V/ 2.0, 6/9/2006

Management of Statistical Metadata at the OECD

1. Introduction

1. The OECD's corporate metadata management system MetaStore and the corporate data warehouse OECD.Stat are designed to improve the efficiency of data and metadata preparation, storage, access, management and dissemination for statistical products produced across the Organisation. Improvements in these areas will enhance both the transparency and accessibility of OECD statistics. Metadata have traditionally been located in a number of production databases and text files maintained by different OECD Directorates, the result being some duplication of effort in metadata preparation, gaps in the metadata available, particularly metadata explaining differences between similar or related series residing in different databases, and in a small number of instances, inconsistent metadata.

2. This document outlines basic metadata principles and management principles relating to statistical metadata management in the OECD. The document also contains a set of common metadata items providing a common structure. It comprises seven Sections:

- MetaStore and other metadata management tools
- Basic metadata principles
- Metadata management guidelines
- Which metadata should be regarded as standard?
- Statistical Information System terminology and attachment levels
- User access to metadata
- Publishing metadata

2. MetaStore and other metadata management tools

3. MetaStore is one of the modules of the OECD corporate Statistical Information System (SIS)[1]. Because it is designed to manage the preparation of metadata, MetaStore is part of the production layer, along with StatWorks and other production systems based on Fame, SAS, etc. Directorates may choose to migrate their metadata to the MetaStore system, or they may decide to continue managing their metadata in their current production environment (SQL, FAME, Word,…). In either case, the metadata are exported to the dissemination data warehouse OECD.Stat and made accessible to outside users through OECD.Stat, allowing it to be shared between production environments and retrieved by outside users. In addition, metadata in MetaStore may also be made available to outside users independent of the statistical data it describes.

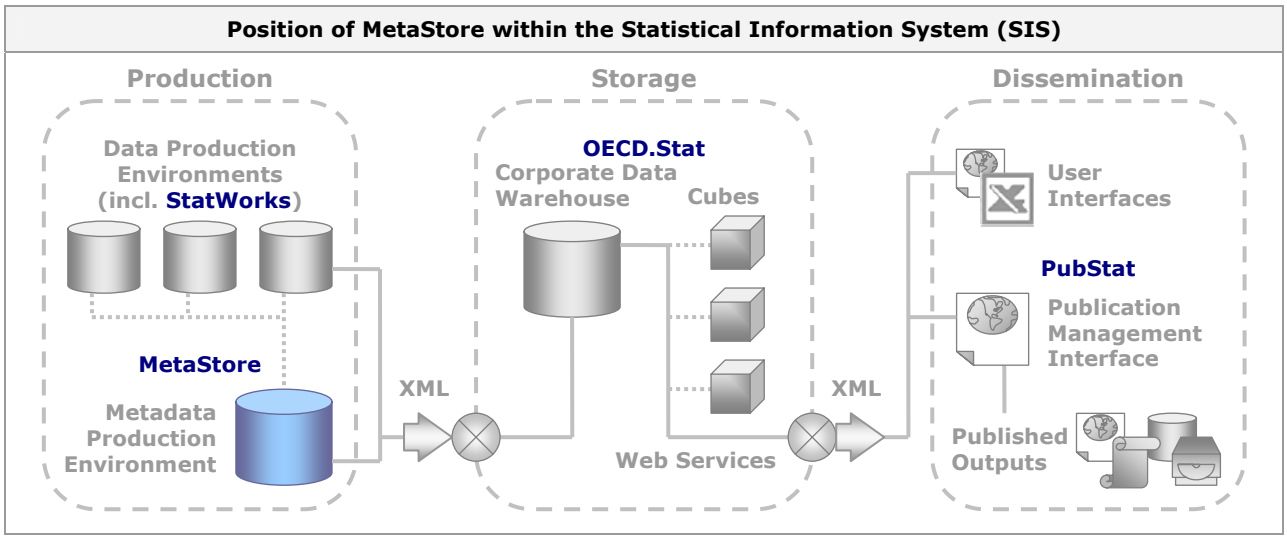
4. Using the MetaStore production tool has a number of advantages. MetaStore and OECD.Stat hold the same types of metadata (see Annex 1) and work seamlessly together, forming a coherent *metadata management and storage system*. Functionality provided in MetaStore includes various functions: (i) advanced search and editing of metadata, (ii) comparison between proposed new metadata and already existing metadata, (iii) the ability to approve the reuse of existing metadata and thus avoid duplication, and (iv) to indicate which version of production metadata is to be incorporated within the dissemination environment, and when this should happen.

5. Internal OECD data managers may, for various reasons, choose to transmit their metadata directly from their production system to OECD.Stat for dissemination. However, because of the aforementioned functionality, it is *recommended* to use MetaStore for metadata management for all OECD statistics. Metadata managed in this way can be utilised for inclusion in publications through the PubStat system. Furthermore, the metadata provides a tool for end-users to locate the data to which it is attached through search facilities within OECD.Stat browser interfaces. Finally, the OECD Statistical Work Programme's (OSWP) online questionnaire and inquiry interfaces manage and provide metadata for statistical activities at the OECD. Since each statistical activity may involve the management of one or more datasets, there exists a logical link between a dataset's metadata in MetaStore and OECD.Stat and metadata relating to its corresponding statistical activity in the OSWP.

5. Whilst the implementation of a corporate metadata facility by different OECD Directorates will improve the storage of and access to metadata, it will not necessarily in itself lead to efficiency gains in the actual

[1] The Statistical Information System (SIS) is a system which supports collection and production of statistics, storage and dissemination.

preparation of metadata and eliminate duplication of work in its compilation, or maximise the sharing of metadata that has already been compiled by the OECD, by other international organisations or by national agencies. The experience of other national or international organisations has shown that the development of corporate IT metadata facilities and its population with actual metadata need to be undertaken in parallel with the development and adoption of a small set of management principles relating to the actual use of the corporate facility by different parts of the organisation. In the absence of such a set of agreed metadata management principles and their effective application across the OECD, the new metadata management environment could simply replicate the existing metadata environment of duplicated metadata and effort in preparation.



3. Basic metadata principles

6. The following basic principles apply to all metadata exported to OECD.Stat for sharing internally or for dissemination internally or externally. Ideally, these metadata principles would also be reflected in the actual creation of metadata as part of the production process (in MetaStore or in some other system - FAME, Word,…), as it is essential that metadata is seen as part of the production process, not as something which providers compile and add in the last minute before publishing, just to fulfil the OECD quality guidelines.

