

SupCom 95 - Metadata and Statistical Databases

Final Report

World Systems (Europe) Ltd.

Chris de Vaney
Sabine Becker
Laurent Plancq

Table of Contents

Introduction	2
Project Summary	3
Background and Problem Description	8
Metadata and Statistical Database Systems within Eurostat	10
Metadata in the Agricultural Price Indices.....	18
Protocol for the Exchange of Metadata	26
Scenarios for use and future developments.....	32
Conclusions	36

Introduction

Conduct of Project

The SupCom 95 Project “Metadata and Statistical Databases” was performed between December 1995 and December 1996 by World Systems (Europe) Limited with the brief of:

- Examining the use of metadata in the EUROSTAT statistical reference and production environments;
- Identifying techniques and methods for improving the representation and flow of metadata between production systems and the reference environments ;
- Design and implementation of a prototype to validate the possible solutions ;
- Installation and validation of the prototype in selected EUROSTAT units

The project activities were undertaken within EUROSTAT, with the co-operation and guidance of EUROSTAT unit A3.

This document describes the objectives, conduct and results of the project, and discusses how the results may be exploited in the future.

Project Summary

Background and Problem Area

The SupCom Metadata and Databases project was initiated in 1995 to consider the problems of metadata definition and management with respect to the informatics activities of EUROSTAT, and to develop a prototype system to demonstrate potential solutions based on the analysis of current and future requirements.

For historical and organisational reasons, the use of metadata in information flow within Eurostat has been affected by the disparate information system infrastructure. Statistical production units have developed their own information systems to support data capture and analysis from the Member States and other international organisations. In addition, they have to a large extent also performed the dissemination activities in the production of documents and datasets for clients. Some directorates within Eurostat have developed reference environments to facilitate harmonisation and distribution of data, but these systems have been “ad-hoc” developments with little or no reference to the use of existing Eurostat systems.

As the need for harmonisation in the Eurostat environment grows, the role of metadata has become increasingly important in data quality and data dissemination issues. From the quality standpoint, the role of metadata as a documentation and harmonisation tool is of particular importance, as the need for detailed comparative analysis of the data grows. Dissemination requires methodological and historical documentation of data to satisfy the needs of clients, and to make it possible to develop new data products for distribution. These requirements have a significant impact on the information systems which, at present, are used in the production and reference environments.

The initial problem framework for the project was stated in terms of the need to explore the uses of metadata in the statistical production units of Eurostat, and the information flows to reference systems. The exploration of these flows to the Eurostat Reference Environment (NewCronos) was identified as a priority in the study. In addition, the role of the departmental reference environments was to be examined in terms of their role as harmonisation and forwarding agents in the information network.

Scope of Project

The specific goals of the project were to:

- Investigate the requirements for, and usage of metadata within Eurostat
- Identify the major reference and production systems currently in use
- Conduct an analytical study of a Eurostat production environment
- Determine the requirements for metadata and devise approaches for informatics support
- Develop a prototype system to meet the most pressing requirements

Because of the wide scope of the project brief and the limited time available for project conduct, a consultative group consisting of responsables from several departments in Eurostat was formed to ensure adequate communication with contacts.

The approach taken to the conduct of the project was to address the practical short-term needs with a view to the longer-term requirements. This required an investigation of the existing systems and production processes considered by Eurostat to be of high priority. The scope of the prototype system to be developed in the framework of the project was to be decided following the investigation and the presentation of results.

Activities Undertaken

The specific activities undertaken in the course of the project were:

- Investigation of the existing Eurostat reference systems
- Metadata use study of Direction F (Agriculture) units responsible for the production of Agricultural Price Indices
- Formulation of options for the development of a prototype system
- Design and development of the prototype system
- Deployment of the prototype system

Investigation of the existing Eurostat reference systems

The investigation of the existing systems within Eurostat concentrated on the metadata management facilities supported by the systems, with particular reference to the representation, use and exchange facilities for metadata. The investigation was conducted through interviews and examination of system documentation, data models and functional specifications. A total of six systems were investigated, namely:

- EUROSTAT Reference Environment (NewCronos)
- SIMONE I and II Nomenclature Systems
- CANDIDE Thesaurus Management System
- Regional and Social Statistics Database
- Directorate B Reference Environment
- COMEXT External Trade system

The majority of effort in the analysis of the Eurostat Reference Environment, COMEXT, Directorate B Reference System and the Regional and Social Statistics database.

The conclusions from the conduct of the study were that the metadata representation for the existing systems was limited, with the emphasis on classification and nomenclature management. During the course of the analysis, it became clear that the production units were not providing the reference systems with metadata unless specifically requested, and that the metadata supplied was very restricted in nature.

Metadata use study of Direction F (Agriculture) units responsible for the production of Agricultural Price Indices

Following the investigation of the existing reference environments for data within Eurostat, a detailed study was conducted of metadata use and distribution in the Directorate F units responsible for the agricultural price indices PRAG, COSA and ZPA1. These were selected due to their importance and existing information flows between the production units concerned and the Eurostat Reference Environment. In each case, the study covered the “production” of metadata, its use within the unit concerned and the onward flow to the Eurostat Reference Environment.

