

Guidelines for OECD.Stat Contents

Version 1, 6 September 2006

Table of Contents

Guidelines for OECD.Stat Contents.....	1
1. Purpose.....	1
2. Structure of OECD.Stat data warehouse contents.....	2
3. Themes and sub-themes.....	2
4. How to delimit a dataset.....	3
5. How to structure a dataset.....	3
6. How to structure metadata.....	5
7. Common dimensions.....	5
7.1 Global common dimensions.....	5
7.2 Domain specific common dimensions.....	8
8. Naming common dimensions.....	8
9. Naming other dimensions.....	9
10. Sequence of variables (in dimension selector).....	9
11. Languages.....	10
12. How to update an OECD.Stat dataset.....	10
13. Making a snapshot dataset.....	10
14. Default views and other queries.....	10
14.1 Deciding on default views.....	10
14.2 Creating default views and other queries.....	11
15. Core Data.....	11
15.1 Deciding on what should be Core Data.....	12
15.2 Creating Core Data.....	12
15.3 Displaying Core Data views on www.oecd.org.....	12
16. How to manage contents in OECD.Stat – DPI user guide.....	13

This paper gives guidance to database administrators (or data providers) on how to efficiently, effectively and properly organise their data and metadata in the OECD.Stat data warehouse to be disseminated as all kinds of outputs, online as well as off-line, electronic or paper format.

In most cases, the data are prepared or produced in a separate production system, which can be either a proprietary system or the generic production system StatWorks. The data structures of these production systems are carried over or mapped to the structure of OECD.Stat, so it is important, when designing production systems, to consider the outputs that are going to be produced.

1. Purpose

Migration of databases to OECD.Stat serves the following objectives:

- Quality improvement
 - Harmonisation of concepts across themes, thus opening the way for cross-cutting analysis

- Good statistical metadata, attached to data, allowing to fully understand their nature and characteristics, and enabling the data to be much more visible on the Internet
- Better coherence of data and metadata across datasets
- User friendliness
 - OECD.Stat offers a true one-stop database, one access method for all OECD statistics and some selected non-OECD datasets
 - Possibility to combine and compare data across domains
 - Possibility to develop alternative ways of seeing the database for different user segments
 - Easy access in very few clicks
 - Common look and feel of all statistical products
- Efficiency
 - Common tools
 - Easy access across OECD
 - One database to serve as a base for all dissemination products (online access to full data or core data, paper publications and electronic publications)

The guidelines given are intended to help the data provider best organise his/her data so as to promote these goals.

2. Structure of OECD.Stat data warehouse contents

OECD.Stat is organised in a number of relatively independent *datasets*. These are multi-dimensional tables (in some writings on statistics also referred to as cubes). 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.

A number of “common dimensions”, described in Section 7 below, provide a link between the datasets.

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.

3. Themes and sub-themes

OECD.Stat builds on a hierarchical thematic structure. The first level is very stable and is decided by SPG; it will normally not be revised more frequently than every two or three years. The second level can be amended as needed. The procedure for doing this is to propose changes to STD/SIMS. Any lower levels of hierarchy can be decided by data providers. It is advised, however, to not use more than three levels, as this could otherwise discourage users from using the data.

At the moment, it is mandatory to use the second level of the hierarchy to categorise datasets.

The datasets are ordered alphabetically within the hierarchy of themes.

4. How to delimit a dataset

The starting point will normally be a (production) database (i.e., 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.

The 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 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 will 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.

5. How to structure a dataset

It is very important that a dataset be structured in a simple and understandable way. However, it is not always easy to structure a dataset, and it's difficult to suggest how this should be done¹. The process involves a lot of personal, subjective judgment for which it is helpful to know the mindset of the dataset's users. There are, nevertheless, a few ideas which can be put forward.

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. An example may help clarify this.

Example Pensions statistics:

The Pensions statistics production database contains 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

¹ This is actually seen by some experts as an art rather than a mere technical exercise, as alluded to in the following article (which cannot be found on the Internet): Ad Willeboordse, Coen van Duin, Jan Willem Altena: Theme building by the art of cubism - Towards a coherent and accessible Output Database. The International Seminar on Statistical Output Data Bases and Marketing, Ottawa, May 2001

- 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)
- 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

It gives a clearer and more understandable structure when each of these is regarded as a dimension.

