata are of particular importance when statistical data are to be reused and combined for other purposes than those for which they were originally collected, so that the "fitness for purpose" of the existing data for new purposes can be responsibly judged.

Nordbotten's statistical file system or the archive-statistical approach

Probability-based sample surveys and use of administrative data for statistical purposes are two important ways of obtaining statistical information about a society in an economical way. Another way of economising with resources, and also speeding up the production of official statistics, is to reuse survey data that have already been collected earlier, possibly for other purposes than those at hand. This approach (in combination with use of administrative data) was advocated by professor Svein Nordbotten in some seminal papers published as early as the 1960's. See Nordbotten (1966, 1967a,b,c).

Nordbotten’s vision included standardised microdata files, systematically documented in a data catalogue[1], and managed by standardised processes supported by generalised software. The top management of Statistics Sweden became interested in Nordbotten’s ideas, and started a number of development projects at the end of the 1960’s with the intention to reengineer the production processes of Statistics Sweden from both a technical and an organisational point of view. The data warehouse would include both microdata and macrodata in standardised form, and the data would be described in a catalogue of variables, both from a technical and from a contents-oriented perspective. Microdata and macrodata processes would be driven by standardised software.

The privacy debate triggered by the 1970 population census in Sweden made it impossible for Statistics Sweden to continue the development of a data warehouse including microdata. The contents of the data warehouse had to be limited to aggregated statistics on a relatively high level. In 1976 Statistics Sweden launched its first online database, available to external users, and including a wide range of statistics, e.g. socio-demographic statistics, economic time series, and regional statistics. All data were managed by the AXIS database management system, developed by Statistics Sweden. The system was metadata-driven, and the metadata model used for that system is still used, in modified form by the current Internet-based output databases, Sweden’s Statistical Databases, which were launched by Statistics Sweden in 1996.

The development work at Statistics Sweden in the early 1970’s also resulted in a generalised, metadata-driven software product for tabulations, TAB68, which could easily be used even by non-programmers. Later developments resulted in a whole family of generalised software products, based on the TAB68 program code, for statistical processes like data editing, file matching, data transformations, and variance computations. A more recent development in the same tradition is the PC-AXIS software for user-friendly retrieval and manipulation of statistical data.

The theoretical basis for these developments is elaborated in Sundgren (1973), where the term “metadata” is used for the first time. It is based on the already mentioned work by Nordbotten, and on the seminal book by Langefors (1966), where the distinction by information and data is made clear, and where a comprehensive theory of information systems is presented. Sundgren (1973) formulates an infological theory of databases and introduces conceptual modelling as a systematic way of describing the contents of databases and information systems. This conceptual framework is further developed for statistical purposes in Rosén&Sundgren (1991).

In summary, Nordbotten’s approach as further developed and practiced by Statistics Sweden, contained a number of features, which are still very relevant:

- a standardised data model for all data in the data archive
- standardised files stored in a standardised way in the data archive
- good documentation of the data in the data archive
- standardised processes and standardised software for processing of the data in the data archive: both pre-planned and ad hoc statistics production

[1] The term ”metadata” was introduced by Sundgren (1973).

The database concept

One important step in the historical development of the concept of a statistical data warehouse was the growing interest for **databases** among researchers and practitioners in information systems and software engineering. The first international conference on database management was held in the small village of Cargese, Corsica, in 1974; Klimbie & Koffeman (1974). Very much of the focus of that conference was on the emerging **data models** that were proposed as foundations for standards for the design of databases so-called **database management systems**, generalised software products for managing databases. The major contendors at the Corsica conference were the network model and the relational data model. The network model was being developed by the ANSI/CODASYL standardisation organisation, and was supported by all major computer manufacturers with the exception of IBM. The relational data model was being developed by researchers associated with IBM, such as Mike Senko, Ted Codd, and Chris Date.

Data models and conceptual data models

Both the network model and the relational data model were data models in the literal sense – models of physical data representations. At the Corsica conference some researchers also presented more abstract, **conceptual data models**, aiming at modelling the conceptual information contents and the real world represented by the data. Thus, for example, Sundgren presented his infological approach, Abrial presented a model called Data Semantics, etc. The ideas of these more conceptually oriented researchers were that a user should not have to know anything about how the data were stored on different media – instead they should be able to communicate with computerised systems in terms of concepts and models that were relevant for them, and these concepts and models would then be mapped into more computer-oriented data models by the database management software.

Conceptual models, like the infological approach, were received respectfully at the time when they were launched, but they were regarded as far too abstract for being feasible for software implementation – a criticism that also hit the relational data model, by the way. However, a few years later the ideas were picked up again by many researchers. These conceptual models had names like the Entity Relationship (ER) model, or the Entity Attribute Relationship (EAR) model, and they were very similar to the infological model, or Object Property Relation (OPR) model, suggested by Sundgren (1973). These conceptual models have since then become mainstream models used by most practitioners in the field when designing databases. Some software developers have also taken up the ideas of using more conceptual data models under the umbrella of object-orientation and object-oriented databases.

Data independence and data/metadata-driven systems

The introduction of the database concept also meant the breakthrough for so-called **data independence**. Until then, computer programs had usually contained within themselves, not the actual data that they were processing, but the descriptions of the data, the metadata or data declarations. If the organisation or the storage of the data was changed, ever so little, modifications of the software source code had to be carried out, and the programs had to be recompiled. This was a rigid, time-consuming, and error-prone process. By placing the metadata outside the computer programs, as part of the database containing the data to be processed by the programs, the processing became much more streamlined and flexible. The computer programs were designed to access the database whenever they needed data, first for the data descriptions, and then for the data themselves. If the data and metadata had changed since last time the program had been executed, the program would automatically adapt to the changes, since it would automatically get the new, updated data descriptions, when it accessed the database for metadata and data.

A more appropriate and exact term for data independence would be software/data/metadata independence. Software applications built on the principles of software/data/metadata independence are sometimes called data/metadata-driven systems.

As a practical example of metadata-driven systems in a statistical environment, we could consider the use of classifications stored in a classification database. Classifications are usually subject to modifications and even more drastic structural changes from time to time, and new versions of the classifications will occur as results of these changes. It is certainly essential that an application program, when processing certain statistical data, will use the right version of classifications with regard to these statistical data (which may be the most current data or historical data). If the software applications are forced to “consult” the classification database, when accessing the data, this could automatically ensure that the right classfication version will be selected.

Database management systems (DBMS)
A database management system (DBMS) is a generalised software product for managing databases. The breakthrough of the database concept among information system practitioners created a huge market for such software, and the need for standardisation was imminent. This intensified the struggles between advocates of different data models that had started within the research community as described above. Although many of the researchers developing the relational data model were financed by IBM, IBM was not immediately ready to give up its investments in old-fashioned hierarchical database management software. Thus small garage companies, like ORACLE, got a chance to establish themselves within the small, academically oriented relational database niche, before it was obvious that relational database management systems would become the *de facto* industry standard for a long time. At this time IBM was ready to give full support for its relational database management system DB2, and later another small garage company called Microsoft was born and (much later) jumped on the bandwagon to compete with ORACLE and IBM for the database management software market with its SQL Server; SQL, pronounced “Sequel” was originally a research prototype, which was presented along with other similar products at the above-mentioned Corsica conference.

Statistical databases and statistical database management

Most databases in commercial environments were, and still are, so-called **transaction databases**, that is, the objects about which data are stored and processed are most of the time business transactions, e.g. a bank customer making a deposit or a withdrawal. Naturally “status-holding” objects, like customers and accounts, are also important objects in business applications, but it is the capacity to manage large volumes of transaction data concerning individual objects (customers, accounts) that determines the performance of a database management system for business purposes.

The requirements of statistical systems are quite different. Individual transactions are certainly important during the data collection and data preparation stages, but even then, the transactions typically occur in relatively well organised batches, and furthermore the batches of transactions associated with a typical sample survey are relatively small, and the batches of transactions associated with big censuses or administrative registers are well known in advance and can be planned for by the statistics producer; there is no customer waiting for money at a cash machine.

On the other hand, statistical systems require database management software that is able to react in a flexible and efficient way to requests for **statistical outputs** (e.g. tables and graphs), which may require dynamic processing of millions of database records, stored in different files that need to be linked together ("joined" as the term is in the relational data model) on the fly, unless the requests have been foreseen in advanced, and the requested aggregated data have been stored in multidimensional structures typical for statistics.