The conclusions drawn from the study were:

- The metadata used internally within the production units was methodological in nature, as the data suppliers in the Member States were not in a position to provide historical information on the data supplied unless specifically requested ;
- The metadata was developed from two main threads, namely as a set of definitions and methodological notes agreed with the data suppliers, and the metadata required for the transfer of data to the Eurostat Reference Environment. The latter objects are the derived metadata for the multi-dimensional tables covered in the NewCronos;
- Organisational communication between the production units and the reference environment administration was limited, and confined to addressing the day-to-day requirements. This is due to the lack of stated requirements and resources for managing and providing metadata on a formal basis.

Formulation of options for the development of a prototype system

Following on from the metadata study in Directorate F, a number of meetings were held with the Eurostat responsables to determine the best approach to the prototype development. A life-cycle metadata management system for use by producer units was considered, but lack of resources for maintenance in the production units mitigated against this option. It became clear that the most practical solution in the short term was to improve the flow of information between the production units and the Eurostat Reference Environment administration. This was seen as feasible given the on-going enhancements to the NewCronos to support documentary metadata in association with the data and hierarchy descriptions.

The requirements for an automated metadata exchange protocol were formulated and agreed, to provide an interface where the production units could access metadata from within the NewCronos, and could notify the reference environment administration of the metadata available within the production units. It was agreed that the main functionality of the protocol system would be provided for use by the information analysts currently responsible for management and terminological validation of the texts associated with metadata within the NewCronos.

Design and development of the prototype system

The requirements and functional description for the protocol were used as the basis for the design of a prototype system to be used within Eurostat to assess the feasibility of the proposed approach. The initial system has been developed in MS Visual Basic for client workstations running under MS-Windows 3.xx and MS-Windows 95. A database to support interim data management has been developed in MS-Access. It is envisaged that the information management support provided by the database will be subsumed over time by the enhancements to the NewCronos data management system.

Deployment of the prototype system

At the time of writing, a version of the prototype is currently under evaluation by the documentalists responsible within Eurostat A3. The system will be evaluated within the Direction F/1 (PRAG) production unit on a longer term basis, following the import of the methodological notes texts for PRAG into the prototype database.

Deliverables and Results Achieved

The main deliverables from the conduct of the project were:

- Prototype system implementing the protocol support tools (Software);
- Preliminary report on metadata use in EUROSTAT unit and departmental reference systems;
- Analysis report on the use of metadata in the Direction F production environments for agricultural price series;
- Functional Requirements and Model for the Metadata Protocol;
- User documentation for the prototype.

In addition, position papers on a number of aspects were developed. A paper describing the protocol was presented at the UN/ECE METIS Working Group meeting in Berlin in October 1996.

The software prototype is currently in evaluation by EUROSTAT, with a projected user base of the documentation team responsible for the population of the NewCronos environment, and the Direction F responsible for the PRAG series.

Background and Problem Description

Historical Background

The issue of metadata management within EUROSTAT has been addressed in several areas over the past few years, with the goals of integrating metadata with the data stored in the EUROSTAT databases and with adding value to the dissemination of EUROSTAT data.

From a historical perspective, the developments have been according to two main areas of activities, namely:

- The development of production and reference information systems by various units within EUROSTAT
- The standardisation of message structures and data exchange protocols

The development of the production and reference information systems, for example CRONOS, COMEXT, Farm Survey Support Reference System (FSSRS) and the earlier generation of EUROSTAT systems, resulted in each system adopting its own definitions and management procedures for metadata, with no reference to other systems. The evolution of these systems has not eliminated the problems in this area.

The standardisation of message structures and data exchange protocols is based on the need to exchange information between administrations, with the metadata required both to describe the structure of the datasets and to explain the concepts stored. This is under the UN/EDIFACT initiative, and has resulted in the development of a number of statistical and economic reporting message structures such as GESMES and CLASET. The use of these messages has not yet reached the mainstream EUROSTAT production and reference systems.

Current Situation

Within EUROSTAT, the changes in the informatics and applications infrastructure have raised again the importance of metadata management. In particular, the management of information flows from production through to departmental reference, EUROSTAT reference and to dissemination has increased the need to collect and manage metadata at all levels. There is also the problem of the incompatibility of the two main reference systems provided by EUROSTAT, namely NewCronos and COMEXT-II.

Project Problem Scope

Within the scope of the Metadata and Statistical Databases project, the problem scope was to examine the use of metadata as it applied to Direction F (Agriculture), to examine the production processes and to examine the information flows from the production units to the EUROSTAT Reference Environment (NewCronos). It was thought initially that the implementation of a metadata management database would be required, but this approach was seen as complicating the issue. The improvement of the metadata flow between Direction F and the NewCronos environment was given a higher priority.

