Distr.
GENERAL

Working Paper No.20
14 March 2008

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2008)
(Luxembourg, 7-9 April 2008)

Topic (iv): User perspective

SHOWING UNITS IN ONLINE DATABASES – THE EXAMPLE OF OECD.STAT [1]

"There is nothing more practical than a good theory" D. Hilbert

This paper analyses the way units are shown in online databases based on multi-dimensional datasets, taking as an example the OECD statistical data warehouse OECD.Stat. It identifies a number of problems that are believed to be quite widespread and proposes a common way, in which unit of measurement would need to be presented in the OECD.Stat Browser in the future, irrespective of how the units are represented in the structure of the relevant datasets.

I. THE PROBLEM

1. As internal and external use of OECD.Stat accelerates, it has become more and more obvious that users lack sufficiently clear information on the unit of measurement employed in individual datasets or parts of datasets, and that units are not easily findable in OECD.Stat. This has given rise to many emails and queries from users of OECD data.

2. This situation is partly because different databases that have largely been developed independently of each other have been merged into OECD.Stat. One result of this is that there is no uniform principle for organising datasets, and consequently, unit of measurement can appear in many ways, as illustrated in Section 4.1 to 4.4 below, where a number of different ways of showing units are highlighted. For a few datasets, the unit does not appear anywhere because of (almost) total lack of metadata.

3. These problems in the OECD.Stat data warehouse are far from unique; similar problems exist in many other organisations' databases that have evolved from decentralised and uncoordinated production systems. One way of addressing the problem is to enforce a standardisation or normalisation of the structures of datasets, e.g. in accordance with the principles described in the OECD.Stat contents guidelines[2]. This may be difficult to achieve, and an alternative is to adopt work-arounds to allow a reasonable presentation of the data.

---

[1] This paper was prepared jointly for the MSIS meeting in Luxembourg 7-9 April 2008 by Bo Sundgren, Statistics Sweden, and Lars Thygesen, OECD

[2] Guidelines for OECD.Stat Contents, http://www.oecd.org/dataoecd/22/3/39137688.pdf

4. The outline of the paper is to first (in chapter 2) describe recommended standards for organising a statistical database or data warehouse like OECD.Stat, based on multidimensional hypercubes (i.e. multidimensional tables following certain rules), and (also in chapter 2) state recommended principles for management of metadata in connection with statistical data organised in hypercubes, then (in chapter 3) move on to look at existing practices in OECD.Stat as regards data structuring and metadata management. It will be studied (in chapter 4) how the existing practices deviate from recommended standards, and what are the consequences of these deviations. Then (in chapters 5 and 6) improvements of existing practices will be proposed; the improvements aim at narrowing the gap between OECD practices and recommended standards, so as to make it easier for users to interpret OECD data and decrease the risks of misunderstandings.

II. Multidimensional databases, metadata, and measurement units

A. Basic statistical hypercubes

5. A basic statistical hypercube is a multidimensional structure, where each *dimension* corresponds to a classification variable, classifying the objects in a population associated with a set of aggregated statistical data (macro data). The cells in the hypercube contain estimated values of a number of *parameters*, where a parameter is a *statistical measure* (e.g. count, sum, average, variance, correlation, index) applied to an observation variable, usually quantitative (e.g. income), or a vector of observation variables[3].

Examples:

(i) Population: Residents in Sweden by the end of 2005. Classification variables: Sex, Region. Parameters: Count, Sum(Income), Average(Income)=Sum(Income)/Count.
(ii) Population: Road accidents in Sweden during 2005. Classification variables: Region, Month. Parameters: Count, Sum(NumberOfPersonsKilled).[4]
(iii) Population: Consumer price measurement transactions in Sweden, June 2005. Parameter: CPI(commodity/service, weight, measured price, base period).

6. The structure and contents of basic statistical hypercubes follow strict rules (as illustrated by the examples above), and they may hence be regarded as **regular hypercubes**.

B Extended statistical hypercubes

7. A basic statistical hypercube may be extended with further dimensions in different ways. This may lead to a situation where all “slices” in the hypercube may no longer be associated with the same population. Some examples are shown in the following.

Time-extended statistical hypercubes

8. A time dimension may be added. In the first example above, this would mean that the hypercube would contain a ”time slice” for each year during a number of years: Residents in Sweden by the end of the year (1990-2005). There would be a different population of residents each year. This changes in no way the cube’s character of being basic, conforming to rules.