Principle 1. In order to improve transparency and interpretability[2] to OECD statistics and to facilitate user access, all OECD statistical data being shared or disseminated, internally or externally, must be accompanied by appropriate metadata.

Statistical "data" are sets of numeric observations which have times associated with them. They are associated with a series of metadata values, representing specific concepts that act as identifiers and descriptors of the data. These metadata values and concepts can be understood as the named dimensions of a multi-dimensional space (or table), describing what is often called a "cube" of data. These are the **structural metadata** needed to identify, use, and process the data "cubes". Accordingly, structural metadata will have to be present together with the statistical data, otherwise it becomes impossible to identify, retrieve and navigate the data. Structural metadata should preferably include all of the following:

- *Variable name(s) and acronym(s)*, which should be unique (e.g. Financial Intermediation Services Indirectly Measured, FISIM). It is an advantage if these names and acronyms correspond as far as possible to entries in the OECD Glossary of Statistical Terms; terms from the Glossary will be clickable from MetaStore/OECD.Stat.

[2] Interpretability is one of the seven quality dimensions in the OECD Quality Framework. For more details refer to Quality Framework for OECD Statistical Activities, available at www.oecd.org/statistics/qualityframework, more specifically, Part 1. Quality Dimensions , Core Values for OECD Statistics and Procedures for Planning and Evaluating Statistical Activities para. 25-27

- *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 OECD data structures and terminology to learn if the Organisation holds some statistics that might suit their needs (e.g. users searching for some statistics related to "inflation" should be given some indication on where to go for a closer look);
- *Technical metadata*, making it possible to retrieve the data, once users have found out that they exist. An example are the coordinates (combinations of dimension members of the dimensions in the data cube), as kept in MetaStore.

On the other hand, **reference metadata** describe the content and the quality of the statistical data. Reference metadata are the focal point of the common metadata items outlined below in this paper[3]. Preferably, reference metadata should include all of the following:

- *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, transformations);
- *Quality metadata*, describing the different quality dimensions of the resulting statistics (e.g. timeliness, accuracy);

These types of metadata are included in the list of common metadata items provided in Annex 1 below.

Principle 2. Statistical metadata must be consistent. This means that:

- the same variable name, definition, and other descriptions 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.

4. Metadata management guidelines

7. The implementation of these basic principles in the context of MetaStore and OECD.Stat requires the adoption of a common set of metadata management guidelines by each author Directorate in the OECD. The initial set of guidelines set out below has been prepared for use in the decentralised OECD statistical environment.

Guideline 1. For each dataset there must be one responsible (unit) who is also responsible for the ongoing maintenance of the metadata. In cases where more than one unit include the same statistics in the data they disseminate, it must be decided who is responsible; if no other decision is taken, the responsible unit is the one who originally collected it.

Guideline 2. All metadata should be administered in the common management environment consisting of MetaStore and OECD.Stat. Relevant existing metadata currently kept in local production environments (production databases or text files) should ultimately be transferred into the common repository, and subsequently be maintained there. The local production environments should subsequently draw on the common repository, preferably by having dynamic access to metadata from there. "New" metadata should be authored in the MetaStore repository from the outset.

Guideline 3. When migrating existing metadata to the common environment, the responsible unit will initially just copy existing metadata into the MetaStore facility. In doing this, the responsible unit should ideally distribute the metadata into the various component elements of the common metadata items provided for and attach them at the level where they are relevant (refer Section 5 below), but they can also choose to just move the existing metadata into the system as one coherent document, at the dataset level[4].

Guideline 4. Where possible, preference should be towards the use of existing metadata – no new metadata elements are created until the proposer has first determined that no appropriate metadata element currently

[3] The distinction between structural and reference metadata corresponds to the terminology used in the SDMX, see *Framework for SDMX Standards (Version 1.0, Pre-Release Working Draft 0.2)* on www.sdmx.org

[4] In this case, all metadata would be stored under the last heading "Other comments" of Annex 1. This will of course mean that there will be no value added as a result of structuring.

exists. The metadata management environment will issue warnings in such cases. All metadata must be created only once, for efficiency reasons and also in order to avoid the insertion of duplicated and/or inconsistent metadata into MetaStore.

Guideline 5. In order to enhance the visibility of the data and metadata on the Internet, the recommended standard metadata outlined below in Section 5 of this document should be provided.

Guideline 6. An essential part of all quality reviews is to review the existence of and quality of the metadata. The responsible unit within the OECD should analyse the links between their data and others, and in cases where identical data are also kept by others, the decision must be taken as to which should be the appropriate metadata to be referenced by both parties. The review should also identify where metadata are missing, and take steps to remedy.

5. Which metadata should be regarded as standard?

8. While it is obvious that all statistics to be shared must be accompanied by appropriate metadata, it is less obvious as to precisely what the "appropriate" metadata should be. In order to achieve an acceptable level of conformity for the metadata to be stored in MetaStore and disseminated or shared via the OECD.Stat, it is necessary to identify some *common metadata items*, which are supported by the metadata systems of the OECD, MetaStore and OECD.Stat. In this way, users will have more certainty about the actual metadata they can expect to be able to find across the different statistical domains and subjects of the OECD. It also provides internal OECD data producers with a useful framework, helping them to focus on metadata items that are perceived as useful. Finally, the existence of the same metadata for different countries facilitates comparisons of national practice and the resulting statistics.

9. The standard list of common metadata items is imbedded in MetaStore/ OECD.Stat. Obviously, this should be aligned with similar lists and standards being prepared in other international organisations, and particularly the Cross-domain Concepts[5] to be agreed as SDMX guidelines . This is especially important because this SDMX standard is intended to be set in accordance with what national statistical organisations can be expected to provide by linking or mapping from their own metadata systems.

10. However, the process of developing a list of common metadata items within the OECD and in consultation with partner international organizations originally took place in 2004 as the new Statistical Information System needed to be populated with metadata at that time. This has led to the development of an interim standard, to be adjusted later on as SDMX standards are finalised. In particular, a definition for each of the common metadata items will be maintained in the Metadata Common Vocabulary (MCV) also being developed within the SDMX initiative. In order to secure a good starting point and avoid too much work when the SDMX standard is fixed, the present guidelines have been drafted in cooperation with Eurostat and IMF, and the SDMX standards setting team has been kept up to date with the developments.

