Distr.
GENERAL

Working Paper
11 April 2013

ENGLISH ONLY

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

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

UNITED NATIONS
ECONOMIC AND SOCIAL COMMISSION
FOR ASIA AND THE PACIFIC (ESCAP)

Meeting on the Management of Statistical Information Systems (MSIS 2013)
(Paris, France, and Bangkok, Thailand, 23-25 April 2013)

Topic (ii): Streamlining statistical production

Streamlining Data Compilation and Dissemination at ILO Department of Statistics Lessons Learned and Current Status

Prepared by Edgardo Greising International Labour Office

I. Introduction

1. In May 2012, during MSIS 2012 meeting in Washington, DC, I made a presentation about the rationale behind the streamlining processes we had started at the ILO Department of Statistics, and made a brief description of the project to redesign the department's approach 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 and dissemination.

2. 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 Country Specialists (formerly 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 ready to be included in a publication.

3. Basically, the new ILOSTAT database compilation procedures aims to have a better response rate from the countries, to reduce the delay of the information received and to improve the overall quality of the data published, with emphasis on comparability among countries.

4. 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. Since the release of the Generic Statistical Information Model (GSIM) we are progressively modifying our reference framework and terminology to its recommendations.

5. After one year, and with more than 80% of the project completed, we have some experiences and lessons learned in the process regarding the implementation of changes in the procedures, IT applications developed, development tools, the reality of data providers, etc., that we would like to share.