Country-extended statistical hypercubes

9. A country (and/or country group) dimension may be added. This is very common in statistics compiled by international organisations like the OECD. In the first example above, this could mean that the hypercube

[3] For example, in the case of a correlation, the vector contains the two variables between which the correlation is estimated.
[4] “NumberOfPersonsKilled” is originally a micro level observation variable for the object type “RoadAccident”. The value set of this variable is the set of non-negative numbers, and the observed values are aggregated by means of the summation operator, “Sum”.

would contain a ”country slice” for each member country of the OECD: Residents in OECD countries by the end of 2005 by Country, Sex, AgeGroup – for example. A hypercube could be extended by time and country at the same time: Residents in OECD countries by Country, Sex, AgeGroup (1990-2005). This changes in no way the cube’s character of being basic, conforming to rules.

Parameter-extended statistical hypercubes

10. As an alternative to viewing the cells of a statistical hypercube as containing estimates of a number of different parameters, one could view the parameters as a dimension in its own right. This is done in some datasets in OECD.Stat.

Example: Residents in OECD countries 1990-2005 by Country, Time, Sex, and AgeGroup: Count, SumOfIncome, and AverageOfIncome

Dimension 1: Country (Member countries of the OECD)
Dimension 2: Time (1990-2005)
Dimension 3: Sex
Dimension 4: AgeGroup
Dimension 5: Parameters (Count, Sum(Income), Average(Income))

Partial sums, marginal sums, and total sums

11. Each dimension in a basic statistical hypercube may be extended with values corresponding to partial sums (subtotals for subdomains of the population) and the marginal sum over all classes in the particular dimension). Then there may also be partial sums and a grand total, where other partial sums are summarised.

12. As long as the extensions (like those exemplified above) of basic hypercubes are done in accordance with well-defined rules, the extended hypercubes may still be regarded as **regular hypercubes**. Other extensions, that is, extensions or adaptations which are not made in accordance with well-defined rules, will result in **irregular hypercubes**, or data structures that cannot be regarded as hypercubes at all.

13. For example, in a basic hypercube the subclasses corresponding to the basic values in a certain dimension are assumed to be mutually exclusive. If a basic hypercube is extended with partial and total sums, and this is done in accordance with certain rules, the hypercube will still be regarded as regular, although the same object will now be counted in several cells of the same hypercube. But the classes corresponding to the same level in the classification hierarchy will still be mutually exclusive.

14. In OECD practice it may occur that the classes mentioned above are overlapping (that is, are not mutually exclusive). Such hypercubes are not regarded as regular, and they are dangerous in the sense that they imply deceptive double-counting and risks of misinterpretations. A special case of this is that a dimension is just a collection of values that are not connected in any logical sense, a kind of concatenation of dimensions

C. Irregular statistical hypercubes

15. In OECD.Stat there are many examples of statistical hypercubes that are irregular in ways that may increase the risk of misunderstandings. Such cubes can be very difficult to interpret, especially if relevant metadata are missing.

Several classification variables combined (concatenated) in one dimension

16. This is a rather common irregularity in OECD.Stat. The reason for having this kind of irregularity is probably that the people who have structured the hypercubes have inherited the structures used in two-dimensional, printed publications. In such presentations it may be practical to let, say, two different variables be

concatenated in the vertical dimension, if they share the same combination of variables in the horizontal dimension. The purpose could be to make maximum use of a paper page.

17. The combination of the variables may take two different forms. The first is that of a logical product, i.e. variable no. 2 takes on all its different values for each value of variable no. 1.

Example Pensions statistics:

The Pensions statistics production database contained[5] a dimension called *Type* with the following members:

A: Total All Funds
- A1: Pension funds (autonomous), total
- A2: Book reserves (non-autonomous), total
- A3: Pension insurance contracts, total
- A4: Other

B: By pension plan type
- B1: Occupational pension plans, total
  - B11: Defined benefit, total
    - B111: Pension funds (autonomous)
    - B112: Book reserves (non-autonomous)
    - B113: Pension insurance contracts
  - B12: Defined contribution (protected), total
    - B121: Pension funds (autonomous)
    - B122: Pension insurance contracts
    - B123: Investment companies managed funds
    - B124: Banks managed funds
  - B13: Defined contribution (unprotected), total
    - B131: Pension funds (autonomous)
    - B132: Pension insurance contracts
    - B133: Investment companies managed funds
    - B134: Banks managed funds
- B2: Personal pension plans, total
  - B21: Defined contribution (protected), total
    - B211: Pension funds (autonomous)
    - B212: Pension insurance contracts
    - B213: Investment companies managed funds
    - B214: Banks managed funds
  - B22: Defined contribution (unprotected), total
    - B221: Pension funds (autonomous)
    - B222: Pension insurance contracts
    - B223: Investment companies managed funds
    - B224: Banks managed funds

It can be seen that this is in reality three independent variables or dimensions:

Pension plan type with the following value set:
B1: Occupational pension plans
B2: Personal pension plans
(B: Total)

Definition type with the following value set:
1: Defined benefit
2: Defined contribution (protected)

[5] See http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=PNN
. It has actually been decided to split this variable up in its component variables in 2008.