5.1. Attachment levels of metadata

11. The list of common metadata items outlined in Annex 1 below should be seen as a kind of structured questionnaire, where the 42 individual "questions" or metadata items to be filled in, presented in column (3), are grouped under 6 headings, presented in column (1). No metadata texts are attached to these high-level metadata headings, and their purpose is exclusively to present the items in an orderly and understandable manner. Some of the items appear to be more relevant for a specific attachment level in a data hierarchy. The OECD data hierarchy is expressed as coordinates in a common data structure (datasets, dimensions, members), which allows different attachment levels (see the description of the various levels in the attachment hierarchy in Section 6 below). However, the metadata items may be relevant at any level of detail.

12. In applying the list of common metadata items, it is recommended to always attach the description of each item at as high a level in the data attachment hierarchy as possible, starting from the "dataset" level at the top down to the "observation" level at the bottom. If a metadata item description is applicable to the whole dataset, it should be attached at the dataset level, but if it is only applicable to a country time series, this is where it should go. The OECD.Stat and MetaStore interfaces will ensure that metadata attached at higher levels in the data hierarchy (e.g. dataset level) are always inherited and displayed at the lower levels (e.g. dimension level, series level, observation level) seamlessly in the sense that all metadata relating to the chosen

[5] See Draft SDMX Content-Oriented Guidelines on www.sdmx.org

combination of coordinates are shown, regardless of some of them coming from the dataset level, and others from lower levels.

5.2. Common metadata items

13. The list of 42 common metadata items specified in Annex 1 below are intended to be general in the sense that they should have a good chance of being relevant to many of the different statistical subject matter areas and domains across the OECD. The experience of other organisations where similar attempts have been made to identify comprehensive lists of discrete metadata items for all domains invariably resulted in very extensive and complex models often incorporating hundreds of discrete items to allow coverage of almost every statistical domain. However, for the OECD to end up with the relatively simple and manageable classification or model provided in Annex 1 the ambition is to be able to place around 80% of the metadata in discrete metadata items or fields. Provision will be made to place the remaining 20% of metadata into "other items", "other aspects".

14. The items are arranged in the list of metadata items in two levels (top level and child level), the intention being to show on the initial screen in MetaStore and OECD.Stat only the top level, to unfold the lower level when the user clicks on the appropriate top level item. Actual metadata text is always stored (in MetaStore and OECD.Stat) at the child level. The function of the top level is to group child level items to facilitate user access.

15. The application of the headings with respect to the actual inclusion of metadata is not mandatory in that all child level items do not have to be filled in. The number of items to be populated with text depends on the metadata objectives of each project, its resource capacity to maintain the metadata[6] and the attachment level in the metadata hierarchy (refer Section 6 below). However, it is essential to include under a heading, and at the appropriate level of detail, all available metadata matching that heading. A definition of each heading is provided in the annex; many of the terms are defined in the SDMX Metadata Common Vocabulary (MCV)[7]. In this way, metadata will be much easier to locate and comparable in terms of content, and thus more useful. It will also enable the mapping of metadata maintained by other international organizations and national agencies and facilitate the exchange of metadata.

16. In the Child level there is another column called "Pre-def." (Pre-defined) for which text or value must be selected from a pre-defined list rather than manually typed or copied in. This is a good way of ensuring their consistent entry across the OECD. It will also make it easier for authors to insert them. The valuesets or code lists of these metadata items are shown in Annex 4.

17. When applying the list of common metadata items, metadata already existing in an authoritative, publicly accessible location (e.g. the web site of another international organisation) should as far as possible be referred to with a link, rather than duplicated in MetaStore and/or OECD.Stat.

18. A special kind of metadata is the *"control codes"*, widely used throughout OECD production databases for flagging peculiarities of individual observations (data points) such as: breaks in a series; missing values; estimated values; etc. They are used in the production systems to manipulate data, to decide which data should be shown to users and which should not, as well as for providing users with information about the characteristics of the data. A subset of the control codes may be exported to OECD.Stat as *observation metadata flags*, normally referred to as *flags.* For performance reasons, it has been decided to import these codes directly into OECD.Stat, together with the observation data themselves, not going through MetaStore. At the moment control codes are not coded in the same way in all OECD production databases. A standard set of values of *flags* to be used in OECD.Stat is provided in Annex 2. When preparing the data transmission to OECD.Stat, the data providers will need to provide a mapping between the control codes used in the production database and the standard set of flag values, as shown in Annex 2. In addition, the data provider may decide that it is necessary to have proprietary additional code values and corresponding texts for their tables in OECD.Stat. It should be noted that there may be three different types of observation level metadata in OECD.Stat: Flags (with standard texts described in Annex 2), proprietary observation codes (with proprietary texts for each code value, given by the data provider), and footnotes, i.e. more special comments for an individual observation; the latter will be reference metadata in MetaStore and OECD.Stat, related to the relevant coordinates.

[6] For example, the OECD's Main Economic Indicator (MEI) project has traditionally only maintained summary metadata in its database within a very simple metadata model comprising: definition; coverage; collection; calculation. The policy of maintaining only summary metadata will continue and existing metadata will be attached to appropriate items in the list of Common metadata items developed for MetaStore.

[7] http://www.sdmx.org/knowledge/document.aspx?id=66

19. The list of common metadata items allows for provision of information on important aspects of the seven quality dimensions in the *Quality Framework for OECD Statistics*. For instance, *Accessibility*: is reported in the item (Child level) "OECD Dissemination format(s)"; *Interpretability*: is enhanced by the provision of appropriate reference metadata in MetaStore; *Coherence* over time is enhanced by the provision of series break information in control codes (Annex 2); Other information and qualitative statements may be found under "Quality statements"

20. Annex 3 provides an example of how existing metadata from a production system can be transformed into the OECD metadata structure.

5.3. Mandatory metadata items

21. The general principle is that the dataset owner decides which metadata items are relevant to the data in question. However, two items are indispensable: Unit of measurement and Power code. Without these, the data become meaningless. The two items can be used by the command-driven tools, such as dotStatGet, that import the data into other systems. The Power code will be set as 0 as default and must be changed by the owner if necessary.

6. Statistical Information System terminology and attachment levels

22. The new OECD Statistical Information System (SIS) is built on a common data hierarchy or structure, based on a set of different levels to which metadata may be attached. It is important for communication concerning data and metadata to have a common terminology regarding the different levels in the hierarchy, in order not to confuse levels, and to ensure that metadata describing different aspects of the data (from broad information about a statistical domain to information describing an individual cell) are inserted into the appropriate level in the data hierarchy. The following terminology is used to refer to the different levels of data to which metadata can be attached.

