



Organisation for Economic Co-operation and Development

## **OECD Expert Group on Statistical Data and Metadata Exchange**

**6-7 April 2006**  
**Geneva**

**Towards an SDMX User Guide: Exchange of Statistical Data and Metadata between Different Systems, National and International**  
By Bo Sundgren, Statistics Sweden, Christos Androvitsaneas, ECB and Lars Thygesen OECD

## Towards an SDMX User Guide:

# Exchange of statistical data and metadata between different systems, national and international

Christos Androvitsaneas, ECB, [christos.androvitsaneas@ecb.int](mailto:christos.androvitsaneas@ecb.int)

Bo Sundgren, Statistics Sweden, [bo.sundgren@scb.se](mailto:bo.sundgren@scb.se)

Lars Thygesen, OECD, [lars.thygesen@oecd.org](mailto:lars.thygesen@oecd.org)

Preface note: This paper is a first preliminary draft. The intended outline is shown, although some of the intended annexes have not yet been elaborated.

This paper is intended as a contribution towards an SDMX User Guide, aiming to enable partners in the international statistical community to know how to proceed towards a standardised and efficient exchange of data and metadata, using the SDMX technical standards. The User Guide endeavours to take into account, as far as possible, the systems existing in international organisations as well as in national statistical organisations, and demonstrate how data and metadata can be exchanged and shared between them. The User Guide puts a strong emphasis on using practical examples that are non-trivial and reflect real metadata as they exist in national and international organisations.

This chapter intends to show mapping between concept schemes of different organisations and the role of Cross-domain Concepts as an intermediary between the organisations. This process leads to a consolidated set of Cross-domain Concepts, building on the limited set included in the SDMX Contents-oriented Guidelines<sup>1</sup> and enhanced to be able to inter-operate with the examples of metadata systems presented. The ambition is thus to present a set of concepts that is more suited for communication between many national and international organisations. Making this communication as easy as possible and minimising the translation or conversion costs would also provide an important service to users of the data, who could then access metadata, across data sources, based on the same modelling structures and common statistical terms.

The draft user guide represents an effort to further strengthen the link between SDMX standards and current working practices followed by statistical organisations. It is, of course, relatively limited in the scope of experience, but it is intended as a basis for further discussion among national and international statistical agencies, adding to its value and eventually gathering a broad consensus in the statistical community.

## 1. Background

Up to the end of 2005, most of the resources allocated to the SDMX project have been allocated on the development of technical standards. Detailed technical standards are documented in SDMX 2.0, which is now finalised and publicly available<sup>2</sup>.

In parallel with this technically oriented work, a set of very preliminary draft SDMX Contents-oriented Guidelines has been elaborated and released for public comment in March 2006. These draft guidelines set out preliminary recommendations for classifications of metadata to be used for international exchange of reference metadata using the SDMX technical standards. It is recognised that these preliminary classifications are not sufficient to cater for all the data exchange taking place between a large number of players and in many subject-matter areas.<sup>3</sup>

---

<sup>1</sup> <http://www.sdmx.org/news/document.aspx?id=146&nid=67>

<sup>2</sup> <http://www.sdmx.org>

<sup>3</sup> However, some of these concepts could be applicable, not only for the definition of metadata, but also (partially) in the definition of data structures.

It has not been possible in the preliminary guidelines to take into account the complete work and reflect the important progress that has been made in the area of contents-oriented management of statistical data and metadata, both in national statistical agencies of the member countries of the seven international organisations in the SDMX consortium, and within the international organisations themselves. This work has mainly been carried out independently of the SDMX initiative. The integration of these efforts with the SDMX project should now be undertaken in close cooperation between experts from national and international agencies. Best practices for integrated data and metadata management, covering both technical and contents-oriented aspects, should be identified and integrated. This chapter represents an effort in this direction.

Consequently, this chapter sets out describing the state of the art concerning statistical metadata systems as part of statistical data warehouses. An important feature is the integration of isolated “stove-pipes” of statistical domains through corporate metadata principles and processes.

In the following, attempts are made to provide concrete procedural advice, illustrated by full-scale examples, on how statistical metadata produced by national systems of official statistics can be conceptually and technically transformed in order to make use of the SDMX technical standards for exchange and sharing of data and metadata. It is essential that these transformations can be done in a rational and efficient way, based upon generic standards, valid for all kinds of official statistics, and supported by generalised software tools.

## ***2. National statistical organisations – the origin of international statistics***

A large number of international organisations strive to provide statistical data that will allow for comparison between countries. The purpose is to enable the organisations to compare the level of development in some area or the efficiency of policy measures taken in the different countries.

The natural source of international statistics is in most cases the national statistical organisations (NSOs), more specifically national statistical offices and central banks who already produce and disseminate the statistical data for national use<sup>4</sup>. International organisations thus spend considerable resources on collecting statistics from these national organisations – an activity which is equally burdensome for the latter. The aim of SDMX is to render this data exchange and sharing more efficient.

The quality characteristics of statistics used for international comparison are similar to those of national statistics (accuracy, timeliness, etc.), but comparability plays an even more pivotal role in international statistics because the difficulties of comparing statistics between countries are considerable.

A major work task of IOs is obviously to agree on common standards for the data, making international comparisons meaningful. This involves setting up standard classifications, common definitions of concepts, handbooks describing conceptual frameworks and guidelines for data collections, etc. In many cases this work is carried out in cooperation between several IOs; for instance the System of National Accounts (SNA) is issued jointly by five organisations. Still, such a handbook leaves room for variation in the data requests from different organisations. The next step in the process of cooperation is to agree among organisations on exactly which pieces of data are needed. Increasingly, agreements are made between organisations of sharing of work, implying that each country only reports data to one organisation, and the organisations subsequently share the data. The ultimate step is to have a general agreement among “all” IOs, saying that in this field we will all be satisfied if countries make these exact tables available on their web sites, using as a common standard SDMX-ML conformant web services. This last step is exactly what SDMX is aiming at, and this is the reason why sponsors see SDMX as the key strategy for developing data collection or sharing.

To the end-users of supposedly internationally comparable statistics, metadata explaining comparability – or lack thereof – are of course crucial. Therefore, a considerable proportion of the work of international

---

<sup>4</sup> There are of course also cases where international organisations collect statistical data directly from individuals or enterprises in countries, or use another secondary data source in the countries such as ministries or NGOs; these cases are not the focal point of the development of SDMX and of this paper

organisations to produce and disseminate statistics is to ensure that it is accompanied by appropriate metadata. The demands for such metadata are discussed in chapter 6 below.

Another dimension of quality of statistics is “accessibility”, in the sense that data should be easily accessible and easily understood by users. The role of harmonisation of metadata also then becomes crucial: the more the metadata concepts are common across the dissemination agencies, the easier and less costly it becomes for users to access them across data sources.

### **3. National statistical systems**

Most countries have one or more national statistical organisations (NSOs), who are endowed with the task of maintaining a national statistical system. Core tasks of NSOs are to collect, process and organise statistical data, and subsequently put them at the disposal of various communities of users, i.e. disseminate the statistics. Obviously, some of the main obligations of NSOs are to make the necessary strategy decisions on what should be measured and how, and to manage and document the statistical system.

A widespread problem is lack of harmonisation across different fields of statistics in a country. This is often related to the statistics production being organised in so-called stove-pipes, or independent production lines. This makes it difficult to use statistics on different subjects in a coherent way, thus impairing the quality of statistics as seen from the user. It also reduces efficiency in the production.

To overcome these problems, there has been a strong tendency in NSOs towards standardisation and integration, breaking down stove-pipes. This leads to the creation of statistical data warehouses, bringing together statistics on different subjects under one system. In this endeavour, the creation of statistical metadata plays an important part. The changes required towards such integrated systems are not only technical, but also organisational.

### **4. National statistical metadata systems**

The character of metadata required by national statistical organisations is highly diversified, as they are intended to serve many different purposes, they emanate from a variety of different processes and sources, and they are produced by and used by a wide variety of experts or users. Also the representation and storage of metadata is often dispersed and incoherent.

The metadata audiences may include:

- Staff with different kinds of responsibility for the production process (e.g. a statistician, a developer, a manager); they will produce and/or need descriptions of the production process or system, as well as other processes related to the statistical data
- Internal or external users of the statistics (e.g. editors of a statistical compendium, news media, analysts, policy decision makers); they will need different kinds of metadata allowing them to identify and locate the data, find out what is the real information content, and what is the quality of the contents.

The structure of the metadata (from a modelling point of view) can be similar and there may also be a smaller or larger overlap between the two categories of metadata just mentioned. For instance, users of statistics may need to look closer at some of the instruments used for their collection (e.g. questionnaires) or process data (e.g. non-response figures) which can contribute to the understanding of the nature and quality of data. And, in other cases, end-users may not need access very detailed metadata that can be used by data producers for operational or specific production purposes.

In SDMX, an additional distinction is often made between structural metadata and reference metadata.

*Structural metadata* are metadata that act as identifiers of the:

- structure of the data, e.g. names of columns of micro data tables or dimensions of statistical cubes;
- structure of associated metadata, e.g. units of measurement.

The structural metadata are needed to identify and possibly use and process data matrixes and data cubes. Accordingly, in the context of a database, structural metadata will have to be present together with the statistical data, otherwise it would be impossible to identify, retrieve and navigate the data. Structural metadata will often include the following:

- *Variable name(s) and acronym(s)*, which should be unique
- *Descriptive or discovery metadata*, allowing users to search for statistics corresponding to their needs; such metadata must be easily searchable and are typically at a high conceptual level, allowing users unfamiliar with the statistical organisation's data structures and terminology; e.g. users searching for some statistics related to "inflation" should be given some indication on where to go for a closer look; for this to be useful, synonyms should be provided
- *Technical metadata*, making it possible to retrieve the data, once users have found out that they exist. These, strictly speaking, may not make part of the "structural metadata" (as long as they are included in the structural metadata for a dataset) but they are necessary elements for the functioning of the database and, thus, they may differ depending on the hosting institution.

*Reference metadata* are the metadata describing the contents and the quality of the statistical data from the user perspective. Thus, as seen by the users, reference metadata should include all of the following subcategories:

### **Box 1. Structural metadata management at the European Central Bank**

In order to exchange or share data (and reference metadata), appropriate *structural metadata* need to be defined for each of the exchanged (or shared) dataflows. These are definitions with respect to the concepts to be used, the structure of the concepts (e.g. in which order should dimensions identifying the data cube appear? at which level are metadata to be attached?) and the concept potential values (which are the relevant code lists for the coded concepts?). So, the structural metadata provide neither the numeric values (observations, aggregates etc.) nor the concrete qualitative information (values to the metadata items), they simply provide background material that allows institutions to subsequently communicate to each other their data and reference metadata. In other words, the structural metadata provide a set of statistical "linguistic vocabulary and syntax rules" to the partners (to be interpreted by their applications) to appropriately understand, store and access the data and related metadata of each particular dataflow.

Due to the reasons mentioned above, the maintenance of the structural metadata is of paramount importance not only for the institution which defines them (structural metadata maintenance agency), but also for the other partner institutions and individuals interested in the data and the metadata made available by the source institution. Usually, the institution defining the structural metadata for a dataflow is also the institution that makes it available or acts as a central hub collecting the corresponding data. For example, the European Central Bank collects data from the national central banks (NCBs) of the European Union, basing this collection on structural metadata, which it defines in its capacity as a structural metadata agency. In the Directorate General Statistics of the European Central Bank (ECB) the "structural metadata maintenance" is a clearly defined function. The responsibilities include the regular liaison with the production units, in order to address their evolving requirements, and the interaction with technical and subject matter Working Groups and other external partners (e.g. Eurostat, Bank for International Settlements - BIS) in order to ensure the co-ordination and synergy at the European and internal level. This is an effective and efficient process for all partner institutions involved, since it allows maximising the use of standards, international classifications and jointly agreed approaches in "describing" data and reference metadata (thus, further reducing the need for "mappings"). The on-going SDMX work towards content oriented guidelines, as discussed elsewhere in this paper, targets this objective in a broader context: increasing interoperability at a global level, improving the means to locate data and metadata, and minimising conversion costs.

The structural metadata also provide information to all essential internal ECB components used throughout the data life cycle: applications supporting data reception, production, compilation, aggregation, production of statistics on the web and on paper, all heavily use the structural metadata; similarly, the browsers, interfaces and search engines used by the ECB statistical data warehouse (SDW) base their functionality on the structural metadata. In SDW and, in general, in the ECB internal dissemination layers (data accessed by end-users), not only the ECB structural metadata are used, but also the structural metadata underlying the data structures of the data and reference metadata coming from other data sources (e.g. BIS, Eurostat, IMF, OECD) in an SDMX compliant and fully integrated manner.

The most up to date version of the structural metadata administered by the ECB is made available to partner institutions through a Eurostat/CIRCA page. All ECB structural metadata (concepts, data structures, code lists), for all statistical subject matter domains, become transparent and can be accessed by partner institutions through a unique file which is available in various formats (e.g. GESMES/TS, html, SDMX-ML soon). It would be ideal for any institution to easily access and use the structural metadata defined by others. Modern technologies and the use of the SDMX standards are expected to contribute to this direction.