3: Defined contribution (unprotected)

Contract type with the following value set:
1: Pension funds (autonomous)
2: Pension insurance contracts
3. Book reserve (non-autonomous)
4: Investment companies managed funds
5: Banks managed funds

18. The other form consists in simple concatenation, mentioning first all the values of the first variable, then of the second, etc.

Example: Dataset: Fisheries[6]

Contains a variable called Products with, among others, these values:
P - PINK SALMON
P - CHUM SALMON
P - SOCKEYE SALMON
...
P - OTHER FISH
P- TOTAL FISH
P - LOBSTER (ROCK OR EUROPEAN)
P - NORWAY LOBSTER (NEPHROPS)
P - SHRIMPS
...
P - OTHER MOLLUSCS (INCL. SEA URCHINS)
P - TOTAL SHELLFISH AND MOLLUSCS
P - OTHER MARINE ANIMALS
...
E - HARVEST SECTOR
E - HARVEST SECTOR INLAND FISHERIES
E - HARVEST SECTOR INLAND FISHERIES MALE
E - HARVEST SECTOR INLAND FISHERIES FEMALE
E - HARVEST SECTOR MARINE FISHERIES (COASTAL)
...
A - ATLANTIC SALMON
A - PACIFIC SALMON
...
F - VESSELS WITH ENGINES UNKNOWN
F - VESSELS WITH ENGINES 0 M
...
F - VESSELS WITH ENGINES
F - VESSELS WITHOUT ENGINES
F - TOTAL VESSELS

It is clear that the variable contains 4 variables, each with aggregations at the end and sometimes partial aggregations in the middle. The 4 dimensions are not relevant for the same statistics; e.g. part P is relevant for part F is only relevant for landings of fish. It is mentioned in the metadata of the dataset that it consists of 6 datasets:
Table 1: Employment
Table 2: Fishing fleet
Table 3-5: Statistics on landings
Table 8: Aquaculture

[6] See http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=AGRFISH
Sse also Example 1 in attachment

D. Hypercube dimensions (n) and presentation dimensions (2)

19. In a multidimensional database or data warehouse, the terms ”vertical” and ”horizontal” have no meaning, of course. Using pivoting, any dimension in a regular hypercube could be presented in any hierarchical order in either the vertical or horizontal presentation dimension. In general n different dimensions in a regular n-dimensional hypercube may be mapped into the two presentation dimensions (horizontal and vertical) in a large number of different ways. The user should have full freedom to use this flexibility, and no particular presentation format should be implied by the logical structure of the hypercubes.

E. Measurement units

20. In a regular basic statistical hypercube every estimated parameter in the cells (or alternatively in the parameter dimension) is associated with a certain measurement unit. If there is only one parameter, or all parameters have the same measurement unit, this measurement unit could be stated once and for all for all figures in the hypercube. Otherwise different measurement units have to be associated with the names of the parameters, when presented, for example by means of explicit text labels, or by means of footnotes, labels activated by moving the cursor over other text labels (“mouse-over”), or similar features.

21. Similarly, every classification dimension in a regular statistical hypercube is associated with a value set (or a codelist, as they are also called).

F. Standard for naming of hypercubes (datasets)

22. In databases based on hypercubes or multidimensional datasets, users are normally offered navigation in several different ways, including hierarchical tree structure navigation by themes and sub-themes and full text search. In either case, the users will get to the point where they see the dataset name and need to be able to decide whether this seems to be the data they are looking for.

23. In order to allow for this, the name of the dataset should be composed of a number of elements, ordered in a standardised way:

1. Population(s), the nature of its units as well as its delimitation; e.g. “Population in Denmark 1 January” ; here the delimitation by country may be implied, for instance if it is in a database covering Denmark.
2. Dimensions by which the data can be broken down, e.g. “by region, age, sex, marital status”
3. Operators, that is which kind(s) of measure is (are) being shown in the dataset; in the case of a count, this may be unnecessary to mention, as in the case of population, but if it is the annual mean population the word mean should be stated
4. Unit(s) of measurement; in case of a count, the unit may be implied and thus omitted; example: “Million USD”
5. The time covered by the dataset; e.g. “1970-2008”

24. Putting all of this together, a dataset name could look as one of the following examples[7]:

- Population in Denmark 1 January by region, age, sex, marital status (1970-2008)
- People acquiring Danish citizenship by sex, age and former citizenship (1979-2007)
- Main accounts per capita (current prices, DKK) by account (1966-2006)
- Harvest in Denmark by region, crop and unit. (2006) ; here unit is a dimension mixing measure and unit with the following values: Area, 1000 hectares; Average yield, hkg per hectare; Production, mill. kg

[7] Adapted from the Danish statistical database http://www.statbank.dk/

III. ORGANISATION OF OECD.Stat

25. The "classical" method of organising a statistical database recommended in the literature[8] is that of a collection of logically separated, well organised multidimensional tables (or cubes) following certain rules, as described in the preceding chapter. This method has been the model for the design of OECD's Statistical Information System, and this is indeed the data organisation principle recommended in the Guidelines for OECD.Stat Contents. This section contains an extract from the principles.