The special requirements on database management systems implied by statistical systems have been analysed and discussed among statistical and scientific organisations for a long time, and a conference series on Statistical and Scientific Data Base Management Systems (SSDBMS) has been running for many years. However, since statistical systems and statistical applications show a requirement profile that is quite different from the typical requirement profiles of traditional business applications, no commercial market for statistical database management systems has existed for a long time. With the growing interest for data warehouses in business companies, this has finally changed.

Statistical data warehouses

Two key concepts in business-oriented information management are "enterprise system" and "data warehouse". Software firms are developing and marketing standard software products supporting information systems based on these concepts.

When the ideas of statistical data warehouses were launched by Nordbotten, and long after that, there was no interest among commercial software developers to support such ideas, at least not to their full potential. The introduction of data warehouses in business environments has changed the situation, since applications based upon data warehouses have many similarities with production and usage of statistics: input data to data warehouses often come from many different sources, e.g. different operational applications; data warehouses again are used for more analytical purposes, like strategic decision-making. All these circumstances, input from multiple, possibly incompatible sources, output for multiple, possibly incompatible purposes, quite different from the purposes of the applications feeding the data warehouses with input data, call for elaborated metadata of similar kinds as are needed by statistical systems – for very much the same reasons.

Enterprise systems and data warehouses in business environments

An enterprise system typically supports a number of common business operations, such as the management of customers, orders, inventory, personnel, accounting, etc. In order to be able to support the business operations in a standardised way, the software products must be based on standardised business models, and the enterprises adopting the software products in their information systems have to adapt to these models.

An enterprise system uses and produces a lot of data. In order to exploit the full potential of an enterprise system, these data (and the concepts behind the data) have to be co-ordinated, so that the different components of the system, corresponding to different business operations, can communicate and share data between themselves, wherever applicable. For example, an order, a product, an employee, and a customer should be defined in the same way in all components of an enterprise system. This may lead to an integrated, corporate data model, based upon standardised concepts and an integrated, corporate business model.

Thus a successful implementation of an enterprise system may lead to a set of well integrated databases supporting the operations of the business. The next step may be to make these data available also for non-operational purposes, notably so-called directive or analytical purposes, such as planning, decision-making, evaluations, etc. This requires the data to be organised in a different way than they are organised in the databases supporting the operations of the business. The business processes typically require information systems that are very efficient in handling individual transactions, e.g. an order from a customer. The data used by operational systems must be correct and up-to-date. Analytical systems, on the other hand, will have characteristics which are very similar to those of statistical systems. Analytical applications are typically based on large sets of data, often current data in combination with historical data in the form of time series. Data warehouses are basically databases based on operational data but adapted to the needs of analytical applications.

A data warehouse may be seen as a part of an enterprise system, but it is more common to see it as a counterpart in the sense that it is a system in its own right, co-operating with an enterprise system. The enterprise system will feed the data warehouse with input, which may be used for many different estimation processes and statistical analyses.

A data warehouse, as it is implemented and supported by commercial software firms, will typically be based on a combination of microdata and macrodata. The microdata, e.g. data about individual business transaction, customers, employees, products, etc, are tapped from the operational systems at certain time intervals, e.g. once a day. Data coming from different components in the enterprise system, and from different points in time, may be inconsistent because of errors and less than perfect integration in the enterprise system. The inconsistencies may require some “cleaning” of the microdata before they are allowed to enter the data warehouse. (Note the parallel between this so-called **data cleaning** and what we call **data editing** in connection with statistical systems.)

Once the data have entered the data warehouse, they are redistributed into another kind of database structure than is typically used in the enterprise system. The data structures of a data warehouse are adapted to the analytical needs. A data warehouse should be efficient in responding to requests for aggregated data, macrodata, based upon large sets of microdata. The requests are often *ad hoc*, that is, they cannot be exactly foreseen in advance, neither as regards contents or time. On the other hand the data in a data warehouse is usually archival in the sense that it is never, or at least very seldom, changed after it has entered the data warehouse (after the cleaning process). The data warehouse is only updated when new generations of microdata are added to the data warehouse, typically at certain time intervals, as was just mentioned.

One important consequence of the differences in requirements between data warehouses, on the one hand, and databases supporting operational systems, on the other, is that the data in a data warehouse do not necessarily have to adhere to the non-redundancy requirements that are usually essential in an operational systems. On the contrary, it may be very efficient to replicate data in a data warehouse in many different way, so as to be able to respond to many different *ad hoc* requests, which may require data to be structured in many different ways.

In particular, a data warehouse will often contain the original microdata, as they were received (after cleaning) from the operational systems, together with some macrodata, aggregated from the microdata. The macrodata are chosen so as to correspond to frequent output requests, and especially to those requests which would otherwise require very complex and time-consuming *ad hoc* processing of microdata.

So what does the growing popularity of data warehouses in business environments mean for statistical systems? Obviously there are a lot of similarities between enterprise systems and data warehouses on the one hand, and statistical systems on the other. One important implication is that there will be a much bigger market for a kind of software products that were previously regarded as “niche products” for the statistical market. We have already seen a rapid development of so-called OLAP products, where “OLAP” stands for On Line Analytical Processing. We may also hope for a growing interest among software firms in developing products supporting more sophisticated metadata management than non-statistical organisations have felt a need for up to now. When statistics production and statistical analysis becomes a more common and routine part of business information systems, the businesses are likely to discover similar kinds needs for metadata and metadata management that we have since long been aware of in the statistical world, but for which we have by and large lacked adequate software support; commercial software firms have not been interested in developing such software, and for statistical agencies such software development has often turned out to be too complex and resource-consuming.

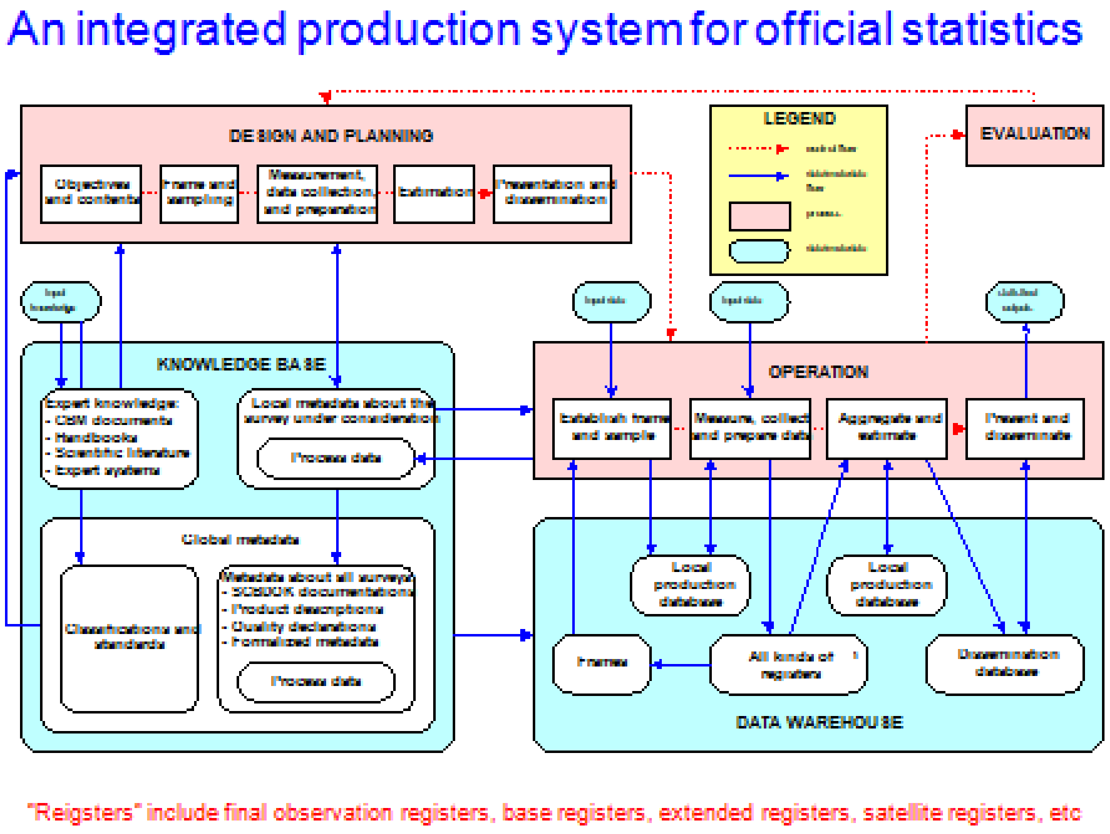
The roles of data warehouses in statistical systems

As was mentioned earlier, Svein Nordbotten launched the data warehouse concept to the statistical community already in the early 1960’s, and he actually lined out in quite some detail, how a technical system based on those ideas could be implemented already with the computer technology that was available at that time. For various reasons, mainly organisational reasons and a general conservatism among statisticians, Nordbotten’s ideas have not started to get implemented on full scale until recently – and probably most people now implementing the data warehouse concept in statistical offices have forgotten who invented it, or they have never even heard about it.

