Distr.
GENERAL

WP.12
25 April 2012

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2012)
(Washington, DC, 21-23 May 2012)

Topic (ii): Streamlining statistical production

Streamlining Data Compilation and Dissemination at ILO Department of Statistics

Supporting Paper

Prepared by Edgardo GREISING, International Labour Office (ILO)

I. Introduction

1. After many years of operating with an old system for the compilation and dissemination of labor statistics, the need to streamline processes and have new tools in the ILO Department of Statistics was clear. High maintenance costs, low coverage and problems of comparability between data were some of the most important gaps that determined the urgent need to redesign the system.

2. The project to redesign the department's approach included not only the development of new applications using updated and appropriate tools to achieve the required functionality, but procedures that could be automatized and allowed to have an auxiliary system for monitoring the flow of information and to assist in the task of data collection.

3. One aspect that was emphasized from the beginning of the new project was the adoption of every possible standard, so as to increase the chance of interaction with our partners. Thus, the process follows the recommendations of the General Statistical Business Process Model (GSBPM), development tools from the Oracle suite (a "de facto" standard) are used, and the means of collection are based on Excel, XML and SDMX (coming soon).

4. The implementation of some simple concepts derived from the lessons learned from the former process, would enable the new ILOSTAT database to have a better response rate from the countries, reduce the delay of the information received and improve the overall quality of the data published. The challenge was to achieve these objectives while simultaneously reducing the TCO of the system.

5. This paper describes the set of new processes and the IT tools developed so far (as well as those forthcoming), to optimize the data compilation and dissemination at the ILO Department of Statistics.

II. The Old System

A. Process Overview

6. Data collection at ILO was made through Excel questionnaires, customized for each country, and then sent to the countries. For the "Yearbook of Labour Statistics", nine Excel books – a total of thirty five spreadsheets - were automatically sent by e-mail. When some of these e-mails were rejected, there was no established procedure to deal with this situation.

7. When answers are received, (See ***Error! Reference source not found.)*** a manual process of reception and internal distribution by e-mail was performed by an administrative assistant. The uploading of the spreadsheets to SAS databases was done thru a semi-automatic process that must be run by each Statistical Assistant (SA) on a per-sheet basis (35 for each country).

8. After receiving the Excel questionnaires manual consistency controls were made by SA on Excel spreadsheets. They have to pass thru all the questionnaires, even those without errors. The loading procedure also obligated the SA to go over each sheet of the Excel book to upload it.

9. When values are received on hardcopy – some countries send publications with the data, or print the questionnaire to fill it in paper - the SA had to use the SAS command line interface to enter the information directly into the SAS database.

10. For those countries with no answer at all, there was no established procedure for retrying. In the best cases, if there was enough time, data used to be grabbed from web sites by the SA to complete the questionnaires not received and uploaded to the system value after value using the SAS command line interface.

11. Once a day, a full consistency check for each table over the whole database was performed. The results were made available for each SA, but erroneous data was not marked, so if nobody fixed the errors (editing and correcting was an optional procedure), they could be published anyway.

12. There was no solution for "false positives" (errors according to consistency rules but not in reality), so errors remained "ad eternum" in the report even if they were not "real errors".

13. When the cause of an error was determined, data editing had to be performed using SAS command line interface, whose operation is not intuitive.

14. In the consistency check process, data was not checked considering it would also be used for other dissemination formats, like in the Country Profiles publication, making it difficult for the SA to arrange the tables produced for this publication when some of the data was missed.

15. The Footnotes System was an attempt from IT Unit to manage the annotations that are added to the tables published. As the uniqueness is not assured in any way, many entries in the system with different codes were conceptually (and sometimes literally) the same.

16. Metadata associated to the statistical activities that take place at the countries to produce the data collected by the Department is compiled and stored in the "Source & Methods" database. This information was systematized but collected as a number of text documents, which content does not

follow any standard regarding metadata documentation, and are organized into “volumes” as it were printed since several years ago.

