Distr. GENERAL
03 May 2013
WP.25
ENGLISH ONLY

**UNITED NATIONS**
**ECONOMIC COMMISSION FOR EUROPE**
**CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION**
**STATISTICAL OFFICE OF THE**
**EUROPEAN UNION (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION**
**AND DEVELOPMENT (OECD)**
**STATISTICS DIRECTORATE**

**Work Session on Statistical Metadata**
(Geneva, Switzerland, 6-8 May 2013)
**Topic (iii): Metadata in the statistical business process**

# DEALING WITH METADATA(S) IN ILOSTAT[1]

**Working Paper**

Prepared by Edgardo Greising, Department of Statistics, International Labour Organization

## I. Introduction

1. ILOSTAT, the new database of labour statistics, is the result after a process of redesigning the department's approach for all its data compilation activities that included not only the development of new applications using updated and appropriate tools to achieve the required functionality, but also changes in the procedures for data compilation, data processing and dissemination.

2. The new system is supposed to be fully "metadata driven", aiming to reduce the costs of maintenance and to make it reusable by different compilation and dissemination projects.

3. During the design and development of the new system, we found different "types" of metadata, according to their usage and role in the statistical production process. The main types of metadata that could be identified in the process were as follows:

   (a) System parameters, often called "paradata"

   (b) Metadata used for table configuration, both for collection and/or dissemination

   (c) Descriptive metadata for the data being collected/disseminated

   (d) Descriptive metadata for the primary sources of the derived indicators collected

   (e) External resources related to the statistical production process, especially in the primary compilation

[1] The author knows that metadata comes from the Greek prefix meta, "beyond", "self", and the Latin word data, plural form of datum, "data". So the word metadata is plural. Nevertheless, the "(s)" intends to highlight the different types of metadata we can find in any information system, in particular in ILOSTAT.

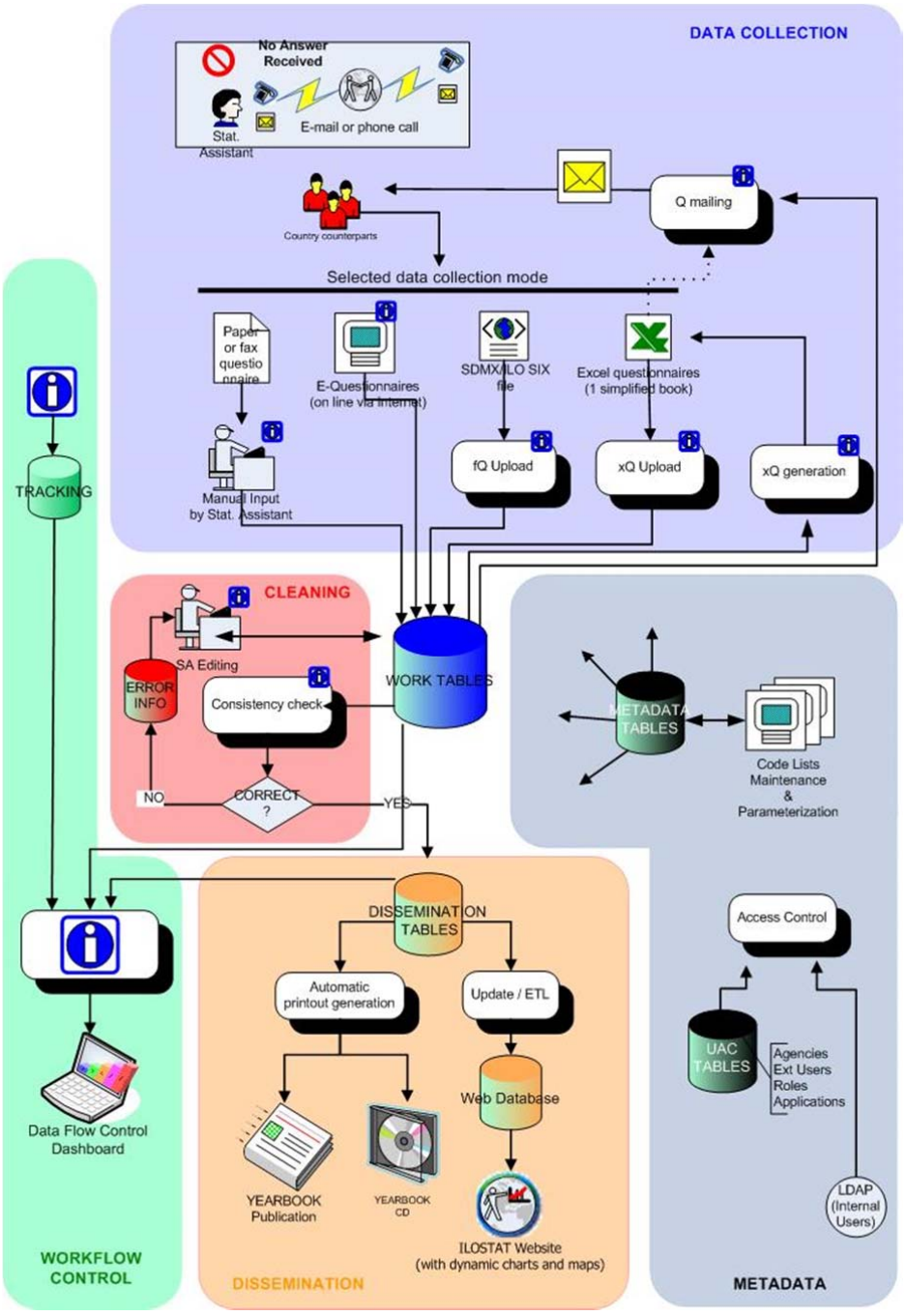
(f) Workflow control information about the statistical process