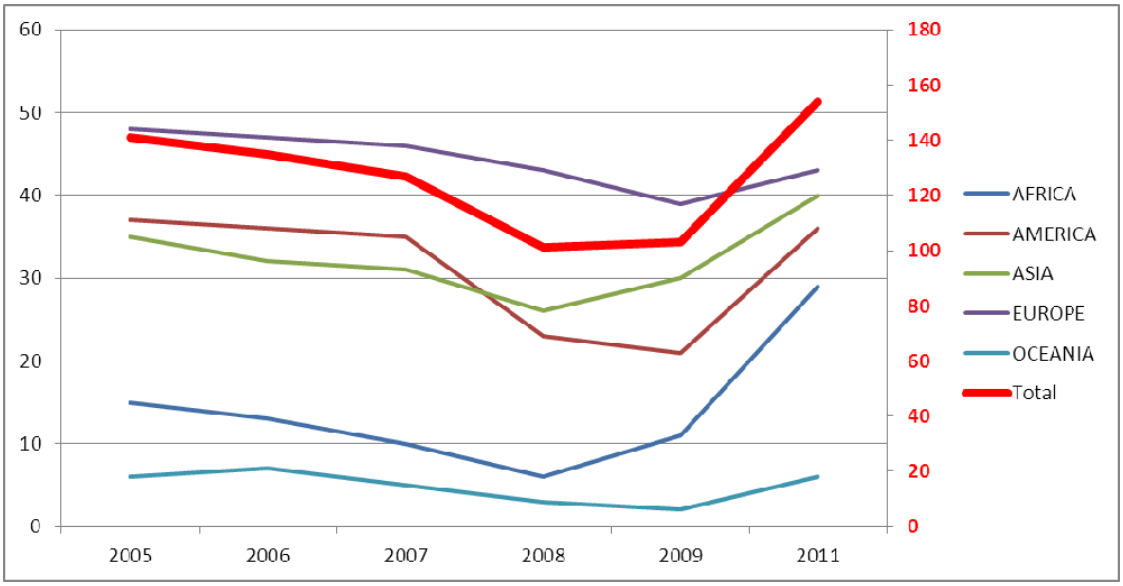
II. Changing the procedures

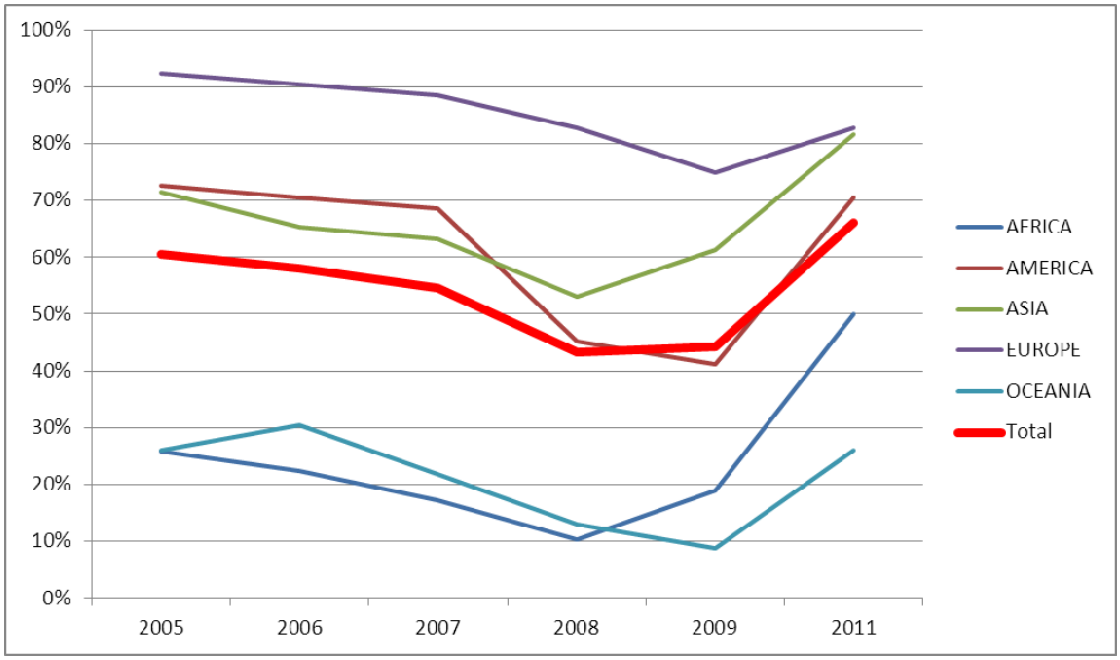
From Topic- to Country-centric approach.

6. Changing procedures that have been in place for more than 15 years is not an easy task. Former compilation was organized by topic, with several people having to interact with their countries’ counterparts, in some cases with language and cultural difficulties to understand each other. The big change consisted in assigning “Country Specialists” (CS) -formerly Statistical Assistants- to a group of countries following criteria of language and cultural affinity whenever possible.

7. These CS would have a better understanding of the country context and establish a closer relationship with their counterparts at the data-providing agencies. At the same time, Topic Specialists (TS) would be backing up the CS in case they receive specific questions beyond their knowledge.

8. After the first yearly collection, that included data for two years and was launched after stopping the compilation for about 18 months, the results are promising. The *coverage* in terms of countries answering the questionnaire has increased compared to the last five years, and reverted a trend of decreasing number of countries.





9. This is even more relevant when it is taken into account that the number of indicators collected has been increased from about 40 in 9 topics to 72 covering 17 topics.

10. Nevertheless, there are still some aspects to be improved. Not all the CS has adopted the new approach with the same soundness. In a recent activity where the team was asked for some feedback and suggestions on how to improve the data compilation process, the demand for training on the different topics, especially the new ones, has been identified as a priority.

11. On the other hand, other suggestion made by the team has been the need to improve the tools for maintaining the contact database, demonstrating its commitment to the new approach and its concern to improve contact with the countries.

Towards timely and comparable data

12. Looking at the *opportunity* of the information compiled, it is clear that this past collection was not what we could call a success. But there are some facts that should be considered as mitigating factors; it was the first collection with the new approach, with new IT tools that in some cases were being developed and improved while in use by the CS. There were 8 new topics, that needed extra-work from the CS to explain to their counterparts what was asked; and moreover, the structure of the questionnaire regarding new descriptive metadata collected with each indicator have also demanded a lot of contacts with the countries to explain how to fill it.

13. *Quality* of the data has been improved, since the new checking procedures ensure the consistency and integrity based on a set of rules that checks for errors inside the tables (typically totals and range) but also among tables, for those indicators with correlation. The consistency-&-correction cycle established with strict controls and automated workflow assures that no data with errors will be published.

14. On the other hand, some breakdowns are to be reviewed for the next compilation since the excessive detail requested for some indicators has led to an important number of values marked as Unreliable by the countries due to sample design. In other cases, countries are not able to provide the data requested at this level of disaggregation. This is a point that is being reviewed very carefully for the next round.

15. The *comparability* among countries, maybe one of the most important issues addressed in the new approach, is still in the must. Unfortunately, many countries do not follow strictly the international recommendations for measuring the different labour indicators, or use proprietary classifications. That conspires against the ability to have series that could be easily compared across countries. Besides, the complexity and/or level of detail of some of the tables requested made it difficult for some developing countries to provide the data the way it was requested.