Dataset – An organised collection of data and/or metadata in a common space: meaning that all values of a dataset share the same *dimensions* (Example: Main Economic Indicators)

Dimension - Axis in the *dataset* space (Examples: Country, Subject, Measure, Version, Frequency, Time)

Dimension Member - Discrete point for an individual (or specific) *dimension* axis. (Example - Subject: Gross Domestic Product [GDP], Consumer Price Index [CPI], Unemployment Rate [UNRT])

Coordinates - A combination of one or more *dimension members* each from a different *dimension*. (Example: AUS.GDP.C.1.A.2003 = Version 1 of 2003 annual Gross Domestic Product at current prices for Australia). Coordinates of data in one dataset consist of an ordered set of places, corresponding to the dimensions of that dataset, separated by point; in each dimension place, there can either be a code value for the dimension member in question, or nothing indicating absence of a value for that dimension.

Coordinate Levels – Refers to where metadata is attached in relation to the data.

1. Dimension Level - Highest *coordinate* level with just one *dimension member*. (Example: AUS..... = Australia)

2. Intermediate Levels - All other combinations of *dimension members* (excluding other coordinate levels 1, 3, 4& 5). (Example: .GDP..1. = Version 1 of Gross Domestic Product)

3. Sibling Level (taken from GESMES) - Has a member from all *dimensions* except for "Frequency" and Time".
(Example: AUS.GDP.C.1 = Version 1 of Gross Domestic Product at current prices for Australia)

4. Series Level - Has one *member* from all *dimensions* except for "Time".

(Example: AUS.GDP.C.1.A = Version 1 of annual Gross Domestic Product at current prices for Australia)

5. Observation Level - Has one *member* from all *dimensions*.
(Example: AUS.GDP.C.1.A.2003 = Version 1 of 2003 annual Gross Domestic Product at current prices for Australia)

7. User access to statistical metadata

23. The metadata should always be presented, whole or in part, along with the statistical data itself through all the different media used for dissemination. Thus, all publications and off-line electronic media will be provided with some edited form of the metadata.

24. The way in which metadata are presented on-line on the Intranet, the Internet and as part of OLIS and SourceOECD is crucial to the usefulness of OECD statistics. There are basically two ways in which users inside and outside the organisation obtain access to the metadata:

- stand-alone metadata that users may want to search in order to learn about potentially interesting data, and
- metadata coming along with the statistical data.

25. Stand-alone presentations of the full metadata of many datasets are gradually being made available on the Statistics Portal on the Internet, see for instance http://stats.oecd.org/mei/default.asp?lang=e. This will make it searchable and increase visibility of the data behind them.

26. Together with the data, metadata is presented in the OECD.Stat Browser. Here the attachment levels will be reflected, so that metadata are shown at the level of detail where they belong. The following principles have been elaborated to present metadata in a way that will be immediately understandable to a wide audience.

- 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.
- In the OECD.Stat Browser, metadata availability is marked in the table view always with a red "i" icon ("i" stands for "information"), clicking this icon will reveal the metadata

27. Besides metadata at the dataset and the dimension level, there exist three different metadata types:

- metadata for single dimension members (most often "definitions", e.g. Population coverage, Key statistical concepts used),
- metadata at higher levels (for any incomplete combinations of dimension members (that is, one value from some of the dimensions, combined with no value for the others); most often "exceptions") and
- metadata for particular observation values (for any complete combinations of dimension members; most often "exceptions").

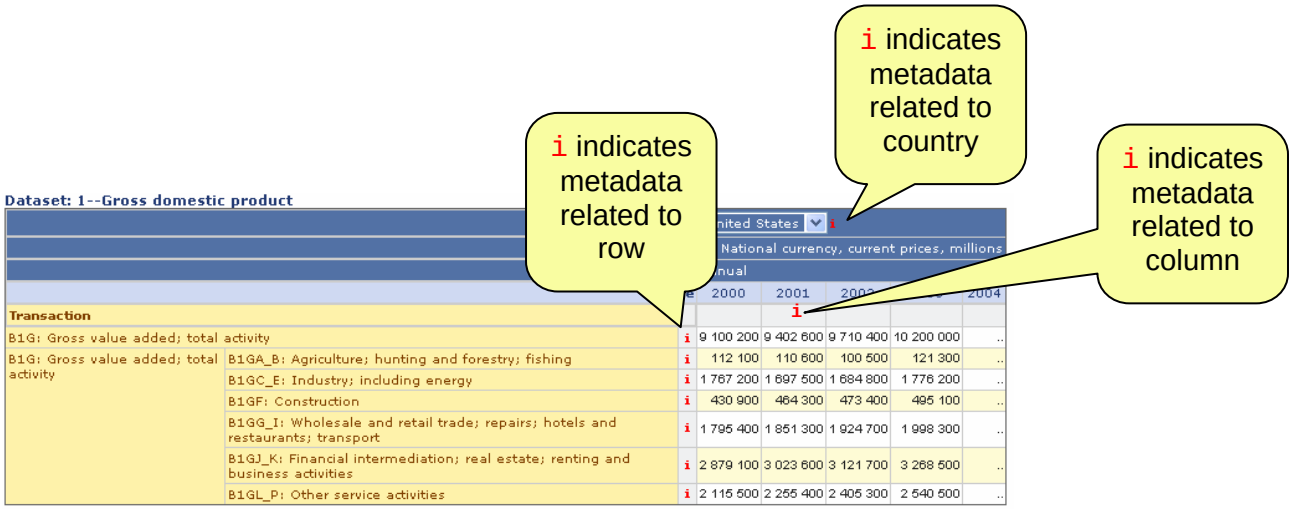
28. These three types will have the metadata availability mark in different places:

1. Metadata for single dimension members have a red "i" in the cell of the dimension member.
2. Metadata at higher levels will be marked in the following fashion (see Figure 1 below): In order to avoid having a red "i" in all cells of a row when metadata pertain to all observations in that row, an extra column (immediately to the left of the data columns) is introduced in the table view, containing a red "i" when there is a piece of metadata pertaining to all observations in the corresponding row. Reciprocally, an extra row is introduced (immediately on the top of the data rows), containing a red "i" when there is metadata pertaining to all observations in the corresponding column. This extra column and extra row will always be displayed independently on the concrete presence of such metadata.
3. Metadata for particular observation values have (as in the past) a red "i" in the cell of the value

29. An attribute "IsInheritable", with the value set "true" and "false", to be set by the data provider is added at Dimension level. For hierarchical dimensions, when this attribute is set to "true", the presence of metadata at parent level is indicated also at all child levels, and this applies for any of the above 3 types of metadata.