- *Conceptual metadata*, describing the concepts used and their practical implementation, allowing users to understand what the statistics are measuring and, thus, their fitness for use
- *Methodological metadata*, describing methods used for the generation of the data (e.g. sampling, collection methods, editing processes)
- *Quality metadata*, describing the different quality dimensions of the resulting statistics (e.g. timeliness, accuracy)

The specific choice and use of reference metadata in the context of a dataset containing numeric data is prescribed through the “structural metadata” of the corresponding dataset. Metadata need to be attached to some statistical artefact: Processes, organisations, particular groups of time series, data collections, surveys (instances of data collections), raw or final data, etc. An important distinction pertains to the different “levels” of statistical data to be described:

- *micro data*: the individual objects or units in the statistics (e.g. persons, households, enterprises)
- *macro data*: aggregated numbers, normally based on micro data (e.g. number of households in a county).

There will typically exist metadata for many different versions of “the same” data, many *intermediate versions* of the data and the “*final*” versions.

Metadata may exist in many different forms and may be difficult to relate to one another. Some of the needed metadata may not exist in any formalised way, perhaps only in the mind of some expert who has produced the statistics for a lifetime (experience shows that, unfortunately, this is the case much more often than one would believe). Other metadata may exist in documents of many different forms, varying from one field of statistics to the next. Ideally, the metadata will be structured according to some general principles.

In recent years many NSOs and CBs have attempted to enhance metadata systems. Firstly, there has been a movement in many organisations towards making it clear which metadata are needed and making an effort to see to it that they are actually produced. Secondly, efforts have been made to standardise metadata within an organisation. The most ambitious attempts have aimed at integrating the metadata completely in the production systems, so that they are partly created automatically by the processes, partly used for the governance of the processes. This has been labelled as “metadata-driven statistical data management systems”<sup>5</sup>. This involves agreement on the metadata components that make up the corporate metadata system, definition of how they are to be generated and presented. Obviously there needs to be a direct connection between the statistical data themselves and the metadata that describe them, as well as links between the different kinds of metadata.

It is evident that international organisations using SDMX must be able to receive the metadata they need – which is just a fraction of the metadata generated in the national organisations – directly from these systems.

## **5. A model for statistical data and metadata**

This chapter presents a general model of the statistical reference metadata that are supposed to be maintained by NSOs. The model concentrates on the aspects of metadata that are of interest to international organisations, describing the contents (concepts) and the quality of the statistics. As mentioned above, the NSOs may gather a lot of other metadata for a number of purposes, for instance internal process control or auditing.

The purpose of this model is to seek a basis for determining, in the following chapters, how this relates with the metadata systems of international organisations and how the transmission or sharing of metadata between them could be best organised.

The model is designed to be able to describe statistical data covering a wide variety of subjects. In order to demonstrate this, it is applied to examples of real data structures and metadata stemming from widely differing fields of statistics, some of the examples being only sketched in annexes.

---

<sup>5</sup> See for instance <http://www.unece.org/stats/documents/ces/ac.71/2004/11.e.pdf>

The metadata model of course relates to the statistical data itself, the structure of which could be described in the SDMX data structure definition (structural metadata). In order to illustrate this, it has been necessary also to show a general model for the representation of the statistical data content itself with its structural metadata. This model is also applied to the same two real-world data collections. The main sample subject-matter field chosen, Population statistics (Birth statistics), is well-known from NSOs in any country of the world, and the same applies to the other examples sketched in Annex.

## 5.1. A general, domain-independent model of statistical data and metadata

SDMX is a standard for exchange of statistical data and metadata. Statistical data (macro data) may be organised in a standardised way as *datasets* structured as *multidimensional cubes*, where the *dimensions* are

- a time variable (e.g. year, year/quarter, year/month)<sup>6</sup>
- one or more space variables (e.g. country, region)<sup>7</sup>
- zero, one, or more *classification variables*, classifying a *population* of observed *objects* (e.g. persons, households, enterprises, goods, events, transactions, relationships)

Each cell in a multidimensional cube contains a value that represents the estimated value(s) of one or more parameters of the subset of the population corresponding to the cell; the subset is defined by a combination of values (coordinates) for each one of the dimensional variables. A parameter is an operator (e.g. count, sum, mean) applied on the values of a summation variable (or vector of variables), e.g. income, of each object in a collective of objects. If there are more than one parameter in the same cube, the vector of parameters may be thought of as a “parameter dimension”, and the elementary cells of the cube will contain estimated values of the different parameters, “side by side”<sup>8</sup>.

Metadata are data about the data in the dataset. In principle, all metadata concerning the data could be seen as attached to the individual data values in the individual cells of the multidimensional data cube, i.e. the estimated values of parameters of object collectives corresponding to the cells in the cube. However, this would cause a lot of redundancy, both from a conceptual and from a physical point of view. Instead one should try to factor out metadata that are common for all data concerning a certain “attachment object” and associate those metadata with that attachment (metadata) object. Examples of attachment objects for statistical metadata are:

- a complete database
- a dataset
- a multidimensional data cube
- a variable (the values of which have been observed, derived, or aggregated)
- a dimension (a variable acting as dimension)
- a dimension member
- a parameter (i.e. the concept that is being measured in the table cells)
- a time series
- an observation or a table cell
- a population (for which values of variables have been, or could be, observed, derived, and aggregated)

---

<sup>6</sup> In exceptional cases there may be more than one time variable, e.g. if the statistical data contain “change matrixes”, where a certain cell (i, j) in the matrix shows the change in a variable between time (i) and time (j). In other cases the time dimension may seem to be missing, but this usually means that time has a constant value, i.e. all data are implicitly associated with the same time (point or interval).

<sup>7</sup> In the case of migration statistics there are (at least) two space variables, the origin location and the destination location respectively. In other cases the space dimension may seem to be missing, but this usually means that the geographical dimension has a constant value, i.e. all data are implicitly associated with the same country, region, etc..

<sup>8</sup> This is what is called in SDMX a *cross-sectional dataset*.

- a survey (by means of which values of variables have been observed and processed, and values of parameters have been estimated for one or more populations)
- a statistical system (comprising a system of surveys, databases, and standards: definitions, classifications, methodologies, etc)

If certain metadata could be factored out to several alternative attachment objects, and these objects are hierarchically related to each other, one should associate the metadata with the object that is on the highest level in the hierarchy, “the highest attachment level”.

Figure 1 visualises a conceptual model for organising the metadata associated with the statistical data stored in the logical form of multidimensional cubes. Both structural and reference metadata can be organised according to this model. The model is general and non-domain-specific, which means that it can be used for all kinds of statistical data, emanating from all kinds of statistical surveys (e.g. social, economical) in all kinds of statistical systems (e.g. national, international).

We may view *Figure 1* as consisting of three major parts:

Part 1: The datasets to be exchanged, viewed as sets of multidimensional data cubes

Part 2: The statistical system and its surveys and processes, from which the datasets emanate

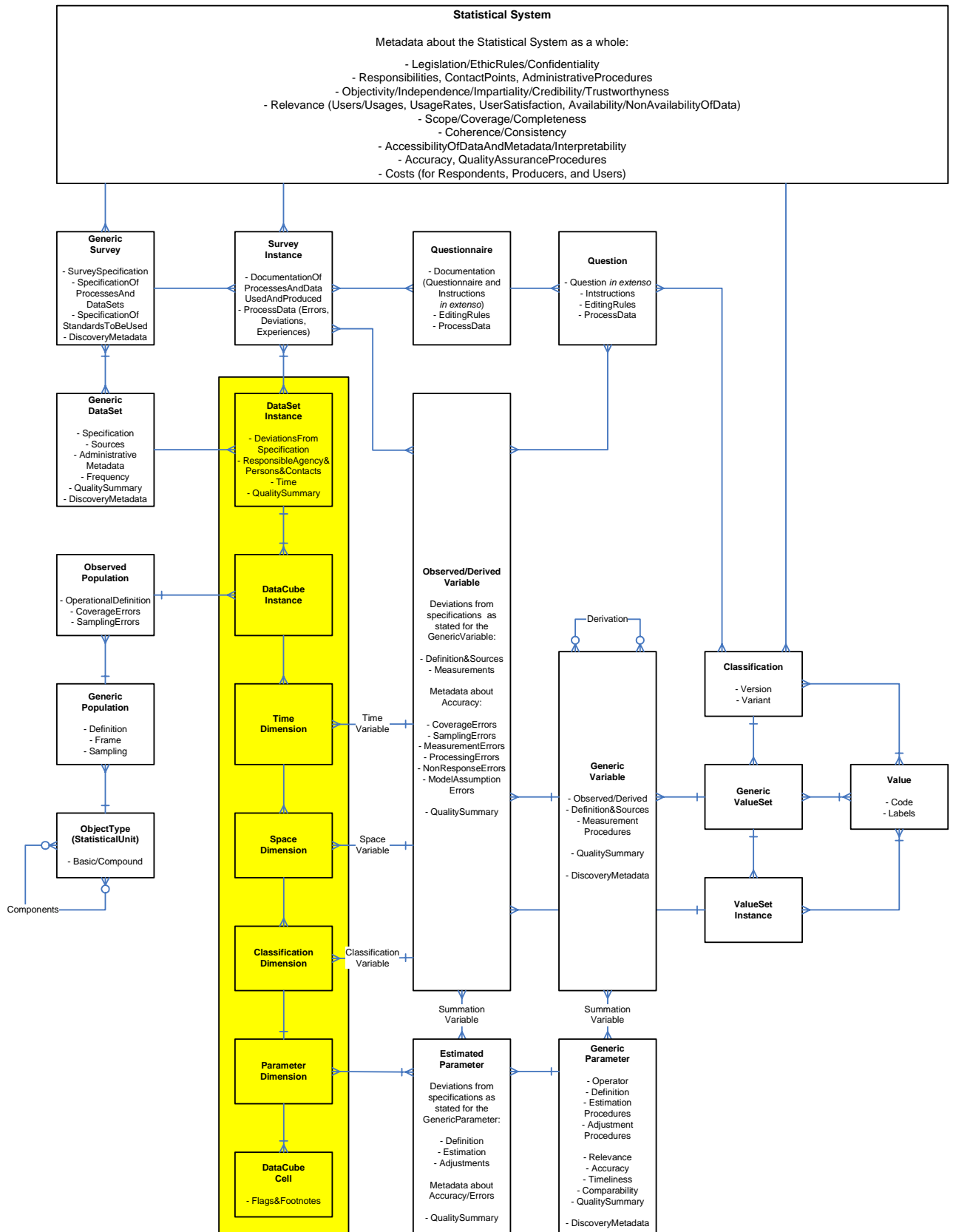
Part 3: Statistical concepts referred to in the structural metadata of the statistical datasets: populations and object types, variables and value sets, parameters

We will now analyse the contents of *Figure 1* in some more detail, indicating also which metadata objects could be suitable attachment levels for which metadata items.

### ***Datasets, multidimensional data cubes, and associated concepts***

There is an object hierarchy in *Figure 1* starting with *DataSetInstance* and ending with *DataCubeCell*. This hierarchy, described in XML, following the SDMX technical standard, is actually the skeleton of what should be exchanged by means of the SDMX standard. In addition to the data in the datacube cells, the dataset should be enriched with metadata properly placed on the relevant attachment levels in the hierarchy.

According to the model that we propose here, a *DataSetInstance* (corresponding to the physical dataset to be exchanged) should always be conceptually broken down into a number of multidimensional cubes, where each cube is associated with one primary (generic) population, e.g. *either* Persons *or* Households *or* MigrationEvents. The primary population is the population whose objects are counted and measured. In the dataset to be exchanged (and in the survey(s) behind the dataset), one may have counted and measured *both* Persons *and* Households *and* MigrationEvents, but, if so, one must break down (conceptually at least) the dataset into one cube for Persons, one for Households, and one for MigrationEvents; cf. the normalisation of relational tables in a relational database. Obviously, the three cubes will be related to each other. Thus data associated with objects belonging to one object type in one cube may be associated with objects belonging to another object type in another cube. For example, household data may be associated with the persons belonging to the household, and vice versa, and the data about MigrationEvents in a MigrationEvents cube may be enriched with data about Persons and Households involved in the MigrationEvents.



**Figure 1.** A general, non-domain-specific model for organising all kinds of metadata concerning a dataset of statistical data.

Thus each *DataCube* belonging to a *DataSetInstance* will be associated with exactly one *ObservedPopulation*. A population is a collective of objects of interest. They share a common property that define them as members of the population, the population property (e.g. persons living in Sweden). A population may be subdivided into subpopulations or domains of interest. One often subdivides a population into subpopulations by cross-classifying it by means of a number of classification variables.

An *ObservedPopulation* (e.g. persons living in Sweden as of 1 January 2006) could be seen as an instance of a *GenericPopulation*, which means that metadata, which are common (as a rule) for all instances of a certain *GenericPopulation*, could (and should) be attached to the latter rather than to the former. However, if there are exceptions to such a “commonality rule”, this should be noted on the instance level.<sup>9</sup>

On an even higher generic level, object instances, object populations, and generic populations may be seen as belonging to an *ObjectType* (also called *StatisticalUnit*). Examples of basic object types in official statistics are *Person, Enterprise, Building, Commodity, Vehicle, Service, FinancialAsset, NaturalResource, Livestock, Crop*. Other common object types in official statistics are compound objects like events, transactions, and relationships, e.g. *PersonMigration, TradeTransaction, RoadAccident, MarriageEvent, MarriageRelationship, DivorceEvent, EmploymentEvent, EmploymentRelationship*. Like basic objects they may be counted and measured, but compound objects are associated with other objects, which may also be counted and measured in their own right. For example, a trade transaction is associated with a seller, a buyer, and a product, and a road accident is associated with a number of vehicles, a number of persons, and a road segment.

