WP.8
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Workshop on the Common Metadata Framework
(Vienna, Austria, 4-6 July 2007)

Topic 3: Organizational, managerial, and cultural issues of metainformation systems.

IMF APPROACH TO STORING METADATA WITH MACROECONOMIC STATISTICS

International Monetary Fund[1]

1. This paper discusses the IMF approach to storing metadata with macroeconomic statistics. The paper summarizes the IMF's metadata work in the context of the Dissemination Standards Bulletin Board (DSBB), IMF.Stat, MetaStore, SchemaLogic, and SDMX. It also reviews issues of institutional culture, organization, management, and next steps.

I. SUMMARY OF STATISTICS DEPARTMENT'S METADATA ENVIRONMENT

2. First, we summarize the IMF Statistics Department's metadata work in the context of A. the DSBB, B. IMF.Stat, MetaStore, and SchemaLogic, and C. efforts to align with SDMX.

A. Metadata work in the context of the DSBB

3. Metadata have emerged as a crucial factor in adding value to macroeconomic data. Two bodies of metadata are currently available in the International Monetary Fund (IMF). The first, and the one addressed in this section, is that pertaining to the IMF's Special Data Dissemination Standard (SDDS) and the General Data Dissemination System (GDDS). These metadata sources are extensive and well managed. The second body of metadata—the metadata used in the data warehouse environment—will be discussed in Section B, under Dotstat and MetaStore.

[1] Prepared by Margaret Ann Salmon, MSalmon@imf.org

4. The SDDS was established in 1996 and the GDDS, in 1997. To promote data transparency, a key feature of the IMF's SDDS and GDDS is to provide information on methods and practices countries use to compile and disseminate economic data. The IMF disseminates these metadata of SDDS subscribers and GDDS participants on its Dissemination Standards Bulletin Board (DSBB), which is accessible to the public at the IMF's external website (dsbb.imf.org).

5. Under the SDDS, subscribing countries are to provide information on their statistical practices (metadata) on 21 data categories covering four sectors of the economy (real sector, fiscal sector, financial sector, and external sector). In addition, subscribers are to certify the accuracy of their metadata on a quarterly basis and to update the metadata to reflect changes in their statistical practices. Countries participating in the GDDS are required to provide metadata pertaining to statistical practices for the economic, financial, and selected socio-demographic data that they produce.

6. The SDDS prescribes good practices for (1) data coverage, periodicity, and timeliness, (2) the public's access to the data, (3) the integrity of the disseminated data, and (4) the quality of the disseminated data. The metadata are to describe practices pertaining to these four dimensions of data compilation and dissemination. The GDDS is a framework that guides countries in developing sound statistical systems as the basis for disseminating data to the public. The GDDS addresses the full range of issues critical to compiling and disseminating data and making explicit plans for developments to align national procedures with best practices. As in the SDDS, four dimensions of data compilation and dissemination are covered in the GDDS framework: (1) data coverage, periodicity, and timeliness; (2) the public's access to the data; (3) the integrity of the disseminated data; and (4) the quality of the disseminated data. GDDS countries' metadata are to describe existing statistical practices, as well as plans for improvements.

7. Currently the DSBB posts metadata on 64 SDDS subscribing countries and 88 GDDS participating countries. The metadata of each country cover general information on its legal and institutional frameworks for statistical production and dissemination, as well as detailed information on coverage, concepts, and methods used to compile and disseminate the various categories of data covered in the SDDS and the GDDS.

8. The metadata shed light on the accuracy, timeliness, and comprehensiveness of the data, allowing users to assess the reliability and limitations of the data. The metadata, when measured relative to international best practices, also guide countries in improving their statistical concepts and methods to enhance data quality.

9. In addition to disseminating the metadata, the DSBB provides hyperlinks to SDDS countries' websites (and those of GDDS countries, if available) where the various categories of data are posted.

10. More recently, the IMF Executive Board has supported the integration of the SDDS and the GDDS into the Data Quality Assessment Framework (DQAF) to sharpen the focus on data quality assessment and promote good statistical practices. The DQAF provides a structure for assessing countries' existing statistical practices against best practices, including internationally accepted methodologies. The DQAF is rooted in the UN Fundamental Principles of Official Statistics and grew out of the SDDS and the GDDS. The DQAF incorporates good practices of the SDDS and the GDDS. The Executive Board formally approved the establishment of the DQAF in July 2003 for use in the IMF's work on the data module for the Report on Observance of Standards and Codes (ROSC). The alignment of SDDS and the GDDS with the DQAF integrates the IMF's work on ROSC, technical assistance, and the SDDS and GDDS, streamlining the IMF's support of its various data quality initiatives.