17. The tables for the publications are produced from the database using SAS to XML and then to PDF, and this documents are sent directly to press. Nevertheless, the “Country Profiles” publication is not produced directly by processing the databases. After producing the Yearbook tables they must be reviewed and temporarily modified by the SA by means of altering the tables definition in SAS using the command line editor to alter columns, years and footnotes, because data is not coherent among subjects combined in some tables of the profile.

18. The LABORSTA web site was updated weekly with the information “as is” in the main database, eventually including those tables with errors listed in the consistency check and not fixed so far.

19. No statistical analysis was performed over the data collected. Not even charts or thematic maps were made, even though all the information relates to country indicators and thus, is naturally geo-referenced.

20. Looking at quality indicators for the data compilation like *veracity*, *coverage* and *opportunity* and taking into account the *direct costs* associated with the process, it was easy to conclude on the urgent need for a complete re-engineering of the data compilation and dissemination process:
    (a) Regarding *veracity* of the data disseminated, it can be considered more than acceptable, but relays on a huge workload for the SA verifying and fixing errors in the database during the data collection process and even after data has been published.
    (b) The response rate is very low and different behaviors can be appreciated when analyzing the data by continent and by topics. But the general trend for the *coverage* is a clear declining, from 63% in 1989 to 43% in 2008.
    (c) As a consequence of the related difficulties, the Yearbook used to be published 11 months later than the end of the reference period, which is not the *opportunity* required by policy makers and other users. The information was not available to give answer to the quick and strong changes observed in the last years.
    (d) Regarding *direct costs* is enough to mention more that USD 150,000 per year paid in concept of software maintenance and updates for the SAS product mainly used as a DBMS.

21. No statistical analysis is performed over the data collected. Not even charts or thematic maps are made, even though all the information relates to country indicators and thus, is naturally geo-referenced.

22. Last but not least, the overall process is not considering that countries have improved, in the last years, the quantity and quality of data produced and disseminated, making most of this information available thru their websites. This discourages them to complete the Excel questionnaires sent by ILO and perceived as a heavy additional work to their routine. Most of the time they suggest that is ILO who is in charge of the implementation of an agile procedure to get this data which has been already published by them. The excess of requirements of information from the different international organisms has to be rationalized, taking into account the lack of resources in the countries, who have to address their effort to improve data quality, rather than answering duplicated requirements in unfriendly formats.