The association of an object with certain fundamental object types are permanent. For example, a person will always remain a person. The association with other objects types may be temporary and dependent on the role of the object for the moment. Examples: sellers, buyers, drivers, passengers, students, patients, employees, etc.

A *TimeDimension* (if it exists as a dimension) of a *DataCube* will refer to a *TimeScale*, which may be regarded as a kind of *Variable* with certain typical types of *ValueSets*.

A *SpaceDimension* (if it exists as a dimension) of a *DataCube* will typically refer to some kind of geographical variable, taking values from a standardised regional *Classification*.

A *ClassificationDimension* of a *DataCube* will refer to a qualitative variable or a quantitative variable, whose values have been grouped into intervals or classes (dimension members). The *ValueSet* of a classification variable may be a standard (national or international) *Classification* or some variant of such a standard.

During a survey one collects values of a number of *observation variables* for each object in one or more populations. Values of other variables may also be obtained for the same objects from other surveys, or by deriving values of new variables from the values of other variables.

In general, a *Variable* takes its *Values* from a *ValueSet*. Standardised value sets are called *Classifications*. New classifications and value sets may be formed from existing ones by combining intervals or groups of values into new values. Classifications are often built up hierarchically in levels in such a way that a group value on one level consists of several more elementary values on the next lower level. Regional classifications are typical examples of hierarchical classifications: a country may consist of counties, which again consists of municipalities. The values in a value set or a classification are often represented both by verbal labels (names) and by codes. For hierarchical classifications the hierarchy is usually reflected in the codes in some way or other.

A *de facto* standard for concepts and terms in connection with classifications can be found in Ehrenström et.al. (2002) “*Neuchâtel Terminology – Classification Database Object Types and their Attributes*”, Version 2.0.

---

<sup>9</sup> In order to fully understand the meaning of the term “generic” one may consider the parallel with human beings as a species, on the one hand, and as instances of this species on the other. Human beings have certain properties in common, such as (usually) having one head, two eyes, one nose, one mouth, two arms, two legs, etc. However, on the instance level there will be exceptions for almost all such properties; a person may have only one eye and be missing arms and legs, but we will still be able to recognise this person as a human being.

A *ParameterDimension* will refer to one or more *EstimatedParameters*. If there are more parameters than one, they must all apply to the same population and subpopulations (domains of interest), as defined by the cross-classification of the other dimensions.

When values for a number of variables are available for (a random sample of) all objects in the population, either as the result of direct observations or by more or less complex derivations from such observations of other variables, the values of a number of parameters of the population (and subpopulations) are estimated by means of an aggregation or calculation process. A parameter summarises the values of a variable (or a vector of variables) for the objects in a population or subpopulation. Examples of parameters are frequency counts of the number of objects, sums, averages, and medians of quantitative variables, variances, covariances, correlation coefficients, etc.

A *DataCubeCell* of a multidimensional *DataCube* will contain the estimated value of a parameter for a subset (domain of interest) of a population; the subset is defined by combining one value in the value set for each one of the classification variables (dimension members)<sup>10</sup>

Thus the number of cells in the data cube will be the product of the number of estimated parameters and the number of values of each one of the classification variables

If the data cube turns out to become very sparse, data compression techniques may be used for economising with storage space and access speed.

Even the individual data cube cells may be appropriate attachment objects for certain types of metadata, e.g. so-called footnotes and flags.

### ***The statistical system as a whole and its surveys and processes***

A statistical system comprises a number of surveys, which collect observations of society that are processed and aggregated into statistics about society. The statistics are organised into datasets that are made available to the general public and exchanged with national and international institutions. The surveys and the statistical system as a whole are supported by an infrastructure containing personnel, methods and standards, data/metadata, and technology.

Ideally the surveys carried out by a national statistical agency complement each other in such a way that one may talk about a complete and coherent national statistical system. Even if a statistical system is not perfect, it usually exists in some sense, at least as a set of surveys that have some kind of official status, and there are standards supporting the statistical system, e.g. standard definitions, standard classifications, standard methods. One may also talk about different international statistical systems, e.g. the European Statistical System, or the United Nations Statistical System.

On this attachment level it is appropriate to store metadata about objectivity and credibility. These quality characteristics have to do with such things as the independence of the statistical system (and the statistical agency) as a whole, in particular its independence from political pressure as regards the contents of the statistics, publishing times, etc. However, if there are particular problems with certain surveys or certain datasets, such exceptions from, or additions to, the general description should be attached at those levels.

Further examples of categories of metadata items appropriate for this attachment level may be found in *Figure 1* above.

---

<sup>10</sup> Often, two of the dimensions are space and time.

National and international statistical systems often maintain some kind of “official listing” of surveys to be executed regularly. Each time a survey is executed, it produces a number of datasets that are made available to the public and exchanged with other statistical agencies, national and international. For example, a labour force survey may be executed according to the same standard specification over a number of time periods. Moreover all member countries of a certain international organisation may follow more or less the same standard specification.

Although there are always some differences between different executions of “the same” survey, the survey executions are seen as instances of one and the same generic survey. As long as the design specification remains more or less unchanged, it can be attached to the *GenericSurvey* metadata object. Deviations from this “standard design” should be documented and attached to the appropriate *SurveyInstance* object. Every execution also generates some unique so-called process data, data about the quality and performance of the survey processes, e.g. non-response rates, and such metadata could also be attached to *SurveyInstance* objects.

A *DatasetInstance* is a concretely identified and defined data entity that is made available for data/metadata exchange, e.g. between national statistical agencies and international organisations, or between international organisations, by means of pull and/or push techniques, via the Internet or in some other way.

In the simplest case, a certain *DatasetInstance* comes from a certain *SurveyInstance*. However, different parts of the dataset (e.g. data concerning different parameters, or even data concerning one and the same parameter) may be based on observations made in different surveys. Some data may be the result of rather complex derivations, based on data from different survey instances and even different generic surveys.

It is common that many datasets can be seen as instances of one and the same generic dataset. For example, “the same” (kind of) dataset can be exchanged regularly, at certain time intervals, or different statistical agencies, in different countries, may exchange “the same” (kind of) dataset with an international organisation, or even with several international organisations.

Since datasets are the concrete entities that are exchanged, it is natural to attach certain administrative data to the dataset objects, in order to facilitate the proper management of the datasets. Both dataset instances and generic datasets need to be associated with unique identifiers, names, descriptions, names of responsible persons, contact persons, etc. As with generic surveys and survey instances, metadata that do not change over time could be associated with generic datasets, and exceptions as well as metadata that change should be attached to the dataset instances.

## ***Our propositions***

The statistical data/metadata model presented here is general and domain-independent. This means that it will cover all kinds of data and metadata to be made publicly available on the Internet and to be exchanged or shared between national statistical agencies and international organisations. This proposition has been verified in a number of cases for public statistical data from representative statistical agencies. So far the proposition has not been falsified in any case.

Furthermore, we claim that most of the components of this generic model can be transformed in a systematic way into an SDMX-compliant generic model expressed in XML. However, this remains to be further validated.

Since cube models, as actually practiced in national statistical agencies and international organisations, differ slightly between themselves and cannot always be said to be standardised, we propose the transformation to take place in two steps:

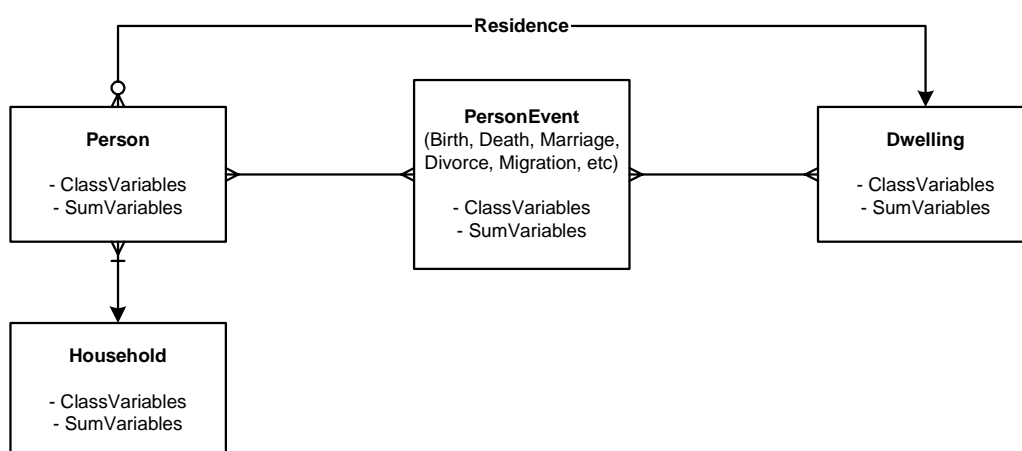
Step1: Non-standardised cubes are transformed into normalised cubes as defined in this paper.

Step 2: Normalised cubes are transformed into standardised SDMX cubes (to be defined).

## Domain-specific applications of the generic cube model

In this section we will demonstrate how a large variety of statistics that actually occur in the databases of national statistical agencies and international organisations may be modelled as normalised cubes as defined in this paper.

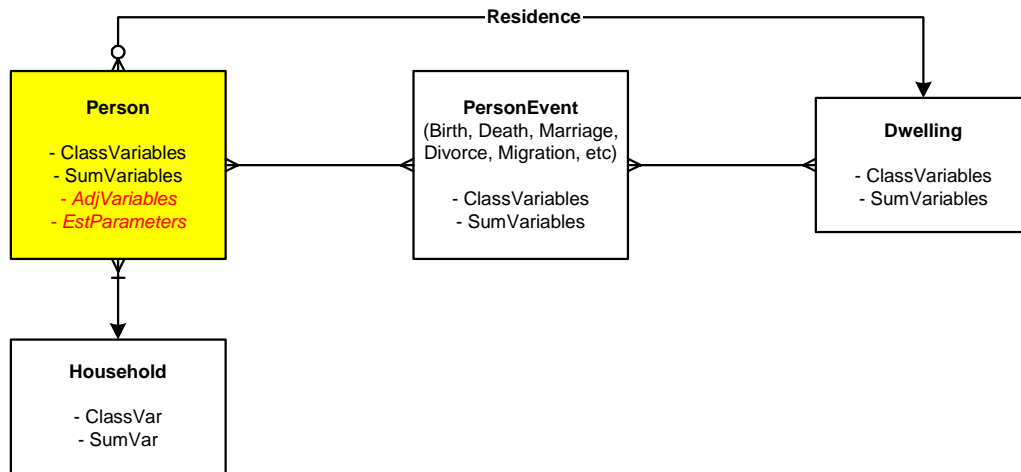
Let us start with population statistics. *Figure 2a* shows a basic conceptual model allegedly covering “all” statistical concepts occurring in population statistics. Three basic object types are **Person**, **Household**, and **Dwelling**. In addition there are compound object types of “event type”, 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.



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

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.

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.



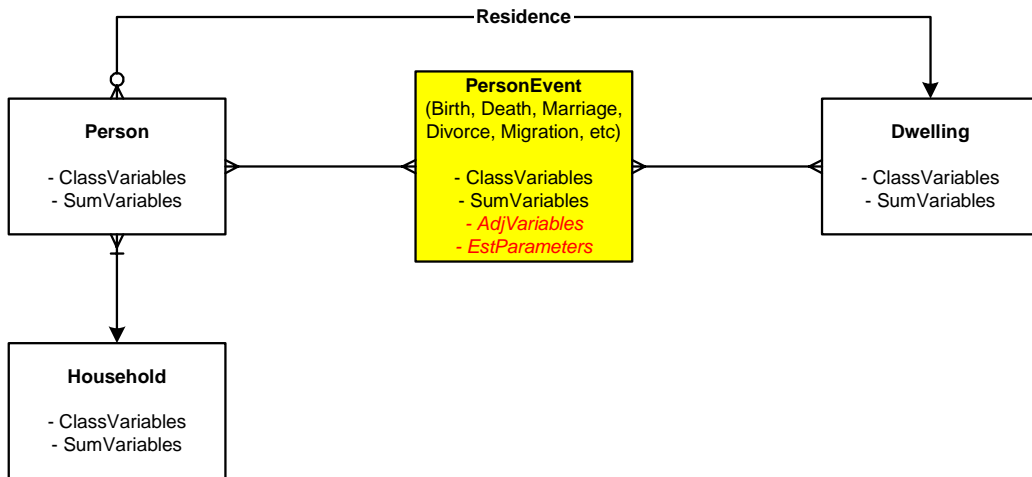
*Figure 2b. Person statistics.*

The **Person** object type is associated with certain *basic* classification variables (like Sex) and certain *basic* summation variables (like Age, Income). Further classification 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:

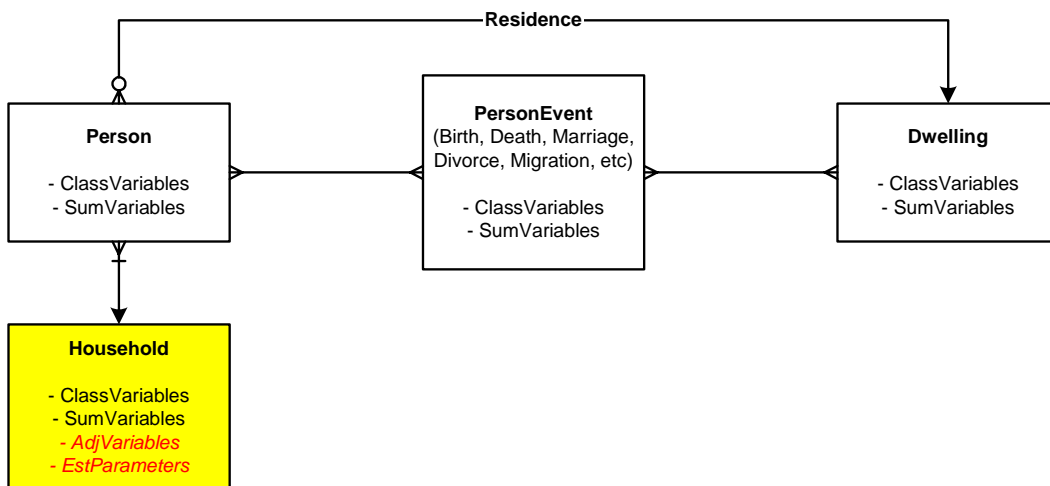
- the Sex of the Person involved in the PersonEvent
- the Size of the Household to which the Person involved in the PersonEvent belongs
- the Region(s) in which the Dwelling(s) associated with the PersonEvent is/are associated



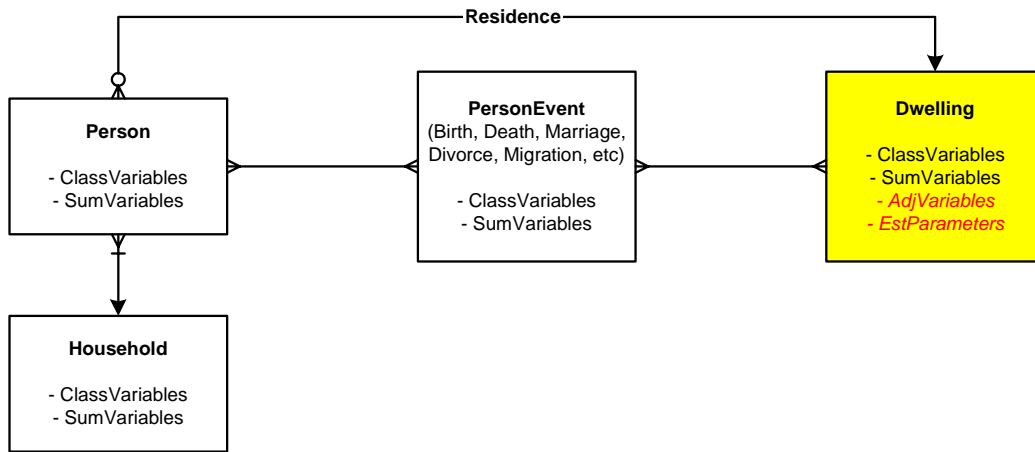
*Figure 2c. PersonEvent statistics.*

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**, respectively, have been put in focus as the basis for normalised cubes.

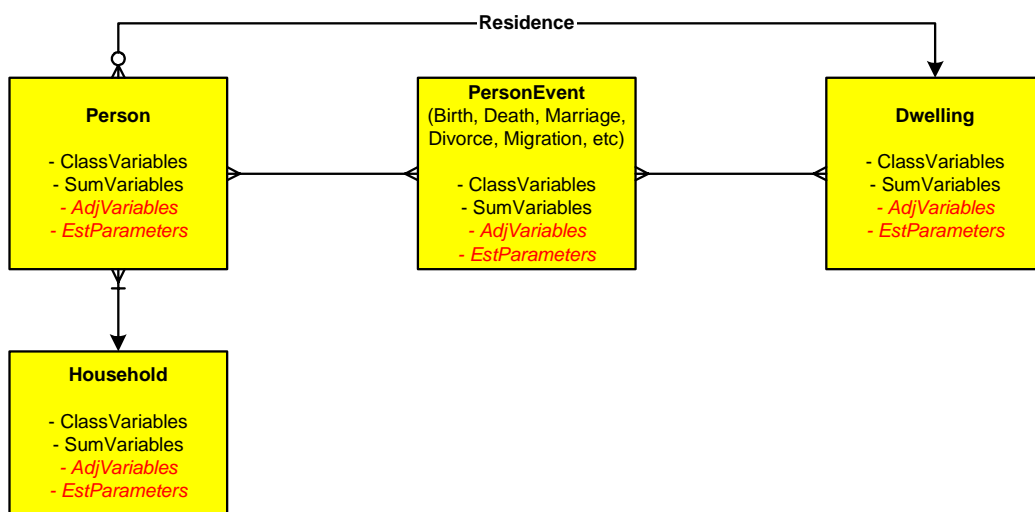


*Figure 2d. Household statistics.*



*Figure 2e. Dwelling statistics.*

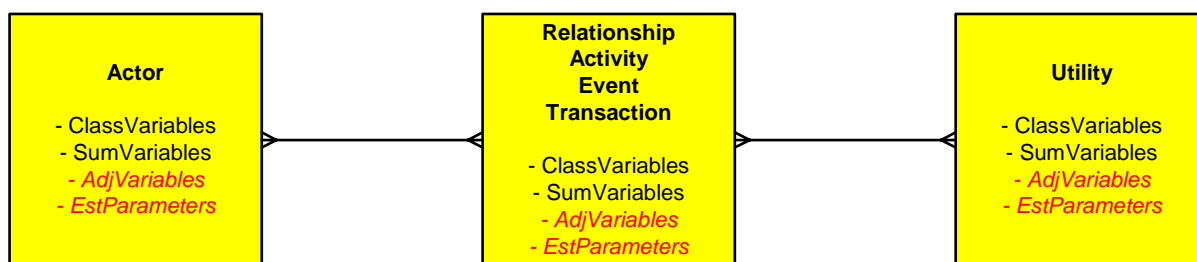
Figure 2f summarises all the preceding models into one, single “split vision” model, indicating by **yellow colour** all the object types that may, one at a time, be focused as the basis for normalised cubes. In the particular case used here for illustration purposes, it actually happens that all object 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 2f. Population statistics: Split vision model.*

In appendix 1 we will show “split vision” models for a wide range of statistics that are produced and made available by national statistical agencies and exchanged with international organisations; see Figure 4-17.

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 annexes 2-4.

## **6. Metadata systems of international organisations**

What are the needs for statistical metadata of international organisations?

The metadata of international organisations basically aim to give the information necessary to understand if it is meaningful to compare macro data between countries, and to understand how much weight can be attached to such a comparison. It must therefore describe what the data mean (concepts), the overall quality of each of the data elements presented, as well as differences in quality and differences in the meaning of the data between different subject matter areas as well as between countries. Moreover, on the dissemination side, international organisations serve with their statistics (data and metadata) several groups of users (e.g. policy makers, individuals and enterprises, financial institutions, data vendors, journalists, researchers, students, etc).

For this paper, we will limit the scope to macro data, so the usually very comprehensive metadata describing micro data are not in focus. This is primarily because at least the international organisations involved in SDMX are almost exclusively focusing on macro data, secondly because SDMX at this point in time only aims to describe macro data. However, when users compare e.g. labour force statistics from different countries, they may wish to drill down in metadata and know more about the instruments for collecting the micro data in the first instance.

The metadata must primarily illuminate the following areas:

- concepts, definitions of concepts, including such phenomena as measurement units, transformations
- delimitation of populations
- dimensions of quality, related to the original production, such as sample sizes, standard errors on estimates, and other kinds of known sources of errors such as non-response

The data and metadata of international organisations are usually collected from the countries' NSOs. This involves a special effort on the part of NSOs, producing them just for the sake of the IOs. In general, this may lead to the quality of metadata being poor – the general rule saying that data which are not produced for a purpose fully appreciated by the producer tend to have a poor quality. Therefore, the aim is to make sure that metadata, as well as data, can be copied directly from the systems of NSOs. Additional metadata and “derived metadata” (based on the national more detailed metadata) may also be produced by the staff of the IOs.

For the sake of this paper, OECD metadata system is taken as an example. The system is based on a set of corporate principles laid down in *Management of Statistical Metadata at the OECD*<sup>11</sup>. The following principles are the most important ones.

<sup>11</sup> <http://www.oecd.org/dataoecd/26/33/33869551.pdf>

**Consistency:** Statistical metadata must be consistent. This means that:

- the same variable name, definition, and other description should be connected to the same statistics, no matter where it is and who is the “owner”;
- the same variable name should not be used for statistics that are not identical;
- terms and concepts should be consistent throughout;
- all OECD metadata, particularly reference metadata, should be made readily and freely available to external users.

**Redundancy:** Metadata on one element (a statistical collection or dataflow, a concept) should only exist as one instance, no matter how many times the same element is reused in different contexts. Ownership to the metadata should be clearly defined.

**Common metadata items:** A set of 41 metadata items are defined. All metadata from different subject-matter areas must be grouped under these headings. These are similar to the SDMX Cross-domain concepts and have been developed concurrently with those. The ambition is to have the closest possible consistency between the two sets. The present list of metadata items is shown in Annex 5.

**Attachment levels:** Metadata can be attached at any level of detail of the statistical data: at the global level (pertaining to all datasets), at the dataset level, at dimension level within a dataset, at dimension member level, time series, individual observations. To ease understanding and avoid repetition of data, it is recommended to always attach metadata at the highest possible (or reasonable) level; exceptions will then have to be stored for those lower levels where they apply.

## **7. Linking metadata systems**

In order for the statistical metadata of the international organisation to be appropriate and always up-to-date, there must be a link between the national system and the international one. As the systems and the concepts they use will differ between countries as well as between international organisations, there must be a linking method that permits the NSO data to be transmitted and be put into the right boxes. This is one of the objectives that SDMX aims to do.

A reasonable way forward for NSOs is to observe the “content oriented” work of the SDMX initiative. Thus, their metadata systems, which probably need to be quite detailed, would have to also cover and serve the concepts and classifications used at an SDMX level in a compatible manner. In the meantime, it is sensible to observe the “structural metadata” defined by the IOs. In this context, IOs should put an extra effort to make their structural metadata easily accessible across IOs and domains and, ideally, in a common place (e.g. a web page or at least providing and maintaining appropriate links). At a next step, if an NSO wishes to make available a dataset on its web site using the SDMX standards, it should first check the corresponding structural metadata defined and used by IOs or other NSOs. Then, a maximum compatibility – at least at the higher level – should be pursued with the existing “content” work (i.e. structural metadata used already by others). This approach maximises interoperability and efficiency, as seen by end-users, and would ensure a smoother transition towards a further harmonised statistical content framework.

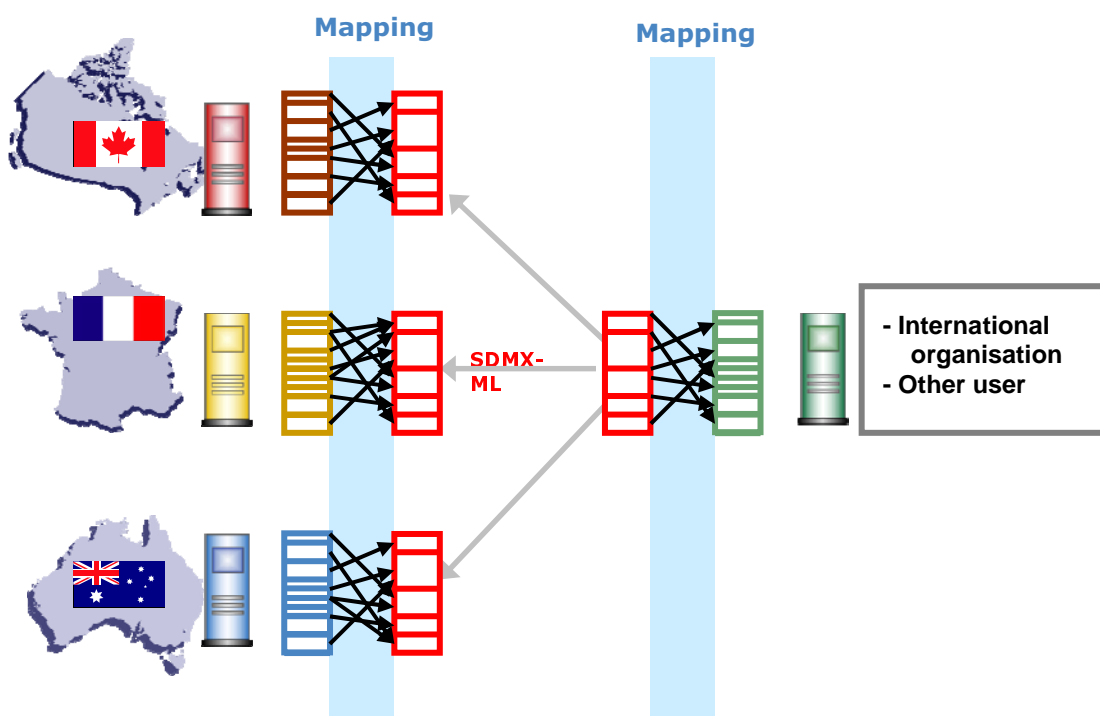
## **8. SDMX cross-domain concepts**

A preliminary list of so-called cross-domain concepts has been defined in the SDMX contents oriented guidelines that are at the moment (March 2006) published for comment from the community of official statistics. The preliminary list is shown in Annex 6.

The SDMX concepts are intended to be a common language that will allow the important features of the different national data structures to be translated into the different international ones, using one common intermediary “language”. It must thus be possible to translate (or map) rather precisely all concepts used in the NSO metadata system to the SDMX concepts, and from those to the international ones. This is what is demonstrated in the following paragraph.