26. OECD.Stat is organised in a number of relatively independent *datasets*, which are multi-dimensional tables (called hypercubes in chapter 2 above). As time is almost inevitably one of the dimensions, a dataset may be viewed as a number of time series, each defined by values of the other dimensions. However, OECD.Stat is not limited to time series data and can accommodate any multidimensional data structure.

27. A number of "common dimensions" (e.g. country, time, frequency, age) provide links between the datasets. Some of the dimensions are used in almost all datasets, while others are only used in a limited number; when such common dimensions exist, it is possible to merge data from the datasets[9].

28. The database structure is compatible with Statistical Data and Metadata Exchange standards (SDMX). Consequently, SDMX-ML messages can be easily derived from the data warehouse[10].

29. OECD.Stat builds on a hierarchical thematic structure. The first level is very stable and is decided by the Statistical Policy Group SPG; it will normally not be revised more frequently than every two or three years. The second level can be amended or otherwise changed by the unit overseeing content coordination[11] as needed; requests for such changes are usually made by the managers of different datasets.

A. How to delimit a dataset

30. The starting point will normally be a (production) database, where the data manager carries out collection, editing and calculations of the data, until the point where the data are ready to be used or shared with other users or analysts.

31. The production database can correspond to one OECD.Stat dataset, or it can be divided into several datasets – for example, to avoid sparse tables, or to create datasets which will be more intuitively understandable to users. As an example of the latter, the production database Annual National Accounts contains a large multidimensional table, where many combinations of dimension members would not make sense. Users, both specialists and non-specialists, are likely to better understand the content if it is broken down in a number of tables, each related to a way in which users tend to look at national accounts. Accordingly, it was decided to break the database into 25 datasets in OECD.Stat. Each of the datasets is rather dense, meaning that all combinations of dimension members make sense and potentially have data.

Example 2: Break-down of Annual National Accounts (see picture in Attachments, example 2.
All of these datasets share some dimensions: Country, Time, Transaction (derived from the System of National Accounts) and Measure, while others also use Activity (industry). Each table may use a subset of these dimensions , e.g. for Transaction in the dataset Table 1:

B1_GE: Gross domestic product (expenditure approach)
P3: Final consumption expenditure
P31S14: Final consumption expenditure of households
P31S15: Final consumption expenditure of non-profit institutions serving households
P3S13: Final consumption expenditure of general government
P31S13: Individual consumption expenditure of general government

[8] Information Systems Architecture for National and International Statistical Offices: Guidelines and Recommendations. UN/ECE, Geneva 1999. http://www.unece.org/stats/documents/information_systems_architecture/1.e.pdf
[9] Using the so-called Merged Query (also known as Multi-Dataset Query).
[10] In fact, an SDMX web service may be called, to generate and deliver SDMX-ML datasets on the fly.
[11] The Statistical Information Management and Support Division (SIMS)

P32S13: Collective consumption expenditure of general government
P41: of which: Actual individual consumption
P5: Gross capital formation
P51: Gross fixed capital formation
P51PI61: Products of agriculture, forestry, fisheries and aquaculture
P51PI62: Metal products and machinery
P51PI63: Transport equipment
P51PI64: Housing
P51PI65: Other constructions
P51PI66: Other products
P52_P53: Changes in inventories and acquisitions less disposals of valuables

In another, Table 2, Transaction looks like this:

Transaction
B1_GS1: Gross domestic product
TGLS1: Trading gain or loss
GDIS1: Gross domestic income
D1_D4NFRS2: Net primary incomes from the rest of the world
D1_D4FRS2: Primary incomes receivable from the rest of the world
D1_D4TOS2: Primary incomes payable to the rest of the world
B5_GS1: Gross national income at market prices
K1MS1: Consumption of fixed capital
B5_NS1: Net national income at market prices
D5_D7NFRS2: Net current transfers from the rest of the world
D5_D7FRS2: Current transfers receivable from the rest of the world
D5_D7TOS2: Current transfers payable to the rest of the world
B6GS1: Gross national disposable income
B6NS1: Net national disposable income
P3S1: Final consumption expenditures
D8S1: Adjustment for the change in net equity of households in pension funds

B. How to structure a dataset

32. It is very important that a dataset be structured in a simple and understandable way. However, it is not always easy for a production person to structure a dataset, and it may sometimes be difficult even for data structuring experts to suggest how this should best be done. The process will sometimes require good judgment for which it is helpful to know the mindset of the dataset's users. There are, nevertheless, a number of good principles and rules, resulting from sound understanding and experiences of the problems involved. Some of these principles and rules are strongly recommended to be followed by OECD.Stat.