It is a good idea to avoid as far as possible dimensions with very long, non-hierarchical code lists, as they are very difficult to view. If possible, introduce a hierarchy that will allow the user to look first at the highest level, and then unfold the values of interest.

Example HS (Harmonised System): The detailed commodity classification in International trade statistics, HS, is a 6 digit code with contains 7,000 values (members). Nobody can look at such a list in order to select values. Instead, a hierarchy of 2 digits, 4 digits, and 6 digits is used: Users unfold one level at a time.

6. **How to structure metadata**

Statistical metadata are key to allowing users to locate the data (on the Internet) and to making the statistics understandable. 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 by SPG for these metadata; see “Management of Statistical Metadata at the OECD” http://web.oecd.org/std_int/quality/metadata_principles.doc.

Some of the main points are:

- Statistical metadata should be arranged under a number of common metadata items.
- The metadata can be attached at any level of detail of the underlying data structure (the attachments level): Dataset, dimension, dimension member, combination of dimension members, observation.
- It is recommended that metadata be managed using the OECD tool for statistical metadata management, MetaStore.

7. **Common dimensions**

Common dimensions may be global (i.e., designed to be widely used across several domains and directorates, or domain specific).

All of them can be viewed and managed via the Data Provider Interface, DPI (choose *Manage OECD.Stat Common Dimensions*).

7.1 **Global common dimensions**

OECD.Stat makes use of a limited number of global common dimensions. These are concepts which in some way occur in several datasets, such as country, frequency and time. The common dimensions are maintained as a common infrastructure that may be used by data providers. Basically, a common dimension consists of a value set (or code list) with attached labels or names of the members, and a name of the common dimension. The common dimensions are essentially comprehensive, meaning that they contain all the values that can occur for the variable, even though some of the values may occur only in a few datasets. For example, the common dimension “country” contains all countries of the world (i.e., not just OECD member countries), and it contains different variants of some countries (such as France including or excluding Monaco, DOM and TOM), countries that do not exist any more but can be found in historical statistics, different kinds of aggregated geographical zones (such as EU12, EU19, EU25)².

The value set of a common dimension may be used in many different “roles” (see examples under Country below). The name attached to the common dimension in the dataset may vary in order to reflect this (e.g., “Country”, “Country of citizenship”, “Partner country”).

² It is possible, however, to extend a common dimension to include values which pertain only to a specific dataset. For example, the Country dimension for the Economic Outlook dataset has been extended to include a number of areas that pertain only to this dataset (Example AFM - Africa and Middle East; ANC - Dynamic Asia; LAT - Central and South America; etc.)

The common dimensions constitute an important harmonisation engine across datasets, as they allow for coherence across datasets and give the possibility of combining datasets, using the Multi-dataset query³; this way, use of common dimensions greatly enhance the possibility of producing horizontal dissemination products, embracing several subject-matter domains. They also ease harmonisation to other statistical organisations (NSOs and international organisations) with whom OECD exchanges or shares data. This is done via a close link to SDMX contents-oriented guidelines. When using SDMX exchange mechanisms, the concepts may be used directly, or links to the code lists used by SDMX partners may be established once and for all, being reused by many partners.

In order to obtain harmonisation between datasets in OECD as well as with data of other organisations, common dimensions should be used wherever possible when creating datasets in OECD.Stat.

It is recommended that a common dimension be used even if definitions of some of the members (codes) differ slightly, as in this case it will still be useful to be able to compare with other datasets; in such cases, metadata of the dataset should be used to explain differences or exceptions.

Example France: France is generally defined as the economic area including Monaco and the French overseas departments (DOM, 'Départements d'outre-mer'), but excluding the French overseas territories (TOM, 'Territoires d'outre-mer'). In the case of International Trade statistics, the country France may (in some parts of the statistics) be defined slightly differently (i.e., including TOM). However, it will still be reasonable to use the common code for France, as it is useful to be able to see the data for this area together with other data for France. The differing definition should therefore be described in the metadata at the country level for France for the dataset (or perhaps only for a combination of France and some values of other dimensions of the dataset)

The code lists underlying the common dimensions may already be used in the database from which the data come, in which case there is no problem. If a different code list is used in the production database, the data provider must translate those values which differ into the corresponding common dimension values when the dimension is created and when the data are exported.

If a new dataset needs values that have not been used before in the common dimension, it is possible (and recommended) to extend the common dimension(s) with non-common (dataset-specific) members as necessary.