## 9. Mapping between SDMX standards and national and international variants

In this section, we show how the metadata from a national metadata system (example Sweden) can be translated into the SDMX cross-domain concepts, and from those concepts to the concepts of an international organisation (example OECD). This is shown in Annex 7. An example of the real metadata and its transformation is also shown in Annex 9.



## 10 Generalisation

This section discusses the important question: Can this model be generalised, can it be used by other organisations and other subject-matter domains?

Regardless of the different platforms or local database systems used, in an ideal situation, and given that all NSOs face more or less the same requirements, it would be sensible to adopt a common internal data model. The model suggested here provides a quite comprehensive framework that could be used as the basis for an internal reference model in NSOs. While it is quite generic (but in certain aspects more specific as compared with SDMX) it actually allows supporting an advanced level of compatibility with the SDMX concepts and content specifications, especially when interaction with the “external world” is needed (e.g. web dissemination, data and metadata exchange, etc).

While the structure of raw data may be quite similar across countries (as viewed by NSOs), the more aggregations are applied, the more probable becomes for some processes to diverge (including also organisational processes). Therefore, it is difficult to expect that in all its layers the same precisely model, in all its details, can be adopted internally in organisations. This would make it quite difficult to generalise the internal model into “one” and support it at a global level through sufficiently generic applications. This is one of the reasons for which the SDMX model adopts already some “simplicity” (but also a slightly more generic and abstract terminology on the other hand) vis-à-vis the model suggested here. Thus, the SDMX model can be easily adopted and used across domains and both in cases of raw and aggregated data.

## **11. Co-ordination of metadata and the need for extensions**

The maintenance of the SDMX cross-domain concepts is expected to be a dynamic and continuous process. First of all, it is expected that, even when a first agreement is reached after the public consultation process and the consolidation of the submitted comments, adjustments would still be needed for a period of time. It is logical to assume that current practices across organisations already differ and the convergence towards some common best practices will need some time to be completed. Moreover, peculiarities of new specific datasets may require a review of the previously followed practices. Considerations from both the production and the dissemination side (including needs identified in the context of the development of statistical data warehouses) will constitute sources of change and of a need for further adjustments.

It has also to be acknowledged that, despite all harmonisation efforts, there may still be a need for some organisations to proceed to temporary, transitional or even (in some cases) permanent “extensions” beyond the standard SDMX cross-domain concepts. These may be jointly administered by very few or even only one organisation. It will be at least needed these extensions to remain transparent in order to be easily used also by others if needed. It could be envisaged that some of the extensions used for cross-domain concepts and classifications may overtime used by wider communities and evolve becoming part of the cross-domain content.

In the context of this work needed, it is obvious that the use of the SDMX model and its advantages are optimised when the administration of metadata and their structural composition is well co-ordinated within an institution (across data reception, production and dissemination; and across subject matter areas) and also across institutions. This requires the establishment of a strong “structural metadata co-ordination function” within each institution (NSOs, IOs) which would also serve the communication and co-ordination across institutions (at least IOs). So, appropriate organisational arrangements and a clear allocation of resources and tasks should be in place. The existing Committees and Working Groups should, of course, play an important role in the work needed for each domain, while the SDMX initiative and IOs can facilitate the co-ordination work across the various communities and domains.

## **12. Exchange of metadata using SDMX**

Metadata related to a dataflow (or to a metadataflow) can be transmitted between different metadata systems. A prerequisite for this is to first exchange or share their structure using (depending on the needs) the SDMX - structural metadata message (or the SDMX – metadata structure definition message) .

In order to demonstrate this, we show in Annex 8 a draft correspondence to the metadata related concepts as used in the SDMX structural data and metadata messages.

# Annex

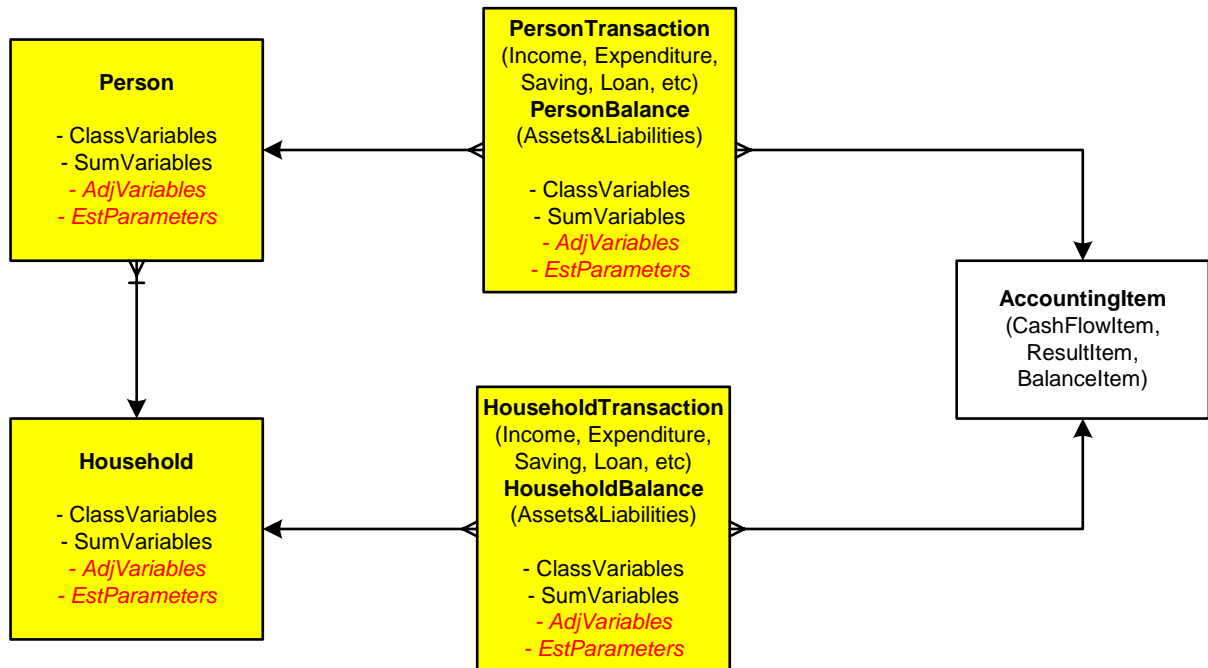


Figure 4. Household economy statistics.

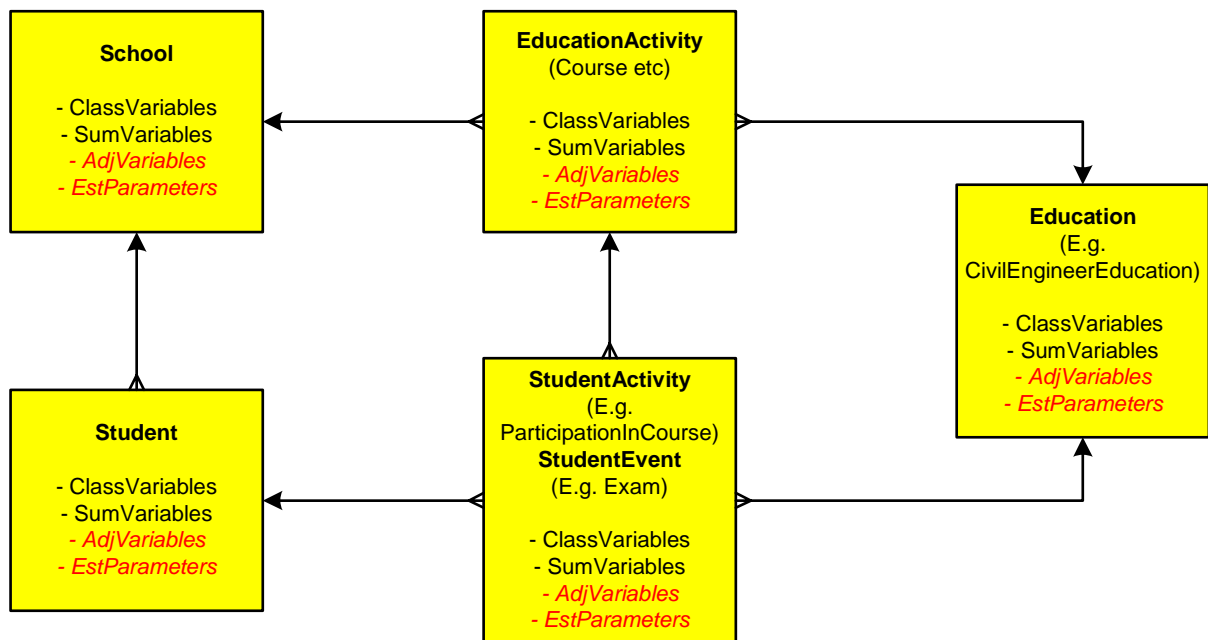


Figure 5. Education statistics.

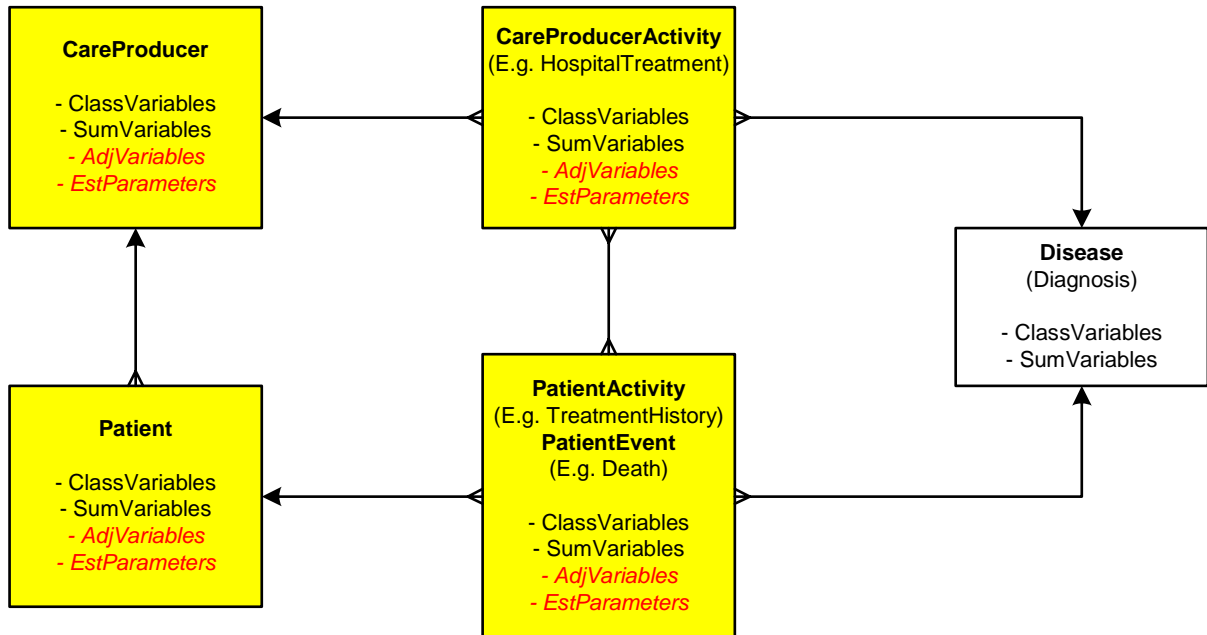


Figure 6. Health statistics.

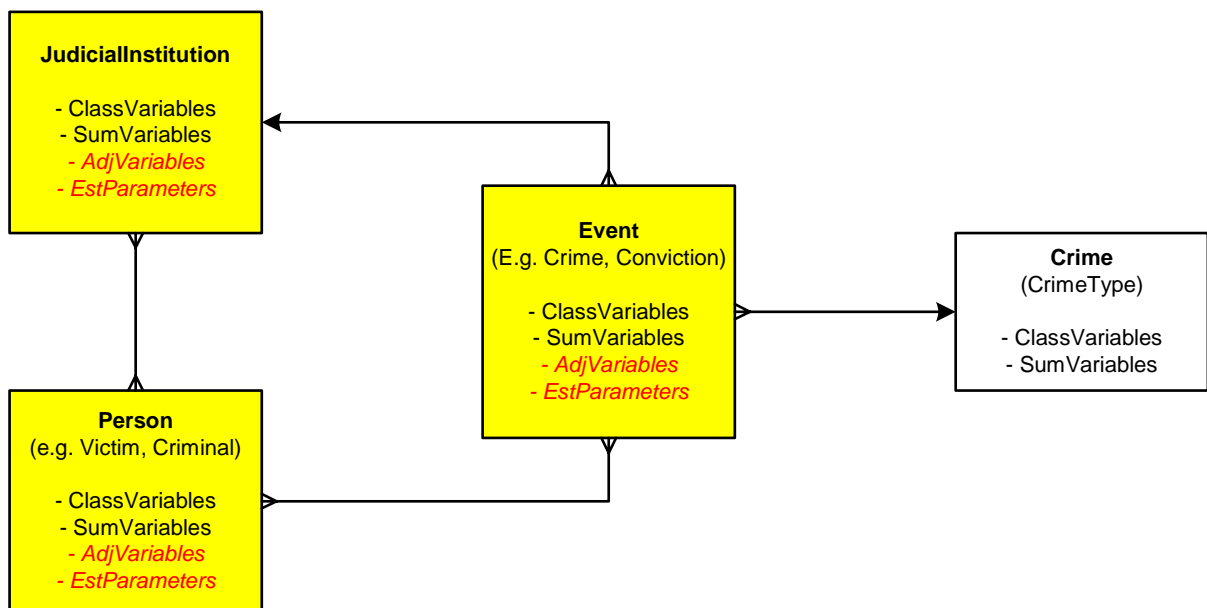


Figure 7. Judicial statistics.