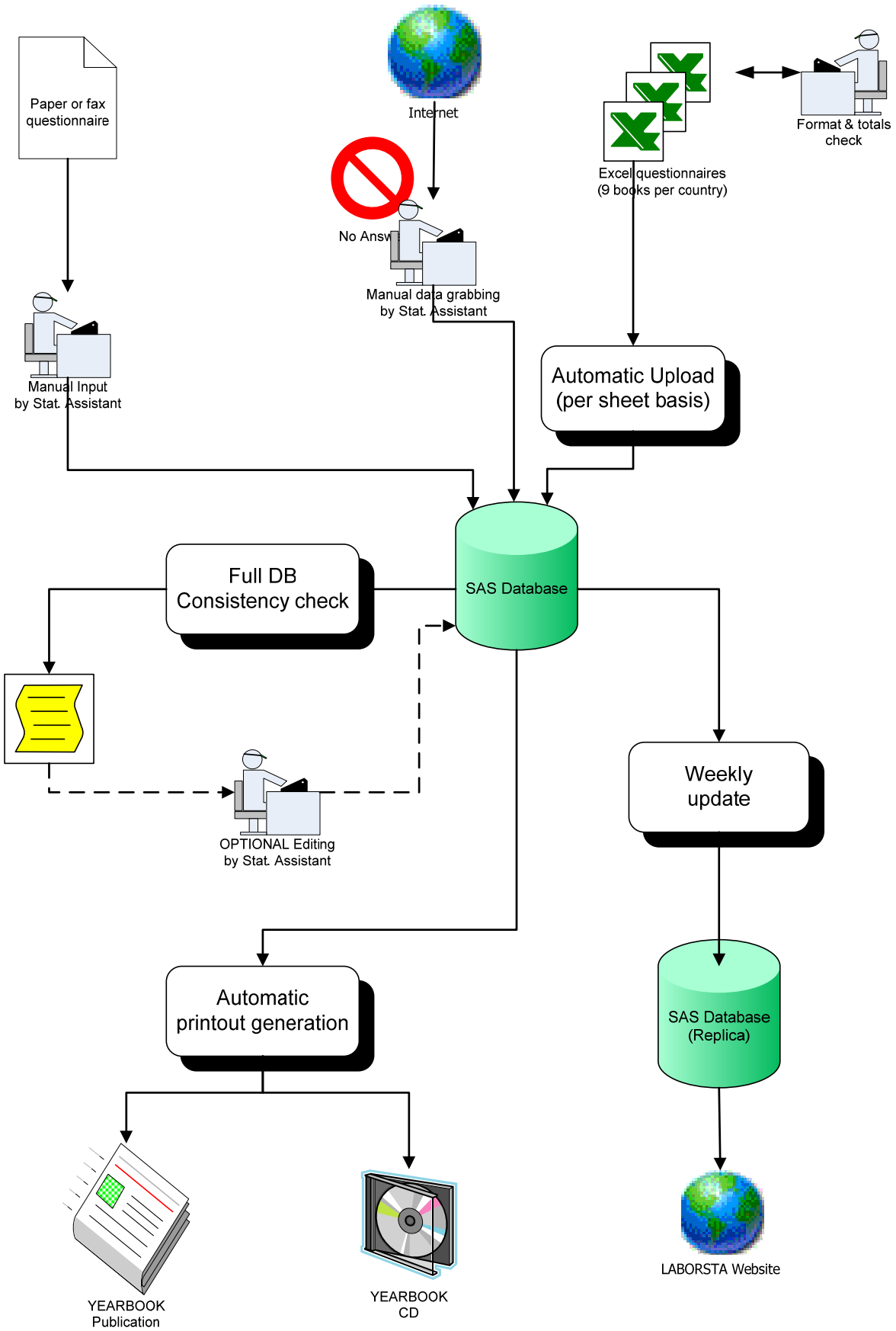

Figure 1: Old process workflow

B. IT considerations

23. All data was stored in a SAS hierarchical database conformed by isolated tables. SAS as a database manager has many limitations, like not offering the maintenance of referential integrity among the tables or the inexistence of a journaling system for data deletion audits. Moreover, a very important amount of money was being paid annually as license/maintenance fee for the SAS software. On the other hand, ILO is a corporate customer for Oracle DBMS and related products, and the Department of Statistics could make use all these products at no charge.

24. The SAS procedure designed for uploading the Excel questionnaires had to be used on a "per-sheet" basis after a visual verification of totals by the SA. This procedure was too much time-consuming for a questionnaire composed of 35 worksheets.

25. The consistency checking program, although very well designed and implemented (in SAS) in terms of the rules applied, turned inefficient due to some characteristics that made it not very friendly to the SA. For example, a single listing was produced for the whole database, with hundreds of errors, making it very difficult to handle. Besides, there was no solution for "false positives" (errors according to consistency rules but not in reality) which remained in the report although they are not errors. All this proved to be very "uncomfortable" for the SA who relied more on their "visual inspection" of the questionnaires than in the consistency report.

26. Erroneous data was not marked in the database, so if nobody corrected the errors, they could eventually be published.

27. The tool provided for data input and editing in the SAS database was based on a command line interface, not WYSIWYG ("What You See Is What You Get") as the graphical user interfaces usually are nowadays. This resulted in a very tedious and time consuming work for the SA.

