WP. 34
ENGLISH ONLY

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

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (iv): Using metadata for searching and finding statistical data in websites and portals

STATISTICAL METADATA - THE KEY ELEMENT TO IMPROVED DATA QUALITY AND DATA EXCHANGE IN THE FAOSTAT2 PROJECT

Contributed Paper

Submitted by Food and Agriculture Organization of the United Nations, Rome, Italy[1]

I. INTRODUCTION

1. Statistical metadata is seen as one of the key elements to the success of the FAOSTAT2 Project which is currently underway at the Food and Agriculture Organization of the United Nations. The FAOSTAT statistical system is one of FAO's most important corporate systems, being a major component of FAO's information system, contributing to the Organization's strategic objective of collecting, analyzing, interpreting and disseminating information relating to food agriculture and nutrition. FAOSTAT lies at the core of FAO's World Agricultural Information Centre (WAICENT) through which access is given to FAO's vast store of information on agricultural and food topics – statistical data, documents, books, images, and maps.

2. This paper provides an overview of work underway in the FAOSTAT2 Project which will result in a complete revision and restructuring of the FAOSTAT statistical system. Statistical metadata will provide the linkages between the various statistical databases at FAO, to country databases as well as the statistical databases of other International Organizations.

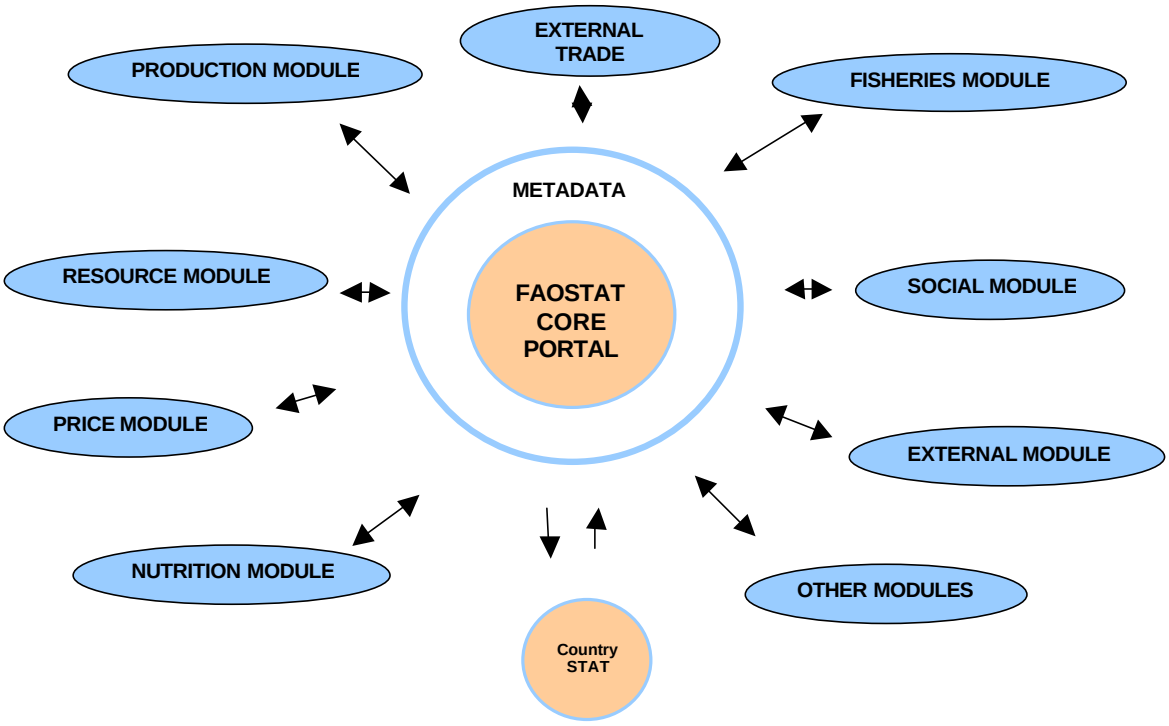
II. The FAOSTAT2 Project and CountrySTAT