30. A red "i" will also be shown, when relevant, in the other windows of the OECD.Stat Browser: theme/dataset selector, dimension selector and dimension member selector.

Figure 1: Extra column and row to indicate higher-level metadata



8. Publishing metadata

31. Any item published needs 'handles' that allow it to be searched for, cited, linked to and catalogued. This is as true for items published online as it is for items published in print or on CD-ROM. In books, publishing metadata includes fields like title, ISBN, number of pages, physical size and blurb. Collectively, these items are known as 'Publishing Metadata'. It follows that statistical outputs, when published in any form (books, serial, CD-Rom, online) will need to have publishing metadata if the output is to be fully useful for users and librarians.

32. Just as less-formal published outputs like Working Papers are more useful to readers if the publishing metadata is well-structured and consistent with publishing norms, so less-formal statistical outputs such as data snapshots and one-off tables will be more useful if the publishing metadata is managed. Publishing metadata makes objects more discoverable (search and inward linking) and more useful to users (ease of citation, confidence that the data is reliable, links to related published items).

33. One of the biggest benefits of publishing metadata for readers is that it can create a virtual path across different content types (e-books, journal articles, sound objects, pictures and, even, statistical objects) allowing the reader to check sources, follow arguments and access accompanying databases. The OECD's StatLink system is an example of how publishing metadata can enable readers to jump from an analytical chapter to a data file. It is planned to extend this so that from the data file the reader could jump to the original database (and conceivably, vice versa). Since more and more publishers are using the same linking system, the pathway can cut across different publishers' content too.

34. In the print world there are standard metadata components such as the ISBN and ISSN numbering systems for books and serials respectively. In order to help publishers, aggregators, librarians and booksellers exchange publishing metadata electronically international standards such as Onix, Onix for Serials and MARC 21 have been established. These systems have already been adapted to cater for e-books and e-serials. However, the advent of online publishing allows items smaller than complete books or serials to be identified separately. Journal articles, for example, when published online, are being numbered in order to make it easier to cite and manage them. The numbering system that has been adopted universally by specialist publishers for published online objects is the DOI (Digital Object Identifier) and this numbering system has already been incorporated into the Onix and MARC systems.

35. PAC (the Public Affairs and Communications Directorate) has run two pilot projects as a first step to bring OECD's publishing metadata up to date. The first has been to introduce DOIs to underpin the StatLink service – each Excel table is given a unique DOI number. The second has been to build a new publishing metadata database for the OECD's working papers. Both pilots have been successful and work is now in hand to extend DOIs and the new publishing metadata system to all OECD published objects. Once done, it will enable OECD to seamlessly exchange information about publications and published online objects with the world's catalogue systems, aggregation websites, search engines and provide the link pathways needed to help readers. This will boost discoverability. Readers will benefit from the 'cite as' tools that will let them import structured citations into their bibliographic management systems like RefWorks and EndNotes.

36. Therefore, in order to maximise the dissemination of statistical outputs, a set of publishing metadata needs to be created for each output (everything from books, CD-Roms, online databases, Core Data snapshots, individual Excel tables and so on). Annex 5 is a near-final list of the fields PAC will be providing.

37. As things currently stand, SIS can only store publishing metadata at the level of datasets. This leaves open the question of how publishing metadata will be stored and linked to sub-dataset outputs (e.g. Core Data, books and StatLink Excel files). Whatever the long-term solution, publishing metadata will be automatically fed into SIS from PAC databases.

Annex 1. OECD.Stat and MetaStore: Common metadata items

Top level (1)	Description derived from the Metadata common vocabulary (MCV) (2)	Child level (3)	Description of metadata requirement (derived from the Metadata common vocabulary (or non-MCV description where no MCV definition exists) (4)	Pre-def. (5)	Notes on possible contents (6)
Source	Source from where data was submitted / extracted	Contact person and organisation	Contact person, title, unit, organisation, phone number, fax, number, email, city, country, postal code [non-MCV]		
		Data source(s) used	List original data source(s) used (administrative data, household survey, enterprise/establishment survey, etc).		
		Name of collection / source used	Refers to full title of the original survey collection, administrative source, database or publication from where the data were obtained. [non-MCV]		Example 1: European Community Household Panel Survey (ECHPS) Example 2: Foreign trade statistics of the National Statistical Office
		Direct source	Refers to the source from where the data was directly collected		Example 1: DataStream [imagining that DataStream keeps the ECPHS mentioned above] Example 2: UNSD [imagining that OECD gets the data from UNSD who has collected it from NSO]
		Source Periodicity	The time distance between observations in source (whether stock or flow). Values: Yearly, Quarterly, monthly, irregular	√	May differ from periodicity of database because of transformations
		Source metadata	Reference or link to metadata from source		
		Date last input received from source	Refers to the date on which the data was last received from the source, e.g. national agency or international organisation. [non-MCV]	√	dd/mm/yyyy
Data characteristics and collection		Unit of measure used	Refers to the unit in which associated values are measured, e.g. USD [non-MCV]	√	
		Power code	Power of 10 by which the reported statistics should be multiplied, e.g. "6" indicating millions of USD. [non-MCV]	√	Natural numbers
		Variables collected	List of variables collected or provision of questionnaire		

Top level (1)	Description derived from the Metadata common vocabulary (MCV) (2)	Child level (3)	Description of metadata requirement (derived from the Metadata common vocabulary (or non-MCV description where no MCV definition exists) (4)	Pre-def. (5)	Notes on possible contents (6)
			[non-MCV]		
		Sampling	Refers to information on sample size, sample frame, sample updating, sample (other)		
		Periodicity	The time distance between observations (whether stock or flow). Values: Yearly, Quarterly, monthly, irregular <other?>	√	
		Reference period	Period of time the data refer to. For business tendency or consumer opinion surveys this field could also refer to the forecasting horizon. [non-MCV].		
		Base period	The period of time for which data used as the base of an index number, constant prices data or other ratio, have been collected.		
		Date last updated	Refers to the date on which the data was last updated. [non-MCV]	√	dd/mm/yyyy To be generated automatically
		Release calendar	Refers to a general statement on the schedule of release of data.		
		Contact person	OECD contact person, title, unit, phone number, number, email [non-MCV]		Fixed format with those fields
		Other Data characteristics and collection			
Statistical population and scope of the data	The scope is the coverage or sphere of what is to be observed. It is the total membership or population of a defined set of people, object or events.	Statistical population	Target population (the statistical universe about which information is sought).		Departures from international guidelines and recommendations
		Geographic coverage	The geographic area covered by the data. [non-MCV]		Information on exceptions and departures from international guidelines and exceptions
		Sector coverage	The range of sectors covered		Information on

