

Case study: United Nations Industrial Development Organization (UNIDO)

Case study: United Nations Industrial Development Organization (UNIDO) United Nations Industrial Development Organization (UNIDO): use of GSBPM

| | |
|-----------------|---------------------|
| Contact person* | Valentin Todorov |
| Job title | |
| Email | v.todorov@unido.org |
| Telephone | |

Summary*

Metadata strategy

Please note this paper is based on a 2010 METIS paper:

Organization Details

1. UNIDO was set up in 1966 and became a specialized agency of the United Nations in 1985. As part of the United Nations common system, UNIDO has responsibility for promoting industrialization throughout the developing world, in cooperation with its 172 Member States. Its headquarters are in Vienna, and it is represented in 35 developing countries. This representation and a number of specialised field offices, for investment and technology promotion and other specific aspects of its work, give UNIDO an active presence in the field.
2. The Service Module "[Industrial Governance and Statistics](#) " assists developing countries and economies in transition to **monitor, benchmark and analyse their industrial performance and capabilities**, and on that basis to **formulate, implement and monitor strategies, policies and programmes** to improve the contribution of industry to productivity growth and the achievement of the [UN Millennium Development Goals \(MDGs\)](#).
3. Building capabilities in industrial statistics: UNIDO provides technical assistance to introduce best practice methodologies and software systems to monitor and assess productivity performance and use it as a guide for policy-making; and enhance the quality and consistency of the [industrial statistics databases](#) so as to provide meaningful inputs for assessing productivity and industrial performance.
4. Most developing countries lack the capacity to systematically gauge productivity performance in such a way as to be able to use it as an effective tool for policy and decision-making in general. Through this service module UNIDO aids developing countries and countries with economies in transition to build capacities in this field by providing technical assistance to:
 - Introduce best practice methodologies and software systems to monitor and assess productivity performance and use it as a guide for policy-making; and
 - Enhance the quality and consistency of the [industrial statistics databases](#) so as to provide meaningful inputs for assessing productivity and industrial performance.

Figure 1: UNIDO Organization Chart: Statistical activities are carried out by the Statistical Unit of the Research and Statistics Branch PCF /RST/STA

5. Statistical activities of UNIDO are carried out by the Research and Statistics Branch.

Overall strategy and metadata management principles:

6. The conceptual development of the UNIDO metadata subsystem was initiated in 1999 with the aim of automation of information production (data and metadata) using latest management technology. Having in mind the inherent structural complexity of the data bodies involved, only a comprehensive metadata-based system re-design approach has been considered promising at all. Thus, the project favoured an **integrated data and data documentation (metadata) framework** emphasizing that, while allowing scrutiny of data documentation (statistical metadata) both individually and jointly with statistical data, any statistical data access always entails the retrieval of associated metadata without demanding specific inquiry measures or actions. This way a rather tight interrelation of data and metadata is both enforced and assured by purely technical means. However, as its major precondition, this principle presupposes a homogenous representation of all pieces of data documentation in order to enable uniform data and documentation access procedures.

7. Moreover, as a change in data representation must not disrupt established UNIDO data services, a **smooth migration policy** is called for, leaving interface requirements of downstream systems and data usage almost untouched. Implying such a great effort to UNIDO, an expected side-benefit of redesigning the INDSTAT system is its potential applicability to further operational data management areas in need of refashion.

8. The concrete design and implementation of this subsystem was realized as a part of an integrated data and metadata system under the name **Integrated Statistical Development Environment (ISDE)**. The development of ISDE was performed in a stepwise manner in the context of a migration project of the complete UNIDO statistical databases from an IBM mainframe to a client/server platform. Further in this document will be given details about the migration project itself, its current status and its relation to the newly developed statistical applications and ICT infrastructure.