33. In order to create a simple and understandable structure, one should analyse the concepts (real life phenomena) appearing in the data structure. One should try to identify "clean" dimensions that do not mix up what are really many dimensions. This means trying to understand if the dimensions used (in the production database) contain in reality different conceptual variables. If this is the case, they should be broken up – "normalised" and "orthogonalised".

34. All the principles and rules recommended for structuring the data in OECD.Stat actually are implications from the principles and rules that were discussed in a general way for so-called regular hypercubes in chapter 2 above.

C. How to structure metadata

35. Statistical metadata are key to allowing users to locate the data (on the Internet) – so-called "exploratory metadata" - and to making the statistics understandable – so-called "explanatory metadata".[12] Unfortunately, metadata for OECD statistics have often been scarce, incomplete and scattered. Therefore a special set of detailed guidelines and recommendations has been designed and agreed for these metadata.

36. Some of the main points in these guidelines and recommendations are:

1. Statistical metadata should be arranged under a number of common metadata items.
2. The metadata can be attached at any level of detail of the underlying data structure (the attachments level): Dataset (hypercube), dimension, dimension member, combination of dimension members, (aggregate) observation[13].
3. It is recommended to attach metadata at the highest level possible (e.g. dataset/hypercube) and attach deviations from the high level description at lower levels.

For OECD statistics it is recommended that metadata be managed using the OECD tool for statistical metadata management, MetaStore, as it fully supports these guidelines.

IV. DEVIATIONS FROM THE RECOMMENDED ORGANISATION OF CUBES

37. If it had been possible in OECD.Stat to thoroughly follow the recommended logical structure of datasets, units would be common to complete datasets and could be easily communicated to users. However, in a number of cases, the way datasets have been organised in OECD.Stat does not strictly follow the principles above. It has been deemed practical to build more directly on the structure which has, over the years, been developed for the production databases from where the data originate. This has entailed a certain incoherence between datasets and in the way units appear in the datasets.

38. In this chapter a number of examples are given. The datasets can be accessed on the Internet using the given URLs (note that refresh screen may be necessary to see the datasets because of caching), or the reader may be referred to the Annex showing pictures from each dataset.

39. **Unit common to a dataset (hypercube)**

- In the dataset **Trade in Services by Partner Country** (see Attachments example 3 or URL http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=TISP), there is only one unit, Millions of US dollars. This has been turned into a dimension called Unit with only one value. This value will always be shown in the header of the table, so in this case there is no serious problem

- In the dataset **Table I Total Support Estimate** (see Attachments example 4 or http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=PSETOT) almost everything seems to be in national currency. Here the mention of units is in the header of the metadata window: **National Currency except Japan and Turkey in Billion.** This is perhaps not where many users would look for the unit. The common metadata item *Unit of measure used* is not used.

[12] Actually the same metadata item may sometimes have both an exploratory and an explanatory role in different contexts. A simple example is the name of a variable. The variable name may be used (for exploratory purposes) by a search function, when a user is looking for potentially relevant data concerning a certain problem or issue. It may also be used (for explanatory purposes), when the user is trying to interpret the meaning of statistics presented to him or her, provided that the name of the variable is chosen in such a way as to give useful hints about the meaning.

[13] The term "observation" is typically reserved by statisticians for the original observations made in the surveys conducted and registers maintained by the organisation collecting micro data, typically a national statistical agency. An international organisation like the OECD will typically receive only aggregations of the original micro level observations. Maybe one could call these aggregated data "observations" from the international organisation's point of view, but the prefix "aggregate" would then be appropriate to avoid misunderstandings.

- In dataset **Creditor Reporting System** (see Attachments example 5 or http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=CRSNEW) there is a dimension called Amount, and in the metadata of Amount there is the following statement: "Amounts are expressed in USD millions. Deflated amounts (i.e. at base year prices including the effect of exchange rate changes) are also available". The labels of dimension members mention whether it is commitments or disbursements, and at the same time if it is current or fixed prices (e.g. Commitment-current).