Top level (1)	Description derived from the Metadata common vocabulary (MCV) (2)	Child level (3)	Description of metadata requirement (derived from the Metadata common vocabulary (or non-MCV description where no MCV definition exists) (4)	Pre-def. (5)	Notes on possible contents (6)
			by the data [non-MCV]		exceptions and departures from international guidelines and exceptions
		Institutional coverage	The range of institutions covered by the data [non-MCV]		Information on exceptions and departures from international guidelines and exceptions
		Item coverage	The range of items covered by the data [non-MCV]		Information on exceptions and departures from international guidelines and exceptions
		Population coverage	The population covered by the data [non-MCV]		Information on exceptions and departures from international guidelines and exceptions
		Product coverage	The range of products covered by the data [non-MCV]		Information on exceptions and departures from international guidelines and exceptions
		Other coverage	Other issues and information concerning the coverage of the data [non-MCV]		Information on exceptions and departures from international guidelines and exceptions
Statistical concepts and classifications used		Key statistical concepts used	A statistical concept is a statistical characteristic of a time series or an observation. This item should define key statistical concepts included in the domain of study		Departures from concepts defined in international guidelines and recommendations
		Classification(s) used	A classification is a set of discrete, exhaustive and mutually exclusive observations which can be assigned to one or more variables to be measured in the collation and/or presentation of data. This item should list the name of all classifications actually used in the compilation of the data.	✓	Departures from international classifications
Manipulation		Aggregation &	Aggregation is the		

Top level (1)	Description derived from the Metadata common vocabulary (MCV) (2)	Child level (3)	Description of metadata requirement (derived from the Metadata common vocabulary (or non-MCV description where no MCV definition exists) (4)	Pre-def. (5)	Notes on possible contents (6)
and dissemination		consolidation .	combination of related categories, usually within a common branch of a hierarchy, to provide information at a broader level to that at which detailed observations are taken		
		Estimation	Estimation is concerned with inference about the numerical value of unknown population values from incomplete data such as a sample.		
		Imputation	Refers to procedures for entering a value for a specific data item where the response is missing or unusable.		
		Transformations	Mention of interpolations, and other transformations, indicating method used including, if relevant, formulas employed for transformation		Examples: a. Creating quarterly data from yearly data using a method (which is then described). b. Unemployment ratio is calculated as (no. of unemplyed)/(no. of persons in the work force)
		Validation	A procedure which provides, by reference to independent sources, evidence that an enquiry is free from bias or otherwise conforms to its declared purpose. It may be applied to a sample investigation with the object of showing that the sample is reasonably representative of the population and that the information collected is accurate. Refers to processes applied for the verification of data, data confrontation, and data reconciliation		
		Index type	Index type,	√	
		Weights	Refers to information on sources of weights, nature of weights, period of current index weights, frequency of weight updates, weights (other)		
		Seasonal adjustment	Seasonal adjustment is a statistical technique to remove the effects of seasonal calendar influences operating on a series. Seasonal effects usually reflect the influence of		

Top level (1)	Description derived from the Metadata common vocabulary (MCV) (2)	Child level (3)	Description of metadata requirement (derived from the Metadata common vocabulary (or non-MCV description where no MCV definition exists) (4)	Pre-def. (5)	Notes on possible contents (6)
			the seasons themselves either directly or through production series related to them, or social conventions. Should provide information to enable users to make an assessment of the validity of the seasonal adjustment applied. Such information would comprise: a short description of the method (software) used; the main parameters of the adjustment (e.g. additive v. multiplicative decomposition) and some of the derived information (e.g. trading-day weights). [non-MCV]		
		Other manipulation & adjustments	Manipulation and adjustments not mentioned under the headings Aggregation & consolidation, Estimation, Imputation, Validation, Index type, Weights, Sampling, Seasonal adjustment		
		OECD Dissemination format(s)	Refers to the different dissemination media used to disseminate the data, e.g. news release, paper publication, on-line or database, CD-ROM or other. [non-MCV]	✓	
		Related publishing	Gives links or references to web sites and publications where the data has been published, used for analytical purposes, etc		
Other aspects		Recommended uses and limitations	To guide users with limited knowledge of the statistics presented and to help them determine whether the product meets their requirements		Could contain examples of the type of indicators that can be constructed and/or inferences that can be made; the types of analyses that can be performed; the type policy questions it can help answer; the 'shelf-life' of the data product etc. These notes could also

Top level (1)	Description derived from the Metadata common vocabulary (MCV) (2)	Child level (3)	Description of metadata requirement (derived from the Metadata common vocabulary (or non-MCV description where no MCV definition exists) (4)	Pre-def. (5)	Notes on possible contents (6)
					explain when caution should be employed, what type of calculations should be avoided and the type of inferences that should not be made.
		Quality comments	Gives the possibility for data managers to insert comments of quality aspects or general evaluation of quality, as seen from a user perspective.		
		Other comments	Other important aspects		

Annex 2. Observation-level metadata (flags) and Control codes (example based on STD code values)

In OECD.Stat, *coded observation-level metadata* will have to be standardised and connected with a precise text, to be presented to users. These standard codes are called *flags.* The code list and its signification are adopted from the GESMES/TS standard, which is part of the SDMX standards[8].
As suggested by the GESMES/TS standard, two standard coded observation-level attributes are used, Flag and Flag_CONF (Confidentiality). Value sets for the two code lists are listed below. In the GESMES/TS model, it is possible to attach more than one "attribute" (metadata item) to the same observation value. Compared to GESMES, the code value set of Flag has been enhanced with one more value (Z) to indicate when an observation is absolute zero, as distinguished from a zero meaning just less than half of the unit precision level of the observation (e.g., an observation stored as 0.0 may mean absolute zero, or for instance 0.041). This flag will allow absolute zero observations to be displayed as "-" in any dissemination product.

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
Z	Absolute zero observation

When more than one "condition" occurs for the same observation, the following table is used: it indicates the level of importance of each specific "event" (for instance, the information that an observation is a break is more important than that it is an estimate, and the B flag should be used rather than E).

