Reality as a statistical construction – Helping users find statistics relevant for *them*

Bo Sundgren
Statistics Sweden

mailto:bo.sundgren@scb.se

Key words: relevance, contents, accessibility

2006-05-10

1 Introduction

"Relevance" is one of six major dimensions in Eurostat's quality concept; Eurostat (2003). The other five are accuracy, timeliness and punctuality, accessibility and clarity, comparability, and coherence. Relevance must always be related to a specific use. Only a user with a specific information need may judge which statistics are relevant in that situation. What can be done in order to help a user find potentially relevant statistics and judge the relevance of retrieved statistics?

A statistical agency can do three things to help users find statistics relevant for *them*: (i) provide an overview over available statistics; (ii) provide search tools; (iii) provide informative metadata. I will discuss how to present overviews of available statistics, enabling users to match their mental conceptualisations of desired information against the agency's formalised view of available statistics.

There are many pitfalls. Different persons have different mental pictures of the real world. Persons living in the same community and working on similar problems may have similar mental frames of reference. There will be bigger differences between conceptual frameworks of people from different communities or working on different problems. General-purpose statistics should meet the needs of different people, with different backgrounds, and different tasks.

Statistical agencies usually provide two access roads to official statistics: (a) through a hiearchical listing of available statistics; (b) by means of a search function. The first method is rigid. It implies one hiearchical conceptualisation of society, one classification of real world phenomena. It is not likely that this view is shared by many users. It is more likely to fit nobody's mental picture of society. Search functions are not perfect either, even if the best ones are chosen, like the Google, since different users use different concepts and terms, and these may again be different from the concepts and terms used by statisticians designing and producing official statistics. Good thesauri may amplify the power of search functions, by making use of synonyms and other related terms, but search functions usually do not take advantage of the inherent structure of statistical information. Ideally structured search methods, free-text searches, and thesauri should be combined in a creative and intelligent way.

Human conceptualisations of reality and the problems associated with information sharing are studied by several disciplines. Philosophers speculate about the relations between the real world

and the world of ideas. Phychologists study how individuals form and develop mental models of reality. Sociologists focus how people in groups and organisations affect each other's conceptualisations; this may lead to intersubjective models of society, social constructions; Berger&Luckmann (1966). Definitions and classifications developed and used by statisticians could be seen as formalised social constructions – reality as a statistical construction. Problems come from the fact that this statistical construction may not coincide with the social constructions of users of statistics.

A simple and flexible structuring of the real world is proposed in this paper, sharable between users and producers of official statistics. It is based on a simple generic model and can be used for giving both overviews and detailed information about official statistics. It is able to accommodate a wide range of views of society. A user may, step by step, cut out a subset of potentially relevant official statistics, without having to follow any specific order or use any predefined set of terms.

2 A generic model of the contents of official statistics

Official statistics are often categorised into different domains, also called topics or subject matter areas, and statistical agencies are often organised in so-called stovepipes or silos based on these domains. Examples: Population, Education, Health, Law, Labour Market, Business Activities, Housing and Construction, Agriculture, Energy, Transports, Environment, National Accounts, Financial markets, Trade.

On a very general level all statistics are estimated values of parameters of populations of objects, where the parameters are summarised values of variables of the individual objects in the populations. Regardless of subject matter domain, a statistical agency counts the objects belonging to a certain population and summarises the values of one or more variables of the objects in the population. Very often the population (e.g. a population of Persons) is broken down into subpopulations, or domains of interest, by crossclassifying the objects in the population by means of a number of classfication variables (e.g. Sex, Region, AgeGroup).

A user of offical statistics may not be able to state exactly which parameters of which populations he or she is interested in, but faced with a short list of object types (or types of populations), and/or topics, and/or parameters/variables, he or she may be able to select a subset of official statistics of potential interest by selecting, step by step, a subset of object types (populations), variables, and parameters.

In order to be able to provide a user with short lists of object types and variables as a starting-point for the user's "drill-down" operations, we must be able to give an overview of the contents of statistics in terms of a small numer of concepts.

The populations occuring in official statistics are based upon a number of **basic object types**. Some of these object types may be described as (conscious) **actors**, objects that are capable of puposeful acting, e.g. persons and organisations (enterprises). Other objects are acted upon by the actors but are not capable of purposeful acting themselves, e.g. natural resources, products, assets; a common label for basic objects of this kind is **utilities**.