11. In this connection, the IMF has been working on presenting SDDS and GDDS metadata under the DQAF structure. The DQAF is structured in six dimensions; for each dimension, the DQAF identifies 3-5 elements of good practice; and for each element, several relevant indicators.

B. Metadata work in the context of IMF.Stat and MetaStore (collaboration with OECD on MetaStore)

12. The IMF commenced a project in 2007 to develop a data warehouse to store macroeconomic statistics used in the Fund. Rather than purchase a commercial software product or commit to in-house development, the IMF chose to enter into a collaborative arrangement with the OECD to use their data warehouse (Dotstat) and metadata management (MetaStore) software. This collaborative arrangement, set out in a Memorandum of Understanding (MOU), is expected to provide a mutually beneficial development path for both organizations.

13. While the IMF utilizes OECD software components, it chose a slightly different set of design parameters with which to begin developing its data warehouse environment. Specifically, the project calls for

    - a single data set,
    - common dimensions,
    - use of authoritative sources, and,
    - provision of rich, revealing metadata attached to warehouse data.

14. These design parameters address the Fund's concern with migrating from a siloed environment to an enterprise environment, providing an opportunity to compare and contrast data, and significantly enhancing the data's context using relevant metadata. The Fund will also evaluate the SchemaLogic suite of tools.

Dotstat

15. Dotstat, known as IMF.stat in the Fund, is the OECD data warehouse product. To create a data environment that encourages comparison and contrast across time series, IMF.Stat's physical data model uses six common dimensions to describe the data:

    - *Country and Group* – an authoritative set of country names and geographic and analytical groupings of countries.
    - *Economic Concept* – listing of economic concepts using the Catalog of Economic Time Series as the authoritative form.
    - *Data Source* – authoritative listing of the data providers.
    - *Unit* – currently treated as an aggregate of unit of measure with scale, e.g., millions of U.S. dollars.
    - *Status* – an authoritative listing of status values.
    - *Time, Frequency* – authoritative lists of values related to observation intervals.

16. Where possible, the intention is to store all data in a single dataset, using these common dimensions to facilitate comparison of data across Data Source and other dimensions. Use of a single dataset implies the need for standardization of structural metadata.

17. A feature of Dotstat is that it provides the ability to store referential metadata with the data. This is seen as a major benefit as it allows the user to easily access information to help explain context relevant to the observations.

MetaStore

18. The metadata that provide critical context to the warehouse data are stored, managed, and sourced from MetaStore. It will store both structural and referential metadata as defined by SDMX.

19. The project considers metadata to be an integral part of the project and places significant value on issues related to the identification, creation, and management of metadata associated with the time series data.

20. In our introduction, we mentioned two bodies of metadata currently available in the IMF. The first, the Dissemination Standards Bulletin Board (DSBB) metadata, was reviewed in an earlier section. These metadata sources are extensive and well managed, but not linked directly to data. The second body of metadata is addressed in this section. It can be described as widely dispersed, individually managed, and not organized within any central scheme. These metadata sources are in the process of being identified and reviewed so they can be standardized where possible and stored centrally in MetaStore. The metadata are

subsequently linked directly to relevant data in the warehouse. Gaps in metadata coverage are being identified, and where necessary, referential metadata are being developed and stored in MetaStore.

21. There is a connection between MetaStore and the DSBB. Currently that connection is in the form of hyperlinks to metadata in the DSBB.

22. MetaStore can be characterized as

- the central repository of both structural and referential metadata,
- the source of all metadata used in IMF.stat
- the store of our structural metadata (authoritative source of metadata for common dimensions),
- the store of mappings that relate production metadata to the authoritative source,
- an inclusion of the forty-two (42) categories or types used by the OECD,
- a use of SDMX Metadata Common Vocabulary (MCV) definitions, where possible, and
- an ability to attach metadata to multiple levels of data (data set, concept, data series, individual observation).