Metadata and Statistical Database Systems within Eurostat

Analysis of EUROSTAT systems

In order to better understand the nature and functionality of the current statistical information systems within EUROSTAT, an analysis of the major reference environments was conducted for various directions and the NewCronos environment. The objectives of the analysis were to:

- a) Identify major information systems for data and metadata management within EUROSTAT;
- b) Determine the scope and nature of the metadata coverage, particularly in reference and data exchange environments;
- c) Identify perceived gaps in the metadata, and potential areas for exploitation and improvement;
- d) Determine potential development options for the SUPCOM metadata project in the areas of metadata modelling and integration.

In the terminology of the analysis, a Production environment supports the collection and analysis of data from external sources within an operational unit. A Reference environment is a pool of data and metadata organised for dissemination of data from several sources, which have normally been harmonised according to common classifications and data formats. Reference environments exist within EUROSTAT at directorate level and centrally with NewCronos. Typically, the directorate level reference systems act as a data feed into the NewCronos environment.

EUROSTAT Systems covered

In the course of the analysis, the following EUROSTAT systems were covered:

- EUROSTAT Reference Environment (NewCronos)
- SIMONE I and II Nomenclature Systems
- CANDIDE Thesaurus Management System
- Regional and Social Statistics Database
- Directorate B Reference Environment
- COMEXT External Trade system

EUROSTAT-Level Reference

At the present time, there are three major components potentially or actually used in the EUROSTAT Reference Environment:

- The NewCronos data management system
- The SIMONE Nomenclature Management system
- The CANDIDE Thesaurus Management system

The NewCronos system is an object-centred data management environment based on the client/server model. It manages multi-dimensional data tables provided by a number of statistical production units within EUROSTAT. The current implementation of the system is organised hierarchically, according to a breakdown of the subject-matter coverage. Third-party applications are interfaced to the NewCronos system via an Application Program Interface (API) library.

The metadata used in the NewCronos environment consists in the main of dictionary files containing code-lists, labels in three languages and a subject/theme hierarchy which is used in data access and management. The current version also has provision for the attachment of methodological notes to various levels in the hierarchy, and a number of methodological notes have been attached to various levels in the NewCronos hierarchy. These notes are available from the NewCronos INTRANET interface.

The acquisition and management of metadata within NewCronos is the subject of agreements with the supplying units. The biggest problem is in ensuring that the metadata used is consistent as far as possible across all subject/theme areas.

It should be noted that the scope of both SIMONE and CANDIDE is wider than reference support. Each of these systems is currently under continued development, or about to evolve to new versions.

The SIMONE database is a general nomenclature development and management environment, developed as a client/server system for use within EUROSTAT and, eventually, by other European administrations.

The SIMONE system provides a strong and comprehensive model for the representation of classifications, and a flexible browsing and retrieval interface. Third-party applications can access the nomenclature information via a dedicated API library. SIMONE also has a strong link to the CANDIDE system for concept management.

SIMONE is intended to eventually contain all active classifications used within EUROSTAT, and to act as a reference server for other administrations by providing an automatic selection mechanism. This will allow other authorities to use the nomenclatures within their own statistical activities, and will support harmonisation at the European level.

The current version, SIMONE-II, contains a number of nomenclatures in several national languages, and these are available to end-users via a world-wide Web interface.

The CANDIDE system is a thesaurus and concept management system for methodological information used in the statistical activities of EUROSTAT, and is intended to support a number of other systems in documentation of data and results. The system is currently under development, and an active link is planned to the SIMONE-II system. As with SIMONE, CANDIDE is an open system which will be accessible by third-party applications needing concept definition support.

The major areas of development are:

In NewCronos:

- Extension of the metadata support to include Flags and Footnotes against the data ;
- Enhancement of the Application Program Interfaces (API's) to standardise access to NewCronos by client applications.

In SIMONE:

- Evolution of the SIMONE data model and API's to the SIMONE II system. The current SIMONE environment is primarily used to maintain the nomenclatures from the SABINE system.

The data coverage of the reference environment is represented in multi-dimensional tables, and the related metadata is primarily based on code-lists or classification sub-sets. The data structure of series or groups of series are determined by the transformation processes, and are explicitly defined in the NewCronos system.

The code lists, stored in NewCronos dictionary files, are currently harmonised in some areas and not in others. This is a methodological rather than technical problem, and needs to be resolved by standards in production units.

Some developments are currently under way in text management under NewCronos, and the results will support the eventual inclusion of footnote references within the data model.

There are several integration possibilities for the systems comprising the Reference environment, which are simplified in principle by the availability of the NewCronos API. One potential contribution would be the development of utilities to maintain the current NewCronos code-lists within the SIMONE environment; the requirements would need to be studied in detail to ensure that this would, in fact, simplify maintenance.