16. To mitigate these issues, the Data Production Unit team is working in a simplification of the questionnaire based on the feedback received from the countries while the CS will be identifying those countries that would need technical assistance in order to help them in the adoption of the international standards.

III. Application development

"The size of the accomplishment can be measured by the obstacles you have to overcome to reach your goals" – Booker T. Washington

17. Looking back to the beginning of the development in early 2011, the road has not been easy, even it was chaotic at some times, but the results obtained according to the resources invested are more than satisfactory.

18. The analysis and conceptual design took too long, leaving short time for development. The compilation had been stopped for a year and there was no way of not taking it back in 2011. At that moment, the only way to start the compilation in the new format was to adopt an iterative-incremental development model so as to initially provide basic versions of the tools, to then complete them and add new features based on feedback received from the users.

19. These led to several situations in which the pressure on the development area was very high, since the time was short and in some cases the CS were not able to advance their work because certain application was not ready or its first release did not work well at all.

20. Additionally, the entire system was mounted on a new software infrastructure (Oracle 11g DBMS, APEX, Web Center) becoming "early-adopters" of this technology within the ILO, with all the consequences this entails. So, several incidents that delayed the development were related to bugs detected in the installed products, problems applying patches, lack of experience in the administration of the new platform in the IT staff, etc.

21. However, the application "back-office" now meets the requirements for which it was designed and its performance is adequate. It is expected that for the next compilation it will not be an obstacle and, on the contrary, it will contribute to achieve the objective of shortening processing times.

"Vamos más despacio, Sancho, que estoy apurado" – *Don Quijote de la Mancha*
(Let's go slower, Sancho, I'm in a hurry).

22. One of the points in the development project that has been deferred was the "multi-modal collection", specifically the availability of the electronic questionnaire (e-Questionnaire) and options for Electronic Data Interchange (EDI) based on SDMX and csv formats.

23. The main reason why we could not implement the e-questionnaire yet is because we have made a great effort to provide extensive features for the Table Editor for internal use in the data correction and metadata coding. Many of those features already implemented in the Editor are actually time gained in the development of e-questionnaire, although we still have to work hard to complete this product in the coming

months since it will be a tool for external users, so many aspects of data security and integrity must be considered.

24. Regarding EDI, we will have for the next compilation an interface to capture SDMX format files and also with the ability to import CSV flat files in a proprietary format that includes data and metadata. One of the aims of ILOSTAT compilation is to reduce the burden to countries, and EDI is one of the ways that may help in that objective.

25. One aspect that it worth highlighting is the fact that in many developing countries they do not have a repository of indicators. In these countries the indicators are calculated as part of a publication plan and then tabulated and the results are published on the website of the NSI, often in pdf or Excel. However, the data is not preserved in a central repository, which makes the generation of SDMX files (or even csv) very difficult, because there is no database from which to take this information.

Knowing how it goes.

26. One of the most controversial modules of the new system has been the Workflow Control Subsystem, which has been seen by some of the internal users as a "spy tool" designed to control how they work. It took a long time to make them understand the usefulness of this tracking system for them to access the whole "history" around a country compilation and for the management to make decisions based on evidence regarding the updated status of the collection.

27. After some time, CS starting to understand its usefulness and they have started to provide some feedback on the dashboard reports and requested for customized reports to fulfil their needs in relationship management with the countries.

To BI or not to BI?

28. But the hardest situation during the project was the definition of the architecture and the tool to be used for implementing the dissemination website. As it was mentioned before, ITCOM (ILO's Department for Information Technology and Communications, services and infrastructure provider), had suggested the use of Oracle technologies for ILOSTAT system. For dissemination the platform selected was Web Center portal manager for holding static and dynamic pages and Oracle Business Intelligence Enterprise Edition (OBI-EE) for building the reports.

29. A proof of concept was done for building reports using OBI-EE, and at that point it was clear it would be very difficult to do an understandable statistical report due to the amount of descriptive metadata we have to display together with the multi-dimensional tables.

30. ILOSTAT metadata is collected and attached at different levels, but normally at the observation level and at the series representation that we call "Qtable"[1] level. There are several "Types" of notes which can be grouped in two sets: those that give information about characteristics of the indicator or the source (actually descriptive metadata) and those that points out exceptions for a particular value or set of values (typically showed as footnotes).

31. The main report for one indicator in ILOSTAT website is a table with the classifications breakdown combined in the rows and one column for each point-in-time. Due to screen-size constraints, twelve columns are displayed at a time, with a slider for moving this 12-columns window alongside the available data.