28. A Data Dictionary system managed two types of codes: the "General" set, updated only by the IT Unit in order to assure the uniqueness of the codes, and the "Specific" of each database, that was maintained by the SA's. In these specific sets, each SA could add new categories as needed to inquiry for new information. This way of administering the Data Dictionary favored the appearing of multiple different categories in variables with the same name in different tables; or categories that are conceptually the same with different values, depending on the table.

29. The footnotes system, intended to rationalize the number of notes added to the tables, had become unmanageable due to the number of entries, which surpassed the quantity of 5000. As it had no indexes of any kind, it resulted useless. Duplications for the same issue and many entries with slight differences in its writing but meaning the same were common problems in the database. It became a "vicious circle" since as more entries were in the system, more difficult was to find one particular note to assign to an observation, and thus, the SA used to add a new entry with the same meaning that one or more pre-existent that he was unable to find.

30. The "Source & Methods" database was related to the series of values thru the "Type of source", but the model did not consider the "specific source" as an entity. So when more than one source of the same type existed in a country, it was not possible to quickly identify which one is each.

31. The dissemination tools comprised the generation of printed publications extracted directly from the SAS database and send to press and the LABORSTA website. In both cases the rigidity of the model (and the software) determined the difficulties to generate reports with different breakdowns or aggregations for the data collected and/or provide more flexible instruments like pivot tables, charts, thematic maps, etc.

32. The workflow management was “full manual”, consisting of the updating by hand of the status of received questionnaires done by an administrative assistant. No information about sent questionnaires, rejected e-mails, consistency checking, errors by country, etc was automatically generated but, in the best case, had to be asked to IT personnel or laboriously calculated by processing the Excel tables downloaded from the LABORSTA website.

III. The New Approach

A. Process Overview

33. The new process for data compilation and dissemination is built on four main ideas:
    (a) The broadening of the ways of interaction with the countries for data collection;
    (b) The full automation of computerized procedures, so as to enable Statistical Assistants to engage more efficiently in non-computerized activities;
    (c) The systematization of the consistency and correction procedure regardless of the way the data was received; and
    (d) The ability to know when and why (or why not) data from the countries is arriving, thus knowing how much information is to be included in a publication.

34. Two new ways of data collecting will be implemented: the e-Questionnaire, allowing the countries to enter the data on line thru the Internet, and the use of Electronic Data Interchange, preferably based on SDMX, the Statistical Data and Metadata eXchange standard, to allow the countries to send the information in XML files automatically downloaded from their own databases.

35. Excel questionnaires will remain as valid options but, countries are expected to discard to use them, thus moving to use EDI or the online e-Questionnaire.

36. The e-Questionnaire system will include the checking for basic consistency rules, like format, range, totals, etc. Warnings for values out of a tolerance threshold regarding last year datum can be issued when saving the form, allowing confirming it.

37. For the Excel questionnaires as for SDMX, fully automated procedures will be developed for uploading of data, including also consistency checking as they are uploaded.

38. All data received, regardless of the way it arrived, will be loaded into the “Data collection database” (Work tables), a common repository where the questionnaires will be stored while not ready to be published, that is, still with errors.

39. The SA will be assigned a set of countries following criteria of language, cultural affinity and previous knowledge of the region. This way, they are supposed to have a better knowledge of their counterparts in each assigned country, so they will be responsible for making their best in order to get the answers. Should the deadline for receiving the data approach, they must get in contact with their counterpart and help them as much as possible to get the information. The online e-questionnaire will be a very useful tool for this contact, since both the SA and his/her counterpart can be seeing the same erroneous or incomplete information. Even the information sent via Excel, once uploaded, can be reviewed, corrected and/or completed using the web e-questionnaire.

40. With this new approach of a single repository of received data, the consistency check is a single process that can be run for the whole database (this is recommended to be done once per day) or by each SA for just one country answers. This gives the SA the ability to interact with their counterparts, complete or correct the information and perform the consistency check, all in line.