All actors may be counted in a straightforward way. Many utilities are also countable, e.g. buildings and vehicles, so-called **cardinal** utilities, but there are also utilities like oil and other substances, wealth, health, etc, which may not be counted but possibly measured in other ways, e.g. by volume, weight, or value; the latter kinds of utilities are **non-cardinal** or **collective**.

In addition to the basic object types there are different kinds of **complex object types** that are counted and/or measured in official statistics. Complex objects involve one of more basic objects. For example, an event, like a road accident may involve one or more persons and one or more vehicles. A trade transaction will involve a seller, a buyer, and a product. An employment relationship will involve a person and an organisation. Etc.

Complex object types may usually be categorised as **events/transactions** (instantaneous, without time extension) or **relationships/processes/activities** (lasting for a certain time period).

3 The statistical reality from a helicopter perspective

Imagine that you are hovering in a helicopter over the world seen through statistical glasses. Or imagine that you have a tool corresponding to GoogleEarth at your diposal to get overviews and zoom in at interesting parts of the statistical reality. Until we have such tools available we could do a lot with simpler surrogates. Figure 1 provides an overview of the world in terms of the simple basic concepts introduced so far.

4 Focused views

Given the simple overview, a user may zoom in on some part of it. Let us assume that the user wants to focus on population statistics. *Figure 2a* shows a basic conceptual model allegedly covering "all" statistical concepts occurring in population statistics. Three basic object types are **Person**, **Household**, and **Dwelling**. In addition there are compound object types of "event type", here collectively referred to as **PersonEvent**, that is, events such as **Birth**, **Death**, **Marriage**, **Divorce**, **Migration**, etc. Some PersonEvents concern exactly one Person, e.g. Birth and Death, others may be defined to concern either one or two Persons (Marriage and Divorce), and Migration is an event that concerns one Person and two Dwellings (or Locations), one ***from*** which the person moves and the other one ***to*** which the Person moves.

For each object type in the model there will be a number of variables defined. For our present purposes it is not necessary to list all these variables concretely, but they will belong to two main categories: classification variables (or qualitative variables) and summation variables (or quantitative variables). Directly or indirectly observed values of summation variables are summarised when estimated values of parameters are calculated in statistical aggregation processes associated with the production of statistical cubes. The classification variables are used for spanning dimensions of the cube.

If the user wants to know more about a certain concept (e.g. an object type, a population, or a variable), he or she should be able to "right-click" on the representation of the concept and get associated metadata directly, or indirectly through chains of dynamic links.

□ ***Actors*** □ *Variables* □ *Parameters*	□ ***Topics*** □ ***ActivitiesRelationsEvents*** □ *Variables* □ *Parameters*	□ ***Utilities*** □ *Variables* □ *Parameters*
□ **Person** □ **Student** □ **Employee** □ **Patient** □ **Client** □ **Household** □ **Organisation** □ **Enterprise** □ **AgriculturalEnterprise** □ **Institution** □ **Establishment** □ **BusinessActor** □ **Producer** □ **Seller** □ **Buyer** □ **Provider** □ **Customer** □ **Subject** □ **CounterSubject** □ **Owner** □ **Possessor** □ **Employer** □ **Employee**	□ *Topic: Population* □ **ActivityRelationEvent** □ *Topic: Education* □ **ActivityRelationEvent** □ *Topic: Health* □ **ActivityRelationEvent** □ *Topic: Law* □ **ActivityRelationEvent** □ *Topic: Labour Market* □ **ActivityRelationEvent** □ *Topic: Business Activities* □ **ActivityRelationEvent** □ *Topic: Housing and Construction* □ **ActivityRelationEvent** □ *Topic: Agriculture* □ **ActivityRelationEvent** □ *Topic: Energy* □ **ActivityRelationEvent** □ *Topic: Transports* □ **ActivityRelationEvent** □ *Topic: Environment* □ **ActivityRelationEvent** □ *Topic: National Accounts* □ **ActivityRelationEvent** □ *Topic: Financial markets* □ **ActivityRelationEvent** □ *Topic: Trade* □ **ActivityRelationEvent**	□ **NaturalResource** □ **Land** □ **Mineral** □ **Oil** □ **CardinalResource** □ **Locality** □ **RealEstate** □ **Building** □ **Dwelling** □ **Vehicle** □ **Product** □ **Commodity** □ **Service** □ **AccountingItem** □ **CashFlowItem** □ **ResultItem** □ **BalanceItem** □ **WelfareItem** □ **Education** □ **Health** □ **Wealth** □ **Security**

Figure 1. Official statistics from a helicopter perspective.