The development of footnote support is a problem due to the need for classification on the textual content. The information analysts working on the text and footnote support for the Reference environment have identified a descriptive framework for classifying and indexing text in external files

The study of the reference environment identified some problem areas which need to be addressed in the future. These were:

- The need for metadata harmonisation by the production units, when supplying data to reference ;
- There did not appear to be much re-use of data stored in the reference environment by producers ;
- Restrictions on the available metadata in Reference, particularly to textual descriptions and methodological notes, may cause problems in any future Dissemination architecture. In particular, there is no clear way of managing the development of historical information on the data, i.e. changes of concept or models over time.

Departmental Reference-Level

Three systems were studied in this area:

- The Regional and Social Database
- The Directorate B Reference Database
- The Direction F ENVSTAT System

The Regional and Social Statistics database is the Direction E Reference system, and is used at the Directorate level to harmonise the data prior to transfer to the NewCronos environment. The data model is a multi-dimensional data manager using the ACUMEN database system. Metadata management is limited to data structure descriptions and classifications, with correspondence tables over the classifications for any necessary conversions.

The system is closed, and provides a reporting interface and data export facilities. Future extensions to extend its use to end-users are not foreseen. The system may, however, be extended with an interface to the COMEXT system to better support the nomenclature management function.

The Directorate B Reference Environment is a Directorate level database system for harmonising National Accounts data from a number of production units. The system has been developed as a multi-dimensional table generator, with some management facilities for a standard nomenclature for the data concerned. The major function of the system is to harmonise the data prior to transfer to the EUROSTAT reference environment. At present, there are no plans for further development.

The ENVSTAT system is the Direction F Environment Statistics reference system, which is used both within EUROSTAT and by administrative units in the Member States. It is a multi-dimensional table manager, with extensive facilities for metadata management of structures, nomenclatures and methodological documentation.

The ENVSTAT environment provides end-users with an interactive client interface, and allows the manipulation of matrices using the metadata provided. At present, the use of ENVSTAT is limited to the domain of Environmental Statistics. However, the system is open and can be applied in other areas of statistics. A prototype database has been developed for ENVSTAT containing structural data from the ZPA1 agricultural price series.

The focus of all three systems is the collection and harmonisation of data from production units for reference purposes. All the systems have strong interfaces to the EUROSTAT Reference Environment.

In each case, a distinct data model and support environment has been implemented to reflect the local requirements. A common factor for the systems is the representation of data as multi-dimensional tables, either physically under ACUMEN or logically under ORACLE.

Each environment has a strong methodological base for the harmonisation of data coming from the production units, and these are enforced by DBA procedures and the developed systems.

At present, each of the systems represents a distinct development with no dependence on metadata support tools developed centrally. This is expected to change, subject to the evolution of the Reference Environment and the availability of other, third-party tools.

The COMEXT System

The COMEXT system is a production and reference framework for the management of external trade series and flows. The system is a custom development, and has several unique features for data and metadata management.

COMEXT's main feature is that it is "data driven"; the system is based on a data dictionary, which allows both data and metadata objects to be defined and processed dynamically. To support this, an application language has been developed to allow client applications to interact with the system.

The data managed within COMEXT is multi-dimensional in nature, but is limited to six dimensions. This is due to the requirements of the subject-matter area. The main metadata objects stored within the COMEXT system are classifications, which are available to statistical users from a dedicated client application. The model for nomenclature processing defined within COMEXT is not documented.

COMEXT has a number of custom interfaces to import data from other systems, and has a link to export certain series to the NewCronos environment. Some work has also been done on exporting data in the GESMES format.

At the time of the analysis, the COMEXT system was discussed at a high level, but it was not possible to examine the complete architecture.

The COMEXT system is characterised by a complete set of production and dissemination tools, based on a unified, object-oriented data dictionary and a generic API which provides an interpreter for a dedicated application language. In conjunction with data management, the system also provides nomenclature development and maintenance tools and a keyword definition and search facility. Because of size and response constraints, the COMEXT system does not provide centralised text-management facilities.

The COMEXT environment is far larger than any other system discussed here, with approximately 70 gigabytes of data under management, and 250 registered users.

The UN/EDIFACT Statistical Message Developments

In contrast to the systems analysed, the EDI message developments are concerned only with data interchange standards, and not with physical implementations. There are two main message groups which are relevant in this area:

- The Generic Statistical Message (GESMES) and its derivatives ;
- The CLASET message structure for the exchange of classifications.

Each message specification identifies a restricted set of metadata for structural and documentary use in generation and re-use. It should be noted that message generation utilities are under consideration for some of the Reference Environment components, but the state of development is unclear.

Results and Conclusions of the study

The development and evolution of the EUROSTAT environment has been disjoint for historical and organisational considerations. The definition and use of metadata remained a methodological problem in most areas, and was aggravated by the pragmatic considerations of meeting current informatics requirements.