41. Questionnaires with no errors according to the consistency rules defined will go to the “Dissemination database”. Those with errors will remain in the Work tables, but marked

accordingly. A report will be sent to the SA for him to know which are the defects that prevent each questionnaire from passing to the Main database.

42. Just in case of what is called a "false positive" (an error for the consistency rules, but not in fact, i.e. an outlier), the SA will be able (and responsible) of issuing an "allowance", that is, make the program to bypass this particular check and allow the questionnaire to be considered OK in spite of this strange or missed value.

43. A full screen editor that includes error message handling is used to fix the errors detected by the consistency check. After closing the editor, those tables of data collected (QTables) that have been edited are marked to be checked for consistency. This cycle is repeated until all errors are fixed or have been "allowed" and the QTable passes to the Dissemination database.

44. Editing of the questionnaire is restricted to authorized users, like the SA assigned or the Supervisors.

45. The Notes system has been totally redesigned. It is now based on a controlled vocabulary of about 400 notes which has been classified by into "Types" and related to the different Topics and Types of Source. The collection methods allow the users to select the notes thru closed lists and, eventually, to add a free text annotation. This annotation is later analyzed and coded by the SA when editing the QTables.

46. From the "Dissemination database", the new ILOSTAT website is composed of dynamic pages built based on the content of the database. For the publication of tables and charts, an ETL procedure will be run periodically to update the datamart for the BI reports embedded in the dynamic pages.

47. Contextual links to different types of documents like Publications, Resolutions, Guidelines, etc. will be available to the user based on the country and/or subject being consulted.

48. A workflow tracking system crosscutting the whole process records the status of the tracked units: e-mails sent to and received from the countries, and the QTables. The Data Flow Control Dashboard will give real-time information about the status of each questionnaire. The count of questionnaires in each stage will be displayed at first; then it will be possible to drill down to see it by geographical area or SA assignment. Then by country, and even go to see the emails interchanged with a country or the error listing associated to a QTable with errors or the content of it, regardless of its status.

49. "Source & Methods" system will be revamped and integrated to ILOSTAT database, making it compatible with the Data Documentation Initiative 2.x (ddi) standard to allow the incorporation of existing metadata.

B. IT considerations

50. Data is stored in a relational database mounted on Oracle 11g DBMS administered by ITCOM (the centralized ILO information technologies service). Two postulates have been established for the design of the new data structure: a) the data structure for the data collection database should be the same for all kind of time series data regardless of the periodicity, units of measure, classification breakdown and way of collection; and b) the main (atomic) unit is the "cell" of each table collected, which will be called VALUE and will keep associated dimensions and other attributes.

51. Although it is a Data Compilation system (and not a proper statistical activity producing microdata), the system has a modular design following the recommendations of the GSBPM, including modules for Data Collection, Data Cleaning, Dissemination, Workflow tracking, Code lists maintenance, User Profiling and Access Control and Source & Methods[1] (See

[1] Not included in the diagram

52. *Figure 3: ILOSTAT Information System modular design*).

53. Program development is based on Oracle **APEX** (Oracle Application Express) for the interactive applications, complemented with some PL/SQL packages and Java classes for specific tasks. Intensive data processing tasks, like consistency checking and Excel questionnaire generation are developed in SAS, accessing the Oracle database. The Workflow control dashboard and dissemination tables and charts are built using Oracle BI Enterprise Edition (OBIEE).

54. The User Profiling and Access Control module, developed in APEX, includes a dynamic menu that lists the applications available for the user based on his user profile. Examples of them are Statistical Assistants, Analysts, Managers and External Users.

55. The Data Collection module kept the automatic generation and upload of Excel questionnaires as in the former system, but it has been redesigned as to make use of a single set of metadata fully parameterized and common to both the collection and dissemination processes. The upload procedure (fully automated) performs basic consistency checking and routes the error report to the assigned SA for correction.