The most obvious role for a data warehouse in a statistical system is as a well organised repository for final observation registers and statistical output databases. From an external user’s point of view, it should be possible to see the whole clearinghouse, or data warehouse, as one common resource, with one common interface, through which the user could put his or her requests for statistical data and receive relevant replies in return to the requests. If a user is only interested in statistical end-products in the form of aggregated statistics presented in tables or graphs, he or she should not even have to know from which internal statistical system the end-products come, or if the statistics are already available in preaggregated form or have to be aggregated from microdata on demand.

Up to now the development in most statistical organisations have focused on this first role of a data warehouse. In particular, the focus has been on making the final statistics (macrodata) produced by a statistical organisation available to external users via the Internet. Such data warehouses are also called **statistical output databases**.

Figure 6 illustrates the architecture of an integrated, data- and metadata-driven integrated system for production of official statistics.


An integrated production system for official statistics

"Reigsters" include final observation registers, base registers, extended registers, satellite registers, etc

Figure 6. An integrated production system for official statistics.

However, in addition to this, a statistical data warehouse may also serve other functions and have other roles in a statistical organisations. Figure 7 *Figure* gives an overview of how an advanced statistical warehouse, consisting of an **input data warehouse** and an **output data warehouse**, both including **metadata**, could be the nave of virtually all activities in a statistical organisation.

For the internal users, the producers of statistics, a data warehouse could be seen as a common clearinghouse for data and metadata, a shared resource, from which both raw data/metadata, semi-processed data/metadata, final products, and other sharable data/metadata resources (e.g. registers and classifications) can be retrieved and combined for further processing, resulting in new and/or updated data/metadata, belonging to the categories just mentioned, which are then returned to the data warehouse to be used by other processes and producers/users.

The statistical output databases have not included **microdata**. At best there have been equipped with certain pre-defined routines, supported by standard software, for tapping off the inputs to the output warehouse from the regular statistical production processes. In some advanced cases this has even been done in such a way that the statistical output database is updated first, and after this the regular outputs, e.g. the statistical reports, are produced by standard software from the statistical output database. This latter solution is of course much more efficient, and is finally becoming more common. However, since it requires some co-ordination and subordination to standards, it has been difficult to achieve in stovepipe-organised statistical agencies.

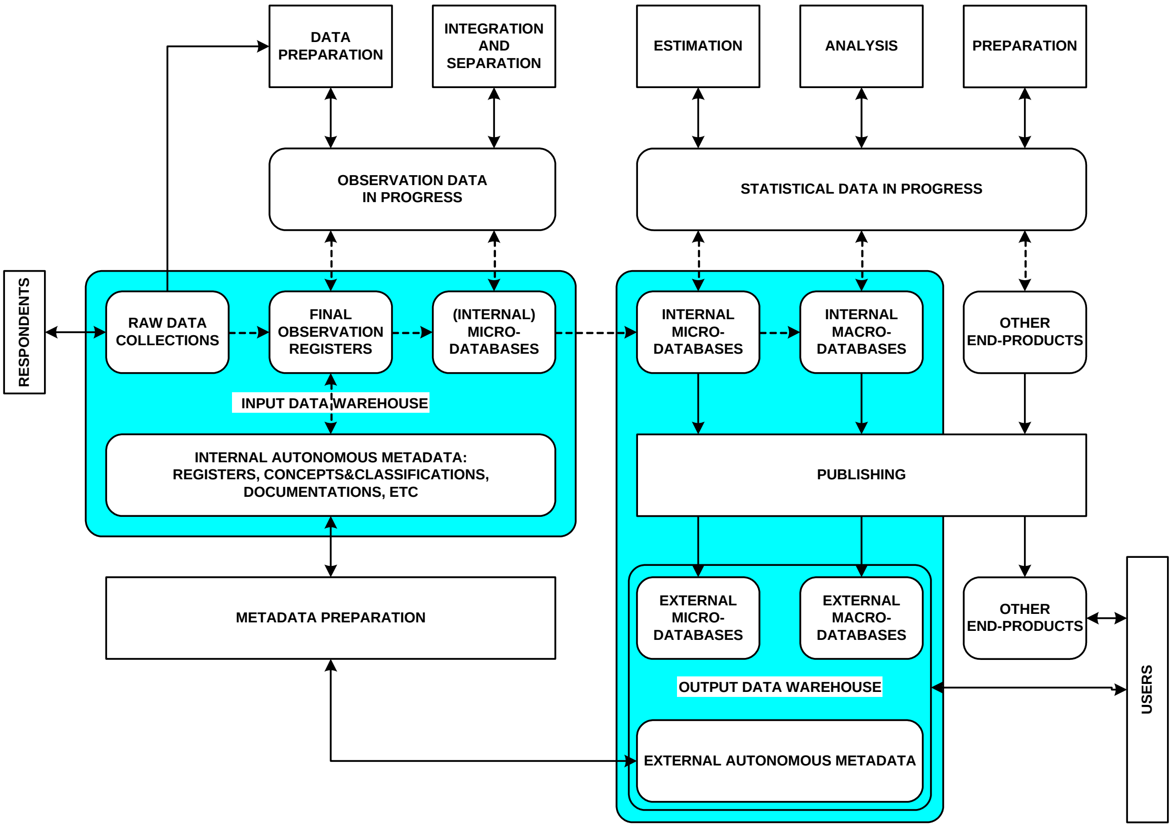


Figure 7. Schematic view of a statistical data warehouse and its components.

There are several good reasons why microdata should also play a role in connection with statistical output databases. One is that it may sometimes be more efficient, and provide more flexibility, to derive macrodata from microdata on demand rather than storing them permanently.

Another reason is that certain categories of users of statistical data, e.g. researchers and advanced analysts, are at least equally interested in microdata as in macrodata. Certainly a lot of precautions have to be taken by the statistics producers, if and when microdata are made available to such users, but assuming that these problems have a solution, there is no reason why a user of statistical data should not have all the data available from the statistical office presented in a uniform way, through one gateway, regardless of whether the data are microdata or macrodata.

Some interesting models for making microdata available to researchers in an attractive way have been launched by several statistical offices during the last few years.

If we return to the internal usages of statistical datawarehouses, there is a recent trend in statistical offices to develop input data warehouses. One purpose of an input data warehouse is to support the more flexible exchange of input data between statistical surveys necessiated by the efforts to co-ordinate data collection processes in a better way from the point of view of respondents. Ideally data providers should be able to provide their inputs to the statistical processes as by-products of the operational work that they are doing anyhow as part of their own business. This will create a need for an input data warehouse functioning as a clearinghouse between the input data collection processes and the statistical production processes aiming at different statistical outputs.

Figures 8 and 9 elaborate further on the subsystems of a statistical system, and on the interactions between them, including both **data flow interactions** and **control interactions**. Figure 8 illustrates a statistical system as seen from a design/evaluation, monitoring, and control point of view. **Metadata** (data about data) and **process data** (data about processes, also called **paradata**) are important components in the control of a statistical system. Figure 9 focuses on the main operations of a statistical system.