A. FAOSTAT2 Project

3. The FAOSTAT2 system (Kasnakoglu and Mayo, 2003) revolves around a core FAOSTAT dataset/portal (see figure 1.) with distributed database modules surrounding the core. This model provides a flexible approach as the satellite databases will have linkages to the core and other modules to enable data interchange. The core portal and other modules will have standard statistical metadata elements to facilitate data interchange. Only selected statistical data will be included in the core FAOSTAT2 portal, the satellite databases will be used to disseminate the detailed statistical data that is not included in the core FAOSTAT2 portal.

[1] Prepared by Robert Mayo, (Robert.Mayo@fao.org) and Sari Jouhki, (Sari.Jouhki@fao.org)

Figure 1. FAOSTAT2 Core portal and Satellite Database System



B. FAOSTAT2 Project statistical metadata related requirements

4. Among the requirements of the FAOSTAT2 statistical system are the following which specifically relate to statistical metadata, data searching and data interchange:

- Improve user access to FAOSTAT data by enhancing and creating new mechanisms for data dissemination, including access to data across domains;
- Enhance data integrity by ensuring that appropriate methodologies and data standards are consistently applied;
- New functionalities will focus primarily on tools to facilitate data capture and improve and monitor data quality: data exchange standards; flexible, yet comprehensive, data editing routines and data estimation tools;
- Sub-national data: The new system (via CountrySTAT) will provide the capacity to store and report on country data that is captured at the sub-national or administrative unit level;
- Geo-referenced data: FAO supports a number of spatial information systems that contain a wealth of information on geographic and climatic conditions impacting food and agriculture production. Integration of FAOSTAT statistical data with the various FAO spatial information systems will enhance FAO's analytical capacity in the food and agriculture arena. To support this requirement, the FAOSTAT2 system will provide the capability to attach a "geo-reference" to data maintained within the system;
- Data from external sources: Data available from external sources is often useful and necessary to assist FAO staff in their analyses. A flexible and easy-to-use mechanism is required to access data from such sources as the International Monetary Fund and World Bank and to merge them with FAOSTAT statistical data;
- Reference data: Data used in supporting calculations and validations, e.g. conversion factors, country codes, editing rules, etc., will be available in and accessible to users of FAOSTAT2.

C. The CountrySTAT sub-project

5. CountrySTAT will be a scaled-down version of FAOSTAT2 available to countries to implement on a modular fashion. CountrySTAT will assist countries in developing a statistical information system containing available data and metadata relevant to agricultural policy, together with data from other extra-national sources, as well as FAO data relevant to the purpose. The major objectives for the CountrySTAT project are:

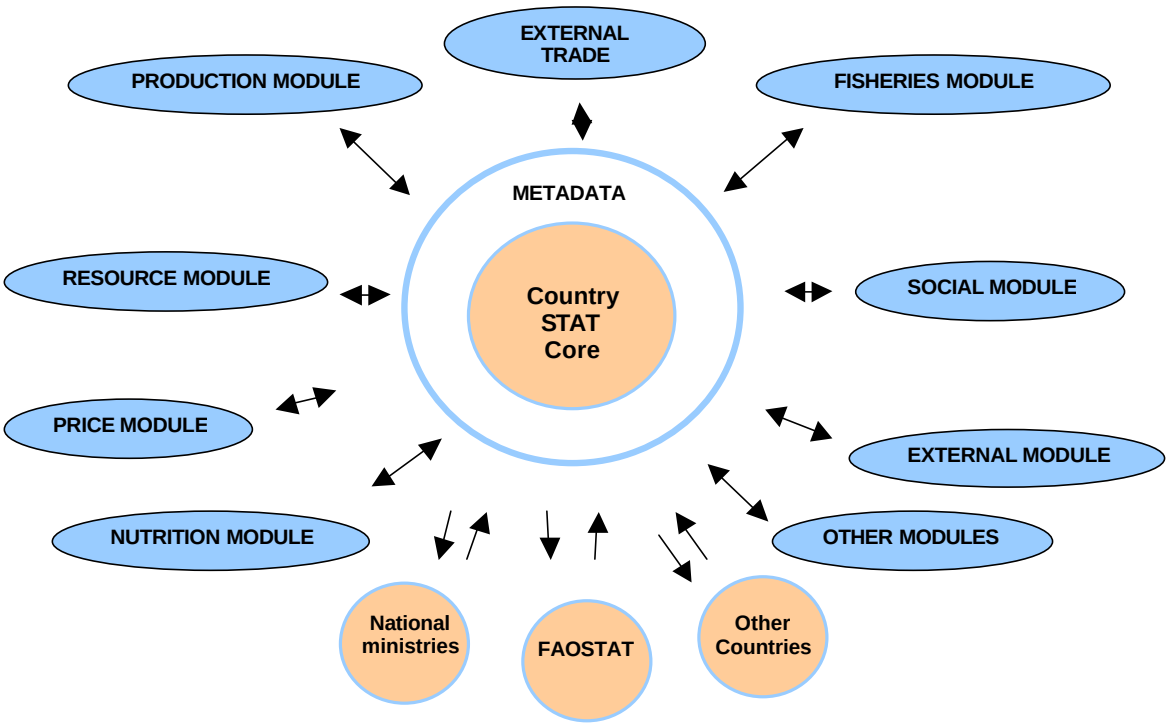
- Capacity building of member countries (coordination, harmonization and enhanced value to data);
- Two-way exchange of data - within countries, between countries and FAO, as well as between countries themselves;
- Facilitate data use by national policy-makers.

6. As with the FAOSTAT2 project, CountrySTAT in addition will have a core set of statistical data series similar to FAOSTAT2 set in a structured framework relevant to the countries specific agricultural situation (see figure 2). The core statistical data series will be surrounded by statistical metadata (classifications, definitions concepts, etc.) which will facilitate the flow of information both within countries, between countries, and between national statistics offices and international statistics offices. The various database satellite modules will be linked via the metadata to the core CountrySTAT.

7. In terms of scope, CountrySTAT will provide a storage, verification, validation, analysis and dissemination system appropriate to a system of food and agricultural statistics at the national level. Thus, it will have the capacity to store sub-national and geo-referenced data. It will also have the basic tools that will be developed for FAOSTAT2 to verify, validate and derive data (food balance sheets/supply utilization accounts).

8. CountrySTAT will be able to integrate within different IT environments of developing countries and to be easily updated, modified, maintained and sustained and provide the linkages to international concepts, definitions, classifications etc., as well as internal agricultural data.