56. The e-Questionnaire application (under development) will be an interactive full screen editor for value and annotations on the data collected developed in APEX and accessible thru the web. It will work on the “Data Collection” work tables and will operate based on the single set of metadata for the QTables.

57. Electronic Data Interchange, probably using SDMX is in the roadmap for 2012, as a way of reducing the overburden to countries due to the request for information they already have in their databases and has to be transcript to offline or online questionnaires.

58. The QTable Consistency process, developed in SAS, can be run as a batch process to analyze all records marked “for consistency” in the “Data Collection database” or can be launched on-demand by the SA. This process will pass the correct QTables to “Dissemination” database and mark those erroneous with the respective error codes, remaining in the repository. The assigned SA is notified of the results, and the status of each QTable is updated in the data management system (See *Figure 4: Workflow status diagram*).

59. The Editor program is used by the SA to correct the errors detected in the data. This program displays the QTable being edited and the error messages related to it. When using off-line data collection methods, the country user can include annotations in the questionnaire that the Editor will display for the SA to code into notes associated to the data at the right level.

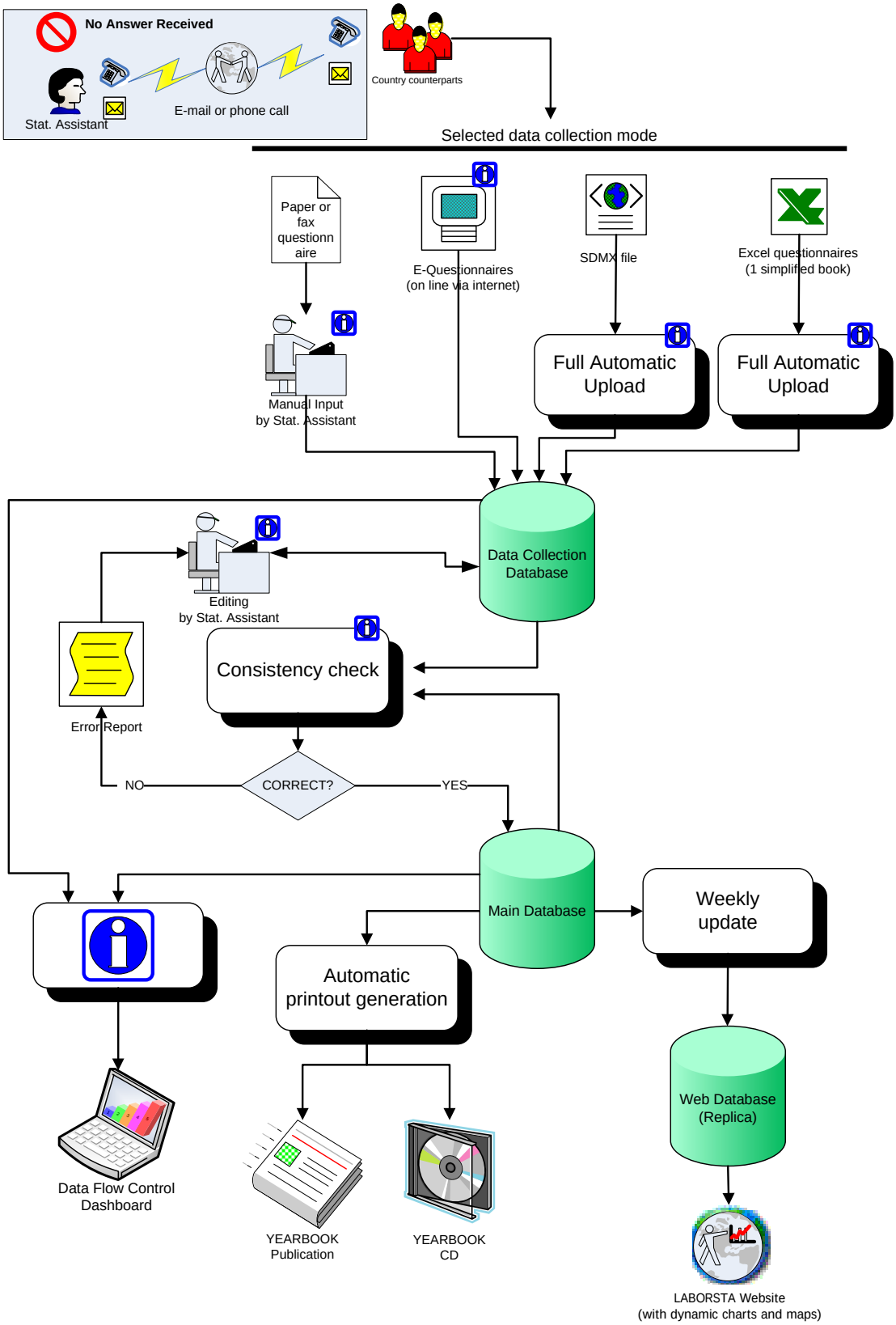

Figure 2: ILOSTAT Information System

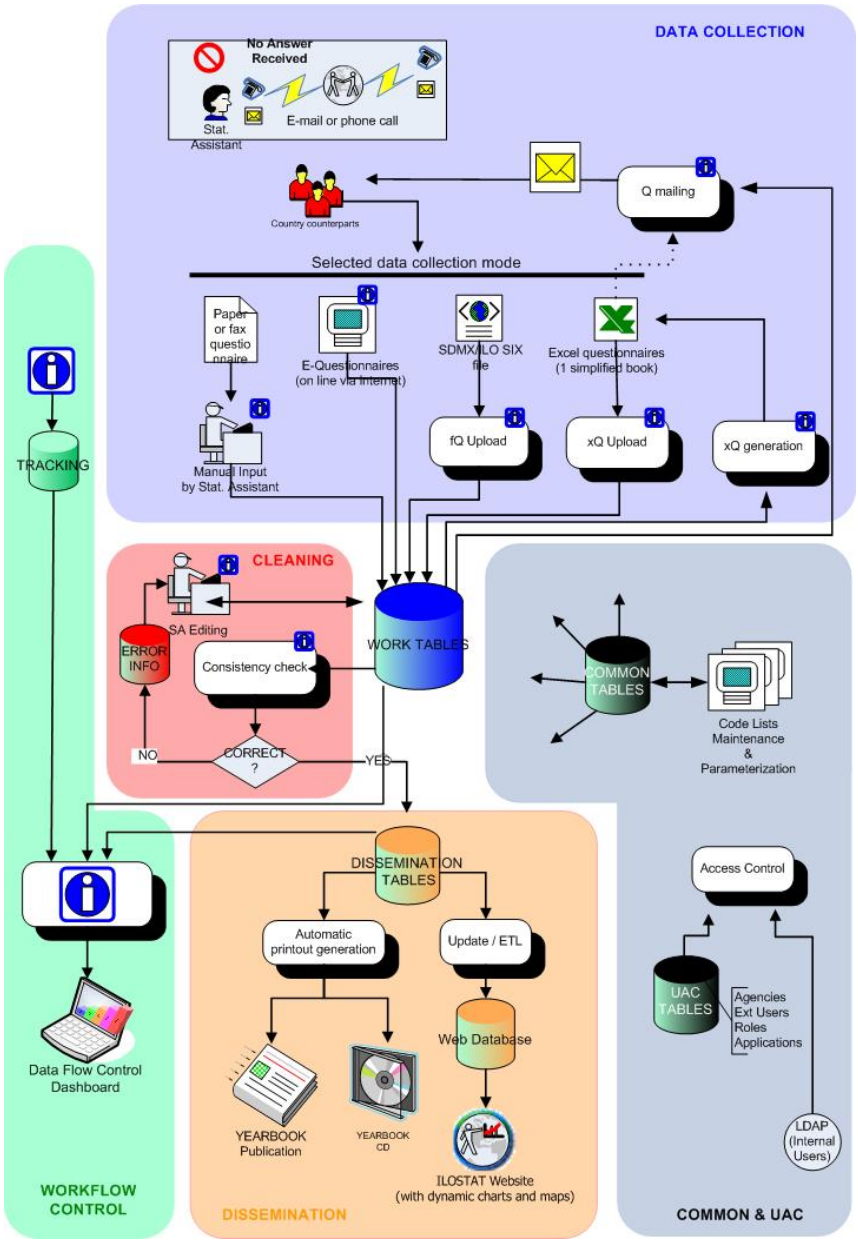

Figure 3: ILOSTAT Information System modular design