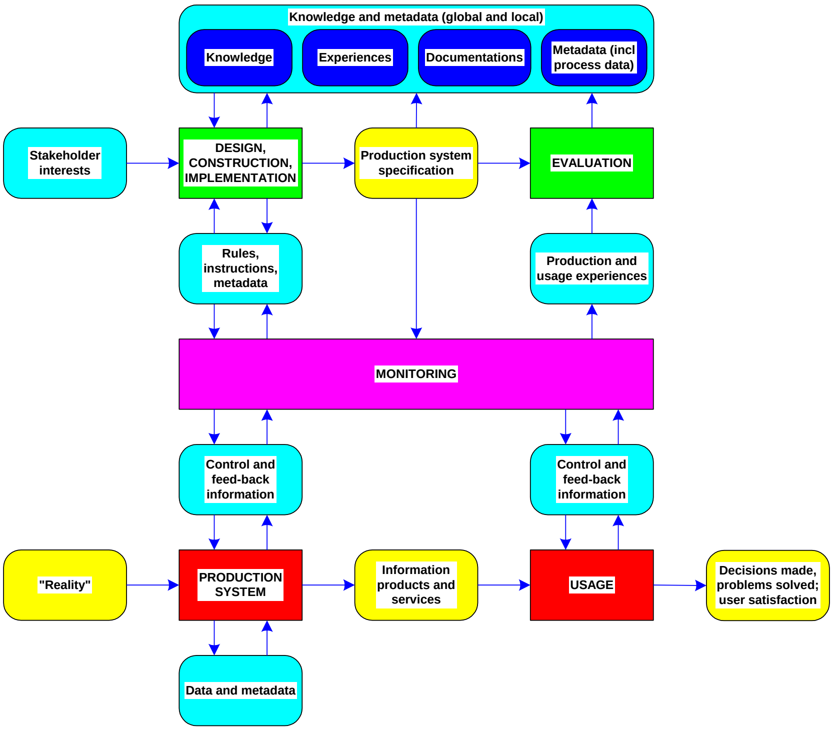

Knowledge and metadata (global and local)
Knowledge
Experiences
Documentations
Metadata (incl process data)
Stakeholder interests
DESIGN, CONSTRUCTION, IMPLEMENTATION
Production system specification
EVALUATION
Rules, instructions, metadata
Production and usage experiences
MONITORING
Control and feed-back information
Control and feed-back information
"Reality"
PRODUCTION SYSTEM
Information products and services
USAGE
Decisions made, problems solved; user satisfaction
Data and metadata

Figure 8. Control and execution of a statistical production system.

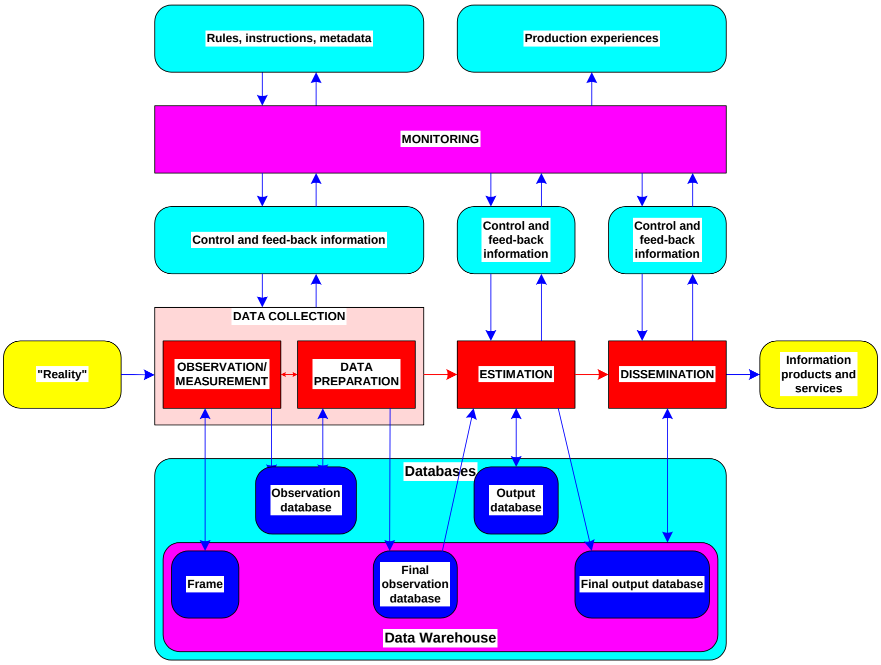


Figure 9. Basic operations in a database-oriented statistical production system.

Standardisation of data: canonical forms and normalisation

The growing interest in databases, both among researchers and practitioners, also led to standardisation efforts as regards the representation and structuring of data.

Microdata

One of the strengths of the relational data model was that it introduced a rigorous standard for data structures in a database. From a practical point of view the standard had great similarities with so-called **flat files**, which had been in use since the childhood of computers in the form of 80-column card decks.

The relational data model can be seen as a theory of flat files based upon the concept of a relation in the sense of the mathematical set theory. A relation in this sense is defined as a set of n-tuples, and such a relation could be viewed as a matrix or table, where the rows correspond to the n-tuples, and the i:th column corresponds to the i:th component of the n-tuples. From a conceptual point of view the rows correspond to objects, and the columns correspond to variables of the objects.

The relational data model puts a number of restrictions on the contents of the rows and the columns and the cells of a relational table. For example, the value in a cell must always be regarded as atomic, no substructure can be assumed or modelled. Furthermore, each relational table should have a column, or combination of columns – the so-called **primary key** – the values of which uniquely identify rows in the table; there must not be two identical rows in a table.

The theory of relational data models formulates a number of **normal forms** (first, second, third, etc) which stipulate stricter rules for the contents of a relational table. The purpose of these rules are to ensure that the database will not contain any redundant data in the sense that one and the same "fact" is stored in more than one place. In addition to wasting space (which is not so serious if space is cheap), redundance in stored data may be a threat to data consistency.

Provided that a database is designed on the basis of a sound conceptual model, like the kind of models suggested by Sundgren (1973) and others, the set-theoretical normalisation exercises recommended by the literature on the relational data model become rather superfluous. Basically a conceptual model transformed into a relational model according to certain simple rules suggested by Sundgren in several papers, will automatically become normalised.

Macrodata

Flat files and relational data tables lend themselves in a very natural way to standardised storing of statistical microdata. Actually the same standard formats can also be used for storing macrodata, and this is done in many statistical database management systems, at least on a basic level. However, another type of standard format, so-called multidimensional cubes, or **hypercubes**, are often used in software for managing statistical macrodata, typically as a user-oriented layer on top of either a relational database or some proprietory non-standard format.

One hypercube format for storing aggregated statistical data is described in Sundgren (1973). This so-called multidimensional box format, or alfa-beta-gamma-tau format is further developed by Sundgren in later papers, e.g. Sundgren (2001).

Metadata

Templates and formats for storing and communicating statistical metadata (including doccumentation and quality declarations) in a standardised way have been discussed for many years, but so far there are no widely accepted standards. However, there is maybe what one could call an emerging consensus about some major concepts and contents components of such standards. See for example Rosén & Sundgren (1991) and Eurostat (2003 a,b).

The data/metadata lifecycle

Figure 10 illustrates what is sometimes called the statistical data/metadata life cycle or value chain; cf Porter (1985), Sundgren (2003a), Sundgren&Lindblom (2004), ECB&Eurostat (2003).

During the life cycle statistical data and accompanying metadata pass through four relatively well-defined stages, corresponding to forms and interfaces:

Stage 1: The input data stage: the input data/metadata as registered on some kind of input form, e.g. a completed (paper or electronic) questionnaire.

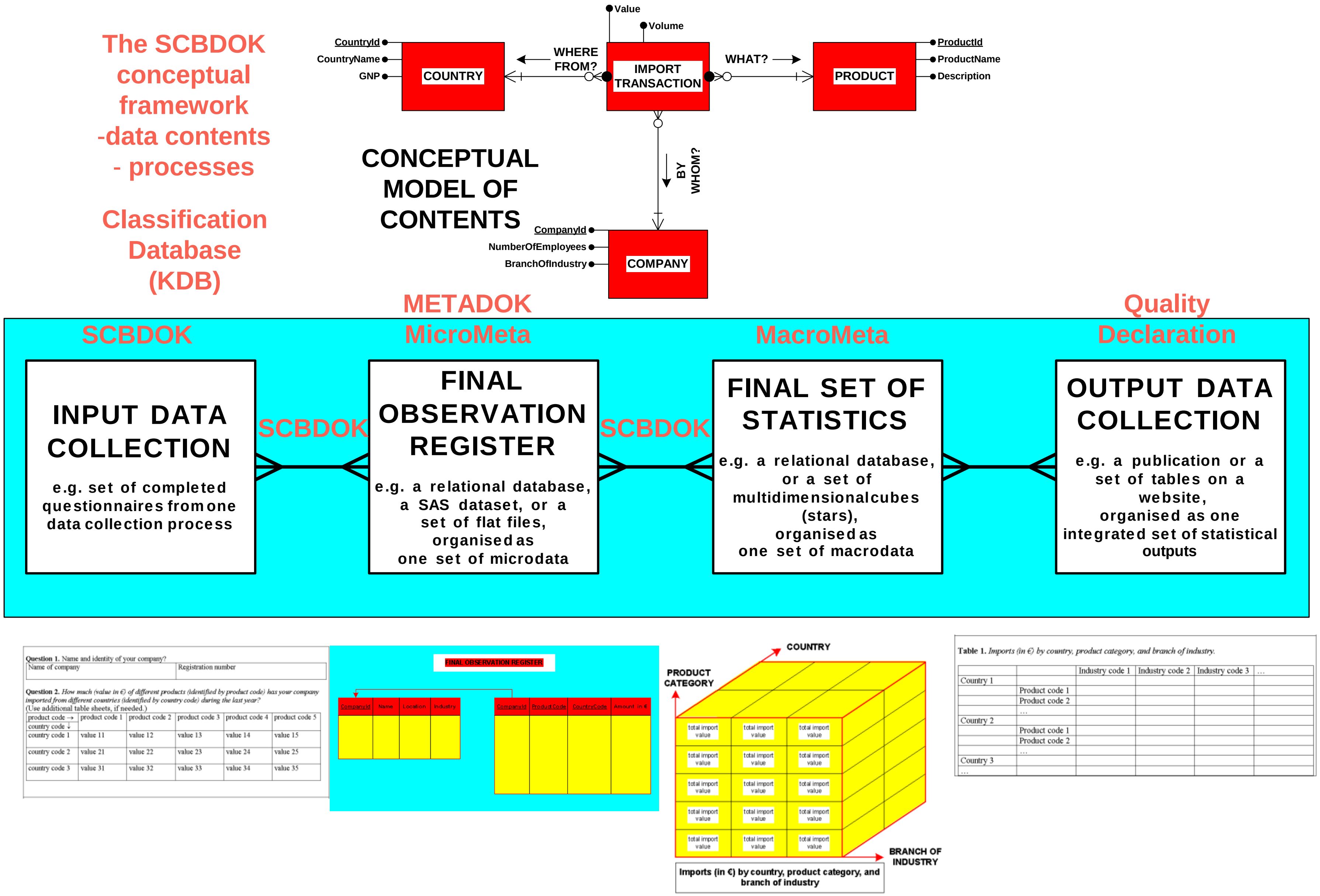
Stage 2: The final microdata stage: the input data/metadata as finally stored in some kind of final observation register, e.g. a relational database, after data preparation operations such as coding, editing, and other transformations (e.g. computation of derived variables).