40. **Unit in a dimension**

- In the dataset **1--Gross domestic product**, (see Attachments example 2.1 or http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=SNA_TABLE1) Measure is a dimension, and unit of measurement is dependent on this and is to a certain extent mentioned in the member name of each measure. However, for measures of National currency, the unit itself (e.g. AUD) is not mentioned.

- In the dataset **Fisheries** (see Attachments example 1 or http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=AGRFISH) (and other agricultural statistics) the situation is somewhat similar, but here the dimension holding the units is called Variables, and it is not clean as it also holds information like a "half time/full time" break down of employment, depending on the selection of another dimension called Tables.

41. **Unit differs for different members of one dimension in a dataset**

- The dataset **Country statistical profiles 2006** (see Attachments example 6 or http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=CSP6) consists of a selection of different indicators in the dimension Subject, each with their own unit of measurement. Information on the unit is stored in the metadata in the common metadata item *Unit of measure used.* Consequently, a red i appears with each value (member) of Subject, and when clicked, the unit is shown.

- In the dataset **Reference Series** (see Attachments example 7 or http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=REFSERIES), the situation is similar, but here the units are mentioned in the name (label) of the dimension members of Subject. The common metadata item *Unit of measure used* is not used.

- There are several other similar examples, e.g. **Population and Vital Statistics** (see Attachments example 8 or http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=ALFS_POP_VITAL) and Migration (see Attachments example 9).

42. **Unit differs for different combinations of members of more than one dimension in a dataset**

- The key example of this is International Trade by Commodity Statistics (ITCS), e.g. **HS 1988**. Here, values are all expressed in current USD and this is mentioned in the common metadata item *Unit of measure used*. But quantities are available at detailed levels of commodity, and the measuring unit may depend on the commodity, the country and the time. This unit is then expressed as observation values for the member "Quantity unit" of the dimension Measure (the other members being Value, Net Weight (kg) and Quantity.
- Another interesting example is **Unit Labour Costs - Annual Indicators** (see Attachments example 10 or http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=ULC_ANN). Here the dimension Measure has the following two members: "Index OECD base year (2000=100)" and "Level, ratio or national currency". The first of these immediately give the unit of measurement, while the second has to be interpreted in context with the dimension Subject, which has the following members: Unit Labour Cost , Total Labour Cost (millions) , Real Output (millions) , Self-employment ratio , Hours Worked: Employment ('000) , Hours Worked: Employees ('000) , Employment ('000) , Employees ('000) , Exchange Rate Adjusted ULC , Labour Compensation per unit labour input , Labour Productivity , Labour Income Share (Real ULC) , Nominal Output (millions). In relation to, e.g. Unit Labour Costs, the metadata declare: "Annual unit labour costs (ULCs) are calculated as the quotient of total labour costs and real output. Time series are presented in both level and index form where the base year of real output is 2000" – so in this case the unit is Ratio.

43. **Unit not mentioned**

This is simply a lack of metadata that should be repaired. One current example is **Strictness of EPL referring to Employment Labour Legislation** (see http://dotstat.oecd.org/wbos/default.aspx?DatasetCode=EPL) where the implied unit is something like "Score between 0 (minimum) and 5 (maximum)".

V. PROPOSAL – THE DESIRED FUTURE SITUATION

44. One way of solving the problem of different attachment levels of units would be to reorganise all datasets in accordance with the basic principles mentioned above. However, that is not seen as a realistic option.

45. It is proposed to agree on a common way to interpret the concept "unit" and to introduce a new metadata item called "Unit code" with a codelist that will encompass all units that are encountered in OECD.Stat. In addition, it will be necessary to introduce two other new items, "Unit multiplier" and "Unit position". It is proposed to show the units in a uniform way in OECD.Stat and use unit codes also in other access methods to OECD.Stat.

A. What is Unit?

46. The precise meaning of the "unit of measurement" concept is not instantly obvious as it often confused with a number of other related concepts such as "unit multiplier", "reference period", "currency". The unit of measurement is also frequently embedded in the concept being measured. In the OECD context an attempt has been made to apply a very strict delimitation of the unit of measurement concept as illustrated in the following examples:

- For the table heading, "The population of Sweden as of 1 January 1990-2007, broken down by municipality, age and sex", the unit of measurement would be "number" or "count" rather than "number of persons", "persons" being part of the concept measured.
- For the unit text, "Data are expressed as indices with average for a reference year equal to 100", the unit of measurement would be "index", the base period would be stored in another metadata item *Base period*.
- For the unit text, "As a percentage of total non-residential fixed capital formation, total economy", the unit of measurement would be "per cent" (pct), while the other parts of the text refer to the concept being measured.
- For the table "Creditor Reporting System; Flow: Official Development Assistance; Amount: Commitment-current", the unit of measurement would be "Million USD"
- For the unit text "National currency units per US dollar" used for exchange rates, the unit of measurement would be the respective national currencies, e.g. "Australian $ (AUD) per U.S. dollar"

B. New metadata items in OECD.Stat

47. As mentioned above, this proposal entails the introduction of three new metadata items.

Unit code, with a predefined valueset (codelist). This is necessary because the existing item *Unit of measurement* is a free text item that has already been used by a number of databases for text. The codelist should be extended to contain all possible basic units used by the OECD. A starting point should be the preliminary lists that has been discussed in SDMX.