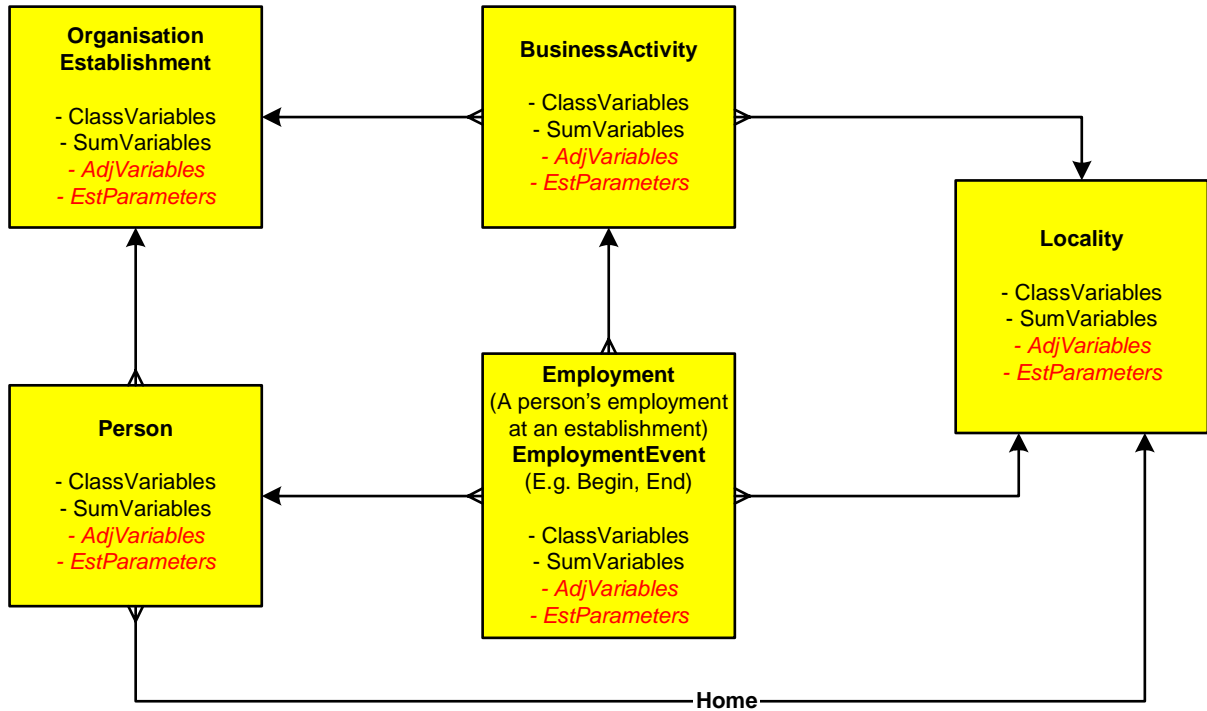


Figure 8. Labour market statistics.

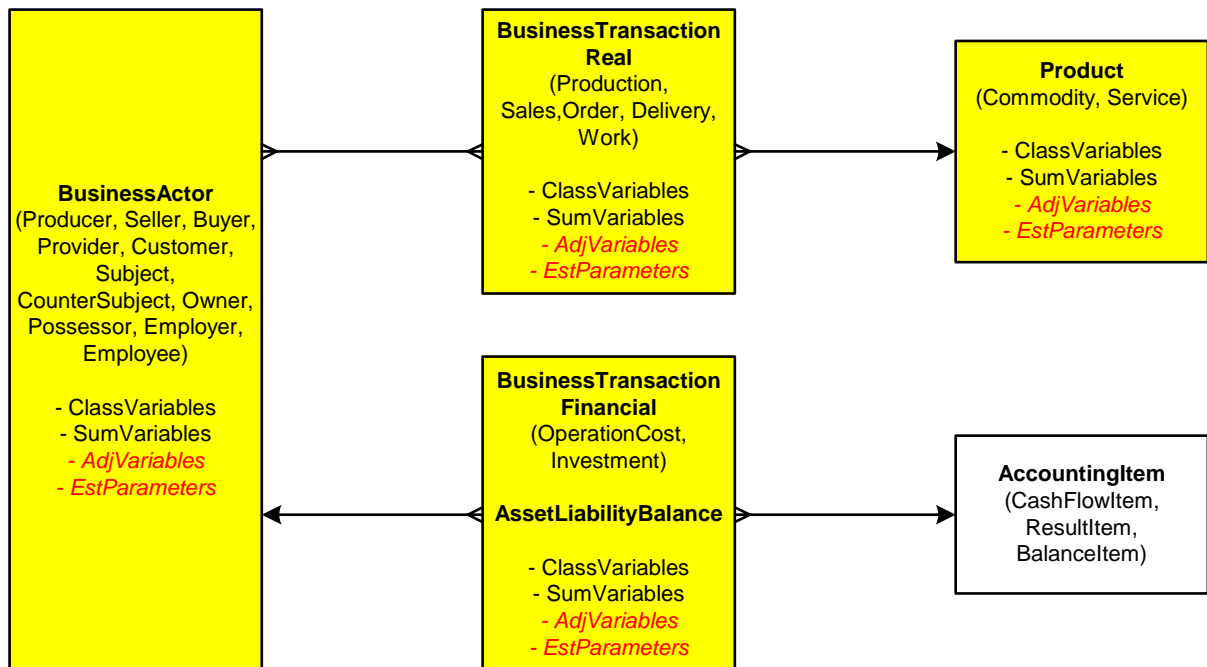
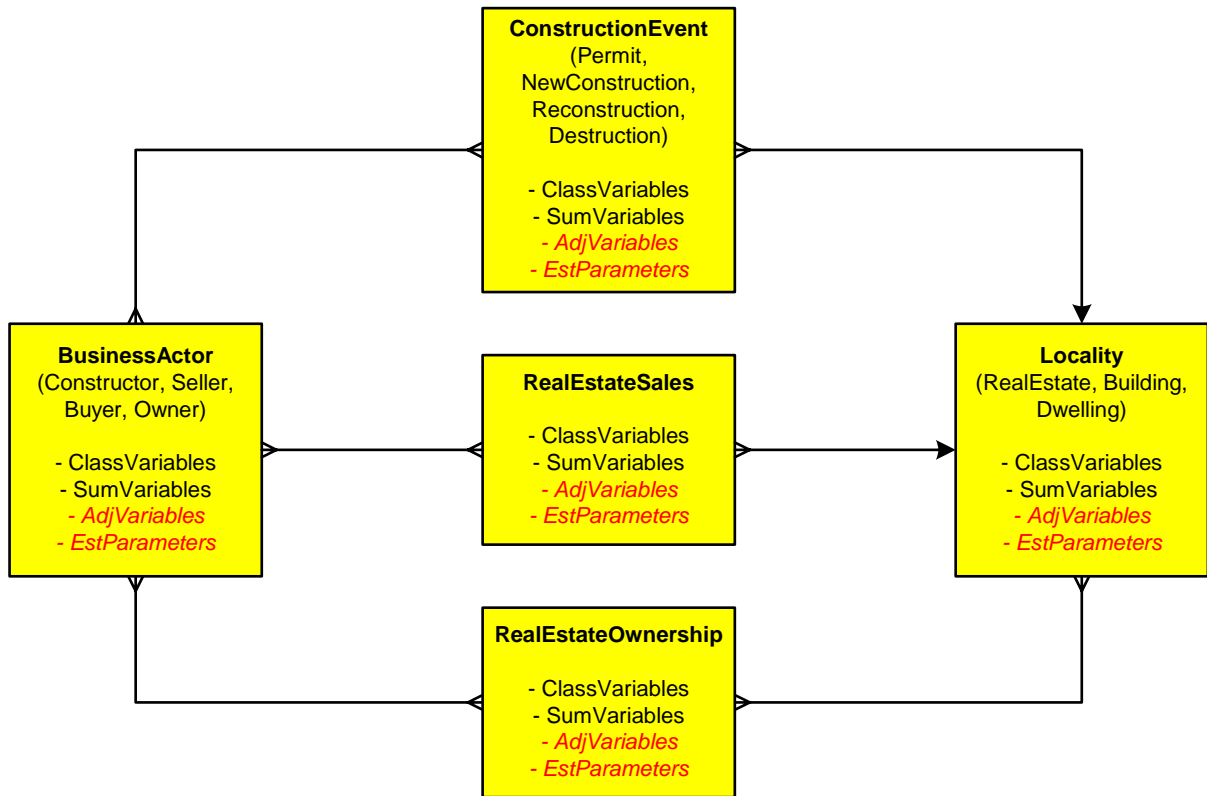
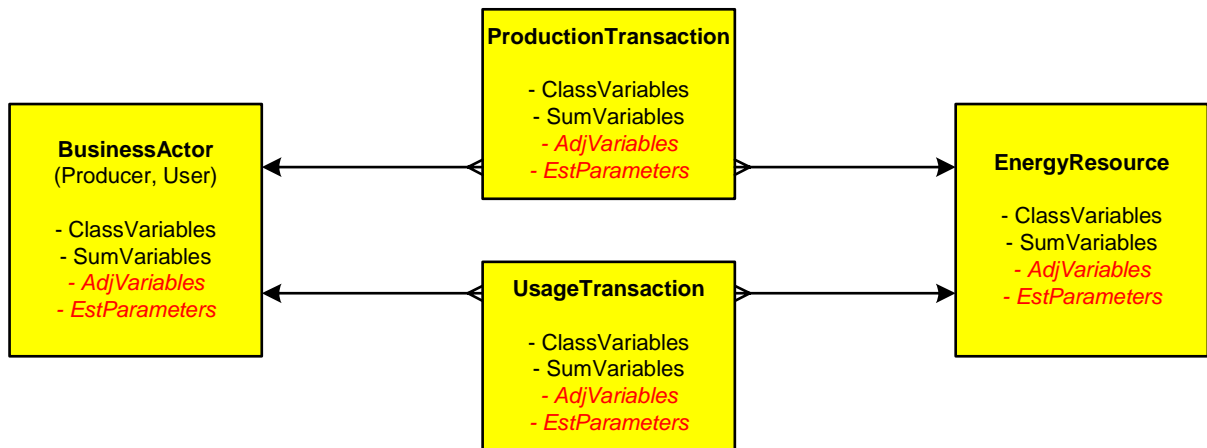


Figure 9. Business statistics.



*Figure 10. Housing and construction statistics.*



*Figure 11. Energy statistics.*

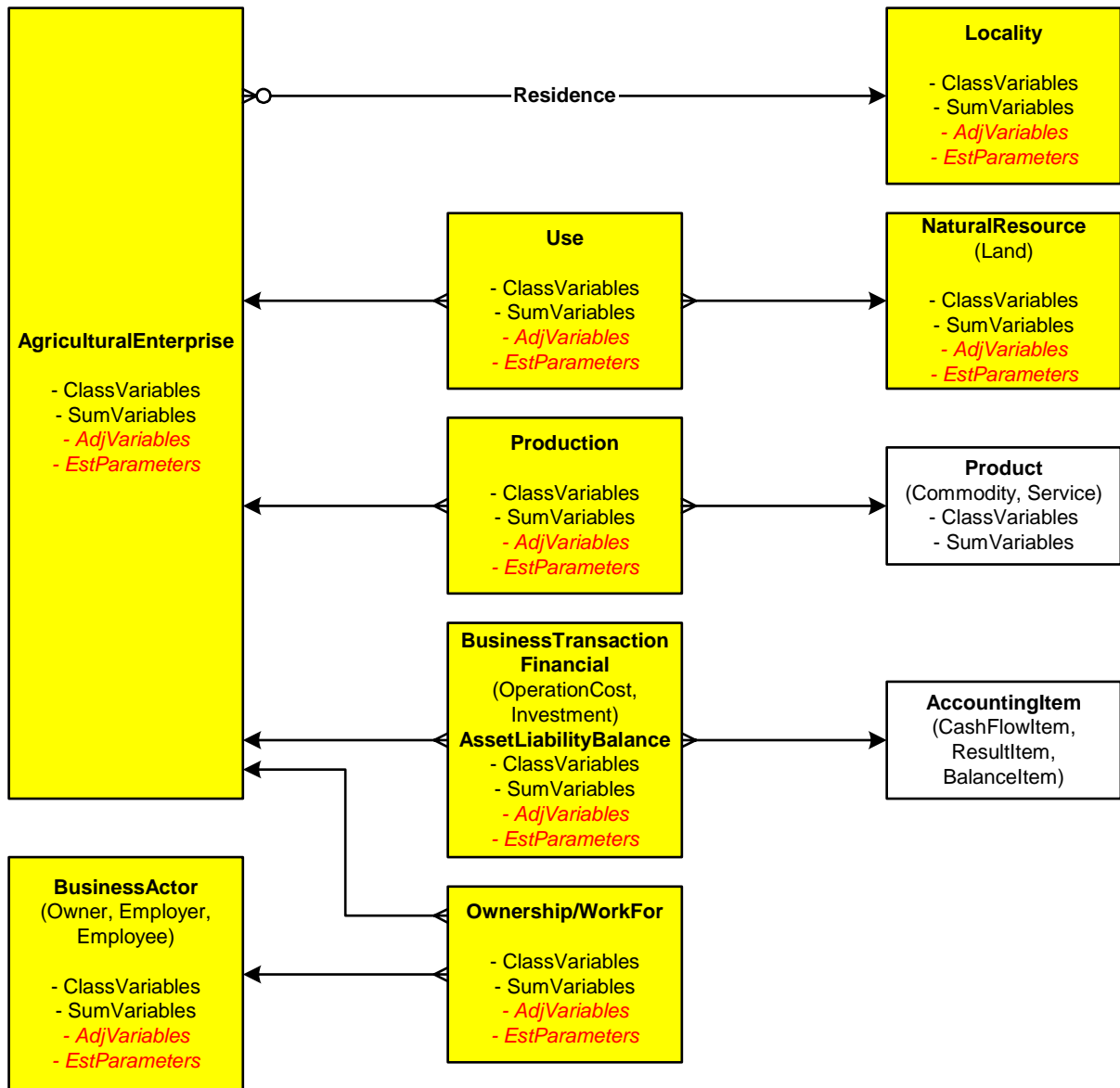


Figure 12. Agriculture statistics.