23. In addition to being used as the source of all metadata for IMF.stat, it is expected that MetaStore will be used as a source of metadata for other systems. It will be released as a tool for users of data to access referential metadata, structural metadata, or authoritative lists as well as mappings, which will reveal original structural information for all authoritative sources.

SchemaLogic

24. The IMF is evaluating SchemaLogic Enterprise Suite, an enterprise tool for managing metadata and business vocabulary at an infrastructure layer. The first application of the tool is likely to be management of metadata associated with SharePoint 2007, which is being introduced as an enterprise collaboration tool. With rigorous testing, it is possible SchemaLogic may prove to be a solution for managing structural metadata and authoritative vocabularies, although the project is likely to continue to use MetaStore for referential metadata. SchemaLogic also has tools for allowing data stewards to manage their domains via consensus, using automated tools and notification alerts.

C. Efforts to align with SDMX

25. In the interest of interoperability, the IMF is keen to ensure standard lists or authoritative sources of structural metadata are used wherever possible. This supports the project’s aim to expose as many series as possible in a standard and comparable way, facilitating comparison of data across sources.

26. We want to ensure our authoritative lists incorporate any known standards including those which will emerge from the SDMX project. This will be an evolving effort as we keep abreast of this work, including work currently in progress to create standards for the balance of payments statistics (BOP). As these standards emerge, they will be incorporated into MetaStore and ultimately IMF.stat. Additionally, we will make existing standards available to SDMX and other projects.

II. CULTURAL ISSUES: INTRODUCING THE IDEA OF MANAGING METADATA WITHIN THE IMF

27. The IMF's data warehouse project highlights a number of areas of transition for the institution. IMF uses data from many sources, only some of which have metadata. While each domain has long collected data in its own areas of expertise, there has not been an overarching effort to bring all the data together and examine it within a shared framework (common dimensions) and with the benefit of detailed context that informs the observation (metadata). There have been previous efforts to manage data and metadata, but the recent emergence of the importance of metadata has intensified the Fund's effort to leverage this resource and to make it available. In the past, when metadata were available, it was as a static resource in hardcopy. The warehouse effort hopes to make metadata a dynamic digital resource.

28. The Fund has experienced a range of issues with its metadata, including the following:

- Other sources of metadata available but often not well managed nor directly linked to data (e.g. published in books, on web sites, in production systems).
- No single repository or point of control.
- No standardized versions, no way to link or harmonize like terms.
- No way to indicate one like term over another (preferred terms) or
- No way to direct user queries to related terms that might also be of interest.
- Multiple versions of classifications across the Fund.
- Similar terms/labels with different definitions (homonyms).
- Terms/labels with different names that actually mean the same thing (synonyms).
- No data stewardship to monitor use of terms/labels.
- Terms/labels that represent aggregates, but cannot be reverse engineered. (The user has no way of knowing what made up the aggregate.)
- Different versions or forms of authoritative lists used by different departments.

29. The Fund is addressing its data/metadata management issues by the following:

- Working with data providers and subject matter experts (SMEs) to locate all relevant sources of metadata.
- Working with data providers, SMEs, and information managers to identify potential warehouse content and analyze it.
- Creating central repositories for data and metadata.
- Reconciling or harmonizing terms/labels that mean the same thing and mapping to a preferred term.
- Gathering, cataloging, and managing authoritative lists identified by data providers and SMEs.
- Working closely with the Fund's information architects in the Information Services Division (ISD) to ensure information management best practice.
- Collecting, recording, and managing classification schemes.
- Assigning data stewards, defining and agreeing on their roles and responsibilities, and giving them tools to automate their work.

III. ORGANIZATIONAL ISSUES: DETERMINING AUTHORITATIVE STRUCTURAL METADATA IN IMF.STAT, IDENTIFYING REFERENTIAL METADATA, AND BUILDING LINKS TO THE DSBB

A. Authoritative sources

30. To facilitate one of the goals of the warehouse—the ability to view data across sources and dimensions—a common framework was required for examining the data. The production data sources were analyzed to determine the structure and vocabulary used to describe time series data. Along with a review of relevant standards, guidelines, and other data models in use internationally, the content analysis produced a set of common dimensions. The dimensions are common in the sense they represent shared ways the data was collected, stored, shared, and disseminated.