48. *Unit multiplier*, with a predefined valueset (codelist). The Unit multiplier indicates the power of 10 by which the unit given by Unit code should be multiplied; e.g., if the Unit code = USD (US dollars) and Unit multiplier = 6, it means that the indicator is measured in millions of USD.

49. *Unit position*, with a predefined valueset (codelist):

1. Dataset with only one unit

2. Dataset with unit in a separate dimension
3. Dataset where unit depends on only one dimension
4. Dataset where unit differs for different combinations of members of more than one dimension

The Unit position is an "internal" code, to be used of the presentation systems, that indicates where and how the unit of measurement can be found in the dataset. Attachment level is dataset

50. The use of *Unit code* and *Unit multiplier* is mandatory in OECD.Stat, so that every dataset has a value of Unit code and multiplier attached to *all* of its observations. This could be at the dataset level or at lower levels, as long as the whole table (dataset) is covered. If there is no indication of unit, there should be a code for unknown (a situation that is of course unacceptable).

51. The introduction of these values should be completed by a certain date, e.g. 1 December 2007. There will be a tool to help people complete it. Whenever possible, the production system from which the data come should generate the codes in Unit code in MetaStore. This kind of automatic generating metadata is already used by ECO. In very complicated cases such as ITCS dataset mentioned under 1.5 above, there might have to be a special, automatic solution incorporated in OECD.Stat.

C. Showing Unit code in OECD.Stat Browser

52. A standard way of showing the units is introduced in the OECD.Stat Browser, no matter which level they are attached to. The following model is proposed and is illustrated by way of mock-up examples in Annex 1.

53. Presentation of metadata items in the OECD.Stat Browser will be rearranged, so that Unit (based on Unit code, Unit multiplier and Unit position) is always shown first (after the name of the measured variable, whenever applicable [which I suppose is always, or at least almost always?].

54. The title line of each table view will, just following the Dataset name, have a label indicating unit (based on Unit code, Unit multiplier and Unit position). One of the following possibilities will apply:

1. When displaying metadata in OECD.Stat, the text corresponding to Unit code is always shown as the first item in the following way: "Unit of measurement: USD"
2. Datasets with only one unit: "Unit of measurement [UnitName]" (example: "Unit of measurement Million USD). A red i attached to the Dataset name points to Dataset level metadata, where Unit is always shown first with the text corresponding to the Unit code) (in the example given, Million USD). See example 1 in Annex 1.
3. Datasets with unit in a separate dimension: "Unit of measurement depends on the dimension [DimensionName]" (example: DimensionName =Measure; the text becomes "Unit of measurement depends on the dimension Measure"; in this case, either the labels for measure must be self-explanatory, or an i must point to the metadata, where Unit code text appears first, as mentioned above). See example 2.1-2.2
4. Datasets where unit depends on only one dimension: "Units of measurement: Click i for member of dimension [DimensionName] to see unit" (example: "Units of measurement: Click i for member of dimension Subject to see unit"). See example 3
5. Datasets where unit differs for different combinations of members of more than one dimension in a dataset: This is a complicated case, so it will have to be explained in the textual metadata item *Unit of measurement*. The line in the table view should be: "Units of measurement: Please see explanation in metadata i ". In this case, whether you go to dataset metadata or click the i, you will see the text in *Unit of measurement,* which could read "Units of measurement differ for different combinations of [DimensionName1], [DimensionName2], ... (example: DimensionName1=Commodity, DimensionName2=Country, DimensionName3=Time). To see the units for a particular observation, click the i attached to it. See example 4

In order to manage this, there will be a code UnitPosition attached to each dataset indicating the position of the units, as mentioned above.

D. Unit codes in other applications

55. In applications (such as DotStatGet and the .Stat/Populator[14]) that access information from OECD.Stat using the web service, there should be an easy way to extract all unit-related information (Unit codes, Unit position, Power codes) associated with any data value in OECD.Stat.

E. Codelist for Unit code

56. The full codelist of Unit code has to encompass all units that are used in any dataset. It will thus have to develop as contents develop. The codelist will build, as far as possible, on ISO standards such as ISO 4217 [15] .

