Distr.
GENERAL

Working Paper No. 9
18 May 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

Changes in business processes and technology tools for taking up increased workloads

Invited Paper

Ann McPhail and René Piché, Statistics Department, International Monetary Fund (IMF) [1]

I. Introduction

1. Like many organizations, the IMF Statistics Department (STA) has experienced a significant increase over the last ten years in the volume of data and metadata that we collect, process and disseminate. This trend is expected to accelerate over the next five years, at an almost exponential rate. The latest and perhaps most important source of demand is the need to fill gaps in the available data that were identified as a result of the financial crisis. Specifically, there are 20 recommendations to fill data gaps that have been adopted by the G-20 Finance Ministers and Central Bank Governors in November 2009, the so called Data Gaps Initiative, or DGI[2]. The DGI has brought increased visibility to the role of official statistics in evidence-based decision making; it provides a golden opportunity for producers of official statistics to increase important "market share" among policy makers and others who rely upon high quality data for high quality decision making.

2. The opportunities presented by the DGI come with equal, if not greater, challenges. In addition to volume increases, we are called upon to improve the quality of our data with respect to timeliness and frequency. On the resource side, we have relied increasingly on temporary sources of financing to hire contractual staff to manage the added work load, but these funds have leveled off and are likely to

[1] The views expressed herein are those of the authors and should not be attributed to the IMF, its Executive Board, or its management.

[2] See *The Financial Crisis and Information Gaps* report, jointly prepared by the Financial Stability Board (FSB) Secretariat and the International Monetary Fund, November 2009, available at: http://www.imf.org/external/np/g20/pdf/102909.pdf

decline in the coming years. Technology has delivered some important pockets of efficiency gains but we have harvested all the "low hanging fruit" and are scrambling for IT capital funds to make more substantial improvements to our systems. These challenges also present an opportunity to evolve our business model and modernize our products. We have launched an initiative to "streamline, standardize and automate" our statistical production processes while also examining the relevancy of our datasets so that we can target areas for improvement and reduce or eliminate low value added products. Perhaps more importantly, we are putting into place the governance structures and processes to ensure that what we change or improve now can be sustained and/or changed again down the road to respond to changes in the external environment.

3. The paper is organized as follows: section II reviews future trends and implications; section III describes our current environment; section IV presents our desired future state; and section V presents some conclusions.

II. Future Trends and Implications

4. An assessment of recent trends in the areas of statistical production and information management has shown that "business as usual" is not a viable option. A few of the key trends could be summarized as follows:

Outside the IMF:

- We hear that National Statistical Organizations (NSOs) are "industrializing" their processes as one means of keeping pace with volume increases and users needs.
- More and more countries are posting their data on the internet in lieu of reporting to national and international organizations, including to STA; the Open Data movement is likely to drive more organizations in this direction.
- Users have greater choices for sourcing their statistical data, and more sophisticated and interactive ways of consuming data.

Inside the IMF:

- The financial crisis revealed some key data gaps related to IMF analysis of spillover effects and interconnectedness.
- STA is under pressure (and has the opportunity) to supply more detailed, timely and higher frequency data as well as an increased number of larger datasets to meet operational needs of other Fund departments.
- STA data are more difficult to access compared to commercial data providers and users are not aware of all the data we collect.

5. STA recognizes that we must adapt our processes and products to keep pace with demands and to remain relevant to our users. Over the past few months, STA's Statistical Information Management Division (STA-SI) explored the barriers and opportunities in its business model and how it could leverage these opportunities in order to deliver "more, better, faster" statistics. A stock taking exercise identified the following areas of improvement in moving the process forward:

- Upgrade STA's data collection methods.
- Streamline and automate data validation processes and target our validation efforts.
- Develop ways to disseminate data in real time.
- Add value to data products by providing more interpretation or storytelling.
- Do a better job of understanding the needs of our priority users.

6. Identifying where changes are most needed and are likely to deliver the greatest benefits is complex but we realize that we need to implement information management practices that are more streamlined, standardized and automated. To achieve those objectives, STA is dissecting and analyzing the three main components of our statistical production processes: collection, validation, and dissemination.

III. Our Processes Today

7. There is significant variation in the data collection and validation systems across STA's data products. STA's organizational structure is one source of variation. STA has four topical divisions: the Balance of Payments Division (which covers all external sector statistics), the Financial Institutions Division, the Government Finance Division, and the Real Sector Division. Each division is responsible for methodological issues, the delivery of technical assistance, and developing and owning data products that fall under their respective topical domains. There are important variations in the rules controlling the collection and validation systems that have been established over the years, which can be traced back to the time when each of these divisions was responsible for all aspects of the statistical production processes within their respective domains.