The major problem identified in the studies of the systems is the relative lack of information exchange possibilities between production, departmental reference and the EUROSTAT reference environment. This is due to the heterogeneous nature of the system developments, and the lack of standard requirements for metadata representation and content common to the majority of reference environments. In consequence, all the available interfaces between the systems are custom-built. This restricts the possibilities for information exchange considerably. It should be noted, however, that the reference environments - particularly NewCronos - have negotiated procedures with the supplier production units concerning metadata content, including new or updated requests.

In order to circumvent this problem, some use could be made of the standardisation work for the GESMES and related EDI messages. The GESMES specification contains a metadata description model which is consistent with the requirements and metadata available with the EUROSTAT systems described above. This can be seen as a “minimum” model, but would be useful as a basis for communications between the EUROSTAT systems. In principle, the metadata covered in the GESMES model is already available in the systems described above. The major goal would therefore be to harmonise the collection and management.

The availability of similar systems and developments in the European Member States should be considered. Some national statistics offices have developed sophisticated statistical data systems, with integrated metadata management, and are using these systems in projects such as the European Reference Environment developments. Such systems include the Statistics Sweden Time-Series database, the INSEE Dictionnaire de Donnees Statistique (Statistical Data Dictionary) and the Statistical Data Warehouse of Statistics Finland.

The opportunities to collect and disseminate metadata in this environment should be exploited as far as possible, in order to improve the flows of metadata from the national offices to EUROSTAT, and to reduce as far as possible the necessity to post-process the data for harmonisation purposes.

In the context of the study, the overriding conclusion was the need to develop solutions which integrate existing systems, rather than develop "new" data models and functionality from first principles. The development requirements framework for the SUPCOM was therefore that it:

- Be based on techniques to abstract and integrate the current solutions ;
- Provide a testbed for empirical integration of metadata structures and functionality, as opposed to development of a closed prototype.

From a conceptual point of view, the definitions and understanding of metadata within EUROSTAT are inconsistent. Some formalisation is needed in this area, covering:

- Identification of the metadata objects ;
- Specification of the metadata structures ;
- Identification of the relevant metadata used at each stage of the EUROSTAT information life-cycle;
- Default functionality associated with the metadata objects ;
- Information flows between the stages of information use.

Metadata in the Agricultural Price Indices

Background to the study

As part of the project activities, it was agreed that an analysis of metadata use and flows in and between production units and the EUROSTAT Reference Environment should be studied. A primary interest for this activity was to determine what the metadata support implications were for the reference environment.

During the discussions leading up to the study, it was decided that the domain of Agricultural Series produced by Direction F was a suitable area for study. This was performed between March and May 1996.

Requirements and Conduct

The specific objectives of the analysis were:

a) Identify Metadata and its use in the production units

The goal was to establish, at a practical level, the metadata used in production units from the points of source, contents, use and future potential. The metadata elements were those that contributed to the methodological and historical use of the data.

b) Establish the metadata flows between the production units and reference environments

The goal was to establish, for each unit, the metadata potentially available for reference and dissemination use, and the actual links to the reference framework for the series. Wherever possible, the requirements for the reference environment were to be noted in detail.

c) Establish the future requirements for the production units regarding metadata management

The goal was to determine the potential and requirements for metadata management and use within the production units, and the informatics and organisational implications.

The analysis was conducted according to:

a) Personal interviews with the responsables for the series covered in the analysis

For each of the series covered, the responsible officials were interviewed to discuss the use of metadata and the strategic issues, both current and for future consideration.

b) Examination of the metadata available and used

Where available, the existing sources of metadata used within the production unit were obtained and examined for structure and content. The potential use in reference was the main consideration here.

c) Examination of metadata flows into the production units from the supplier institutions

In each case, the metadata available from the external organisations supplying the data was examined, and the uses of the information available was studied.

d) Examination of metadata flows to reference environments.

The flow of metadata from within the production units to the EUROSTAT Reference Environment was studied, and the potential requirements detailed on a case-by-case basis.

Analyses

The analysis was conducted for the series:

- Agricultural Price Indices (PRAG)
- Agricultural Production and Food Balances (ZPA1)
- Agricultural Accounts (COSA)

For each series, interviews were held with the responsables for the series, and copies of the available metadata were examined. The potential and actual future requirements were discussed in all cases.

PRAG

a) Production Process

The production process for the agricultural price series is:

Data is collected by the unit from the member states

The collected data is harmonised according to the established procedures

The normalised data is placed in a FAME database for processing

The processed results are exported to the EUROSTAT Reference Environment.

At present, there is no local reference system for the PRAG data. Local access to data within the unit is provided by the local FAME databases.

b) Metadata Use.

i) Existing Metadata

The main metadata instrument used to support the Price Indices data is a Catalogue of Characteristics which documents the individual series.

Within the catalogue, each series is documented on a country-by-country basis according to:

- Definition of the agricultural product
- Marketing Stage and Sales Channel
- Marketing Conditions
- Place of Recording
- Recording Procedure
- Statistical Processing of Prices
- Representativeness of the Series
- Other Characteristics

The content is strictly methodological, and changes over time are recorded within the characteristics notes directly. There are no changes submitted or recorded with the data provided by the member states, and there is therefore no historical metadata associated with the series observations.