Figure 2. CountrySTAT Core Satellite Database System

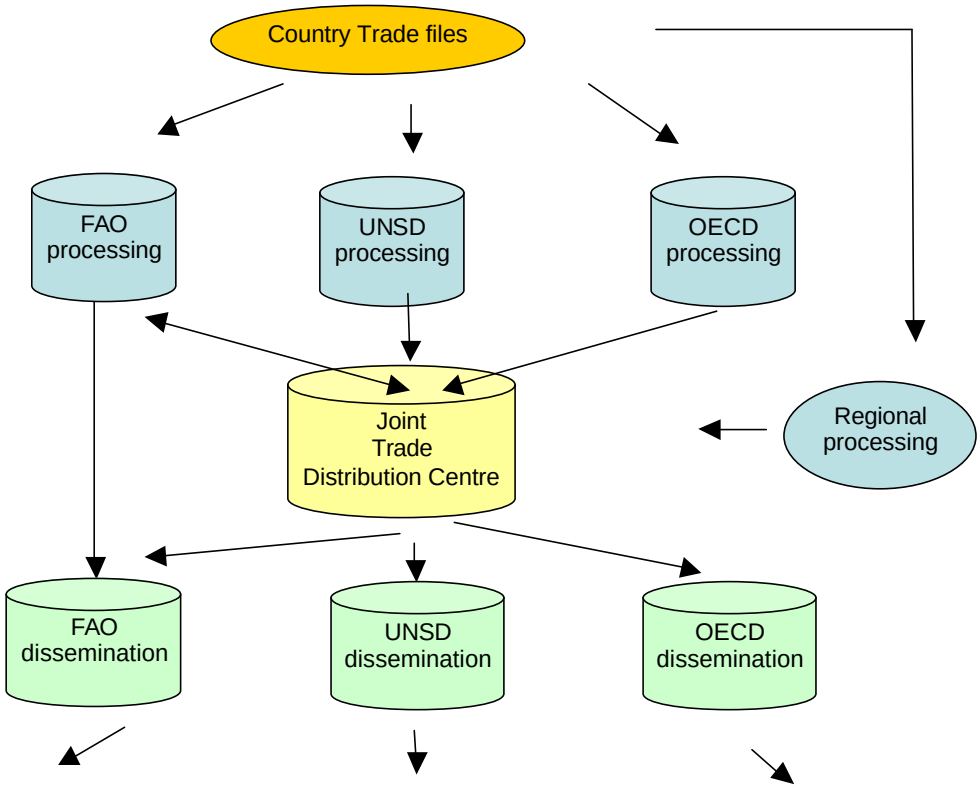


III. Statistical Metadata and FAOSTAT2

A. International Trade Statistics and FAOSTAT2

9. The External Trade Module of FAOSTAT2 is part of a collaborative international trade statistics system. FAO, as well as the OECD and UNSD collect, process and disseminate international merchandise trade statistics. It is essential that relevant statistical metadata (classifications, conversion factors, treatment of confidential data, quantities, adjustment of quantity data, estimation of missing commodity data, validity checks etc). are recorded and exchanged between the entities in order for unambiguous processing and dissemination of data. The proposed data flow is shown below in figure 3. Initial tests using an adapted SDMX methodology will be explored in the Spring of 2004 as a possible data interchange standard. The establishment of standardised statistical metadata between these three international organizations will provide the user with a more efficient access to the disseminated trade statistics via improved searching of the various dissemination database. Fewer problems in understanding differences between datasets will also be a major by-product due to the development of consistent statistical metadata.

Figure 3. Exchange and Flow of International Trade Statistics



B. Metadata models and FAOSTAT2

10. The review of metadata models undertaken in the MetaNet Project – Methodology and Tools (2002) presented some 22 metadata models and noted that they differ in many aspects generally due to the purpose and the amount and nature of the metadata contained. As metadata models are being adopted by international organizations into their statistical systems there needs to be careful evaluation of the needs both within the organization and with external partners.

11. The final metadata models employed in FAOSTAT2 will have to take into account the major metadata domains in the project: CountryStat; FAOSTAT2 Working System; FAOSTAT2 Dissemination system (Core FAOSTAT2 Portal and FAOSTAT2 Data Warehouse). Each of these major metadata domains have

requirements for various levels of metadata which cover the main areas of: General information; Data sources and methodology; Concepts and definitions; Classifications; Statistical footnotes; Dissemination and Accessibility; Contact information; Units of measurement and other symbols and abbreviations.

12. The level of metadata (detail) needed in each of the major domains CountryStat; FAOSTAT2 Working System; FAOSTAT2 Dissemination system will vary. The CountryStat metadata will focus mainly on providing linkages to country based metadata systems and therefore will not have the detail required in FAOSTAT2 Working System. It may not be desirable to develop a model of data/metadata relevant for all parts of the statistical production process and all types of statistics in FAOSTAT2. It may be better to develop data/metadata models specifically based on the underlying data structures.