4. Each one of these types of metadata adds value to the data itself, and therefore is crucial information for the correct interpretation of the figures.

5. In the following paragraphs we are going to see how each of them is related to the phases of the Generic Statistical Business Process Model (GSBPM) and which specific tools and procedures have been implemented in ILOSTAT to deal with them in the different modules of the statistical information system.

# II. ILOSTAT and GSBPM

## A. ILOSTAT modules



6. The system can be split into three main modules that maps with the three main stages of the data compilation process: Data collection, Data processing and Data dissemination.

7. The Data collection module comprises the design and build of the data collection instruments, which vary according to the data channel to be used. Currently ILOSTAT collects data through Excel questionnaires and csv files. The SDMX connector will be released very soon and will allow for uploading data through SDMX data flows. An electronic questionnaire (on line web form) is in the roadmap for this year as well.

8. Some of the activities in this stage include sending e-mails with questionnaires and reminders, answering questions from countries, uploading the data, etc.

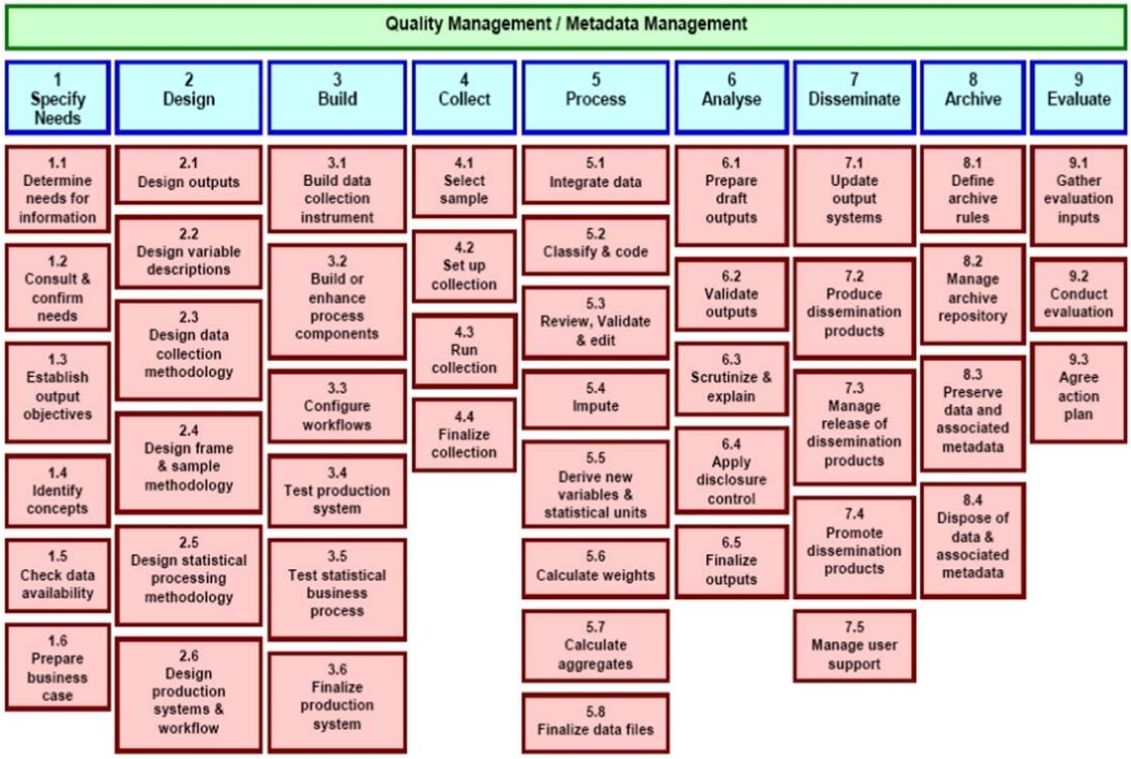
9. Once data is collected, regardless the mean used for, it is processed by an exhaustive consistency checking and correction process. At this stage is where some descriptive metadata (in the form of footnotes and annotations) is coded based on free text provided by the countries.