The catalogue is published in document form on an infrequent basis. The contents of the catalogue have been converted to Microsoft Word files on electronic media.

ii) Future Requirements

The future requirements concerning the PRAG metadata are primarily concerned with the maintenance and dissemination of the existing metadata.

Maintenance

The contents of the catalogue are stored as on-line document files, and are used only in the preparation of the Catalogue document. Given the hierarchical structure defined over the catalogue contents, the information could be more easily maintained in a database structure with a user interface to support entry, update, report and documentation generation of the catalogue elements.

Dissemination

There is a requirement expressed by existing suppliers and users of the data for on-line access to the methodological notes used in the compilation of the Catalogue of Series Characteristics. A world-wide Web HTML approach was being considered. A prototype system was developed for dissemination using an extended version of the CUB.X software.

c) Metadata development options for PRAG

The potential development options for the PRAG series metadata were mainly in developing a database structure for management of the catalogue contents. This required developing a data model based on the characteristics structure, and importing the document contents into the structure.

This could then be used as a basis for the development of dissemination interfaces, and be eventually extended to cover specific historical information on the series. The resource costs of metadata maintenance would have to be carefully considered before such a development.

ZPA1

a) Production Process

The process for the agricultural production series is:

- Data is collected by the unit from the member states
- The collected data is harmonised according to the established procedures
- The normalised data is placed in a FAME database for processing, using FAME Level 1 to harmonise the data with the classification plan
- The processed results are exported to the EUROSTAT Reference Environment.

At present, there is no local reference system for the ZPA1 data. Local access to data within the unit is provided by the FAME databases.

b) Metadata Use.

i) Existing Metadata

The main metadata instrument used to support the Price Indices data is the classification plan for the ZPA1 series, revised periodically. This is available in MS Word files, in an MS EXCEL workbook, and as a MS Windows help file providing an interface to the data stored in CUB.X.

There is little or no metadata flow from the member states to the production units. At present, there are resource constraints on the processing of metadata not directly related to the production process.

ii) Future Requirements

The future requirements concerning the ZPA1 metadata were primarily concerned with the maintenance and dissemination of the existing classification plans. These are edited manually in MS Word files which is time-consuming and error prone. A tool was therefore required to improve the checking and version control between revisions of the classification plan.

c) Metadata development options for ZPA1

The ZPA1 metadata development options were restricted due to resource problems, and the need to improve the quality of information from the source institutions.

COSA

a) Production Process

The production process for the series is:

- Data is collected by the unit from the member states
- The collected data is harmonised according to the established procedures
- The normalised data is placed in a FAME database for processing
- The processed results are exported to the EUROSTAT Reference Environment.

The data is submitted by the member states twice a year, in September and January. The majority of data is supplied by entry into MS-EXCEL spreadsheets provided by the COSA administration.

At present, there is no local reference system for the COSA series, with all local access to the data being provided by the FAME databases.

b) Metadata Use

i) Existing Metadata

The main metadata instrument used in the collection of data in COSA is a Manual of Agricultural Accounts Statistics, developed in collaboration with the contributing Member States. This manual describes the detailed methodology of concepts, classifications and data requirements for the domain. It is maintained in collaboration with the member states, and re-published on an ad-hoc basis. The manual may change in future to reflect the requirements imposed by the European System of Accounts (ESA-95).

The structure of the data collections covered in the COSA series are described in the classification plan used by the production unit, and within the EUROSTAT Reference Environment.

The production environment does not collect historical notes on the data, and all contacts and requests for explanation are handled on a personal contact basis with the national representatives.

ii) Future Requirements

The future requirements concerning the COSA metadata are concerned with refining and completing the methodological information, and with implementing any changes required for the ESA conformance.

Dissemination of the COSA accounts have been the publication of summary tables, prepared from the FAME database in MS-EXCEL format. Since 1995, the data is also available on disk, under the CUB.X interface. It would be of interest in the future to provide some of the conceptual metadata to users within the same framework.

A major problem seen by the administration for COSA was the lack of feedback on the user population for the Agricultural Accounts data within the EUROSTAT Reference Environment.

The development of an interface for extracting metadata from the NewCronos environment was of interest for the documentation of the data, and the maintenance of the COSA Classification Plan.

Conclusions

Within the units covered in this analysis, the use and flow of metadata suffered from a number of methodological and organisational problems which needed to be addressed. These were:

- Lack of Standards and Models

The interpretation of what actually constitutes metadata is subjective, given the lack of any formalised definitions and examples of the specific items concerned. Metadata management is therefore an ad-hoc issue, and can only be addressed as a priority when enough user interest has been expressed.

- Weak flows from the supplier institutions to the production units

In all cases examined, methodological and historical metadata is normally only made available by suppliers when specifically requested, and is of variable quality when supplied. The metadata flow from supply to production is therefore negligible, and leaves the methodological notes as the only active metadata.