31. Because the production sources do not share a common organization framework or a common vocabulary, it was necessary to reconcile or harmonize the various terms to a preferred term designed for use in the data warehouse. The resulting mapping of production terms to a preferred term (common dimension) created semantic consistency for the data warehouse without changing the terms used by the production sources. It was equally important that wherever an authoritative set of terms existed to define a common dimension, those terms would be presented as an authoritative list or a controlled vocabulary, so only authorized terms would be used to generate values for the common dimensions. These authoritative lists are presented to the user as picklist, drop down lists, etc. Both the authoritative lists for the common dimensions as well as the mappings are stored in MetaStore.

B. Referential metadata

32. Referential metadata to be used to add value to data stored in the warehouse are being gathered from many sources and stored in MetaStore. While many datasets have metadata available, they have often been stored separately from the data they are intended to describe, and often in the form of paper publications. A time-consuming process is in place to identify the metadata and, in some cases, revise them to a form more suitable for digital environments. A key feature of MetaStore is the ability to link metadata to any level of the dataset. While this is seen as an important feature that is being implemented, it is necessary to carefully consider at which level the metadata are most useful. Metadata linked to many individual observations may prove to be repetitive and distracting.

C. Building links to the DSBB

33. As discussed earlier, there is a large body of very useful metadata stored in the DSBB. While it would be technically feasible to move these metadata to MetaStore, the current plan is to leave it where it is and where it adds value to the data in the warehouse, include links. MetaStore, and by definition IMF.stat, currently include links to the Advance Release Calendar, SDDS, and GDDS in referential metadata linked to various levels of the data. For instance, SDDS and GDDS links have been included for the relevant countries in the Country and Group dimension.

IV. MANAGERIAL ISSUES: ESTABLISHING GOVERNANCE FOR METADATA ACROSS IMF DEPARTMENTS AND ASSIGNING ROLES AND RESPONSIBILITIES

34. It has been recognized that if the Fund is to achieve the benefits expected from identifying, storing, and managing its metadata in a more organized way, we need to ensure that suitable governance arrangements are in place. Work is in progress to establish groups or individuals who will have ownership or other responsibilities for metadata.

35. The Economic Data Advisory Group (EDAG) is comprised of representatives from many departments across the Fund with an interest in data and metadata. Several working groups have been established to focus on certain areas of interest such as the data warehouse and data management practices.

36. The Information Services Division (ISD) staff members are key stakeholders in the project, responsible for provision of metadata. They analyze data provider's content and related referential metadata, identify all sources of metadata, and identify authoritative sources and preferred terms. They map production dimensions and values to the common dimensions and predefined values, where applicable. They work

with project team members on information management best practices as well as relevant standards and guidelines.

37. Organizational arrangements have been changed to create new areas responsible for data management as well as a small group devoted to addressing key metadata issues. Initially this group is focusing on issues specifically related to data and metadata being loaded to IMF.stat.

38. There are still many issues to be addressed including how the many versions of classifications and lists will be identified, who should be the owner, how to determine when a new value should be added, or whether it should be mapped to an existing value. There will need to be decisions made about how decentralized the management of metadata should be and how to identify the appropriate level of resources to establish an agreed framework for metadata and for its ongoing maintenance.

V. NEXT STEPS: SOME ISSUES

39. Work has commenced to define and agree on roles and responsibilities. This will include identification of work practices that will need to change, as well as new processes designed to ensure all metadata are well managed.

40. Identify a data steward for each common dimension with the authority and responsibility to enforce standards.

41. For structural metadata, ensure standardization where possible, mapping to an authoritative source and utilizing standards and guidelines. This will include following sound information management practices, (for example, normalize data, separate content and presentation, design for agility, etc.). It will also mean keeping in touch with work on other standards as they evolve such as those being developed for SDMX.

42. As new systems are developed, ensure they use standards consistent with the warehouse environment, including use of existing metadata where possible.

43. Raise awareness across the Fund of the value of quality referential metadata attached to data (manage metadata with the same care you would any other asset).

44. Ultimately tie together basic schemas (business vocabularies for structured and unstructured data), so that when users look for relevant content, they get a comprehensive overview of relevant content, whether that content is data from a warehouse, a policy analysis, or a mission report.