Stage 3: The final macrodata stage: the output statistics (estimated values of statistical characteristics) and accompanying metadata as finally computed and stored in some kind of output database.

Stage 4: The output product stage: the statistical data/metadata as published and disseminated via printed and electronic media.

As indicated by the “fork arrows” (>—<) between the four boxes in figure 1, there are “many-to-many”-relationships between the four stages, i.e. the same input observations used in several observation registers, each one of which may be used in the production of several output data sets, which again may be combined into several statistical end-products; and vice versa: a certain statistical end-product may be based upon several output data sets, each one of which may be derived from several observation registers, which again may be the result of combining input data from several sources.

It should be noted that these complex relationships between inputs, throughputs, and outputs already exist to a high degree in modern statistics production, although most statistical offices are still organised according to the traditional stovepipe model; the production system logic is not necessarily isomorphic with the organisational structure. This is something that has to be carefully considered when designing statistical production systems.



The SCBDOK conceptual framework -data contents - processes

Classification Database (KDB)

CONCEPTUAL MODEL OF CONTENTS

INPUT DATA COLLECTION

e.g. set of completed questionnaires from one data collection process

FINAL OBSERVATION REGISTER

e.g. a relational database, a SAS dataset, or a set of flat files, organised as one set of microdata

FINAL SET OF STATISTICS

e.g. a relational database, or a set of multidimensional cubes (stars), organised as one set of macrodata

OUTPUT DATA COLLECTION

e.g. a publication or a set of tables on a website, organised as one integrated set of statistical outputs

Question 1. Name and identity of your company?

Name of company	Registration number

Question 2. *How much (value in €) of different products (identified by product code) has your company imported from different countries (identified by country code) during the last year?*
(Use additional table sheets, if needed.)

product code → country code ↓	product code 1	product code 2	product code 3	product code 4	product code 5
country code 1	value 11	value 12	value 13	value 14	value 15
country code 2	value 21	value 22	value 23	value 24	value 25
country code 3	value 31	value 32	value 33	value 34	value 35

FINAL OBSERVATION REGISTER

CompanyId	Name	Location	Industry

CompanyId	ProductCode	CountryCode	Amount in €

Imports (in €) by country, product category, and branch of industry

Table 1. *Imports (in €) by country, product category, and branch of industry.*

		Industry code 1	Industry code 2	Industry code 3	…
Country 1					
	Product code 1				
	Product code 2				
	…				
Country 2					
	Product code 1				
	Product code 2				
	…				
Country 3					
…					

Figure 10. The data/metadata lifecycle in a statistical production system.

More about the technical aspects of a statistical system

When designing information systems, one may use standardised structuring methods and architectures, such as database orientation, process orientation, client/server architecture, and service orientation.

Today's applications are often **database-oriented**, that is, different functions of the system interact with each other via a common database, including both data and metadata.

Until recently, database-orientation has often been combined with a structuring of the information system according to the **client/server** principle. In its original form, the client/server architecture consists of two types of subsystems: (a) user-oriented **client systems**, which are served by (b) **server systems**, handling common resources like printers and databases. There are developments of the client/server architecture, using three or more types of subsystems, called tiers. In a **three-tier client/server architecture** there is a distinction between

- subsystems for user interactions
- subsystems for business logic
- subsystems for data management

With the rapidly growing importance of the Internet and web-based information systems, the client/server architecture is becoming replaced by **service-oriented architectures** (SOA), based on well-defined, standardised services, which can be used in a standardised way, via standardised messages and communication protocols, by other services.

Service-oriented architectures are based on the following design principles; Erl (2005):

- **Loose coupling** – Services maintain a relationship that minimises dependencies and only requires that they retain an awareness of each other.
- **Service contract** – Services adhere to a communications agreement, as defined collectively by one or more service descriptions and related documents.
- **Autonomy** – Services have control over the logic they encapsulate.
- **Abstraction** – Beyond what is described in the service contract, services hide logic from the outside world.
- **Reusability** – Logic is divided into services with the intention of promoting reuse.
- **Composability** – Collections of services can be coordinated and assembled to form composite services.
- **Statelessness** – Services minimise retaining information specific to an activity.
- **Discoverability** – Services are designed to be outwardly descriptive so that they can be found and assessed via available mechanisms.

More briefly and concretely expressed, a service is a piece of reusable software, smaller or bigger, which performs a well-defined function, described in a standardised way. The service can be requested by other pieces of software, which may themselves be services, through standardised messages. The service requestor should not have to know anything about the internal functioning of the activated service, and the latter should not have to know anything about its external environment, but only perform its function and (possibly) provide a standardised response message in return. During its execution a service may itself request the execution of other services in the same way.

Service-orientation can be seen as a further development of earlier software design methodologies like modular programming and object-orientation. It is obviously well in line with the general systems approach and systems thinking; cf the description of services above with our earlier discussions of the systems concept and about how to manage complexity and unperceivable systems.

Service-orientation, as defined above, has the great advantage that it can be introduced step by step in an organisation, e.g. a statistical agency. Any large organisation today has an enormous burden of legacy systems that cannot quickly and easily be redesigned and redeveloped. A legacy system that has not been developed in accordance with modern design principles can be encapsulated into a large black box component, which is not internally consistent with service-oriented principles, but which interacts with its environment according to such principles. Of course it requires some work to develop the "sarcofag" surrounding the black box, making it look and behave like a true service to the other services in the system, with which it interacts, but this is a small effort in comparison with a total make-over or redevelopment of the whole legacy system.

Service-orientation often goes hand in hand with process-orientation. On the business level – for example the business of statistics production – the employees interact with customers, suppliers (respondents and data providers in the case of statistics production), colleagues, and external and internal service systems (typically computerised), in order to provide services, demanded by the customers, to the customers. This work may be organised into processes, preferably standardised processes, so as to ensure that the work is done according to best methods and best practices and will give the same good quality results to the customer, regardless of which individual persons are executing the processes.

Another recent trend is to replace in-house software developments, and even in-house licensing and installation of commercial software packages, with software components that are provided as services, for free or for a fee, via the Internet. This is called "cloud computing" or "Software as a Service", SaaS, and is also consistent with service-oriented architectures and process orientation.