- Restricted requirements for reference and dissemination metadata

No concrete requirements concerning metadata for reference and dissemination environments had been expressed, other than those necessary for the NewCronos system. It was therefore not possible to estimate the tool or resource requirements accurately, or to describe the necessary information flows at all stages.

Protocol for the Exchange of Metadata

Background to the Development

In the discussions with Directorate F concerning the use of metadata in the production process and the interface to the Eurostat Reference Environment (NewCronos), it became clear that one of the major problems regarding the exchange of metadata was the lack of an automated mechanism for notifying the interested parties of the metadata available at all levels. From the production level, there was no mechanism for determining the basic metadata resources available within the NewCronos system. Similarly, the administrators responsible did not know what metadata resources were available within the production units, that may be of benefit to the reference data.

At the time, some agreements between the production units and NewCronos administrators were available as documented memoranda. These were developed after negotiation between the production units, the responsible agents for the Eurostat reference environment, and various user groups acting as clients for the data. Changes to these agreements was a slow process, and consequently the documents did not accurately reflect the metadata actually available. This limited the use of this metadata within the general framework of the reference environment.

Conversely, the substance of the documented protocols was mainly centered on the classification plan for the data. The production units had to check this manually, and negotiate potential changes in the structures used in the reference environment. There was no interactive mechanism for the production units to check the contents and structure of the dictionary files from which the classification plan documents were generated, and to advise the reference environment administrators of potential changes.

In order to address these problems, it was proposed to develop an automated interface to allow the production units to identify the available metadata to the reference environment, and to facilitate the checking and change notification of the classification plans used.

Functional Requirements

The system was to provide functionality in the following areas:

- Classification of Metadata Resources
- Identification of Metadata
- Description of Metadata
- Interactive access to Dictionary Files
- Change Request Control
- Metadata Request Control

Classification of Metadata Resources

The system provides a facility for the classification of the available metadata resources, both within the production units and the Eurostat reference environment. The classification covers the nature of the metadata and the potential uses. It is anticipated that the classification will be agreed between the production units and the reference environment information analysts.

Identification of Metadata

The system provides interfaces for the identification of current and future metadata resources available between the production and reference environment. The identification shall include version information where relevant, and a change control history.

Description of Metadata

Once metadata elements have been identified, the system provides for the detailed description of available metadata in terms of:

- Content

The content description should explain the methodological nature of the metadata, its organisational sources and use in the statistical process. If any structure has been defined over the metadata, this will also be described in the content. The categories defined for the documentation stored in the EUROSTAT Reference Environment will be provided as a default.

- Format and Sources

The Format and Sources description give details on the physical format and nature of the metadata resource, and the location of the information if available in electronic form.

- Availability

The Availability description describes the access restrictions in force for the metadata resources, and the change control applied for each instance. It will also describe the update frequency for the resource.

Interactive access to Dictionary Files

The system provides for the storage of dictionary files describing the structure of the data within the protocol framework, with facilities for browsing the plans and for exporting the plans to document files where necessary. For the reference environment information analysts, a facility has been implemented to allow the plans to be edited and updated, subject to change control. This facility is not available to the production units who will use the change request mechanism.

Change Request Control

The system provides an interface to permit the production units to request changes or updates to the dictionary files. The interface will allow the units to relate the requests to the specific classification items to be changed. The reference environment administrators will have a similar facility to frame responses, and if necessary to log the changes to the classification plan.

Metadata Request Control

The system provides an interface to permit both the production units and the reference environment administrators to request metadata for particular purposes. Each request should clearly specify the required metadata in terms of content and format. The responses to the requests are logged within the protocol framework.

Implementation Issues

The system was required be implemented as an interactive user interface with shared access to the metadata across the network, with a number of small client interfaces facilitating the use of the system. The main NewCronos interfaces are those provided by the NewCronos API. Each production unit would have one or more local databases, and the reference environment administrators would have the facility to connect to several databases.

The proposed configuration is shown in Figure 1.

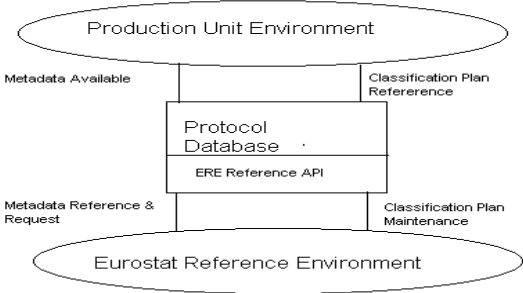


Figure 1 : General Configuration of the Protocol system

System Components

The initial version of the system was to provide the following components:

- API functional equivalents and local databases supporting the protocol

Client interface for:

- Data entry and update (Maintenance)
- Browsing the database
- Report production on the database contents
- Change control on the classification plans
- Request control for metadata
- Import and export of classification plans between formats.

The system was to be developed in such a way that future extensions can be made without compromising the basic data model.