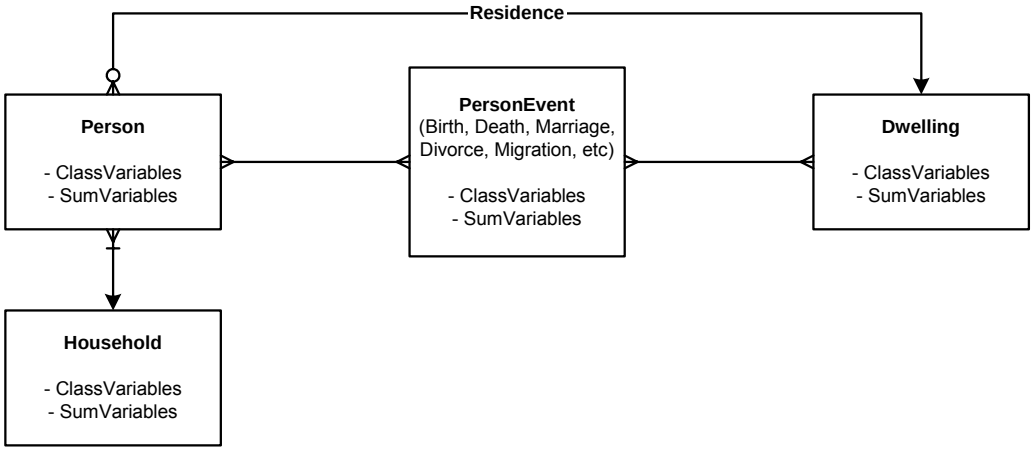

Figure 2a. *Population statistics: Basic model.*

On the basis of the simple model in *Figure 2a* many normalised cubes may be defined, covering all kinds of population statistics. **Each normalised cube is defined by putting exactly one of the object types in *Figure 2a* in focus, and by selecting variables for dimensions and parameters associated with the cube.** In the following figures we indicate the object type in focus by giving it yellow colour. Thus in figure 2b the object type **Person** is in focus. All cubes formed with **Person** in focus are associated with **Person** populations. Persons are the objects, or "statistical units" that are counted and measured, and person populations are the populations for which values of parameters are estimated. The dimensions of these cubes are spanned by classification variables of Persons (in addition to the Time and Space dimensions), e.g. Sex, HomeRegion, and the cells contain estimated values of parameters of a Person population; the parameters summarise values of summation variables of Persons, e.g. Age, Income.