Summary of a modern data warehouse approach to official statistics

Some important characteristics of a modern version of the data warehouse approach to official statistics would be:

- **standardised and coherent conceptual models** for all statistical data – as discussed earlier in this document

- **a standardised input data warehouse**, storing all microdata obtained from different data sources through standardised data collection procedures: traditional statistical surveys, administrative registers, and other administrative data sources

- **a standardised output data warehouse**, storing all macrodata produced through pre-planned and ad hoc aggregation and estimation processes

- **standardised input, throughput, and output processes** for the collection, transformation, aggregation, and communication of statistical data

- **standardised documentation and metadata/paradata** for supporting users' needs, and for monitoring and evaluating the quality and efficiency of the statistical production processes

- **standardised software** developed and maintained according to best practices at any point in time (service-oriented architecture, cloud computing etc)

- **communication interfaces corresponding to the needs of all important stakeholders**: respondents and data providers, production statisticians and system operators, designers and evaluators, managers, different categories of users of statistical outputs, developers and providers of additional services

- **standardised storage formats and communication procedures** for microdata, macro-data, and metadata (including process data, or paradata)

Organisational aspects of a statistical system

Systems for official statistics are organised differently in different countries. It is not easy to determine objectively which kind of organisation is best. However, in peer reviews centralised systems usually come out better than decentralised systems. Also in this area more research is needed. Sundgren&Nordbotten (2003) made an attempt, in connection with an evaluation of Statistics Denmark, ordered by the Danish government, to compare the efficiency and quality of national statistical agencies, using regression analysis, after having "normalised" the outputs of the agencies to make them comparable.

Also internally statistical organisations may be organised in different ways. Still the traditional, so-called **stovepipe** or **silo** organisation, is by far the most common one, but at least modifications of this scheme occur more and more often, and in some cases, e.g. in the U.K. a radically new type of organisation has recently been introduced, so far with good results, as reported by Stephen Penneck (2009) from the UK Office for National Statistics.

Different ways of organising a statistical organisation are discussed in Sundgren (2004a), chapter 6, and Sundgren (2004b), chapter 3.

Figure 11 provides a two-dimensional space, spanned by two dimensions: production system integration and organisational integration. Statistical organisations existing in the real world may be placed into this space. Traditionally almost all statistical agencies were in the bottom left quadrant, the stovepipe organisation with very independent subject matter areas. Figure 12 provides a more detailed illustration of the organisational space.

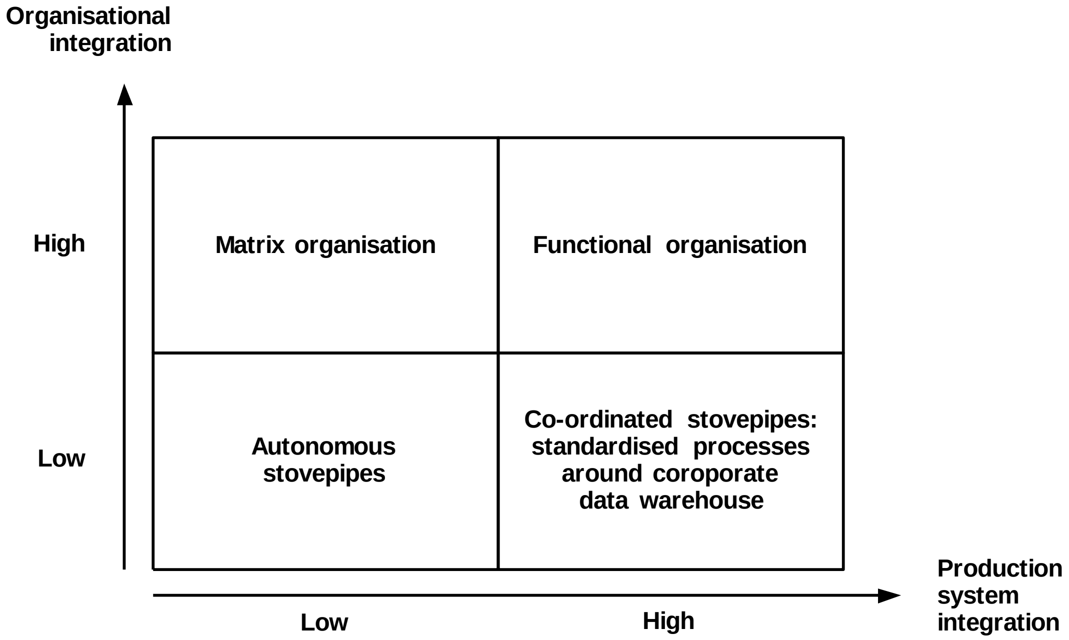

Organisational integration
High
Low
Matrix organisation
Functional organisation
Autonomous stovepipes
Co-ordinated stovepipes: standardised processes around coroporate data warehouse
Low
High
Production system integration

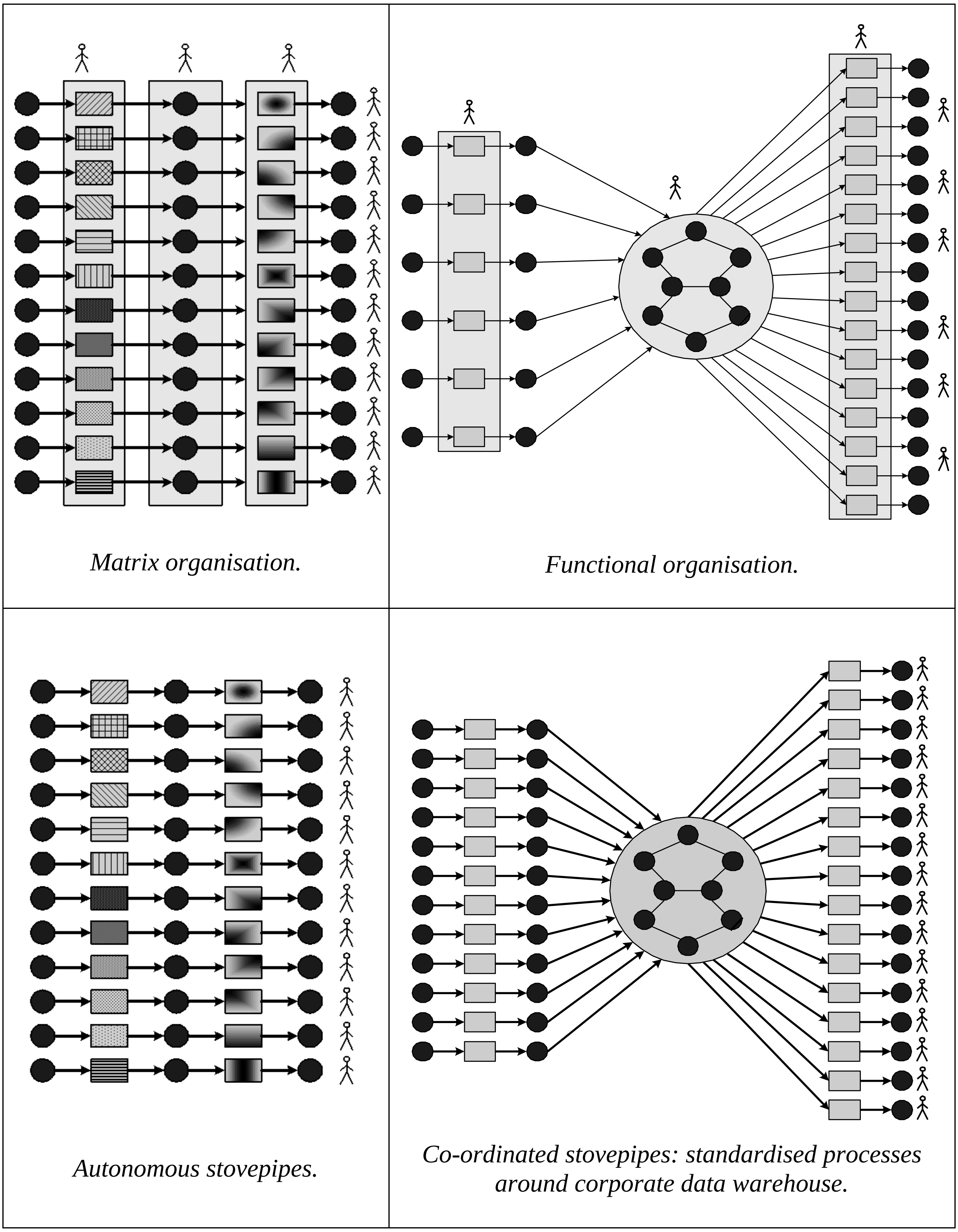
Figure 11. Different ways of organising a statistical organisation.

We will now discuss the typical four types of organisation corresponding to the four quadrants in figure 11 and figure 12.