9. The basic metadata management principles are backed by the **UNIDO Quality Assurance Framework** which is targeted to ensure that the statistical activities of UNIDO are relevant and the data compiled and disseminated are accurate, complete within the defined scope and coverage, timely, comparable in terms of internationally recommended methods and classification standards and internally coherent to variables included in the datasets. While these generally accepted, broad dimensions of quality of statistical data may be defined in each NSO's own quality assurance framework, UNIDO makes maximum effort that data produced from the statistical operation undertaken with the UNIDO technical cooperation are accurate, internationally comparable and coherent. UNIDO has been a forerunner among international organizations in using a Statistical Metadata System as a tool for observing and evaluation the quality of statistical data, especially for completeness and cross-country comparability. For further details see Yamada (2004) and Upadhyaya (2008).

10. Following the **International recommendations for Industrial Statistics**, the development of metadata is given a high priority and their dissemination is considered an integral part of dissemination of industrial statistics. Moreover, it is recommended that in consideration of the integrated approach to compilation of economic statistics development of a coherent system and a structured approach to metadata across all areas of economic statistics be adopted, focusing on improving their quantity and coverage. Further, the dissemination of statistical data and metadata using web technology and SDMX standards is recommended as a way to reduce the international reporting burden (the Statistical Data and Metadata Exchange (SDMX) technical standards and content-oriented guidelines provide common formats and nomenclatures for exchange and sharing of statistical data and metadata using modern technology).

11. An essential requirement was that the metadata is available in three languages (English, French and Spanish). This allows to pre-fill each questionnaire in the preferred language for the country and then to process it accordingly.

12. The integrated system is based on a formal framework, described in detail in Froeschl et al. (2002), Froeschl and Yamada (2000). The proposed information system architecture comprises two cubes, one for statistical data and another for the metadata interrelated by a set of shared dimensions

Current situation

Metadata Classification

25. The metadata is classified according to their usage and their role in the statistical production process.

The main types of metadata according to this criteria are as follows:

- **Definitional metadata** - The definitional metadata refer to metadata that act as identifiers and descriptors of the data. They are prior to the data, are created and maintained independently from the data and are used to define the data structure. Examples of definitional metadata are country names and codes, currency names and codes and their relation to the countries, definitions of the indicators, classifications like ISIC Rev. 2, ISIC Rev. 3, etc. Through these core data are defined also some basic metadata elements like metadata classes, stages, sources and methods, etc. Historically this metadata type was the first to be established (ported from the Mainframe, re-factored and formalized) in ISDE. The definitional data are maintained by the statistical staff using the tool Nomenclature Explorer (NE) following strictly the user authorisation and ownership.
- **Implicit metadata** - The implicit metadata are a special class of metadata arising throughout the specific usage of other metadata. Typical example are the ISIC combinations. For example several industry categories can be combined and reported together by a given country for a given indicator and years. In the questionnaire returned by the NSOs such a combination is expressed in the following way (see - *Figure 6*):

The codes 1511, 1512 and 1513 are combined and reported as a single number '1234'. The combined industries are linked by the footnote a/. This is resolved by the system as a dummy ISIC code 1511A defined as "1511 includes 1512 and 1513" which is used throughout the production process and appears accordingly in the publications as well as in the pre-filled Questionnaire. In a similar way can be solved other country specific classification discrepancies like industry codes at 3-digit level that exclude one or more specific 4-digit industry codes. The implicit metadata can be used also for defining of synonyms - for example '040' is the country code of Austria and this is the same as, i.e. substituted by the ISIC code 'AUT'. Or for specifying of aggregation e.g. the aggregation code 'EU' is composed by the codes of the single countries. The keywords substitute, included, excluded used in the above described context are called operators.

