**Working Paper**
ENGLISH ONLY

**UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION**

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

**Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality**
(Ottawa, Canada, 28-30 October 2013)

Topic (iv): The trade-off between quality, utility and privacy

# DATA CONFIDENTIALITY, RESIDUAL DISCLOSURE AND RISK MITIGATION

Prepared by Raja Hettiarachchi, Statistics Department, International Monetary Fund (IMF)

# Data Confidentiality, Residual Disclosure and Risk Mitigation:
## Challenges in Managing the Demand for full Disclosure and the Need to Safeguard Restricted Information

Raja Hettiarachchi[1]
Statistics Department, International Monetary Fund, Washington, DC (ghettiarachchi@imf.org)

**Abstract:** The paper will discuss essential components of the policies, procedures, and IT implementations put in place to safeguard confidential data, and reduce the possibility of accidental disclosure with specific examples from the ongoing Data Management procedures within the IMF Statistics Department. The primary topics to be covered will include: Levels of Data Confidentiality - with respect to the commitments with National Authorities, and internal Fund policies; Procedures – omissions and suppression of data to limit residual disclosure, and safeguard information derived only for the purpose of internal analysis; Technology – data management and dissemination system and user-level restrictions to limit the access to confidential data; Challenges – in terms of balancing transparency of data dissemination and the need to safeguard dissemination of unauthorized data.

## 1. Introduction

The Statistics Department (STA) of the International Monetary Fund (IMF) collects, validates, and disseminates financial statistics provided by National Central Banks, Statistics Departments, Finance Ministries, and International and Regional Organizations. These data are compiled and reported following established domain specific methodologies which result in internationally comparable data. In turn these data are managed and disseminated for public consumption, and used by national authorities, international organizations, researchers, and Fund staff for internal analysis. This paper will discuss the ongoing work by STA on methods of statistical disclosure controls and challenges brought on by the demand for full disclosure.

## 2. Business Context

STA manages and disseminates restricted information, defined as information that may not be made public and may be disclosed only under limited circumstances. Strong non-disclosure work practices are required due to the restrictive nature of some of the financial statistical data. In collaboration with national authorities, STA has embarked on efforts to re-evaluate the disclosure controls of sensitive data. As a global institution entrusted with highly confidential data by national authorities, IMF has little or no margin for error in regard to

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

disclosing sensitive information. STA has broadly categorized the levels of sensitivity/confidentiality of the data it manages and has established systems, procedures, and access level controls to safeguard sensitive information from misuse. All these are done while increasing the efforts to improve data utility by disseminating granular level data to the highest extent possible along with analytically useful aggregates.

## 3. Overview (Levels of Confidentiality)

This section describes the data confidentiality scenarios STA manages in its day to day operations. Confidentiality issues arise due to the sensitivity of data, and the statistical disclosure controls applied by countries, as well as by the IMF. Clearly there is a need to protect accidental disclosure of unauthorized data.

These data confidentiality levels are:

**3.1 National level data reported to STA only for the purpose of internal analysis, and/or calculation of Global and Regional aggregates**
National level data are considered confidential and can be reviewed only by authorized staff.

**3.2 Reported data series suppressed by the authorities to protect data confidentiality**
Due to statistical disclosure controls applied at the national level, some of the series (all data points) are considered confidential within a data report form. In this scenario, a data submission will have partial data due to suppression of some data series, where the sum of components will not equal to aggregate values.

**3.3 Individual data observations suppressed by the authorities to protect confidentiality**
This is done due to statistical disclosure controls applied at the national level.

**3.4 IMF staff estimates treated as confidential data**
Data not reported by authorities but derived by staff through manual or automated processes are at times considered confidential and used only for internal analysis.

**3.5 Global and Regional aggregates treated as confidential and suppressed from dissemination**
This is due to dominance of individual country(s) in a group or the limited size of the reporting sample.

For most of the above scenarios, apart from omissions and/or suppressions of primary data observations, it is necessary to suppress secondary observations to prevent residual disclosure

where a data user may be able to calculate the value of a suppressed confidential number by deduction from other available information.

## 4. Policy

As a general rule, even though there is high demand to publish as much information as possible, STA has to comply with data control policies implemented at the national level as well as IMF internal statistical disclosure controls. For some of the data sets collected by STA, there are written agreements with national authorities regarding the extent of the use of the data and the access controls within the Fund and the general public. For data sets deemed as not confidential, but that contain certain data observations deemed as confidential, STA encourages authorities to suppress these data observations prior to reporting to the Fund. Only on rare occasions, at the request of national authorities, STA will omit data series and/or suppress individual data observations to protect the confidentiality of national level data.

## 5. Procedures

STA receives data sets from national authorities and other International Organizations with varying degrees of sensitivity. To balance the need to safeguard restricted information from misuse and the desire to disseminate as much information as possible, STA has developed several data management systems and procedures to safeguard confidential data while minimizing IT costs and staff review time.

Within the information security framework of the IMF and STA data management guidelines, data confidentiality is managed by using a combination of system designs, operational procedures and access controls. System designs take into consideration the required access controls and dissemination controls depending on the sensitivity of the data set. Operational procedures are developed for careful information reduction (i.e., suppression of primary and secondary data observations, and suppression of aggregates if needed) and data validations. Access controls and statistical disclosure controls are used to ensure that restricted information is only available to authorize staff, and to prevent misuse of confidential data or the accidental dissemination of restricted information.