13. The FAOSTAT2 Project is currently evaluating the various models such as: Dublin Core; ISO/IEC 11179; The Data Documentation Initiative–DDI; IMF SDDS-GDDS for use in FAOSTAT2. These models generally have been developed on a country basis. There adoption and use by an International Statistics Office would in most cases require some restructuring of the model in order to meet the metadata needs of the each of the major FAOSTAT2 domains.

14. The External Trade Module of FAOSTAT2 is a good example of how internal databases are part of a wider statistical process which have linkages both within an International Statistics office and between countries and other International Statistics offices. Metadata models developed for such modules need to take into account these linkages.

C. Data Quality

15. Several mechanisms for ensuring, monitoring and reporting on data quality are envisioned for the FAOSTAT2 system. All data entered into the system will be submitted to a series of rigorous editing and consistency checks. System editing will be performed in real-time to identify data errors early in the process and increase data entry efficiency.

16. Where a country has not provided complete times series data, the system will provide appropriate tools for estimating missing values. These tools will include interpolation and extrapolation algorithms as well as statistical models developed using analytical software packages.

17. Provision of sufficient information to users on the different aspects of the life-cycle of the data from the collection through processing to dissemination is one of the fundamental quality requirements of statistical data. In the FAOSTAT2 system all statistical datasets will be documented according to a standard pre-described methodology and quality descriptions which are concerned both with data submitted by countries and those that have undergone processing within FAO. Like all international organizations that aggregate and publish national data, FAO depends on the quality of the national statistics it receives, but also on the quality of its internal data processing procedures. The documentation will identify, among other things, how FAO procedures influence the overall quality of the data, which are the areas in need of improvement, which are the methods employed and quality checks that the data have been through as well as the statistical metadata standards applied to ensure the quality of the data, etc.

18. On the basis of the quality descriptions, a set of indicators will be developed and incorporated into data processing systems to monitor progress and development over time and to provide feedback to data providers. The quality indicators will follow the definition of quality developed in the context of the European Statistical System (relevance and completeness, accuracy, timeliness and punctuality, accessibility and clarity) but adjusted to the specificities of agricultural data. Initiatives and work currently under way on quality issues in statistics on international level, including quality indicators, will be closely followed.

D. Organising statistical data (data definitions and glossaries)

19. As noted by Sundgren (2003) "Developing and implementing statistical metadata systems Analysts specialising on international comparisons face, in principle, the same kind of problems as regional analysts – but on a higher level, and with even more challenging obstacles, since there is less common

administrative and statistical infrastructure on the international level than between regions in a single country. One remedy to overcome some of the difficulties is to promote the use of international standards as regards classifications and definitions of variables.

20. Both regional and international analysts need instruments to overcome inevitable differences between countries and between regions within a country, e.g. best methods, best practices, and tools for supporting such methods and practices. Experiences and "good examples" from colleagues may also be of considerable help. These are things to be taken into account by designers of statistical systems and metadata systems."

21. One of the main objectives of the FAOSTAT2 project is to create procedures and mechanisms to avoid redundant collection and maintenance of metadata, as well as inconsistencies caused by multiple sources. To this end a central repository will be created for concepts and definitions, and classifications respectively. A common repository is a necessary tool in the preparation of consistent and comparable data according to agreed standards by multiple countries. The repositories will be able to serve data producers at different stages of data production process, providing them with a reference against which to check concepts and classifications used in questionnaires, data compilation and dissemination, e.g. to check that the definitions of the national data conform with FAO recommendations.