Autonomous stovepipes

The classical way of organising the tasks of a statistical organisation is to let the individual surveys carried out by the organisation be the basis for organising staff and control. Thus there will be one organisational unit for each survey (or a number of closely related surveys), and this unit is responsible for all work that has to be done in association with the survey – from data collection to presentation and dissemination of final statistical results. Although all the stovepipes in figure 12 look the same when you look at them casually, if you take a closer look, all the process boxes are shaded in different ways, illustrating that the same things are being done in slightly different ways in different surveys. There is no exchange of data between the stovepipes. This is an exaggeration, of course. In reality some surveys will get some data from other surveys, but this is not something that can be done automatically or easily. Any exchange of data between surveys will have to be negotiated from case to case, and a survey that wants to use data from another survey will have to ask the person responsible for the other survey for permission – and most likely also for quite a lot of practical assistance, since the data are not prepared for reuse by others.

The strength of the stovepipe organisation is that the staff responsible for a survey has really full responsibility for everything that has to be done in connection with the survey, and as a result of this, they become very knowledgeable about all aspects of the survey. In this kind of organisation, there are many dedicated professionals who have worked very long on a particular statistical survey and know everything about it.



Matrix organisation.

Functional organisation.

Autonomous stovepipes.

Co-ordinated stovepipes: standardised processes around corporate data warehouse.

Figure 12. Four ways of organising tasks and responsibilities in a statistical organisation.

The stovepipe organisation is also easy to manage in the sense that the top manager may delegate responsibilities in a straightforward way. Since there are very few contacts between the stovepipes, the conflicts between them will be rare. Every stovepipe is more or less self-sufficient. In fact it is not unusual for staff members in this kind of organisation to think that they would do even better if they were completely left alone. They are often convinced that the do not have much to gain from belonging to the statistical organisation, which to them means only unnecessary bureaucracy, which causes disturbances and delays in the “real” work that they are doing.

A general weakness of a stovepipe organisation is that there are very few incitements for compliance with standards and for contributing to a coherent statistical system managed by the statistical organisation as a whole. A user who is only interested in some particular survey may find the stovepipe organisation excellent, but most users will need to combine statistics from different surveys, and for such users the stovepipe organisation does not have much help to offer, and even if such an organisation has very service-oriented staff members, they cannot do very much, since the statistics and the processes behind them are often ill documented, and the data from different surveys are ill co-ordinated both as regards contents and form.

Co-ordinated stovepipes: standardised processes around corporate data warehouse

So what can a manager of a stovepipe organisation do in order to overcome its shortcomings and disadvantages? The lower right part of figure 12 illustrates some steps towards integration that a stovepipe-based statistical organisation may take without changing the formal organisation very much. The production systems are developed to become more standardised and compatible with each other.

It is probably the most important task for a managing director of a stovepipe organisation to recognise the inherent weaknesses of such an organisation and to try to do something about them. One possibility is for top management to encourage and enforce all kinds of standardisation that will make it easier for the users to overview available statistics, and to combine statistics from different surveys. From an efficiency point of view it is also important to encourage and enforce functional and technical standardisation of the statistical production processes, including the use of standardised tools and interfaces. The very first thing that has to be done is typically to make sure that all processes are well documented. Otherwise it is very difficult for anyone, however technically and methodologically competent, to tell what needs to be done to improve standardisation and coherence of the statistical system as a whole.

Another initiative from top management that encourages both contents-oriented and technical standardisation is to establish a statistical data warehouse, where all statistical surveys in the stovepipe organisation have to deposit “their” statistical data and make them available to the organisation as a whole and to the external users.

Matrix organisation

The upper left part of figure 12 illustrates an organisation model, where the integration process has started with the formal organisation rather than with the production systems. The stovepipes are still very present in this model, and there is one responsible manager for all processes belonging to a certain survey stovepipe. But a number of functional units, with their own managers, have been added to the organisation. For example, there may be unit and a manager responsible for data collection, and another unit and manager for data dissemination. There may be a function responsible for a corporate data warehouse with a manager in charge.

This type of organisation, where the basic stovepipes are combined with orthogonal functional structures, is called a matrix organisation. The functional units will tend to become service units only, doing whatever the managers of stovepipes tell them to do, hopefully in an efficient way, since similar tasks will now be performed on a larger scale. But the standardisation, co-ordination, and integration effects are often quite limited. The functional units and managers will have certain responsibilities, but there is a risk that their authorities will not be in balance with their responsibilities.

An advantage with the matrix organisation is that it will make inefficiencies and lack of uniformity in the traditional stovepipe organisation more visible and tangible. If the top management is alert, this may be the starting point of organisational changes that go deeper and lead to radical changes in the organisation of the production systems along the lines that we discussed above in connection with the so-called "co-ordinated stovepipes", with standardised processes around a corporate data warehouse.

There are several risks associated with a matrix organisation. The organisation becomes complex, and the distribution of authorities and responsibilities between stovepipe units and functional units will always be under debate and conflict. The top management will have to be quite active in order keep the momentum in the development towards a more integrated statistical organisation, producing more, and more coherent statistics.

Functional organisation

The fourth organisational model that we present here is the one in the upper right corner of figure 12, labelled "functional organisation". In this type of organisation the production systems are integrated in a carefully designed way, and the organisation has some functional units with real power and responsibilities. The production processes are standardised and interact with a corporate data warehouse via standardised interfaces. The data in the corporate data warehouse is conceptually and physically well co-ordinated, in such a way that data put there by different data collection processes may rather easily be used and combined, resulting in a more coherent and flexible statistical system, seen as a whole. The number of outputs from the system will grow. On the other hand, the number of input processes will be reduced, since the production processes will be better co-ordinated and more standardised. All data needed from a certain respondent by any part of the statistical organisation will be collected by the same data collection process, which reduces the efforts required from respondents.

The stovepipes are no longer visible in this model. Instead the statistical data warehouse will act as a clearinghouse function, enabling input data from different sources to be combined according to a "many-to-many" pattern into a wide range of statistical outputs, corresponding to a multitude of user needs, many of which were not known or identified at the time when the data collection processes were designed.

Network organisations

On the international level it is urgent that different national statistical systems are compatible and able to communicate smoothly. There is a movement towards network solutions, where the statistical data stay within the systems, where they are originally produced, but where clear and well-defined standards are used in order to ensure that data from different systems are comparable and consistent, and may be easily combined with each other in the production and analysis statistical data on the international level.

For political entities, like the European Union, integration of the national statistical systems is a must, but even here, or maybe in particular here, the network approach with loosely coupled systems, coordinated by standards, seems to be feasible and efficient. The SDMX standard for statistical data and metadata exchange, developed by seven international organisations in cooperation, could become, after some further developments, an important component in this on-going endeavour. For further information, see www.sdmx.org, where information about SDMX is available, and relevant documents and tools can be downloaded. See also Sundgren&Androvitsaneas&Thygesen (2006).

Implementation considerations for "Vision 404"

We will discuss some alternative strategies for implementing the "Vision 404" in the context of a future European Statistical System. First we will discuss the choice between a "big bang" implementation strategy, and an incremental implementation strategy. Then we will discuss the choice between implementing the vision on the European level first, and then on the national level, or implementing the vision first on the national level, and then on the European level, or, possibly, implementing the vision on both the European and the national level in parallel.

"Big bang" or incremental, "step-by-step" implementation

There is a choice between two major alternatives for implementing "Vision 404": (i) some kind of "big bang" strategy, and (ii) some kind of incremental, step-by-step strategy.

A "big bang" strategy

A "big bang" strategy for implementing a data warehouse approach to the European Statistical System in accordance with "Vision 404" would mean

(a) Making first a firm decision about a complete plan for the development and implementation of a data warehouse solution for the European Statistical System, followed by

(b) a systematic development and implementation of the solution, as defined by the plan. The plan would define the target system in detail, and it would contain a budget and a time schedule for the whole undertaking.

It is rather obvious that a "big bang" strategy, as just defined, would be completely unrealistic. "Vision 404" outlines a future European Statistical System which is indeed very much different from the present European Statistical System. Even under very favourable circumstances, it would take more than 10 years to develop and implement all the desirable and necessary features of the envisaged system on both the national and the European level.

An incremental, "step-by-step" strategy

An incremental, "step-by-step" strategy for implementing a data warehouse approach to the European Statistical System in accordance with "Vision 404" would mean

(a) defining and operationalising in some detail what a data warehouse solution for the European Statistical System would mean more concretely, on both a national and a European level, and what it would mean for different kinds of official statistics, e.g. national accounts;

(b) defining and operationalising a number of development and implementation steps, some of which could be carried out in parallel, and some of which are dependent on the previous, successful development of other steps;

(c) ensuring that early developed and implemented steps become useful on their own merits, even before steps to be developed and implemented later have actually been started or finalised;

(d) ensuring that later development and implementation steps can be reconsidered and revised on the basis of the results and experiences from earlier development and implementation steps.