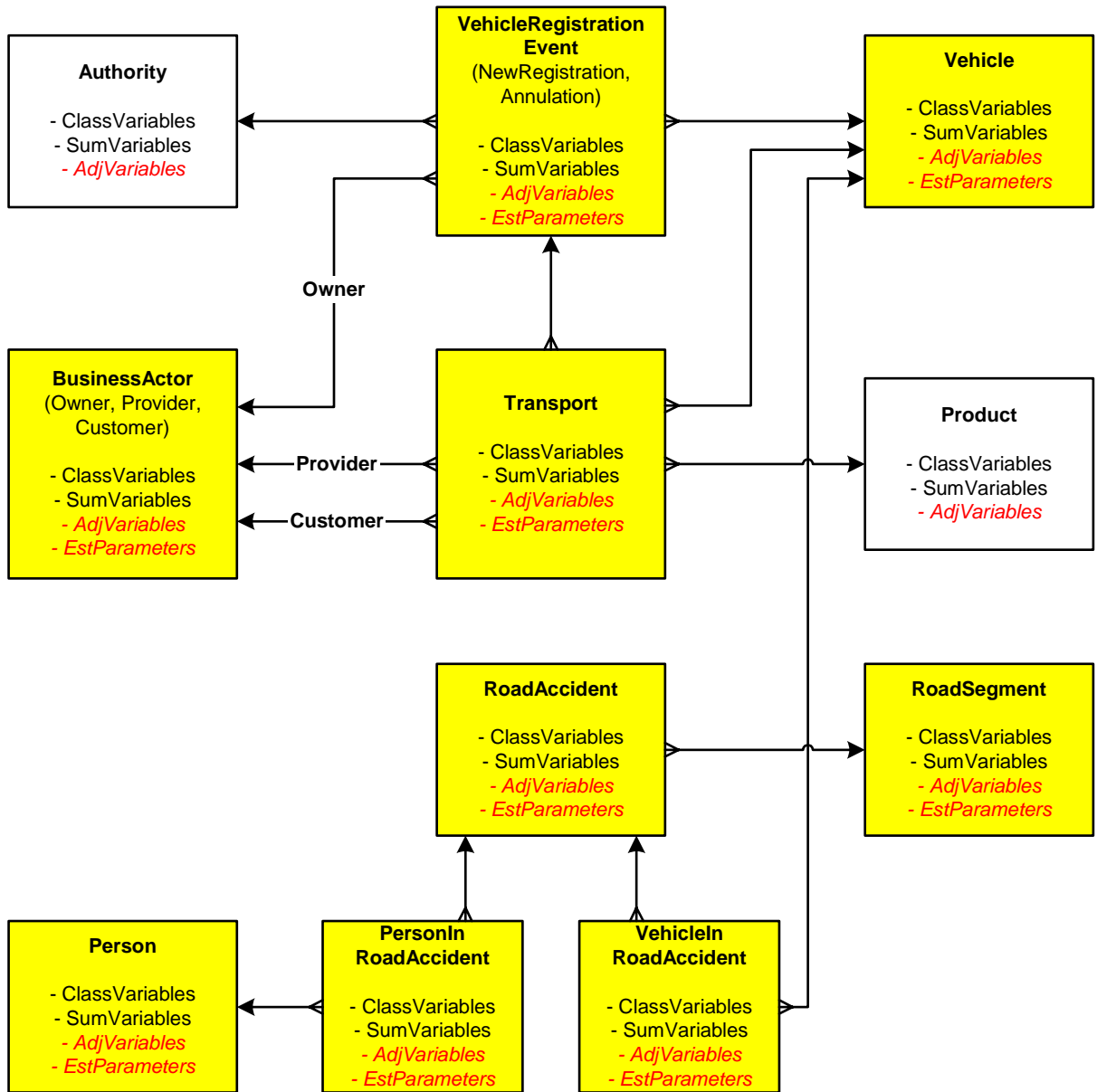


Figure 13. Transport statistics.

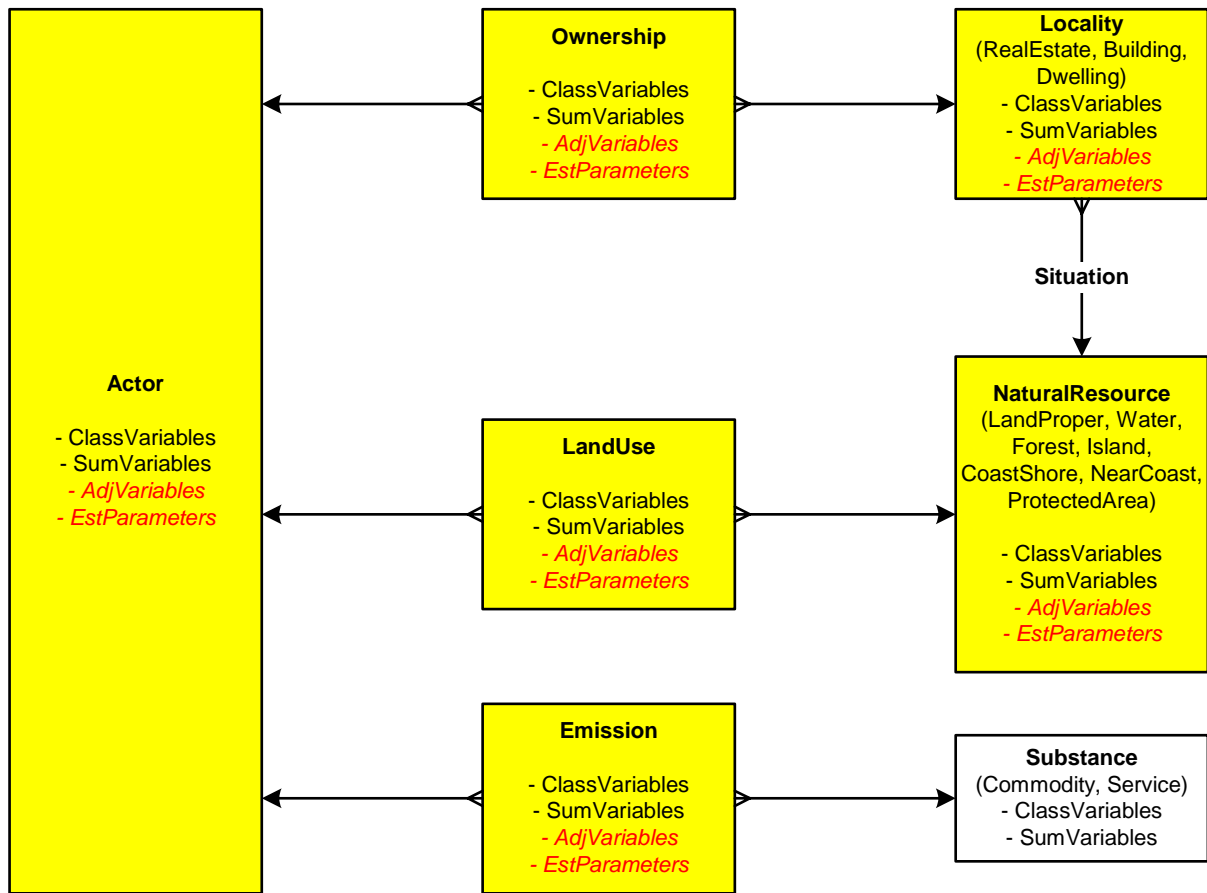


Figure 14. Environment statistics.

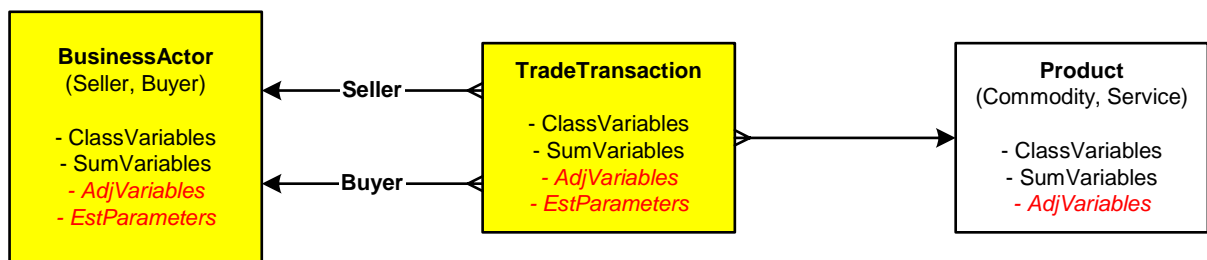
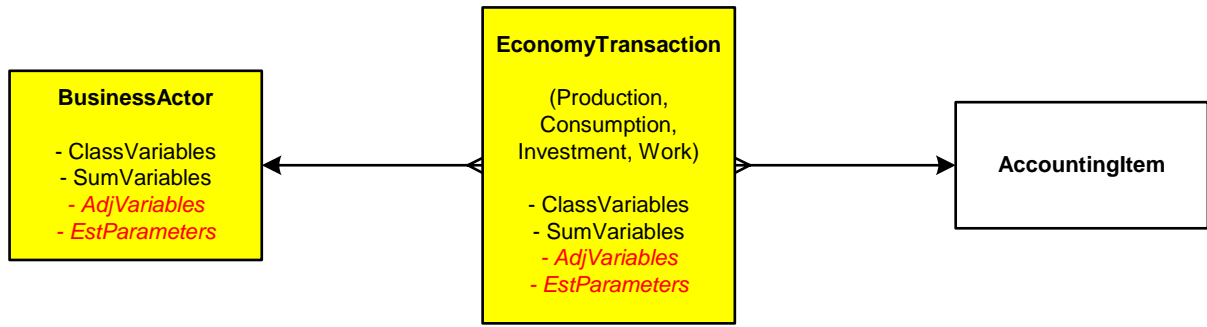
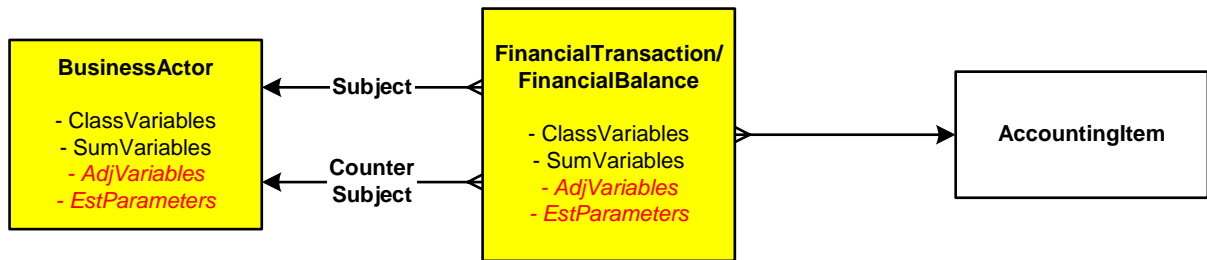


Figure 15. Trade statistics.



*Figure 16. National accounts.*



*Figure 17. Financial statistics.*

## **Annex 2. Contents analysis for SDMX**

### **Method of analysis**

#### ***Counted and measured objects in official statistics***

In this analysis, the contents of official statistics have been modelled in terms of three major types of objects:

- actors: persons and organisations
- activities: states and events
- utilities (and evils): cardinals and masses

#### **Actors: persons and organisations**

Actors perform human actions, either directly, by themselves, or indirectly through organisations (including enterprises, governmental institutions, and non-profit, non-government organisations). Groups of actors, e.g. households (groups of persons) and groups of companies, may also be regarded as actors, as may more or less independent parts of enterprises, e.g. establishments.

#### **Activities: states and events**

Activities are typically initiated, performed, and terminated by actors. Alternatively they may be so-called “acts of God or Nature”, where no personalised actor can be identified (although actions by some actors may have influenced the probability of the action). An activity will usually have a certain duration in time, and it may involve a number of other objects: actors and utilities, as well as other activities. Some activities are more or less momentary and are regarded as taking place in one point in time, such as events and transactions (exchanges of other objects).

#### **Utilities (and evils): cardinals and masses**

Utilities are used and produced in activities, initiated by human actors (person/organisations) or by God/Nature. Some utilities are cardinal in the sense that they are individually identified and counted. Others are masses in the sense that they are not counted by heads, but measured in some other way, e.g. by volume, weight or value. One and the same type of objects may alternatively be regarded as cardinal or mass. For example, a car is individually identified as an object in a car register, but may be regarded as a non-identified part of a mass of cars in production statistics, like a barrel of oil or a kilogram of butter.