Country

The common country dimension has the internal name LOCATION. It contains codes and names for all countries and aggregates of countries (zones) that have been used in OECD statistics. The codes are based on ISO three digit alphabetic codes, to which codes have been added for countries not included in that nomenclature, as well as for zones. The names reflect the official OECD way of naming and spelling each country. The system is hierarchical, as it shows which countries are part of which zones. In order to correctly identify ISO country codes, please refer to the United Nation Statistics Division internet page, 'Countries and areas, codes and abbreviations' at the following address:

<http://unstats.un.org/unsd/methods/m49/m49alpha.htm>. The names given for OECD member countries are the official names used by the OECD

This dimension is normally represented as "Country" in dimension selector as well as in tables. However, the value set country can have different roles. The most common role is that of "reporting country", which

³ This will allow users to combine data from two or more datasets, merging them by common dimensions.

most frequently coincides with the “reference area” (i.e., the economic area to which the data refer). This role of the country value set is referred to simply as “Country”. Other roles include cases where there is reference to more than one country in a dataset, such as vis-à-vis or partner country – for example, in international trade (e.g., “Partner Country”), country of citizenship of persons (e.g., “Country of Citizenship”) or country of birth.

In all cases, it is recommended to use the same value set (LOCATION). This will ensure that data can be combined across datasets also using this dimension (i.e., look at trade partners in relation to their economic size). The LOCATION dimension holds all of the correct code values as well as labels and metadata. It facilitates the use of aggregates of countries, and the metadata of the aggregates show which countries are included

The value set of LOCATION can be extended (specifically to each dataset) to include any country or any aggregate zone that is missing.

Time and Frequency

The common dimensions Time and Frequency are two separate, albeit interlinked, dimensions. If only one frequency is used in a dataset, such as annual data, the Frequency dimension will have only one value, and it should be omitted altogether.

Currently the system accommodates time/frequency specifications for monthly/quarterly/semi-annual and annual data.

Common dimension name FREQUENCY. Contains the following value set:

<i>Code</i>	<i>English_Name</i>	<i>Nom_français</i>
A	Annual	Annuelle
D	Daily	Quotidienne
M	Monthly	Mensuelle
Q	Quarterly	Trimestrielle
S	Semiannual	Semestrielle
W	Weekly	Hebdomadaire

Common dimension name TIME. Contains the following value set:

<i>Code</i>	<i>English_Name</i>	<i>Nom_français</i>
1900	1900	1900
1900M1	Jan-1900	Janv-1900
1900M10	Oct-1900	Oct-1900
1900M11	Nov-1900	Nov-1900
1900M12	Dec-1900	Déc-1900
1900M2	Feb-1900	Févr-1900
Etc.		

Sex

Common dimension name SEX. Contains the following value set:

<i>Code</i>	<i>English_Name</i>	<i>Nom_français</i>
W	Women	Femmes
M	Men	Hommes
T	Total	Total

Age

This value set contains all kinds of age classes, defined from a one-year classification: 1-year classes, 5-year classes, etc.

Common dimension name AGE. Contains the following value set:

<i>Code</i>	<i>English_Name</i>	<i>Nom_français</i>
1	1 year	1 an
1_10	1 to 10	1 à 10
1_100	1 to 100	1 à 100
1_11	1 to 11	1 à 11
1_12	1 to 12	1 à 12
1_13	1 to 13	1 à 13
1_14	1 to 14	1 à 14
1_15	1 to 15	1 à 15
Etc.		

Industry classifications (ISIC rev. 3)

This common dimension has not yet been defined but is in preparation.

7.2 Domain specific common dimensions

In addition to the “real” common dimensions, domain specific common dimensions – used in several datasets within a domain – can and should also be loaded as common dimensions; they are managed by the corresponding directorate and not by STD/SIMS. The domain specific common dimensions include a reference to the domain as a prefix to their name.

Example Labour Force Statistics: ELS have created specific dimensions that are common to several datasets within their area, such as "Employment status" (name LFS_EMPSTAT), or "Programme" (name LFS_PROGRAMME).

8. Naming common dimensions

The following common naming convention applies for common dimensions:

- Dimension names should use the singular form instead of the plural (i.e., Country is always called Country and not Countries). Of course one will have to take into account that Country can play different roles, as in foreign trade. Therefore it may be necessary to specify this role.
- Time & Frequency likewise (sometimes called Year, sometimes Time but when you click it you get Time / frequency).