60. The Dissemination website (ILOSTAT) has been developed on Oracle WebCenter building reports with OBIEE. The ETL procedures that copies the data in the Dissemination tables to the BI Datamart is run three times a week or whenever a substantial addition of data has been made and deserves to be published immediately.

61. Every program and procedure that has to do with data in the questionnaires records information regarding changes in the status of Questionnaires (prior to receiving the data) and QTables (after data is in the database). Using the BI Analyzer, a control dashboard has been developed, showing real-time information regarding responses, errors and efficiency indicators for each operation, and allowing drilling down to the country and even questionnaire level, to see detailed information like emails interchanged with the country or the history of errors and corrections related to a QTable. This automated system will reduce the risk of errors or misses due to human factors.

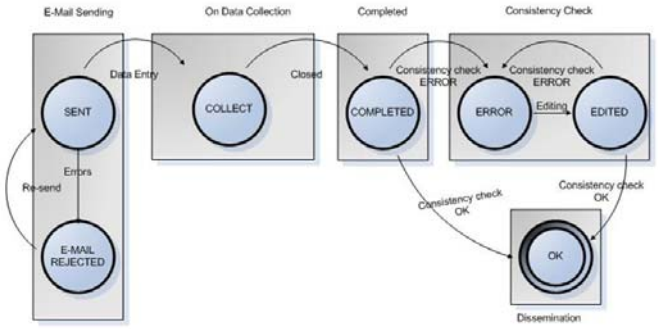

Figure 4: Workflow status diagram

IV. Conclusions

62. *Increased coverage:* With the implementation of the new Workflow control module in the ILOSTAT Information System, Statistical Assistants count with tools that are aligned with the new approach in setting the relationship with the countries for data compilation. SA are expected to be proactive, know about the reality in their assigned countries and establish a relationship with their counterparts in order to take the necessary actions towards increasing the number of countries and indicators answering to the enquiries.

63. *Improved opportunity*: Having more automated procedures and much information on the compilation workflow will make SA work more efficient, reducing the time needed to process the data compiled and being able of publishing it earlier.

64. *Improved quality*: The new Data Cleaning module including required cycles of Consistency Checking and Data Editing will minimize the possibility of erroneous data being published.

65. *Reduced overburden*: The addition of new methods for data collection like EDI and online questionnaires will help countries to reduce the time allocated to satisfy ILO data collection requirements.

66. *Standards based*: The adoption of standards make it easier to establish agreements with other supra-national organizations to compile and share data in a coordinated way avoiding the duplication of efforts to the countries.

67. *General Purpose*: The system has been designed to serve as a general purpose information system able of collecting, processing, storing and disseminating any type of time-series data and associated metadata. This has the clear advantage of reducing the learning curve for any new user since a single set of tools will be used for all the data.

68. *Reduced TCO*: The Total Cost of Ownership of the new system will be zero thanks to the significant reduction in the cost of software maintenance licenses for the next five years. The migration of the databases from SAS to Oracle and a careful review of the products and number of users contracted will generate economies for more that USD 500,000 for the period 2011 – 2015. This amount will largely cover the costs of implementation of the whole system.