European or national level first – or both in parallel

The "Vision 404" document contains the following passage about the so-called European approach to statistics (Art. 16 of the Regulation on European Statistics**)**:

> "The underlying idea is that having data that are reliable at the national level is a sufficient condition for having reliable data at the European aggregate level, but it is not a necessary condition. If the only purpose of the data is to provide information at EU level, there is no need for a full set of national data, and there is therefore a potential efficiency gain in the system. EU sampling is a possible approach for making this gain real. In areas where there is no need to have national data, EU sampling may lead to a reduced burden on respondents, better timeliness, and improved quality. The European approach to statistics could also include the production of European statistics by the use of non-published national contributions or contributions from a subset of Member States, as well as the use of partial information by modelling techniques."

One could drive this question a step further, and ask whether it would not be feasible for the new European Statistical System to become more like the United States Statistical System, where the design and production of official statistics is driven by statistical agencies on the federal level (top-down) rather than by coordination and aggregation of statistics produced on the state level (bottom-up); the latter approach is taken in another federal state, Germany.

Considering that the member states of the European Union are very different in many respects – administrative, cultural, etc – and are likely to remain so for a long time ahead, maybe forever, it seems wise and most practical to continue to emphasise the bottom-up approach for European official statistics, but to strengthen the European support and enforcement of the coordination and harmonisation of both administrative and statistical procedures in the member states. However, at the same time it may be interesting to make some limited experiments with truly European design and production of official statistics – especially in cases where the European level quality of the statistics produced is much more important than quality on the level of member states.

Bibliography

Achenwall, G. (1749). *Staatsverfassung der heutigen vornehmsten europäischen Reiche und Völker im Grundrisse*.

Berger , P.L. & Luckmann, T. (1966). *The Social Construction of Reality: A Treatise in the Sociology of Knowledge*. Anchor Books, Garden City, New York.

von Bertalanffy, L. (1968). *General System Theory: Foundations, Development, Applications*. George Braziller, New York.

Bruneforth, M. & Sundgren, B. (2007). *Conceptual analysis of UOE and UNESCO education surveys*. UNESCO Institute of Statistics.

Churchman, C. West (1968). *The Systems Approach*. Delacorte Press, New York.

ECB&Eurostat (2004). *CMFB Issues paper on the IT tools for the European monetary, financial and balance of payments statistics*. Joint paper by the ECB Directorate General Statistics and Eurostat for the 27th meeting of the Committee on Monetary, Financial and Balance of Payments Statistics, Luxembourg, 29 - 30 January 2004.

Erl, T. (2005). *Service Oriented Architecture – Concepts, Technology, and Design*. Prentice Hall.

European Commission (2009). *Communication from the Commission to the European Parliament and the Council on the production method of EU statistics: a vision for the next decade*, Document COM(2009)404.

Eurostat (2003a). *Definition of quality in statistics*. Eurostat/A4/Quality/03/General/Definition.

Eurostat (2003b). *Standard Quality Report*. Eurostat/A4/Quality/03/General/Standard_Report.

Hox, J. J. (1997). *From Theoretical Concept to Survey Question*. In L. Lyberg et al (eds). Survey Measurement and Process Quality. Wiley.

Klimbie, J. W. & Koffeman, K. L., editors (1974). *Data base management*. Proceedings of the IFIP Working Conference on Data Base Management. North-Holland.

Langefors, B. (1966). *Theoretical Analysis of Information Systems.* Studentlitteratur, Lund.

Nordbotten, S. (1966): *A Statistical File system*. Statistisk Tidskrift, Stockholm. Available for free downloading from http://www.nordbotten.com/frame.htm.

Nordbotten, S. (1967a): *On Statistical File System II.* Statistisk Tidskrift. Stockholm. Available for free downloading from http://www.nordbotten.com/frame.htm.

Nordbotten, S. (1967b): *Automatic Files in Statistical Systems*. Statistical Standards and Studies. Handbook No. 9. United Nations. N.Y. Available for free downloading from http://www.nordbotten.com/frame.htm.

Nordbotten, S. (1967c): *Purposes, Problems and Ideas Related to Statistical File Systems*. Proceedings from the 36th Session of the International Statistical Institute. Invited paper. Sydney. Available for free downloading from http://www.nordbotten.com/frame.htm.

Nordbotten, S. (2009). *Use of Administrative Data in Official Statistics – Past, Present and Future – With Special Reference to the Nordic Countries*. Official Statistics in Honour of Daniel Thorburn, pp 205-223.

Penneck, S. (2009). The Office for National Statistics (ONS) Statistical Modernisation Programme: What went right? What went wrong? In Proceedings of *Modernisation of Statistics Production*. International conference organised by Statistics Sweden. Papers and presentations available for downloading from the website of Statistics Sweden.

Porter, Michael E. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance"*. The Free Press, New York.

Rosén, B. & Sundgren, B. (1991). *Documentation for reuse of microdata from the surveys carried out by Statistics Sweden*. Statistics Sweden. Available for free downloading from http://sites.google.com/site/bosundgren/.

Statistics Netherlands (2004). *The Dutch Virtual Census of 2001 – Analysis and Methodology*. Referred to in Wallgren & Wallgren (2007), pp 62-63.

Statistics Sweden (2009). *Modernisation of Statistics Production*. International Conference in Stockholm. Papers and presentations available for downloading from the website of Statistics Sweden.

Sundgren, B. (1973). *An infological approach to data bases*. Stockholm University and Statistics Sweden. Available for downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (1996). *We want a user-friendly and flexible system! Designing information systems with partially conflicting and unknown purposes*. In ”Advancing your Business - People and Information Systems in Concert”, Stockholm School of Economics, edited by Mats Lundeberg and Bo Sundgren.

Sundgren, B. (2001). *The αβγτ-model: A theory of multidimensional structures of statistics*. MetaNet conference, Voorburg, the Netherlands. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. & Nordbotten, S. (2003). *Review of Statistics Denmark*. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren (2003a). *Developing and Implementing Statistical Metainformation Systems*. Deliverable D6 from EU project “MetaNet” (IST-1999-29093), June 2003.

Sundgren (2003b). *Strategies for Development and Implementation of Statistical Metadata Systems*. Invited paper for the ISI session in Berlin, 2003.

Sundgren, B. & Lindblom, H. (2004). *The metadata system at Statistics Sweden in an international perspective*. Invited paper for the conference “Statistics – investment in the future”, Prague, Czech Republic. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2004a). *Statistical systems – some fundamentals*. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2004b). *Designing and managing infrastructures in statistical organisations*. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2005). A conceptual model of society as reflected by official statistics. Statistics Sweden. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. & Androvitsaneas, & Thygesen, L. (2006). *Towards an SDMX User Guide: Exchange of statistical data and metadata between different systems, national and international*. Meeting of the OECD Expert Group on Statistical Data and Metadata Exchange, 2006. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2006). *Reality as a statistical construction – Helping users find statistics relevant for them*. European Conference on Quality in Survey Statistics (Q2006), Cardiff, U.K.

Sundgren, B. (2007a). *Process reengineering at Statistics Sweden*. MSIS 2007, Geneva. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. (2007b). *Navigating in a space of statistical surveys of society*. ICES-III, Montreal. Available for free downloading from http://sites.google.com/site/bosundgren/.

Sundgren, B. & Thygesen, L. (2009). *Innovative approaches to turning statistics into knowledge*. Statistical Journal of the IAOS, 26(2009), pp. 1-10.

Sundgren, B. (2010a). *A systems approach to official statistics*. Official Statistics in Honour of Daniel Thorburn, pp. 225–260, http://officialstatistics.files.wordpress.com/2010/05/bok18.pdf

Sundgren, B. (2010b). *Designing surveys and statistical systems - complex decision processes.* Paper submitted to the Scientific Council of Statistics Sweden.

Sundgren, B. (2010c). *Statistical file systems and archive statistics*. Paper presented at the Nordic Statistical Meeting in Copenhagen, 2010.

Zarkovic, S. (1966). *Quality of statistical data*. Food and Agriculture Organisation (FAO) of the United Nations, Rome.

Zloof, M. (1975). *Query by Example*. AFIPS NCC 1975, pp 431-438.

Wallgren, A. & Wallgren, B. (2007). Register-based Statistics: Administrative Data for Statistical Purposes. Wiley.