## 6. IT Implementations to manage confidential data

The following IT implementations solutions were developed to manage the confidentiality scenarios described in Section 3 Levels of Confidentiality:

### 6.1 Access level restrictions

Depending on the sensitivity level of the data set and the agreements with national authorities, database access is restricted only to authorized staff. In the case of

extremely sensitive datasets, dissemination spaces are hard coded to receive only authorized data series to prevent accidental dissemination of restricted information.

In the case where national data is not restricted but there is a need for IMF staff to estimate data observations due to non-availability of data, the working database with estimates has restricted access for internal use only and the reported database is available without restrictions. Dissemination for these data sets will be from the reported database for country level data and the working databases for global and regional aggregates.

**6.2 Global and regional aggregate data cell suppressions**

Data for individual country reported data are protected from disclosure in published totals by using a threshold rule and a dominance rule.

Totals will not be disseminated if the sample of the data reporters for the particular concept is very small. Likewise, totals will not be disseminated if the largest reporters for a particular concept amount to a very large percentage of the calculated total. For example, if the largest reporter has 80 percent of the total or the largest three reporter's amount to 90 percent of the calculated total, these totals will not be disseminated.

**6.3 Primary data cell suppressions**

According to national statistical disclosure policies and procedures, data cells deemed confidential are provided to STA as null values in the data report forms or a "C" notation will be entered in the report forms to denote the suppression of national level data. Accordingly a null value will be entered in the database with a "flag" to denote confidential data.

In rare occurrences, as instructed by national authorities, STA would suppress previously reported data observations due to confidentiality.

**6.4. Secondary Suppressions or Omissions to protect confidential data from residual disclosure**

To reduce the risk of residual disclosure of primary suppressions by deductions from available information, it may be necessary to omit or suppress secondary data observations. These secondary suppressions are done mainly by using templates to suppress entire concepts based on historical reporting patterns, or by using automated routines to analyze the impact of a suppressed data observation and suppress all cells that have a risk of residual disclosure of a primary suppression.

**6.4.1. Reported data which has established patterns of suppressing entire data series**

In the case where there is an established pattern of entire data series being deleted by national authorities to protect confidentiality, the practice is to use country specific data templates to omit loading data series at the same hierarchy as the suppressed series. This practice preserves the next higher level aggregate while eliminating the possibility of deducing confidential data series by subtracting other available data observations from reported aggregates.

**6.4.2. Reported data which has no established data suppression patterns, and ad-hoc data cells are suppressed due to confidentiality**

When there are no established patterns on data suppression methods by national authorities and data suppressions are done at the individual cell level, it is difficult to manually assess the impact of these suppressed cells and the extent of the need for secondary suppressions. For ad-hoc primary data suppressions, a two stage process using automated data deleting capabilities within the data management system combined with additional data validation procedures and re-edits are used.

Within the STA database management system called DMXPlus, in addition to direct data deleting options, there are two additional delete functionalities used to prevent recalculation of primary suppressions from remaining information.

These functions are called “D+” and “D*.” A data cell with D+ will delete all data observations that are calculated (resultants that use the original data suppression as a component. This is done through the analysis of the equation graph (the resultant tree) of the dataset. The second component “D*” will delete all data observations that have consolidated (Months to Quarter to Annual) data observations that have used the primary data suppression. The system will analyze whether a reported cell has calculated a lower frequency value (e.g., A March value has calculated a first quarter data observation based on the consolidation rules of the system), and delete all such consolidated values which were calculated from the original deleted observation.

Hence the combination of “D+” and “D*” which is the “D+*” entry in a data cell will delete all calculated and consolidated values derived from a data observation, and reduce the residual disclosure risk of the primary suppressed value. While this process eliminates the residual disclosure issues, data utility will be affected because large sets of data will be deleted. Hence a two step process is implemented to improve data utility. An additional validation process that compares the original reported dataset and the edited data set (after the use of D+, D* or D+*) will highlight the secondary deletions and will provide the opportunity to re-enter data observations at higher aggregates that do not result in residual disclosure.

Since the data deletes and the subsequent validation and re-edits are not fully integrated, currently these procedures are costly in terms of data processing resources.

## 7. Challenges

In the context of the current global financial situation, there is growing demand for full disclosure of financial statistical data. With national authorities having specific statistical disclosure control policies, it is difficult to coordinate a concerted effort to release sensitive information.

A major challenge is to develop and maintain IT and operational procedures that guarantee the protection of confidential data while disseminating maximum available granular information. Although automated routines could be developed for large scale secondary suppressions, it may take additional review processes to assess the impact of the deletes and to provide opportunity for re-edits to optimize the data utility.

For system upgrades and testing, at times, it is necessary for IT professionals who are not authorized to handle sensitive information to have access to restricted databases. Another layer of data protection procedures, which include data perturbation for the entire database, is required to protect restricted data during system upgrades.

It is also difficult to manage the expectations of the data user community due to the specific nature of national policies on the release of data previously identified as confidential. Also, when authorities suppress previously released numbers due to changes in confidentiality assessments, data users can recover these suppressed values from previous data releases.

## 8. Conclusion

Proper management and access controls of restricted information, and the protection of confidential data from residual disclosure, are major concerns for the IMF Statistics Department. In order to address these confidentiality concerns, IT solutions and operational procedures have been developed to disseminate analytically useful data while mitigating disclosure risks, to fulfill the IMF's obligations to national authorities, and to comply with internal disclosure control policies. The practical solutions discussed in this document are in place to manage sensitive data in STA's ongoing efforts to disseminate as much information as possible in a timely manner.