[1] Qtable (Questionnaire table) is a flat representation of the cube defined by the combination of the different classification breakdowns established for an indicator, considering just one point-in-time (a single column).

32. Normally, for a given country, a note attached to a value for one month (for example an exception on the classification item) can be the same for several months, provided the data source remains unchanged. But this is not known when data is collected month after month and the note is attached at the value level. Nevertheless, when the dissemination report is build, the objective is to consolidate all these occurrences of the same note at the highest level possible and show the footnote reference just once.

33. One additional difficulty is the fact that the consolidation has to be done "on-the-fly" after displaying the report, since it is not possible to know in advance which rows are to be displayed (due to the filters on the breakdowns that can be applied) and which timespan will be displayed in the 12 columns. And has to be re-built each time the 12-columns window moves back or forward by means of the time slider.

34. This consolidation resulted very complex to be done with OBI-EE reports since there's no way of controlling the cell display or doing any kind of post-processing once the table has been defined. The report builder has very powerful features for filtering dimensions, but every computation is supposed to be done previously, during the Extract-Transformation-Load (ETL) process of the datawarehouse.

35. Another problem was the code and label display for the different breakdowns, since the multi-dimension tables and the concatenation of codes and labels for building a row or column header, so common in statistics reports, are not "typical" BI reports.

36. Alongside with all that constraints in the report generation, using BI demanded the creation of a "proper" datawarehouse, a "star-schema" database which had to be maintained thru complex ETL procedures and, in fact, added no value to our application. This data does not come from heterogeneous sources, there's no need for complex filters or aggregations. On the contrary, the statistical tables to be displayed can be easily obtained with a simple SQL pulling the data from two flat tables, one for the observation values and other for the notes (metadata), which compose our "datawarehouse".

37. With all that constraints, we decided that a development tool which provided a better "control" at the moment of building the reports, which allowed for post-processing the table for computing the notes on-the-fly and that could query the data from a flat view instead of a star-schema datawarehouse, would be a better option. The Oracle Application Development Framework (ADF) with some extensions in Java was the perfect solution for our needs.

38. It is worth to mention the great help we received from the official statistics community when, in the middle of the discussion about the convenience of abandoning OBI-EE and the datawarehouse to go to a flat dissemination database and an ad-hoc application for displaying the reports, I sent an e-mail to some colleagues asking for similar experiences and advice. I received plenty of answers that helped me to reinforce the idea that statistical data processing is a different "business", and as such has its own rules, and sometimes standard tools designed to other environments not necessarily fits our needs.

39. And very special thanks to OECD colleagues who told me the story of their own information system redesign a couple of years before, in a very similar situation. Knowing that experience was very relevant in our final decision.

All you need is data…

40. The ILOSTAT dissemination application has been conceived as a fully metadata driven website. Every navigation menu is contextual and is built dynamically based on the information stored in the back-office database. For example, for a given country, only indicators with actual data values are presented in the

selection panel. On the other hand, only countries with data are displayed when a particular topic or indicator has been selected.

41. This characteristic of the system is great in terms of avoiding duplication of efforts, since it is not necessary to edit a web page to reflect changes in the database; if a new topic is added, it is automatically reflected in the website. This is also a great advantage in terms of minimizing errors due to missed updates.

IV. Current status

42. *Increased coverage:* With the implementation of the new Workflow control module in the ILOSTAT Information System, CS count with tools that are aligned with the new approach in setting the relationship with the countries for data compilation. They have started to be more proactive, knowing about the reality in their assigned countries and establishing a better relationship with their counterparts. After the first round of collection, the figures regarding the number of countries and indicators answering to the enquiries are promising.

43. *Improved opportunity*: This is still a pending issue, but we are confident the new collection will be much more efficient and timely completed.

44. *Improved quality*: The Data Cleaning process implemented with required cycles of Consistency Checking and Data Correction, and the totally redesigned notes and metadata system, with coded notes have raised the quality of our data and minimize the possibility of erroneous data being published.

45. *Reduced overburden*: We hope that 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. This new data channels are expected to be released for the compilation in the second quarter of 2013.

46. *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. We have one agreement with Eurostat and we are willing to do the same with countries or partner organizations.

47. *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. We are in the process of integrating three new datasets and have another three in the queue. This is a great achievement for the Department of Statistics and the ILO in general since it’s a real implementation of the “*Delivering as one ILO*” vision statement made by our Director General.