10. Each weekend (or whenever the amount of data incorporated justifies it) an automatic process computes a set of derived or calculated indicators based on those indicators that have passed the consistency rules and are considered "ready for dissemination". After this process, all these indicators (collected or calculated) are moved to the dissemination database.

11. The last module comprises the tools for data dissemination. It includes a dynamic website, data download in csv, pdf and Excel formats and SDMX web services (very soon).

12. The Workflow control module tracks the evolution of the questionnaires and questionnaires' tables through the overall process, and the Metadata module provides the tools and procedures for general metadata maintenance.

## B. GSBPM phases



13. The GSBPM is structured in nine phases: 1. Specify needs, 2. Design, 3. Build, 4. Collect, 5. Process, 6. Analyse, 7. Disseminate, 8. Archive and 9. Evaluate. Each of these phases is divided into different number of sub-processes which not need to be followed in a strict order since GSBPM is not a linear model. There are also several overarching processes among which Quality management and Metadata management are the most relevant.

### C. ILOSTAT compliance

14. ILOSTAT complies with the model and the different processes' tasks and activities map to the model's sub-processes of each of the nine phases. The Workflow update modules and the Metadata management tools provide compatibility with the overarching processes for Quality and Metadata management. Detailing which sub-processes are covered by ILOSTAT's processes is out of the scope of this paper, but is summarized at the first level in the following table:

| GSBPM | ILOSTAT |
|---|---|
| 1. Specify needs | Specification of requirements prior to collection launching |
| 2. Design | Design of Questionnaires, Dissemination tables, charts, DSD, etc |
| 3. Build | Questionnaire production, Tools development |
| 4. Collect | Data collection |
| 5. Process | Consistency check, Editing, Calculated indicators |
| 6. Analyse | Data analysis, Estimations and Projections |
| 7. Disseminate | Data dissemination (different channels) |
| 8. Archive | Data and metadata preservation |
| 9. Evaluate | Process evaluation (continuous) |
| Quality management | Data and metadata quality evaluation (continuous) |
| Metadata management | Metadata maintenance |

## III. Types of metadata

### A. Metadata classification

15. A generally accepted definition of "Metadata" is "data about data". It is so relevant in statistics (as in many fields) that without proper metadata the information delivered by a statistical information system is likely to be misinterpreted.

16. Many types of metadata can be recognized in an information system, each with different characteristics and purpose. This allows classifying metadata following different criteria. For the purpose of this paper we will define a hierarchical classification to cover the different types recognized in ILOSTAT as well as in most of the statistical information systems.

17. Ralph Kimball[2] refers to three main categories of metadata: Technical metadata, Process metadata and Business metadata. The first one, often referred as System metadata, is definitional, and comprises system parameters and rules that permit altering the behaviour of a program at run time. It is sometimes called "Paradata". The other two are mainly descriptive metadata. The Process metadata describes the results of process execution. Note that while the System metadata rules the process execution, the Process metadata provides information about it. Finally, the Business metadata is always related to the domain. In our statistical domain, the SDMX information model recognizes two main types of metadata: Structural metadata needed to search for and display data and Reference metadata that gives more information on definitions, methodologies, processes and quality[3].