9. Naming other dimensions

The following recommendations apply to the non-common dimensions:

- Try to avoid giving a dimension name without any reference to its contents (i.e., avoid "Series"); the name should relate to the concepts that are specified by the dimension.
- Don't use "Type". If the dimension means something, use it (i.e., Residential Status instead of Type)
- Don't use Classification, but tell what it is (i.e., Assets and Liabilities)
- If the concepts in the dimension can not be described by a short name (i.e., because it really comprises many different aspects): Don't use Variable, but use rather Subject or Topic (Variable doesn't mean anything to a non-statistician)
- Measure should only (and always) be used to describe the concept measured inside the table; if there is only one measure, there should be no such dimension; measure should only be described in title and metadata
- Dimension member names should never include a code value for the member, it should only contain a name making it clear (to the extent possible within a short label) what the conceptual content is. If the data administrator wishes to present codes for the members of a dimension together with the names when displaying them in tables on the web browser, she/he should use the function "Change Code Prefix setting:" in the Data Provider Interface; this will make the code and name appear separated by a colon and a space.

Example:

Dimension member name "Consumer price index",

Code value "CPI"

Presented as "CPI: Consumer price index"

10. Sequence of variables (in dimension selector)

It is easier for a user to understand if the sequence of the dimensions of different datasets resemble one another. Therefore it is recommended to organise datasets in a similar manner unless there are good reasons to do otherwise. The main recommendations are the following:

- Frequency & Time/Time should always be the last dimension(s); This will assure high performance because of the OECD.Stat data warehouse structure
- The first variables should always be the other common ones (to the extent they are represented) in the following order: Country (reporting), Industry (ISIC rev. 3), Age, and Sex.

Deviations from this order may be made in cases when the database administrator finds it useful to mention certain dimensions before others (i.e., because of the structure of the dataset). This could make it easier for users who use the Dimension selector in the Browser, and start with the variables mentioned first.

11. Languages

All labels (dataset labels, dimension labels, dimension member labels, flag descriptions) and metadata should be provided in both English and French.

12. How to update an OECD.Stat dataset

OECD.Stat datasets should be updated with all the data that the manager intends to share with users outside his/her own unit. This means that the dataset will reflect the data and the related metadata that have been cleared for use.

The updating can be made manually or on request (i.e., the data provider initiates the update from DPI), or it can be done automatically, daily or hourly. In the latter case, a routine for doing this has to be developed. However, the Data Provider Interface (DPI) can facilitate this by creating a batch file to be run on a routine basis.

13. Making a snapshot dataset

In addition to the most recently updated dataset (dynamically updated datasets), OECD.Stat can hold snapshot datasets reflecting the contents at a certain point in time. This kind of snapshot can be seen as a counterpart to a paper publication.

At present, snapshots can be made by creating a new dataset as a copy of the current contents of the dataset. A new facility is being developed in connection with PubStat whereby a database administrator can create a snapshot corresponding to a forthcoming paper publication.

In order to be able to clearly identify the snapshot datasets, the following naming convention is proposed : name should consist of the name of the live dataset combined with the date (publishing edition) of the snapshot in brackets (i.e., the Main Economic Indicators snapshot version corresponding to the May 2006 edition of the publication would be called MEI [May 2006]).

14. Default views and other queries

The Browser shows a “default” table when a user clicks on a dataset name. The default view is a table defined by the dataset administrator; it should give a good first impression of what can be found in the dataset and make the user interested in digging deeper into the dataset, changing the dimensions or dimension members shown if the view does not, by chance, correspond to needs.

14.1 Deciding on default views

It is recommended that each database administrator creates a table which gives a good overview of the subjects inside his/her dataset, illustrating it by selecting one country, preferably Australia (first in the alphabet). Alternatively, and perhaps better, the country dimension (or another dimension) can be put as a “page dimension” on the top of the table, presenting the user with a drop-down box of countries at the top of the table with Australia chosen as the default; in this way, the user is able to select other countries very easily.

The main subject dimension (e.g. Activity) could be shown as the rows; vertical scrolling may be required to see the whole table, but the number of rows should not be excessive. Try to make the table viewable within the central, main area of the browser screen so as to avoid horizontal scrolling, using the recommended screen setting 1024 X 768 and having the metadata area in the right hand side open. This will often mean that it is best to show only one or two time periods. Remember that the user can easily redefine the table.