22. In the process of creating the repositories, standards and definitions currently employed by FAO will be reviewed in order to identify discrepancies or inconsistencies, and to see to what extent concepts and definitions conform with international standards. Possibilities for harmonization and promotion of the use of standards can be identified and any deviations from the standards, due to operational purposes etc., explicitly explained and described. Consistency with, and possibilities of making use of existing international metadata standards, models and tools, such as the SDMX glossary, will be studied

23. In addition to the review of the FAO concepts and definitions as regards national and international practices, the linkages between FAOSTAT2 metadata module and other existing FAO metadata systems developed somewhat independently in the organization. AGROVOC (http://www.fao.org/agrovoc/) is a multilingual, structured and controlled vocabulary/thesaurus designed to cover the terminology of all FAO subject fields of agriculture, forestry, fisheries, food and related domains (e.g. environment) in order to describe documents and other information resources in a controlled system language. FAOTERM (http://www.fao.org/faoterm/) is a reference point for multilingual terminology used in the organization and in the international community. Although initially designed as a major reference for translators, it is also targeted to all authors and technical writers wishing to check and validate specific FAO terminology in the production of documentation and information products. Integration of the above-mentioned systems greatly enhances the internal conceptual and methodological harmonisation and standardisation, and creates a powerful search tool for various FAO systems and products, including FAOSTAT2, one that utilizes a structured thesaurus.

E. New Mechanisms for Data Dissemination

24. The FAOSTAT2 dissemination system will place emphasis on active linkages between data and metadata; the system should contain all the necessary information for the user to be able to make correct interpretations of the data and judge the usefulness of the data for his purposes. To serve users with different information needs, metadata need to be organized to provide all the necessary information but without forcing the user to browse through unnecessarily detailed data descriptions.

25. The dissemination system is the visible portion of FAOSTAT, the component that most external users are familiar with. It is through the dissemination system that data can be accessed for further use. Enhancing the functionality of the dissemination system is a means to strengthen the capacity of internal and external users of statistical data to perform more substantive analytical work. Opportunities exist to enhance and extend the functionality of the dissemination system via by using statistical metadata which will allow better access to the data as well as the ability to visualize FAOSTAT2 data through graphs and maps, combine datasets which will enable users to highlight trends, anomalies and areas of concern.

IV. CONCLUSIONS

26. Current international metadata standards focus on country level metadata. In the development of the FAOSTAT2 it is apparent that more attention needs to paid by International Organizations on the basic metadata models employed at the international level and how they relate to each other and those employed at the country level. Particular attention needs to be paid to the consolidation and elaboration of international concepts and definitions as well as the interchange of statistical data thus providing the end user with easier access to higher quality statistics.

References

Food and Agriculture Organization of the United Nations, *AGROVOC* , <http://www.fao.org/agrovoc/>.

Food and Agriculture Organization of the United Nations, *FAOTERM,* <http://www.fao.org/faoterm/>.

International Monetary Fund, *Information on dissemination standards and metadata* <http://dsbb.imf.org/Applications/web/dsbbhome>.

Kasnakoglu, H. and Mayo, R., *FAOSTAT2 and CountryStat*, Statistics in Transition, Volume 6, Number 2, October 2003.

MetaNet, *A training manual for the adoption of metadata systems and standards,* 2003, <http://www.epros.ed.ac.uk/metanet/>.

MetaNet, *Proceedings of the Final MetaNet Conference, Samos, Greece, May* 2003.

Statistical Data and Metadata Exchange (SDMX) – BIS, ECB, European Community, IBRD, IMF and OECD, 2003, Metadata Common Vocabulary, <http://www.sdmx.org>.

Sundgren, B., *Developing and implementing statistical metadata systems* <http://www.epros.ed.ac.uk/metanet/deliverables/D6/IST-1999-29093-D6.doc>, 2003.

UNSC and UNECE, *Guidelines for Statistical Metadata on the Internet*, Conference of European Statisticians Statistical Standards and Studies – No. 52, United Nations, Geneva, 2000,

UNSC and UNECE, *Guidelines for the Modelling of Statistical Data and Metadata*, Conference of European Statisticians, United Nations, 1995.

UNSC and UNECE, *Terminology on Statistical Metadata*, Conference of European Statisticians Statistical Standards and Studies – No. 53, United Nations, Geneva, 2000.