8. In January 2007, STA established the Statistical Information Management Division (STA-SI) and centralized the operational aspects of data collection, validation, dissemination, with the aims of achieving a more consistent approach to its data operations, to improve data quality through standardization, and to realize efficiency gains and achieve staff reduction targets. While incremental improvements have been realized and some processes have become more standard across more data sets, many of the existing data production processes still reflect the siloed development approach that pre-dated the establishment of STA-SI. We lack an end-to-end view of our business processes and further change has been difficult to realize.

A. Data Collection Processes

9. The approach to data collection differs by data set. A review of the practices across topical domains suggests too many variations in the collection processes. The standard collection practice for most datasets is the reliance on data submissions by country authorities. Data are submitted directly to STA on an on-going basis for many data collection exercises, and at some pre-agreed schedules for others. The preferred way for data to arrive at STA is via the *Integrated Correspondence System (ICS)*, a secured web-based data reporting system developed by STA for use by our member countries.

10. Due to the high demands for more timely data, STA-SI has started substituting data reporting with web scrapping of country websites using commercially available tools. In addition, we are working towards increased usage of the Statistical Data and Metadata Exchange (SDMX) standards, particularly for data sharing with regional and international organizations. STA-SI is also investigating options for using data from data resellers for selected series, to the extent that licensing permits such reuse. In addition, some country data still come to STA through such channels as fax or surface mail.

11. Data transmitted to STA via the ICS are in standard Excel report forms structured by STA, and for the most part, are loaded automatically into the data processing environment. However, there are plenty of occasions where a country data reporter has made a change to the structure of the report form which necessitates pre-processing clean-up on our end before a file can be processed. The data reporting requirement are not identical across domains; in one domain, empty cells must be filled with zeros, in order for the business rules for aggregation to be applied successfully. In another domain, zeros must be replaced with empty cells for aggregation routines to run successfully.

12. In addition to ICS, we have various tools for batch processing and for loading data files to our production environment that vary according to the dataset. The variation in processes increases the maintenance costs and results in a steep learning curve for our data processing staff-- many of whom are on yearly contracts with a steady pace of turnover. A key benefit to standardizing and automating our collection processes would be to reduce the amount of time we spend on staff training and in maintaining these processes.

13. In the past several years, STA has adopted standard report forms that align with the updated statistical methodologies developed by the IMF, such as the *Balance of Payments Manual* 5th edition (*BPM5*), the *Monetary and Financial Statistics Manual 2000*, and the *Government Finance Statistics Manual 2001*, while work is underway to update the BPM5 report form to the *Balance of Payments and International Investment Position Manual* 6th edition.

14. As a result, the majority of STA's datasets are standard, which means that the same time series, aggregation and validation rules apply to all individual countries in the dataset. However, for all datasets, there remain a number of instances where the time series, aggregation, and validation rules are country-specific. This is the result of STA's attempt to address incomplete reporting of data by countries that have not yet adopted the recent statistical methodologies.

B. Metadata Collection Processes

15. Reference metadata are collected across all datasets, with the most comprehensive metadata residing in the Dissemination Standards Bulletin Board (DSBB). The DSBB comprises the metadata that IMF member countries subscribing to the Special Data Dissemination Standard (SDDS) or participating in the General Data Dissemination System (GDDS) are required to submit. These country metadata cover 93% of Fund membership. However, apart from SDDS/GDDS, countries are encouraged, but not required, to submit metadata in support of the data report forms they regularly submit to STA.

16. The metadata submission process for the SDDS/GDDS is highly standardized and automated and this automation lead to significant reduction in the metadata processing cost incurred to STA to provide this information on the DSBB. However, the metadata collection, validation, and dissemination processes for STA's data products are independent from the DSBB metadata and vary significantly across statistical domains. This high level of customization makes it difficult to develop and maintain the multiple different processes that allow the reference metadata to flow together with the data all the way to the end users.

C. Data Validation Processes

17. While data processing and development of new data products are now centralized in STA-SI, most established data validation processes are still domain and topic specific. This is the result of legacy systems that, until 2007, were managed independently by the topical divisions within STA that were solely responsible for their own datasets.

18. Streamlining and standardization are needed to free up the resources that should be assigned to meeting the demand for the increased data volumes. The data validation process takes place at various stages of the data work flow, but duplication of data review and the number of manual steps involved during the review increase significantly the cost of data processing.

19. Basically, the data go through a five-step review and validation process. The first three steps look for format issues and check for internal consistency, the last two steps look at more subject elements such as do the data make sense.