- **Operational Metadata** - The operational metadata are generated by the process of data transformation and attributed to the respective data items. As described in the presentation of the Data Transformation phase, each data item is stored in the database with a stage indicator reflecting its credibility. Also the transformation process generates "Source" and "Methods" metadata, describing the source of the data item and methods applied for its generation.
- **System metadata** - these metadata are used to drive automated processing throughout the phases of the life cycle. These can be layout definitions for the yearbook (for each country, for each edition of the yearbook) as well as country lists, etc., used in the automatic generation

of the PDF output; Installation and packaging lists, directories, templates, etc. for creation of the CD product. These metadata are specific for the application where they are used and do not relate to the data, therefore, although stored in the centralized repository, are maintained by each application separately and are called "Properties" of the respective process, i.e. Yearbook properties, Questionnaire properties, etc.

- **Descriptive and Methodological metadata** - these form the main bulk of metadata. They are received from the primary data reporters, using the UNIDO Questionnaire and then are further processed together with the data. During this processing additional metadata can be added by the UNIDO statistical staff. Descriptive or methodological metadata can be attached to all possible levels ranging from the complete data set down to individual data items. This is done by assigning to the metadata same dimensions as those of the data.

Metadata system(s)

18. The metadata system is a part of the Integrated Statistical Development Environment, provides end-to-end metadata services throughout the statistical production process and was developed in the context of the migration from Mainframe to a Client Server Platform. Figure 2 presents the overall structure of ISDE and its relation to the statistical production life cycle. The client part of the system is presented to the user as a desktop application, the **ISDE shell** that serves as a container for the rest client/side applications. These applications are described briefly below.

- **ADMIN** - provides administrative services, like user and authorisation management, logging and auditing of the system, backup and restore management;
- **Nomenclature Explorer** is the tool for maintenance of the core definitional metadata, which is not related to particular data items but rather serve for defining the structure of the data and metadata. These first two applications are outside of the life cycle. **Figure 3** and **Figure 4** show examples of the ISDE shell;
- **Questionnaire** is the application for management of the pre-filling and distributing of the questionnaires to the member countries (i.e. used in the *Initialisation* phase);
- **Data Wizard** is the main data and metadata maintenance tool used in the Data Collection and Transformation phases of the life cycle. It provides services for:
 - i. Reading in the data and metadata from the returned back Excel questionnaire
 - ii. Initial validation of the read in data and storing in the database (at stage 1)
 - iii. Maintenance of the metadata
 - iv. Screening
 - v. Aggregation and further data validations and transformations
- **Presentation Wizard** is mainly a visualization tool which can be used in the Dissemination phase for answering ad hoc requests, but because of its versatile functionality it finds a wide usage also in the *Data Transformation* phase
- **Publication applications** - these are the applications used in the Dissemination phase for generating the different publication products
 - i. **Yearbook** - this is a complex set of applications for production of the Industrial Statistics yearbook including aggregation, layout, PDF file generation according to pre-defined templates and other tools. The final result is a publication ready PDF file of about 700 pages;
 - ii. **INDSTAT CD** - used to produce the INDSTAT type of CD products;
 - iii. **ISDB CD** - used to produce the INDSTAT type of CD products;
 - iv. **WEB** - used to generate the necessary data and metadata for updating the WEB dissemination database (this database is outside of the ISDE system, managed by the computer section);
- **Other applications** - in this category are included any other applications used in the process, like SAS, R, tools for compilation of Production index numbers and National Accounts data (which are outside of the scope of this document) and others.

19. As already mentioned the ISDE was developed in the context of migration from Mainframe to a Client/Server platform. For the migration a stepwise approach was chosen because of the following reasons:

- The project was not urgent, since the discontinuation of the mainframe was postponed because of other important services still running on it
- The software test and sustaining of the created system has to be done in-house
- Only limited resources were available
- The staff was very willing to participate in the project
- The goal was not only to migrate the system but rather to develop a completely new one and the requirements were not yet completely specified (because of the limited resources)
- A key requirement was that the established UNIDO data services must not be disrupted

20. The first step was a rigorous analysis of the existing system and development of a data model which was as generic as possible in order to be able to accommodate any changes. Based on this model a loader application was developed which allowed in any moment to synchronize the data in mainframe and in the Sybase database of the new Client/Server system. The development of the new metadata subsystem was initiated by implementing a tool for maintenance of the definitional metadata [2005-2006]. Thus a kind of proof of concept was successfully completed.