Observation status hierarchy	Relevant in connection with	
	numeric values	missing values
B / Break	√	√
M / Missing value; data cannot exist		√
L / Missing value; data exist but were not collected		√
H/ Missing value, holiday or weekend		√
S / Strike	√	√
F / Forecast value	√	
E / Estimated value	√	
P / Provisional data	√	
A / Normal value	√	

Flag_CONF code list (Observation confidentiality)

Code value	Code description
C	Non-publishable and confidential
F	Free
N	Non-publishable, but non-confidential
R	Confidential statistical information due to identifiable

[8] See http://www.sdmx.org/Data/GesmesTS_rel3.pdf, p. 177. These codes are also included in the proposals for the SDMX standards

In the production databases which feed into OECD.Stat, many different code lists have been applied. In order to fit into OECD.Stat, these must be translated, as indicated in the example below, where the control codes have been taken from across STD databases.

Control code	***Explanation***	***OECD.Stat code value***	
		Flag	Flag_CONF
U	Not available (and never will be)	M	C
Z	Not yet available	L	F
S	Non-publishable estimate (not used in calculations)	M	C
N	Non-publishable estimate (can be used in calculations)	M[9]	C
C	Publishable estimate for zone (indicated in zone)	E	F
E	Publishable estimate for zone (not indicated)	A	F
P	Publishable estimate not indicated	A	F
B	Break, signalled and propagated everywhere	B	F
K	Break, not signalled but propagated everywhere	B	F
R	Break, signalled propagated but not in zones	B	F
A	Break, not signalled but propagated but not in zones	B	F
X	Strange figure	A	F
L	Link period	A	F

All others: A and F

[9] Data will not be transferred to OECD.Stat

Annex 3. Example of transformation of existing metadata into the new structure

This example shows how existing metadata (in this example from MetaStore) can be transformed from the current metadata in MetaStore (part 1 below) into the new structure (part 2 below). The example relates to all time series for one country relating to one subject.

The transformation has largely been carried out by cut-and-paste, and all current metadata have been reused.

In practice, no dataset will contain metadata under all headings in the metadata structure. In order to show how a full set of metadata could look, the metadata of Business Tendency Surveys for USA in this example have been supplemented in part 2 by partially artificial metadata. Sometimes these metadata reflect the real situation of the series in question, while it has been necessary in other cases to be more creative because the type of metadata indicated by the heading were not relevant to these statistics. The artificial metadata are shown in **blue bold type**.

1. Current Metadata - MetaStore

Country: UNITED STATES, Subject: Business tendency surveys (manufacturing)

(USA.BS.....)

Definition

The tendency survey of manufacturing activity is often used to obtain indicators of business confidence. Along with the consumer indicator, indicators of business confidence allow comparisons to be made of the business and consumer moods. The overall index is a key gauge of manufacturing activity. Although its correlation with manufacturing output on a month to month basis is limited, it is usually an excellent gauge of the underlying trend. Moreover, despite the simple system of collection used, the series is much less volatile than other manufacturing indicators. In the survey, companies are asked if activity in each category is better, higher or greater; worse, lower or less; or the same compared with a month ago. Categories include New Orders, Backlog of Orders, New Export Orders, Imports, Production, Supplier Deliveries, Inventories, Customers' Inventories, Employment, and Prices.

Coverage

Over 400 companies in the manufacturing industry participate in the survey. Twenty industries in the manufacturing sector from various U.S. geographical areas are represented. The 20 manufacturing Standard Industry Classification codes are: Food; Tobacco; Textiles; Apparel; Wood & Wood Products; Furniture; Paper; Printing & Publishing; Chemicals; Petroleum; Rubber & Plastic Products; Leather; Glass, Stone, & Aggregate; Primary Metals; Fabricated Metals; Industrial & Commercial Equipment & Computers; Electronic Components & Equipment; Transportation & Equipment; Instruments & Photographic Equipment; and Miscellaneous (a preponderance of jewelry, toys, sporting goods, musical instruments).

Collection

The Institute for Supply Management (ISM) carries out the monthly survey. The ISM is the new name of the formerly National Association of Purchasing Management since January 2002. The Manufacturing ISM Report On Business® is based on data compiled from monthly replies to questions asked of purchasing and supply executives in over 400 industrial companies. Membership of the Business Survey Committee is diversified by Standard Industrial Classification (SIC) category, based on each industry's contribution to Gross Domestic Product (GDP). The full text version of the Manufacturing ISM Report On Business® is posted on ISM's Web site at www.ism.ws on the first business day of every month after 10:10 a.m. (ET).

Calculation
The ISM provides two figures for each category: "net" figures, corresponding to the balance of the percent of positive responses over the percent of negative responses and a diffusion index calculated as the seasonally adjusted value of the sum of the percent of positive responses plus one half of those responding "the same".

Adjustment
Manufacturing ISM Report On Business® data is seasonally adjusted except for Backlog of Orders, Prices, and Customers' Inventories

Source
Institute for Supply Management

2. Transformed Metadata – New Structure

Country: UNITED STATES, Subject: Business tendency surveys (manufacturing)

(USA.BS.....)

Source Organization

Contact person and organisation: US Institute for Supply Management, **www.ism.ws, John Doe, e-mail: doe.John@ISM.ws**

Data sources used: The Institute for Supply Management (ISM) carries out the monthly survey. The ISM is the new name of the formerly National Association of Purchasing Management since January 2002.

Name of collection / source used: The Manufacturing ISM Report On Business® (posted on ISM's Web site at **http://www.ism.ws/**)

Direct source: -

Source Periodicity: **Monthly**

Source metadata: **http://www.ism.ws/**

Date last input received from source: 11 April 2004

Data Characteristics and Collection

Unit of measure used: % balance

Power code: 0

Variables collected: The Manufacturing ISM Report On Business® is based on data compiled from monthly replies to questions asked of purchasing and supply executives in over 400 industrial companies. Membership of the Business Survey Committee is diversified by Standard Industrial Classification (SIC) category, based on each industry's contribution to Gross Domestic Product (GDP).

Sampling: **The data are compiled by ISM from monthly replies to questions asked of purchasing and supply executives in over 400 industrial companies, sampled from a frame of member of ISM, using proportional sampling from kind of activity strata.**

Periodicity: Monthly

Reference period: **Middle of month**

Base period: **Same period last month**

Date last updated: 21 June 2004

Release calendar:
http://www.oecd.org/document/50/0,2340,en_2825_293564_1837362_1_1_1_1,00.html
[Note: This is actually the calendar of OECD News Releases on Standardised Unemployment Rates, as no calendar exists for BTS]

Contact person:
Olivier Brunet,
Statistics Directorate, OECD,
2 rus André-Pascal,
75116 Paris,
France,
olivier.brunet@oecd.org
Tel. +33 1 4524 7877

Other Data characteristics and collection: None

Statistical Population and Scope of the data

Statistical population: The business manufacturing sector of the United States of America.