[2] Kimball, Ralph (2008). The Data Warehouse Lifecycle Toolkit (Second Edition). New York: Wiley.
[3] SDMX Content Oriented Guidelines, Annex 4. (2009).

18. According to that definition the Structural metadata is definitional, and gives information about data containers; while the Reference metadata is descriptive and provides conceptual information about the data itself. We can recognize at least three sub-types of conceptual Reference metadata: the Statistical Metacontent that gives information about the data, the Methodological metadata regarding primary data collection and production methodologies, and Quality metadata describing different quality dimensions of the data.

19. There's a fourth group of Reference metadata that is comprised by the so called External resources that consist in additional artefacts and documents related to the data and its sources like questionnaires, methodological guidelines, maps, etc.

20. The following schema describes the hierarchy among these types of metadata:

<table>
<tr><td rowspan="6">Metadata</td><td colspan="4">Technical</td></tr>
<tr><td colspan="4">Process</td></tr>
<tr><td rowspan="4">Business</td><td colspan="3">Structural</td></tr>
<tr><td rowspan="3">Reference</td><td rowspan="2">Descriptive (conceptual)</td><td>Metacontent</td></tr>
<tr><td>Methodology</td></tr>
<tr><td>Quality</td></tr>
<tr><td colspan="2">Ext. Resources</td></tr>
</table>

# IV. ILOSTAT metadata

## A. Technical metadata

21. There are several examples of general system's parameters that govern automatic updates and scheduled executions of batch processes. In the Data Collection module the default settings for structural validation of data collection instruments and several parameters and thresholds belong to this category. All the information in the User Access Control module (roles, credentials, access rights, etc.), the consistency rules and the formulas for calculated indicators are examples of "paradata" as well.

22. All these parameters are normally managed by system administrators or high level specialists on the substantive area.

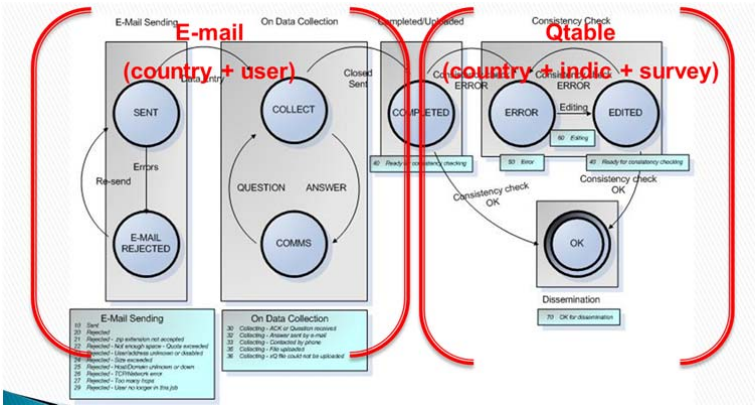
23. In the dissemination website, the context information like language, country or subject are examples of technical metadata that is self-defined according to the navigation.

## B. Process metadata

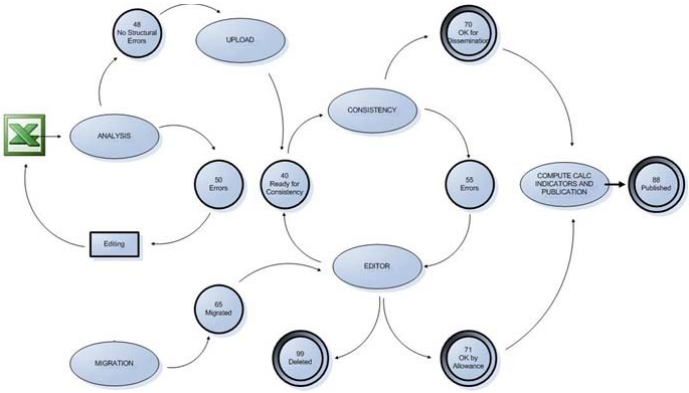
24. The different sub-processes in ILOSTAT compose a workflow for the questionnaires and collection tables (Qtables). This workflow comprises different statuses which are recorded automatically by every program in the system as the actions are completed.

25. This process metadata provides crucial information that is used by the Country Specialists to manage their data collection activities, by the Supervisors to have real time information about the amount and quality of the data compiled and by the different applications to drive the data workflow.