20. The first-tier data processing and review takes place prior to loading the data in the data processing environment. Data submitted using STA's report forms are processed using a "compare tool" which compares the data in the report form with existing data in the processing environment. The tool performs a number of operations, including identifying revisions through color coding of data cells based on topic-specific thresholds. Newly reported data are also identified, as well as growth rates to last previously reported data. All these validations are performed in an instance of the (Excel-based) report form, prior to the data being uploaded in the database.

21. Additional validations are applied through the use of checks inherent to and defined specifically for each report form. There is a great deal of variation in the extent to which these additional validations are included in the report form; some forms have very few and most data validation takes place after the data are loaded in the database, while for others, a significant portion of the data validation takes place directly in the report form (or more precisely, an instance of the report form that gets generated after the form is processed using the "compare tool").

22. These validations are essentially performed manually; the system identifies issues using color coding of data cells or by compiling tables of validation checks, but the data processing officer is responsible to visually review the report form and go over flagged cells or validation tables, identify whether there are data issues or not, and whether and how the issues should be addressed. The optimal way of reviewing issues, identifying those that require addressing, and the methods for addressing them are report-form, and sometimes country-specific and require extensive training of the data processing officer for a quick turn-around time.

23. Once the reported data are loaded in the processing environment, more tools and reviews become available for their processing. For most data collection exercises, the validation rules built-in the databases are similar, but usually more detailed, then the ones available from the report forms. In addition, the system generates reports for further review and analysis of the data. They include trend reports and tables reproducing publication-format data, as well as presentations developed for internal data users.

24. For most STA's data gathering exercises, collection/validation is centralized within two teams in STA-SI and data officers in those teams are expected to develop expertise in the processing of the various report forms under their responsibility. For each statistical domain, a "domain manager" within STA-SI is assigned to oversee the production of the data for its domain. He/she will help develop the domain expertise of the data officers, help them address specific data reporting issues, and provide methodological and other expert advice to continuously improve the domain data.

25. The domain manager is also responsible for the second-tier data review, which is usually, but not always, supplemented by a third-tier review prior to the dissemination of the data to the public. The validation role assumed by the domain manager varies by domain and is in part guided by the importance of the third-tier review, which is performed by economists in the topical divisions. The amount of data review performed by topical divisions varies; some data collection exercises are under great scrutiny, while other less important datasets are rarely reviewed. The role of the domain manager in the validation of the data is adjusted to take that diversity into account and ensure a proper final review and clearance of the data.

26. As indicated, STA's topical divisions' economists also review data submitted by member countries. Data submitted to this third-tier review have passed an extensive series of data checks intended to guarantee the internal consistency of the data. However, at this juncture, STA-SI has no tools available to ensure that the story told by the reported data (e.g., trends and outliers) accurately reflects the reality of the reporting country. That type of review is currently assumed by topical divisions' economists. To facilitate the work of economists and improve their productivity, STA has recently deployed a new software solution that provides a platform for performing trends and outliers analysis, which includes cross-checks with data from other collection exercises and data visualization.

27. The validation processes vary greatly across the datasets maintained by STA. Notwithstanding these variations, STA is able to continue delivering high quality statistics but it spends a large amount of time and resources in doing so. The lack of process standardization, the large number of manual steps, and the duplication of efforts are inefficiencies of the current legacy environment that increases the cost of producing the data.

C. Data Dissemination Process

28. Shortly after STA-SI was established, STA management decided to adopt a data warehouse approach for the dissemination of all its electronic data. Development took place during 2007-2010 and the data warehouse-supported IMF eLibrary was launched on March 28, 2011. The data warehouse provides a repository for all data produced by the IMF Statistics Department and for a number of other IMF-produced and externally produced datasets, which are restricted to internal users.

29. However, the data warehouse does not represent an end-to-end solution to data collection and dissemination, but adds another layer to the network of processes that move data from the processing environments to the dissemination environments. The data processing environments include a "production space" and a "dissemination space", and data for dissemination are usually taken from the "dissemination space", but for various legacy and work practice reasons, there are a number of instances where data are fetched from the "production space". There are three independent processes that take data from the processing environments and move them to three separate data dissemination environments. The implementation of these independent processes varies across STA's datasets, resulting in a large number of customized "dissemination" processes.

30. The Economic Data Sharing System (EDDS) is a legacy data sharing system that was launched in 1998 to improve data sharing in the IMF. Data are fed to EDDS daily using a replication process of the data processing environments pulling data from either (or both for some datasets) the "production" or "dissemination" spaces.

31. The data publication system is a separate system that includes a series of processes that generate the paper publication pages for the various STA's print products. For some of these print products, the processes are supplemented by a CD-Rom creation process. Data in the CD-Roms have a coding structure that usually aligns with the one adopted for the processing environment, while the print publications have adopted a print-specific shorter coding structure.