57. For reasons of clarity of presentation, the codelist is broken down in sub-categories (or the list has a hierarchy where the first level is only for presentation of the list)[16]:

1. Value (currency)
2. Ratios (indexes, percentages, other ratios)
3. Count (“number”)
4. Physical measures (volume, weight (mass), length, area, speed, energy)
5. Time
6. Other

58. Initially, a first list will be drawn up on the basis of (i) all values in any ISO standard pertaining to the measure, (ii) all values or codes that have been identified and agreed in the ongoing SDMX content oriented guidelines work (if any), and (iii) all units that can be found in OECD.Stat. The two sets will be integrated, and in cases where a text in OECD.Stat corresponds to an existing SDMX code, they will be merged.

VI. PROPOSAL – HOW TO GET TO THE DESIRED SITUATION

A. Creating the Unit codes in OECD.Stat / MetaStore

59. At present, data providers have used many different kinds of texts in the Unit of measurement metadata item. The texts contain a number of dimensions that are not really the unit. Some examples illustrate this diversity:

Pure units designations:
AUD
1991=100
2000/2001=100

[14] DotStatGet is a tool allowing analysts using FAME for their analytical work to include and link to data from OECD.Stat in their analytical applications. The .Stat/Populator is a tool allowing from Excel to link to OECD.Stat data by referring to the structural metadata

[15] See codelist on ISO web site: http://www.iso.org/iso/support/faqs/faqs_widely_used_standards/widely_used_standards_other/currency_codes/currency_codes_list-1.htm

[16] This classification is inspired from Gunther Schadow, Clement J. McDonald: The Unified Code for Units of Measure http://aurora.regenstrief.org/UCUM/ , as this was the most comprehensive overview of units we could find. However, the list has been simplified to suit the purpose of statistics.

Year 1995 = 100
Head counts
Numbers
Number per 100 000 population
Number of students
Number of years
Hour worked
Per million population
Annual growth in percentage
Area in km2
Deaths per 1 000 live births
m3/capita
Mean score
National Currency [?]
National currency units per US dollar [?]
OECD = 100, at 2000 price levels and PPPs
Terawatt hours (TWh)

Units and Power code:
AUD millions
1000 Turkish Lira (TRL) per U.S. dollar
Million m3
Thousand tonnes

Units and the concept measured:
Area in km2
As a percentage of total manufacturing value added, change in share 1995-2001
Percentage of households with access to a home computer
1999 PTE euro, EUR from 1999
Total annual hours worked for all jobs
BEA's Full-Time and Part-Time Employees (FTPT). Number of Jobs - underlying source is BLS's CES establishment survey.
Quantity index: 1993=100

Something else:
A pension plan by which benefits to members are based solely on the amount contributed to the plan by the sponsor or member plus the investment return thereon. This does not include plans in which the employer that sponsors the plan guarantees a rate of return
Idem 1.1.1.

60. The first task is to decide if the text corresponds to a unit, and how the other elements of the texts should be represented. How should, e.g. the conceptual descriptions of current or fixed prices, be stored and shown. Then establish a preliminary translation table, translating the texts into Unit code and Unit multiplier values. A first version of this is shown as the last folder in Annex 2.

VII. DISCUSSION

61. The authors believe the problems mentioned in this paper are widespread in statistical databases today, and we know from OECD experience that they give rise to considerable problems of understanding, and to erroneous use of the contents, maybe leading to bad decisions.

62. The solutions proposed here are quite heavy to implement, as they involve all database owners in the organisation. On the other hand, they should create a common and rather robust way of representing units. This is even more important when data are to be presented not only by a dedicated tool designed together with the

database such as the OECD.Stat Browser, but also in other connections. We expect that the data in the database can be presented by other data providers, downloading them from OECD.Stat e.g. in SDMX-ML format. Some data providers may present them in their own graphical tools. This could make the chance of misunderstanding even bigger. Therefore, in this time of extensive data sharing and reuse, a good solution of the units problem becomes even more crucial.

63. We hope that the solutions proposed here could be more widely applied, or at least give inspiration to other organisations operating online statistical databases

References:

Guidelines for OECD.Stat Contents. OECD 6 September 2006,
http://www.oecd.org/dataoecd/22/3/39137688.pdf

Management of Statistical Metadata at the OECD. OECD, 6 September 2006,
http://www.oecd.org/dataoecd/26/33/33869551.pdf

Annex 1: Mock-up examples of showing units (in red)

See http://www.oecd.org/dataoecd/28/35/40284208.pdf

Annex 2: Codelists for Unit code and translation table from texts to Unit code and Unit multiplier

See http://www.oecd.org/dataoecd/28/57/40284040.xls