Geographical coverage: **All US states and territories**

Sector coverage: Over 400 companies in the manufacturing industry participate in the survey. Twenty industries in the manufacturing sector are represented. The 20 manufacturing Standard Industry Classification codes are: Food; Tobacco; Textiles; Apparel; Wood & Wood Products; Furniture; Paper; Printing & Publishing; Chemicals; Petroleum; Rubber & Plastic Products; Leather; Glass, Stone, & Aggregate; Primary Metals; Fabricated Metals; Industrial & Commercial Equipment & Computers; Electronic Components & Equipment; Transportation & Equipment; Instruments & Photographic Equipment; and Miscellaneous (a preponderance of jewelry, toys, sporting goods, musical instruments).

Institutional coverage: **All institutions undertaking manufacturing**

Item coverage:

Population coverage:

Product coverage:

Other coverage:

Statistical Concepts and Classifications used

Key statistical concepts used: The tendency survey of manufacturing activity is often used to obtain indicators of business confidence. Along with the consumer indicator, indicators of business confidence allow comparisons to be made of the business and consumer moods. The overall index is a key gauge of manufacturing activity. Although its correlation with manufacturing output on a month to month basis is limited, it is usually an excellent gauge of the underlying trend. Moreover, despite the simple system of collection used, the series is much less volatile than other manufacturing indicators. In the survey, companies are asked if activity in each category is better, higher or greater; worse, lower or less; or the same compared with a month ago. Categories include New

Orders, Backlog of Orders, New Export Orders, Imports, Production, Supplier Deliveries, Inventories, Customers' Inventories, Employment, and Prices.

Classifications used: **NASIC**

Manipulation and Dissemination

Aggregation & consolidation: The ISM provides two figures for each category: "net" figures corresponding to the balance of the percent of positive responses over the percent of negative responses, and a diffusion index calculated as the seasonally adjusted value of the sum of the percent of positive responses plus one half of those responding "the same".

Estimation: **Quarterly and yearly data are calculated as arithmetical means of the months**

Imputation: **Missing observations are imputed using moving 6 months centred averages**

Transformations: **Quarterly figures are calculated as means of monthly figures**

Validation: **Graphical inspection tool indicating outliers (MEI Data Capture, Data Compare)**

Index type: Diffusion

Weights: **Total for the whole manufacturing sector is calculated using US production values**

Seasonal adjustment: **X12 ARIMA**

Other manipulation & adjustments: **These series are key input to OECD's leading indicators**

OECD dissemination formats: Paper, CD-ROM, Web (B2020 on WDS)

Other Aspects

Recommended uses and limitations: **The main purpose of this data is to identify turning points in the economic cycle of the country. The data can also show where in the economic cycle the country currently is. These data can not be used in measuring actual manufacturing economic output or manufacturing production levels.**

Quality comments: **Data is made available 2-5 days after the reference period**

Other comments: **The data is available back to 1952**

Annex 4: Valuesets for common metadata items with pre-defined values

Source Periodicity: **Yearly, Quarterly, monthly, irregular**
Unit of measure used: **[unspecified]**
Classification(s) used: **[unspecified]**
Index type: **[unspecified]**
OECD Dissemination format(s): **[unspecified]**

Annex 5: Publishing metadata

Publishing Metadata for published content	**Examples of corresponding XML element(s)**
Series Title *(in both English and French) with* **Series ISSN number**	<Series> <ISSN> <SeriesTitle Lang="FR"> <SeriesTitle Lang="EN"> <Series>
Main Title *(in all available languages)*	<MainTitle Lang="EN"> <MainTitle Lang="FR">
Sub Title (i*n all available languages)*	<SubTitle Lang="EN"> <SubTitle Lang="FR">
Author(s) **-** *first name* *- last name* *- affilitiation* *- order (when several authors)*	<Authors> (unique element) <Author> (multiple element) <FirstName> <LastName> <AuthorOrderNumber> <AffiliationId> </Author>
Publication date *- Year* *- Day (when available)* *- Month (when available)*	<DayPublish> <MonthPublish> <YearPublish>
Order Number *(when available)*	<OrderNumber>
Year number *(when available)*	<YearNumber>
Language(s) of the content *- English* *- and/or French* *- and/or othe language* *(2 languages mean the content is bilingual)*	<Languages> <LanguageCode> <LanguageCode> ... </Languages>
Long abstracts *(in both English and French)*	<DescriptionLong Lang="EN"> <DescriptionLong Lang="FR">
Pages number *(when relevant)*	<NumberOfPages>
Keyword(s) *(in all available languages) when available*	<Keywords Lang="FR"> <Keywords Lang="EN"> …..
JEL classification *(in all available languages) when available*	<JEL Code ="value is JEL code"> value is *the JEL classification label* </JEL>

Theme(s) *(in all available languages) when available*	**Multiple element** **<Themes>** <Theme ISSN="" Lang="EN"/> <Theme ISSN="" Lang="FR"/> **</Themes>**
DOI *(digital object identifier of the content)*	<DOI>
Link to content file *(whatever iis its format)*	<Filename>
Related Link(s) to content available in other language(s). *Different types of links can exist* *- links to the same content in another language* *- links to related content* *- inherited links to parent content (print publication, database, etc.)* The type attribute specifies the type of cross reference	<Xrefs> <Xref ID="value is the DOI of the related content" Type=" "/> value is the filename of the content </xref> <Xref ID="value is the DOI of the related content" Type=" "/> value is the filename of the content </xref> ….. </Xref>
Comparative table flag *This flag determines whether a table is a cross country table.*	
Related Countries *composed of* *- Country ISO Code* *- "Country specific" flag (yes/no): This flag determines whether a table is specific to one country, for example "France", (as opposed to a table related to several countries which may or may not include "France").*	
Time range such as: yyyy : a single year yyyy-yyyy: a range of years (EVERY year between the two specified values) […], yyyy: the previous range AND the specified values. *Examples:* 1980, 1995-2003 means: 1980, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003. 1970, 1980, 1990, 2000-2005 means: 1970, 1980, 1990, 2000, 2001, 2002, 2003, 2004, 2005.	
Variable *(concept of statistics "variable" to present content on the Statistics homepage)*	