32. The data warehouse dissemination system includes a number of extract-transform-load (ETL) processes that would point to the data processing environments, either the production or dissemination spaces, and perform a series of transformations prior to loading the data in the data warehouse. These transformations include mapping the native coding structure of the processing environment to the "Catalog of Economic Time Series" (CTS), which is the Fund-wide mnemonic-based coding structure. The ETLs also transform the time-series based production data into the dimensional model adopted for the data warehouse, and populate reference metadata fields using attributes from the production data.

Additional reference metadata are added by a series of processes that uses the reference metadata developed for the print publication to populate the data warehouse.

33. STA strongly believes in the potential of the data warehouse for improving the dissemination of and the access to its data. However, in the current implementation, the entire data dissemination process has become an intricate network of operations. The multiple independent data dissemination processes make it costly to keep the data in the multiple dissemination environments fully in synch with the production data. In most situations related to the usual data production cycle, this is a non-issue, as data flow seamlessly from one environment to the next. However, whenever a new series is added in a dataset, non-relevant series eliminated, or a change is made to the dissemination tables, it requires thorough follow-ups with the various business and IT teams responsible for the EDSS, the print publication, the CD-Rom creation, the CTS management, the ETL management, and the data warehouse dissemination. As each of these teams has specific requirements, it is costly to maintain the current dissemination processes.

IV. Where we need to be

34. The main challenge faced by STA-SI in its endeavor to streamline, standardize and automate our processes is to identify the priority requirements for each of our data/metadata products and re-engineer the processes to accommodate these requirements while adopting leaner and more manageable work practices that reduce complexities and lead to cost savings.

35. The most important first step has been to take stock of our current processes to identify the extent and types of variation across our business processes, and then to question the need for variation. In the process we are creating a fact base about our products and services that will support future decision making and are developing a shared view of priorities and opportunities. We are developing an end-to-end view of our business processes for statistical production to guide our efforts at standardization. The Generic Statistical Business Process Model (GSBPM) and Generic Statistical Information Model (GSIM) are key frameworks that will help us in this process.

36. STA is also taking stock of the evolution of our user base over time and the recent shift to meet more of the operational data needs of the Fund through the provision of timely, official statistics. At the same time, we are reexamining our portfolio of data products to identify high value-added products or where we have a comparative advantage.

37. Further, to assist in taking stock of our existing data sets and publications, STA has recently adopted a centralized governance model for our products and publications and formalized this function in two cross-departmental STA committees that will guide the portfolio of data products and the publication practices and formats, respectively. STA management saw this approach as important to break down silos and move to more cross-departmental development efforts and increased standardization of processes.

A. Change Management Initiative

38. STA-SI is taking a lead role in these initiatives and, in addition, has established three internal working groups tasked with identifying opportunities for change in data collection, validation, and dissemination, respectively. The challenges STA faces, which have been fleshed out at two recent staff retreats, will assist the working groups in identifying concrete actions that will be implemented within twelve months and other activities that would require more gestation and resources to implement over a longer time period.

39. The future vision of STA's data production activities could be summarized by Table 1 below.

Table 1. What's In and What's Out for Future Data Production Activities

What's In	What's Out
Real-time Data, data harvesting	Report Forms
Continuous dissemination, Open Data	Paper publication
Standardization	Stovepipes
Industrialization	Cottage Industries
Automation	Manual processing
Data visualization, Analysis (story telling)	Books of data tables
Collaboration with commercial providers	Area departments collecting their own data
Data quality	Data collection
Validating sectoral consistency	Validating each number
International comparability	

V. Conclusion

40. Recognizing the future challenges of increased workload with a flat budget, STA-SI has challenged itself and STA management to a vision – to achieve more, better, faster data processing by evolving our business model and taking a prioritized approach to data validation. Further to the commitment of the vision, we have taken stock of our existing key users, we are adjusting our data priorities to focus on the needs in the organization for multilateral surveillance while also serving the needs of the international statistical community. This change initiative was well received by STA management and additional efforts are now underway to elaborate an implementation plan for year 1 and proposals for the medium term. In addition, STA has adopted a centralized governance model for its products and publications and formalized this function within the Task Force for Methodology and Data Initiatives (TFMDI) – for data sets and products – and Data Publication Transformation (TFDPT) for electronic and print publications.

41. We will continue to scrutinize our information management practices and put into place processes to ensure that these processes keep pace with our needs. We will initiate some changes to our processes to begin the longer term effort to streamline, standardize and automating, focusing initially on collection, validation and dissemination practices. STA-SI has three working groups who are wrapping up their initial reports and recommendations for concrete actions to take in the current fiscal year.

42. We have learned from our colleagues at ABS and elsewhere that starting small but planning big is a good approach.