When looking at the world from an economic perspective, one may distinguish between real and financial utilities (assets), where a financial utility represents a potential for its owner to acquire a real utility of any kind, or exchange it for another financial utility. A liability is a negative financial utility.

Real utilities may be completely physical (e.g. commodities) or more or less abstract (e.g. services). Undesirable utilities, “unutilities”, may be called “evils”, e.g. diseases, disasters, accidents, and crimes (acts of human beings). Note that it is the type of evil that is the utility object; the individual cases or incidences of the evil are activity objects. It is usually the cases/incidences that are counted in official statistics, but the types are used for classification purposes. One may also view the count of cases of an evil as a measure of the frequency or severity of the evil as such. Desirable utilities like “education” may be analysed analogously.

TOPIC	ACTORS	EVENTS, PROCESSES, RELATIONS	UTILITIES
<b>Population</b>	<p><b>Person</b></p> <ul style="list-style-type: none"> <li>.Sex</li> <li>.Age</li> <li>.HomeLocation</li> <li>.MaritalStatus</li> <li>.WorkStatus</li> <li>.WorkLocation</li> <li>.Income(by Kind)</li> <li>.Wealth(by Kind)</li> <li>.EducationLevel</li> <li>.HealthStatus</li> <li>.CriminalStatus</li> <li>.SocioEconomicGroup</li> </ul> <p><b>Household</b></p> <ul style="list-style-type: none"> <li>.Size</li> <li>.Income(by Kind)</li> <li>.Wealth(by Kind)</li> </ul>	<p><i>Events:</i></p> <p><b>Birth</b>[Person]  <b>Death</b>[Person]  <b>GetMarried</b>[Person]  (or <b>GetMarried</b>[Person,  Person])  <b>GetDivorced</b>[Person]  (or <b>GetDivorced</b>[Person,  Person])  <b>SeekAsylum</b>[Person]  <b>GetAsylum</b>[Person]  <b>SeekCitizenship</b>[Person]  <b>GetCitizenship</b>[Person]  <b>Migration</b>[Person,  FromDwelling, ToDwelling]</p> <p><i>Relations:</i></p> <p><b>Membership</b>[Person,  Household]  <b>Residence</b>[Person, Dwelling]  <b>MarriedTo</b>[Person, Person]  <b>ChildOf</b>[Person, Person]  <b>Commute</b>[Person,  FromRealEstate,  ToRealEstate]</p>	<p><b>RealEstate/Building/Dwelling</b></p> <ul style="list-style-type: none"> <li>.Location</li> <li>.Type</li> </ul>

TOPIC	ACTORS	EVENTS, PROCESSES, RELATIONS	UTILITIES
<b>National accounts</b>	<b>Actor</b> .Sector .TypeOfActor .KindOfActivity	<i>Events:</i> <b>ProductionTransaction</b> [Actor, Utility] .Amount <b>ConsumptionTransaction</b> [Actor, Utility] .Amount <b>InvestmentTransaction</b> [Actor, Utility] .Amount <b>WorkTransaction</b> [Actor, Utility] .Hours	<b>Utility</b> .Type/Purpose .Type/Durability

## Annex 3. Data/metadata structures for SDMX

This Annex contains a pseudo-formal proposal for data/metadata structures that would be compatible with existing data/metadata structures in national and international statistical agencies, and which can hopefully also be mapped easily into existing and future SDMX standards.

### GenericDataSet

- .Source\* (one or more surveys and registers)
- .ReferenceTimeInterval (Begin – End)
- .Frequency (Regular(Year, Quarter, Month, ...), Irregular, Continuous)
- .QualityInformation
  - TimeVariable** (if applicable)
    - .GenericDefinition
    - .GenericValueSetReference
  - SpaceVariable** (if applicable; e.g. Country, Region)
    - .GenericDefinition
    - .GenericValueSetReference
  - ObjectType** (counted and/or measured statistical unit)
    - .GenericDefinition
      - ComponentObjectType\*** (if applicable)
        - .GenericDefinition
  - Population**
    - .GenericDefinition
  - DomainOfInterest\*** (if applicable)
    - .GenericDefinition
  - CrossClassification\***
    - .GenericDefinition
  - ClassificationVariable\*** (if applicable)
    - .GenericDefinition
    - .QualityInformation
    - .GenericValueSetReference
  - Parameter\***
    - SummationVariable\***
      - .GenericDefinition
      - .GenericValueSetReference
    - SummationMeasure** (Count, Sum, Average, Median, Variance, Covariance, Correlation)

## **DataSetInstance**

- .GenericDataSetReference
- .DataSetInstanceIdentification
- .Source\* (one or more survey instances and register snap shots)
- .ResponsibleSender (Country/Agency, Department, Person)
- .QualityInformation
  - TimeVariable** (if applicable)
    - .ValueSetReference
    - .Value\*
  - SpaceVariable** (if applicable; e.g. Country, Region)
    - .ValueSetReference
    - .Value\*
  - ObjectType** (counted and/or measured statistical unit)
    - .DefinitionUsed
      - ComponentObjectType\*** (if applicable; for complex objects)
        - .DefinitionUsed
  - Population**
    - .OperationalDefinitionUsed
    - .QualityInformation
  - DomainOfInterest\*** (if applicable)
    - .Definition
    - .QualityInformation
  - CrossClassification\***
    - .Definition
      - ClassificationVariable\*** (if applicable)
        - .Definition
        - .QualityInformation
        - .ValueSetReference
      - Parameter\***
        - .Description
        - .ValueSetReference
      - SummationVariable\***
        - .Definition
        - .QualityInformation
        - .ValueSetReference
      - SummationMeasure** (Count, Sum, Average, Median, Variance, Covariance, Correlation)
      - Value\*** (cube spanned by classification variables)
        - .QualityInformation

## **GenericSurvey** (including on-going registers)

### .General

- ..Name
- ..Identification
- ..Topic\* (according to one or more classifications)
- ..Legislation (e.g. EU regulation)
- ..Frequency (Regular(Year, Quarter, Month, ...), Irregular, Continuous)
- ..ReferenceTimeInterval (Begin – End)

### .PurposesAndUsages

- ..IntendedPurposesUsersAndUsages

### .Contents

- ..MainContentsOverview (object types / statistical units, populations, variables, parameters)

### .Processes

- ..MainProcessesOverview (input – thrupt – output)

### .GeneralQualityConsiderations (according to some dimensional quality concept, e.g. Eurostat's:)

- ..Relevance
- ..Accuracy
  - ...CoverageErrors
  - ...SamplingErrors
  - ...MeasurementErrors
  - ...ProcessingErrors
  - ...NonResponseErrors
  - ...ModelAssumptionErrors
- ..TimelinessAndPunctuality
- ..Accessibility
- ..Comparability
  - ...InSpace
  - ...OverTime
  - ...BetweenDomains
- ..Coherence
  - ...Provisional\_Final
  - ...Annual\_ShortTerm
  - ...WithNationalAccounts
  - ...WithOtherStatistics

## **SurveyInstance** (or RegisterSnapShot)

### .General

- ..SurveySeriesNameAndIdentification
- ..ReferenceTime (survey instance identification)

### .PurposesAndUsages

- ..ActualUsersAndUsages (as observed; deviations from intended users and usages specially noted)

### .Contents

- ..MainContentsOverview (deviations from survey series description)

### .Processes

- ..MainProcessesOverview (deviations from survey series description)

### .SpecificQualityData (e.g. process data for this particular survey instance)

- ..Relevance
- ..Accuracy
  - ...CoverageErrors
  - ...SamplingErrors
  - ...MeasurementErrors
  - ...ProcessingErrors
  - ...NonResponseErrors
  - ...ModelAssumptionErrors
- ..TimelinessAndPunctuality
- ..Accessibility
- ..Comparability
  - ...InSpace
  - ...OverTime
  - ...BetweenDomains
- ..Coherence
  - ...Provisional\_Final
  - ...Annual\_ShortTerm
  - ...WithNationalAccounts
  - ...WithOtherStatistics

**GenericValueSet** (GenericClassification)

**ValueSetInstance** (ClassificationInstance)

## Annex 4. Data/metadata structures for SDMX: Subject-matter domain *examples*

This annex contains a pseudo-formal proposal for data/metadata structures that would be compatible with existing data/metadata structures in national and international statistical agencies, and which can hopefully also be mapped easily into existing and future SDMX standards.

### GenericDataSet: **Births in Sweden (yearly) by region and sex**

.Source: **The Swedish Population Register**

.ReferenceTimeInterval (Begin – End): **1960 -**

.Frequency (Regular(Year, Quarter, Month, ...), Irregular, Continuous): **Year**

.QualityInformation

**TimeVariable** (if applicable)

.GenericDefinition

.GenericValueSetReference

**SpaceVariable** (if applicable; e.g. Country, Region): **Country=Sweden**

.GenericDefinition

.GenericValueSetReference

**ObjectType** (counted and/or measured statistical unit): **BirthOfPerson**

.GenericDefinition

**ComponentObjectType\*** (if applicable): **Person**

.GenericDefinition

**Population**

.GenericDefinition: **Live births that have occurred in the country during a certain year**

**DomainOfInterest\*** (if applicable)

.GenericDefinition

**CrossClassification\***

.GenericDefinition

**ClassificationVariable:** **Region(County/Commune)**

.GenericDefinition

.QualityInformation

.GenericValueSetReference: Ref to generic value set of Region

**ClassificationVariable:** **Sex**

.Definition

.QualityInformation

.ValueSetReference: Ref to generic value set of Sex

**Parameter:** **NumberOfBirths**

.Description

.ValueSetReference: **The set of non-negative integers**

**SummationVariable:** **1**

.GenericDefinition

.GenericValueSetReference

**SummationMeasure:** **Count**

**DataSetInstance:** Births in Sweden 2005 by region and sex

.GenericDataSetReference: Births in Sweden (yearly) by region and sex

.DataSetInstanceIdentification: 2005

.Source: The Swedish Population Register at the end of year 2005

.ResponsibleSender (Country/Agency, Department, Person)

.QualityInformation

**TimeVariable** (if applicable)

.ValueSetReference

.Value\*

**SpaceVariable** (if applicable; e.g. Country, Region): Country=Sweden

.ValueSetReference

.Value\*

**ObjectType** (counted and/or measured statistical unit): BirthOfPerson

.DefinitionUsed

**ComponentObjectType\*** (if applicable; for complex objects): Person

.DefinitionUsed

**Population**

.OperationalDefinitionUsed: Live births that have occurred in the country during 2005

.QualityInformation

**DomainOfInterest\*** (if applicable)

.Definition

.QualityInformation

**CrossClassification\***

.Definition

**ClassificationVariable:** Region(County/Commune)

.Definition

.QualityInformation

.ValueSetReference: Ref to a value set of Region(County/Commune)

**ClassificationVariable:** Sex

.Definition

.QualityInformation

.ValueSetReference: Ref to a value set of Sex

**Parameter:** NumberOfBirths

.Description

.ValueSetReference: The set of non-negative integers

**SummationVariable:** 1

.Definition

.QualityInformation

**SummationMeasure:** Count

**Value\*:** (Cube of values)

.QualityInformation

## **Annex 5. OECD's common metadata items (March 2006)**

1. Contact person and organisation
2. Data source(s) used
3. Name of collection / source used
4. Direct source
5. Source Periodicity
6. Source metadata
7. Date last input received from source
8. Unit of measure used
9. Power code
10. Variables collected
11. Sampling
12. Periodicity
13. Reference period
14. Base period
15. Date last updated
16. Link to Release calendar
17. Contact person
18. Other Data characteristics and collection
19. Statistical population
20. Geographic coverage
21. Sector coverage
22. Institutional coverage
23. Item coverage
24. Population coverage
25. Product coverage
26. Other coverage
27. Key statistical concepts used
28. Classification(s) used
29. Aggregation & consolidation
30. Estimation
31. Imputation
32. Transformations
33. Validation
34. Index type
35. Weights
36. Seasonal adjustment
37. Other manipulation & adjustments
38. OECD Dissemination format(s)
39. Recommended uses and limitations
40. Quality comments
41. Other comments

In addition to this list, OECD also uses a small set of coded standard metadata attached to the individual observation values, so-called flags:

Flag code list

Code value	Code description
A	Normal value
B	Break
E	Estimated value
F	Forecast value
H	Missing value, holiday or weekend
L	Missing value; data exist but were not collected
M	Missing value; data cannot exist
P	Provisional data
S	Strike

## **Annex 6. SDMX cross-domain concepts (March 2006)**

1. Accessibility of Documentation
2. Accounting conventions/basis
3. Accuracy
4. Classification systems \*\*
5. Comparability/Coherence
6. Concepts and definitions
7. Confidentiality
8. Contact
9. Data presentation \*\*
10. Date of update \*\*
11. Dissemination formats
12. Frequency and Periodicity
13. Legal authority and reporting requirements\*\*
14. Professionalism and ethical standards\*\*
15. Quality management (including resource management)\*\*
16. Release calendar
17. Relevance
18. Revision policy and practice
19. Scope / coverage \*\*
20. Simultaneous release
21. Source data \*\*
22. Statistical processing \*\*
23. Supplementary data \*\*
24. Timeliness and punctuality
25. Transparency \*\*
26. Validation \*\*

## ***Annex 7. Mapping of concepts***

To be elaborated.

***Annex 8. Examples of mapping of metadata***

***To be elaborated***

## ***Annex 9. Examples of SDMX-ML metadata messages***

To be elaborated