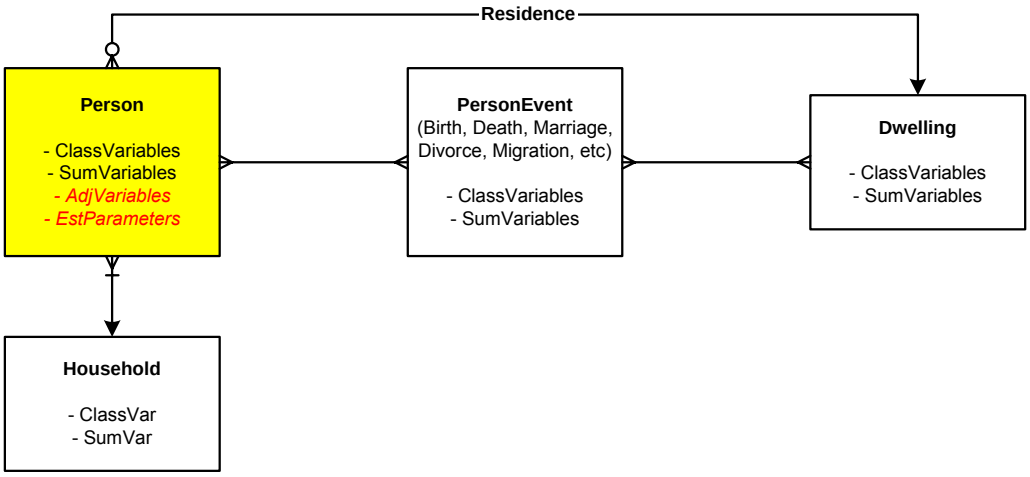

Figure 2b. *Person statistics.*

The **Person** object type is associated with certain ***basic*** classification variables (like Sex) and certain ***basic*** summation variables (like Age, Income). Further classfication variables may be defined by ***grouping*** summation variables, e.g. AgeGroup, IncomeGroup. We may also define so-called ***adjoined variables*** of the object type in focus (here Person) by adjoining variables of object types that are related to the object type in focus in a well-defined way. In this case, where Persons are in focus, the following categories of variables may be adjoined to Persons: (i) variables of the Household to which a Person belong; (ii) variables of the Dwelling in which the Person resides; and (iii) variables of the PersonEvents in which the Person is involved.

Figure 2c illustrates the situation when some kind of **PersonEvent** is made the object type in focus. Normalised cubes based on this model will be associated with PersonEvent populations, that is, it is PersonEvents like Births, Deaths, Marriages, Migrations, etc., that are counted and measured, and it is populations of PersonEvents for which parameters are estimated, e.g. the estimated number of migrations between different regions, where Region is primarily a classification variable of the dwellings or location from and to which the migrations take place. This variable (and others) will have to be (logically) adjoined to the Migration objects, before the cube can be properly defined and the requested parameter can be estimated.

In the case of normalised cubes based on PersonEvents, the originally specified basic variables of PersonEvents may be extended by adjoined variables from related object types, in this case Persons, Households, and Dwellings; for example:

- the Sex of the Person involved in the PersonEvent
- the Size of the Household to which the Person involved in the PersonEvent belongs
- the Region(s) in which the Dwelling(s) associated with the PersonEvent is/are associated

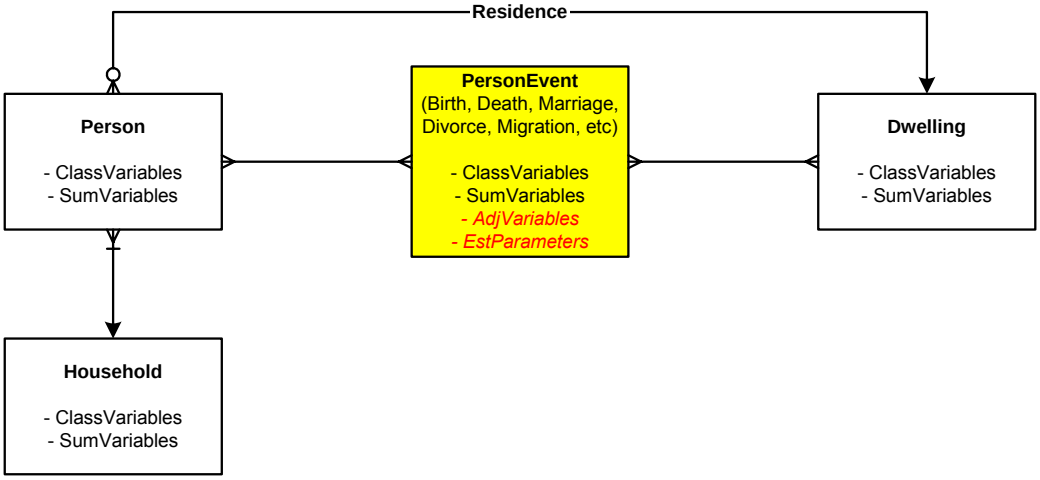

Figure 2c. *PersonEvent statistics.*

The dimensions and parameters associated with the normalised cube may of course be both basic and adjoined variables. For example, the parameter of a population of Deaths may be the average Age of the Persons who have died, and the parameter of a population of Migrations may be average Income of the Persons who have migrated, or the average change in Size of the Dwelling to which the Person moves in comparison with the Size of the Dwelling from which he or she comes.

Figure 2d and *Figure 2e* illustrate variations of the basic population statistics model where the object types **Household** and **Dwelling**, repectively, have been put in focus as the basis for normalised cubes.