21. A capture/maintenance tool for reference metadata was developed and the description/methodological metadata, which existed so far in the form of Word documents or Excel worksheets, were entered into the system. The mainframe footnote database (data-item level metadata) was imported too. Thus the complete process of maintenance of the available metadata was migrated to the Client/Server platform

22. In the next step the data dissemination applications were developed which allowed to produce the recurrent statistical publications/products from the mainframe system and from the Client/Server platform in parallel which was an ideal acceptance test for the new applications by just comparing the results [Q4-2006 - Q4-2007].

23. As an example of the migration-to-new development relation can be noticed that while the International Yearbook of Industrial Statistics was produced from the main frame as a cameraready line printer output which was glued together with many MS Word and MS Excel documents, the output of the Client/Server system was an automatically generated page numbered PDF file of about 700 pages.

24. In the third step the pre-filling of the questionnaire was implemented using the new Client/Server data- and metadata-base [Q1-2007]. The data capturing as well as the data maintenance tools were developed and are now in the phase of final testing. The questionnaires, which are expected to start arriving in June, will be entered only in the Client/Server system. This will be the ultimate decoupling of the new system from the mainframe.

44. The overall structure of the Integrated Statistical Development Environment is presented in Figure 2. The system utilizes a 3-tier architecture build on .Net technology. The data and metadata are stored in centralized database, and the user interacts with the system through the ISDE shell which is a desktop application serving as a container for the other ISDE applications. The commonality of the system is achieved through using shareable component libraries.

45. The development of the entire Integrated Statistical Development Environment has been carried out inhouse, taking international standards (ISO /IEC 11179) into consideration.

Database layer

46. The database consists of two identical but physically separated databases - a test and production databases - running on Sybase ASE RDBMS under Linux. A sample of the data model is shown in Figure 14 and the complete data model is presented in an attached Erwin diagram.

47. The access to data and metadata from the client applications is performed through component libraries. These would allow replacing for example the Sybase database by an MS SQL Server or Oracle without any modification of the applications.

Component libraries

48. The object oriented component libraries are developed also in C# and are used to unify many common tasks like database access, file access, printing, access to common data structures, etc.

Client applications

49. The client applications are developed using MS Visual studio in C#. They connect to the database and interact with each other using component libraries developed also in C#.

Other tools

50. Table 1 lists some other tools integrated in the ISDE system.

Costs and Benefits

.

Implementation strategy

.

IT Architecture

.

Metadata Management Tools

.

Standards and formats

.

Version control and revisions

.

Outsourcing versus in-house development

.

Sharing software components of tools

.

Overview of roles and responsibilities

51. No specialized metadata roles are necessary, since the processing of the metadata is tightly coupled with the processing of that data and the responsibilities are organized by country, i.e. each statistical staff member is responsible for a given number of countries throughout the complete statistical production process (of course the assignment of countries to statisticians is metadata itself and is stored and maintained in the same way as the rest metadata). 52. The introduction of the new Client/Server platform including the new metadata system also did not require any new roles related to the metadata since the same people are maintaining the metadata, but using the modern tools instead of the previously existing clumsy methods.

Metadata management team

.

Training and knowledge management

53. No special training for the staff was necessary since all statisticians participated actively in the specification and the development of the system. As already mentioned, the main part of the system testing was performed by parallel runs on the Client/Server and Mainframe (one very important advantage of the stepwise approach) and the found problems and issues were entered into a simple bug tracking system. 54. A one-week SAS-PC training was given to the staff members in order to facilitate the transition from the mainframe to the Client/Server platform.

Partnerships and cooperation

.

Other issues

Lessons learned

| |
|--------|
| Links: |
| |

Attachments