26. The main stages in the workflow are:




27. A more detailed look into at the Qtable workflow let us find the following stages:




28. A workflow dashboard with several reports and charts provides comprehensive information for the users.

## C. Structural metadata

***Business → Structural metadata***

29. The Structural metadata is the “heart” of the metadata driven system, since based on it the different modules “know” which data and in which format is to be processed. This metadata comprises the code lists for all the concepts in use by the model and the definition of all the artefacts used for data collection (questionnaires, DSD’s, etc.) and dissemination (tables, charts, navigation menus, etc.).

30. All the modules take the structural information from a single place, the only metadata repository to be maintained. For example, defining a new indicator in this repository is enough to have it ready to be collected and disseminated by every data channel. Technical metadata should be added for the consistency checking rules. Practically, that means no need to create Excel worksheets, DSD, or to edit web pages.

## D. Statistical metacontent

***Business → Reference metadata → Descriptive metadata → Statistical metacontent***

31. This first type of Descriptive metadata is the typical “data about data”. ILOSTAT has two classes of metacontent: the observation value status (sometimes called flag) which is attached at the observation value level, and the notes which can be attached at any level in the concepts hierarchy. Typically, notes are stored attached to an observation value or to a qtable. Every note in the system is part of a controlled vocabulary and is coded at collection time.

32. The consistency process includes rules for checking the existence of mandatory notes and warns when contradictory notes are found at the same level.

33. In the dissemination website, according to the type and the content of a note, it can be displayed as part of the Table metadata or as a footnote.

34. To avoid an "overcrowding" of notes in the table displayed, a specific algorithm consolidates footnotes that appear in all cells of a row or column and "promote" it to the header. This process is dynamic and is run every time the table is re-displayed due to changes in the time span or filters selected since the disposition may change.

## E. Methodological metadata

***Business → Reference metadata → Descriptive metadata → Methodological metadata***

35. This metadata consists on information about the methodology used for primary data collection and processing, and in ILOSTAT it forms a sub-system called "Source & Methods".

36. It provides detailed information about the source of the primary data from which the compiled indicators come (different types of surveys, censuses, administrative registers, etc.) and is connected to the series by the "survey" key.

37. The new version of "Source & Methods" (still to be developed) will collect and disseminate the information in DDI 2.x standard format. (Only the Document- and Study-description sections).

## F. Quality metadata

***Business → Reference metadata → Descriptive metadata → Quality metadata***

38. The information kept in the Workflow tables provides information about the quality of the data. For example, according to the result of the consistency and correction process, a Qtable is signalled as "Error" or "Ready for dissemination". Periodically (once a week by default), all data "Ready for dissemination" is used to compute the calculated indicators and is published.

39. Other quality related information is attached to the questionnaires and Qtables by the country specialists and analysts during the process as "comments".

## G. External resources

***Business → Reference metadata → External resources***

40. The "Source & Methods" sub-system attaches artefacts and documents related to the studies like questionnaires, methodological guidelines, maps, etc. which can be accessed together with the Methodological metadata.

41. If available, full study metadata and microdata files will be available from the ILO Central Microdata Repository, which is being implemented.

## V. Summary

| GSBPM | ILOSTAT | DEFINES/PRODUCES | USES |
|---|---|---|---|
| 1. Specify needs | Specification of requirements prior to collection launching | Structural, Metacontent | Process, Quality (previous), Methodology |
| 2. Design | Design of Questionnaires, Dissemination tables, charts, DSD, etc | Technical, Structural, Metacontent | Quality (previous) |
| 3. Build | Questionnaire production, Tools development | Process | Technical, Process, Structural |
| 4. Collect | Data collection | Metacontent | Technical, Process, Structural |
| 5. Process | Consistency check, Editing, Calculated indicators | Metacontent, Quality | Technical, Process, Structural, Metacontent |
| 6. Analyse | Data analysis, Estimations and Projections | | Metacontent, Quality |
| 7. Disseminate | Data dissemination (different channels) | | ALL |
| 8. Archive | Data and metadata preservation | Ext. Resources | ALL |
| 9. Evaluate | Process evaluation (continuous) | Process, Quality | Process, Quality |
| Quality management | Data and metadata quality evaluation (continuous) | Technical, Quality | Reference |
| Metadata management | Metadata maintenance | ALL | ALL |