Figure 2f summarises all the preceding models into one "split vision" model, indicating by yellow colour all the object types that may, one at a time, be focused as the basis for normalised cubes. In the particular case used here for illustration purposes, it actually happens that all object types may be used (and are used in actual population statistics produced by statistical agencies) as the basis for normalised cubes. However, in other cases, as we shall see examples of further on in this paper, there may also be object types that are not actually used as the basis for normalised statistical cubes, even if, theoretically they probably could be in most cases.

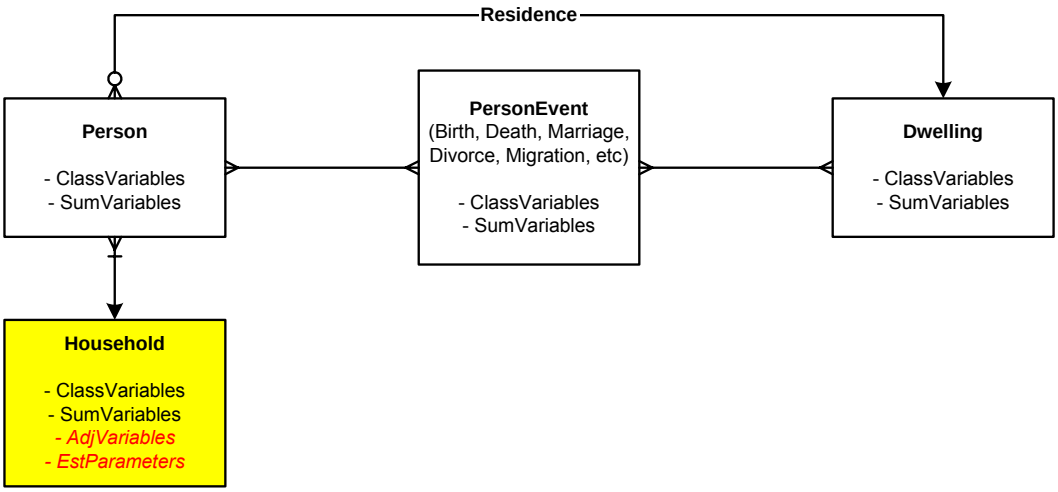

Figure 2d. *Household statistics.*

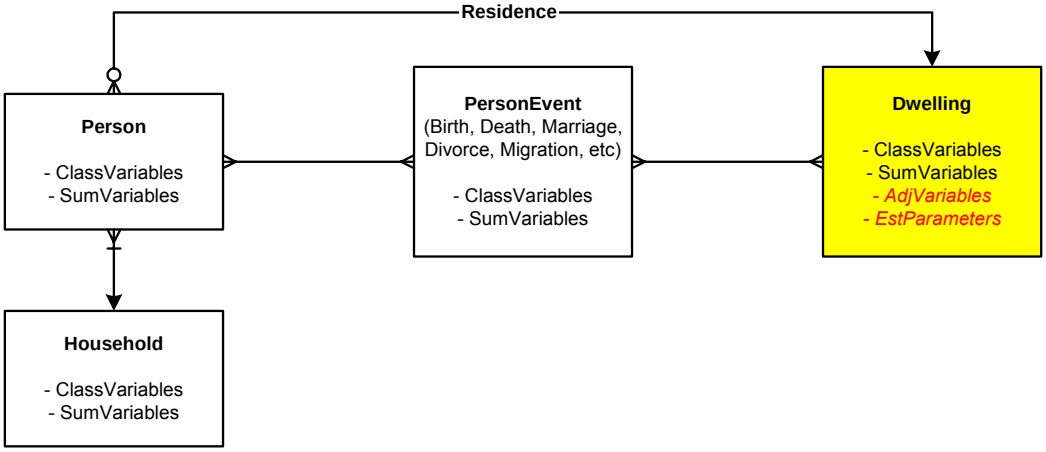

Figure 2e. *Dwelling statistics.*

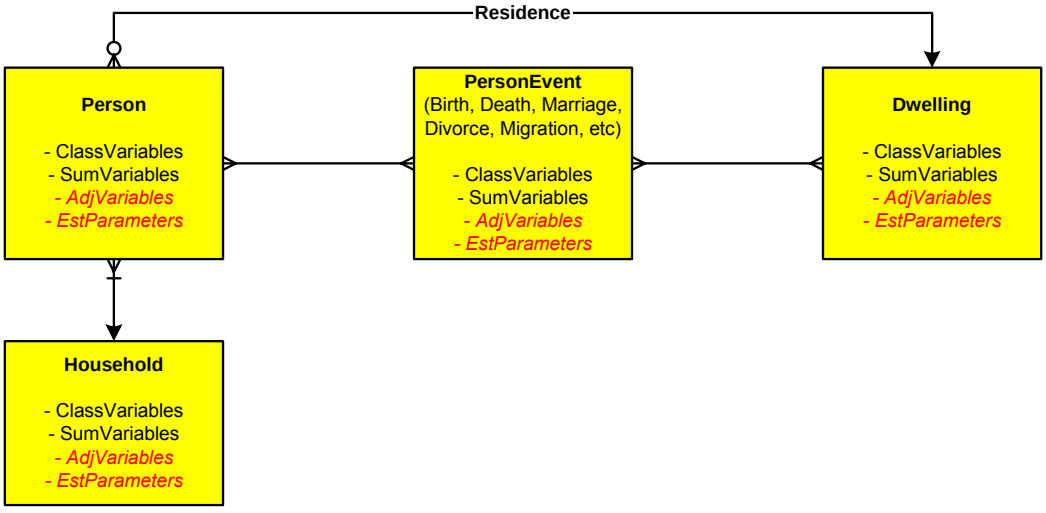

Figure 2f. *Population statistics: Split vision model.*

Figure 3 is a simple generic model that actually covers all these domain-oriented models, that is, all subsequent models can be seen as more specific interpretations of this model.

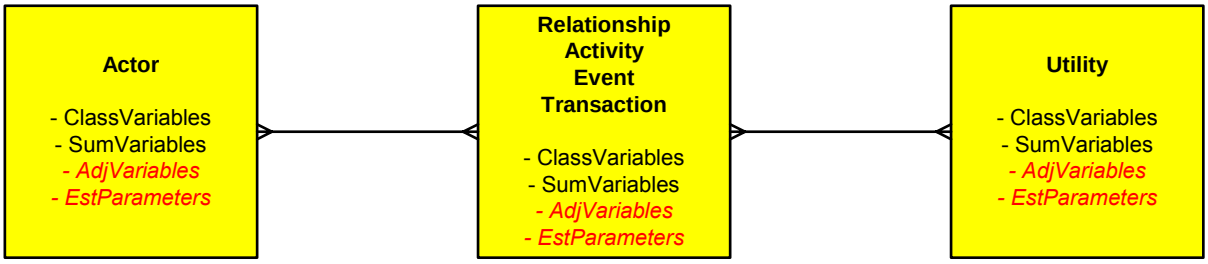

Figure 3. *Generic model, covering all statistical domains.*

The general model presented above is described in more detail and the examples elaborated in Androvitsaneas&Sundgren&Thygesen (2006).

5 Contents By Example (CBE)

Contents By Example (CBE)[1], as proposed here, is a technique for

- giving an overview of the contents of available official statistics (microdata and macrodata)
- allowing a user of official statistics to select a subset of available official statistics that he or she would like to investigate further, e.g. by
  - selecting object types of potential interest
  - opening lists of available variables for selected object types
  - opening lists of available value sets (classifications) for selected variables
  - getting definitions and explanations by moving the cursor over terms used on the screen
  - drilling down for more metadata, also about underlying survey processes, by right-clicking terms and selecting from lists of available metadata and metadata links
  - selecting time intervals and populations of interest

Once again, as often as the user wants to know more about a certain concept (e.g. an object type, a population, or a variable), he or she should be able to "right-click" on the representation of the concept and get associated metadata directly, or indirectly through chains of dynamic links.

Figure 4 (a, b, and c) provides three examples of how a data set (microdata or macrodata) could be selected in a very intuitive way. In the first example the user indicates (by #) that he or she wants to count Births of Persons. The count should be broken down according to the Person's Sex and HomeLocation (the latter variable is made more precise by qualifying it by means of a certain RegionalCode, presumably a standard classification). Furthermore, the count should also be broken down by the Age of the Person's Mother according to a certain AgeGrouping (probably also a standard classification, otherwize it has to be separately defined by the user, who may be promted to do so).

In the second example it is Migration events that are counted, crossclassified by the Migrant's Sex, Age, and Occupation, as well as by the respective Localities from and to the Migrant moves.