14.2 Creating default views and other queries

Only the dataset owners are authorised to create default views on their tables. This can be done from the Browser in the following way:

- In the Current Query window, select the dimension members to be shown (see below)
- When satisfied with the table, the selection, or query, can be saved by clicking the Save Query button in the Current Query panel, and then choosing the option *Set this query as default*, and finally clicking the Save Query button in the centre area of the screen.
- Other queries can be created and referred to by their URL on the Statistics Portal.

15. Core Data

In accordance with the OECD dissemination policy, each data provider (unit) can and should make some basic part (the so-called Core Data, formerly called 10% views) of his/her dataset accessible to the public in order to show what OECD data are available, give some basic knowledge and demonstration of the usefulness of OECD statistics, and point to where more complete access can be obtained. In some cases, it has been decided to give access to the whole dataset.

Core Data should be defined as parts of OECD.Stat. These will progressively replace the 10% data available in various media formats which are accessible on www.oecd.org as WDS tables, Excel sheets, PDF, etc.

This means that:

- publicly accessible dimension members for each dimension should be identified, together defining the data space or part of cube that will be available to the public
- the database administrator should define a default view (if different from the default view to be seen by users with global access) and possibly other views each of which is given a URL
- the database administrator will need to give this url to the division/directorate webmaster who will reference the default view on www.oecd.org by creating a Vignette DocumentLink. This will enable the default view to be found on a directorate's website, the Statistical Portal, via our internal search engine, but also via external search engines such as Google.

The Core Data views are simply views on parts of the full dataset. If that dataset is regularly updated (which will normally be the case), the Core Data view will correspondingly be (automatically) updated.

For each theme/sub-theme, there may be 0, one or more such views (each one referring to one dataset, and each dataset having only one core data view). The Core Data views should be presented in the same way as default views (using the Internet version of the Browser), and users should be allowed to modify the Core

view dimensions if they so choose (although no data outside the range defined for the Core Data may be added). Users may download the data in Excel or with Bulk Export.

15.1 Deciding on what should be Core Data

Core Data are extracts from the original dataset (10% at present). A joint PAC/STD paper on the future dissemination of OECD Statistics has recently been presented to CSTAT and CPAC and will be presented to the Budget Committee in the fall. The paper proposes to increase the dissemination of OECD statistics free of charge in parallel with the provision of improved products and services for subscribers, thereby ensuring sustained revenues for the Organisation. The means to implement this are currently under discussion.

The data chosen for the Core View should:

- show what OECD data are available,
- give some basic knowledge and demonstration of the usefulness of OECD statistics,
- point to where more complete access can be obtained,
- adhere to the OECD publishing policy

As previously noted, in some cases it has been decided to give free access to a whole dataset.

15.2 Creating Core Data

The Data Provider Interface (DPI) allows data providers to define the Core Data by providing the relevant dimension members in a CSV file⁴. This addition is in test and should be released soon.

15.3 Displaying Core Data views on www.oecd.org

When presented on a directorate's website or in the Statistics Portal, each Core Data view (whether it is the default view or another query) will be presented with 3 different links or options: Open view in OECD.Stat, open CSV file, and open Excel file:

1. OECD.Stat view corresponds to the way default views are presently shown on the Internet, but adjusted where possible to accommodate presentation issues raised by PAC.
2. CSV files will be bulk downloads of the Core data and all metadata, including publishing metadata
3. Excel extracts should be the full Core Data, not just what can be seen on the screen. If there is a dimension at the page level, there should be one sheet per member of this dimension. Metadata must be linked as comments to the relevant levels.

⁴ Later enhancement of functionality will allow ticking the dimension members to be included, allowing the administrator to see and manipulate the resulting view before accepting it.

16. *How to manage contents in OECD.Stat – DPI user guide*

The creation and updating of datasets in OECD.Stat can be carried out using the "Data Provider Interface" (DPI). The DPI is a web-based interface which data providers can use to assemble CSV-type files exported from their production databases and to generate the XML file format required by the OECD.Stat Entry Gate. A command-line version of the DPI is also available and allows data providers to create automated export procedures. A Data Provider Interface Guide will be available in autumn 2006.