[1] Cf *Query By Example* (QBE) in relational database theory; Zloof (1975).

In the third example, it is the average Income (marked by an m in front of the Income label) of a population of Persons that is requested, and the figure should be broken down by the Person's Sex and Education.

In all three examples the population could be more precisely defined by adding properties to the population object type selected, and the user could be interactively assisted in doing this. Furthermore it should be noted that microdata could have been requested (instead of macrodata) just by avoiding marking any variable by a summarising function (like # or m).

	□ *Topic: Population*	
x Person x Sex □ Age x HomeLocation .*RegionalCode* □ WorkLocation □ MaritalStatus □ Income(byKind) □ Wealth(byKind) □ EducationLevel □ Occupation	**# Birth[x Person]** x AgeOfMother.*AgeGrouping* □ **Death[Person]** □ **GetMarried[Person, Person]** □ **GetDivorced[Person, Person]** □ **Membership[Person, Household]** □ **MarriedTo[Person, Person]** □ **ChildOf[Person, Person]**	□ **Locality** □ Location □ **RealEstate** □ **Building** □ **Dwelling**
□ **Household** □ Size □ HomeLocation □ Income(byKind) □ Wealth(byKind)	□ **Residence[Person, Locality]** □ **Commute[Person, Locality.From, Locality.To]** □ **Migration[Person, Locality.From, Locality.To]** □ **Immigration[Person, Country.From, Locality.To]** □ BirthCountry □ **Emigration[Person, Locality.From, Country.To]** □ BirthCountry	

Figure 4a. Example 1 of a selected data set: Number of births by sex and home location of the child and age of the mother. (4-dimensional frequency macrodata – cf underlying microdata.)

	□ *Topic: Population*	
x Person x Sex x Age.*AgeGrouping* □ HomeLocation □ WorkLocation □ MaritalStatus □ Income(byKind) □ Wealth(byKind) □ EducationLevel x Occupation .*OccupationCode*	□ **Birth[Person]** □ AgeOfMother □ **Death[Person]** □ **GetMarried[Person, Person]** □ **GetDivorced[Person, Person]** □ **Membership[Person, Household]** □ **MarriedTo[Person, Person]** □ **ChildOf[Person, Person]**	**xx Locality** □ Location .*RegionalCode* □ **RealEstate** □ **Building** □ **Dwelling**
□ **Household** □ Size □ HomeLocation □ Income(byKind) □ Wealth(byKind)	□ **Residence[Person, Locality]** □ **Commute[Person, Locality.From, Locality.To]** # **Migration[x Person, x Locality.From, x Locality.To]** □ **Immigration[Person, Country.From, Locality.To]** □ BirthCountry □ **Emigration[Person, Locality.From, Country.To]** □ BirthCountry	

Figure 4b. Example 2 of a selected data set: Number of migrations by sex, age, and occupation of migrant and place moved from and to. (5-dimensional frequency macrodata.)

x **Person** x Sex □ Age □ HomeLocation □ WorkLocation □ MaritalStatus m Income(byKind) □ Wealth(byKind) x EducationLevel □ Occupation □ **Household** □ Size □ HomeLocation □ Income(byKind) □ Wealth(byKind)	□ *Topic: Population* □ **Birth[Person]** □ AgeOfMother □ **Death[Person]** □ **GetMarried[Person, Person]** □ **GetDivorced[Person, Person]** □ **Membership[Person, Household]** □ **MarriedTo[Person, Person]** □ **ChildOf[Person, Person]** □ **Residence[Person, Locality]** □ **Commute[Person, Locality.From, Locality.To]** □ **Migration[Person, Locality.From, Locality.To]** □ **Immigration[Person, Country.From, Locality.To]** □ BirthCountry □ **Emigration[Person, Locality.From, Country.To]** □ BirthCountry	□ **Locality** □ Location □ **RealEstate** □ **Building** □ **Dwelling**

Figure 4c. Example 3 of a selected data set: Average income of persons by sex and education level. (2-dimensinal summarised macrodata.)

References

Androvitsaneas, C. & Sundgren, B. & 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, Geneva April 6-7, 2006.

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

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

Zloof, M. (1975): "*Query by Example*" AFIPS, 44, 1975.