Development and Deployment

As the initial system was a prototype, the development was achieved within the project time-frame using the facilities of an interface library simulating calls to the NewCronos API. The library functions were implemented in Visual Basic.

The initial implementation is general, and will ultimately be installed in three Direction F production units, as well as the Eurostat Reference Environment administrator's environment.

Cost/Benefit Analysis

The costs associated with the use of this system are:

Organisation Issues

The actual use of the protocol database will still be an organisational matter for discussion between the production units and the reference environment administrators. This particularly applies to the classification of resources, update mechanisms and access rights to the metadata.

Maintenance

Both the production units and the reference environment administration will need to provide the resources necessary to maintain the protocol databases. The effort clearly depends on the amount and nature of metadata available, and the active use of the protocol interface itself.

The main benefit expected from the production of the prototype is as a starting point for the formal exchange and re-use of metadata within Eurostat, and the documentation and monitoring of metadata flows and sources. In addition, other projected benefits are:

Extensibility

The basic facilities of the prototype system can be enhanced as required to meet future requirements, without compromising the functionality provided.

Use by other units

As the protocol is “open” and makes no assumptions about the metadata available, the basic system can be used to provide the protocols for other production units without change.

Scenarios for use and future developments

The Role of the Protocol

At the time of writing, the protocol system has two main goals: to provide a practical solution to the metadata flow and management problems between the EUROSTAT Reference Environment, and to determine future needs in the field of metadata management within EUROSTAT. The prototype itself will be used for approximately two years within the EUROSTAT Reference Environment, as a stop-gap solution until the projected extensions to the NewCronos environment have been implemented and deployed.

Extensions To Use

The extensions to the use of the protocol depends on how the overall concept meets the requirements in practice. The extension of its use within EUROSTAT would be as a means of communication between statistical production units other than Direction F, and the EUROSTAT Reference administration. This is a feasible approach if it is used as a preparatory environment for the extended NewCronos system.

The software for the protocol, as described, covers the basic metadata structures which would be used in the implementation of the metadata components of a GESMES message. The system could therefore be used to collect the basic metadata objects for such messages. Some further development would be required to extend the support necessary for modelling data relationships within the same framework, that is the mapping between the NewCronos metadata and the GESMES metadata requirements.

Given the problems of metadata collection from external data suppliers experienced by the EUROSTAT production units, some future extension of the protocol to support communications between the suppliers in the Member States and the EUROSTAT production units may be useful. It should be noted, however, that this approach may only work if the prototype is radically modified, and if the supplier organisations are willing to provide the local resources necessary. The Internet would provide a good basic framework for this approach.

Future Developments

The future directions of the tools developed within the SupCom metadata project framework depend on two main factors, namely their benefit to the production unit and reference environment users, and the evolution of the reference environment itself. In the medium term,

the tools may be used not only to support the day-to-day activities, but may also be used as a benchmark to specify further requirements in terms of metadata exchange.

One of the key factors in the longer term will be the ability to exchange metadata in standard formats, intended to act as pivots between several co-operating systems. This will be of particular concern in reducing the time for quality information to flow from production to reference and then to dissemination, where the demand for metadata sources for products and publications is highest.

Wider coverage of metadata types

At present the protocol and other systems in EUROSTAT cover only a small subset of the possible range of statistical metadata types. The metadata requirements in functional support of GESMES and CLASET message processing may require the extensions of the recognised metadata types. A standard set of metadata type definitions, possibly derived from the METIS (Meta-Information In Statistical Offices) reports of the UN/ECE may be developed as a basis for this activity.

The results of the IMIM and IDARESA Projects, currently active under the DOSIS research program, may provide a useful solution to the metadata structures problem. Both projects are working to a common metadata model covering all stages of statistical activities from production to dissemination of final results.

The results of the ESPRIT Projects IMIM and IDARESA will be of interest to EUROSTAT in future metadata development issues. These projects have been established to explore the specification and use of statistical metadata in conjunction with data within the context of the production and distribution activities of national statistical administrations. Owing to the nature of the projects, a formal collaboration has been established between the teams to ensure that the metadata structures used in the developments of both projects are compatible between them..

The basis of the metadata exchange between the two projects is the Common Application Platform (CAP), which allows metadata objects to be defined, populated and exchanged between statistical systems. The CAP resides on an independent server, and provides all management, search and conversion facilities for the metadata used in the various systems. The main framework for the CAP is a "pivot format" object, which is generated from specifications and populated by third-party applications. This ensures the free interchange of metadata at levels appropriate for the intended use.

To date, the CAP architecture has been defined and an initial set of metadata objects developed for metadata management and exchange. These objects are based on the methodological specifications of the Swedish Bureau of Statistics for documentation of surveys and their results (SCB-DOK). The metadata content has been specified and

developed for the European Labour Force Survey and three related national surveys in Sweden, Denmark and the United Kingdom. In future, coverage of the European and national Education and Training surveys will be added.

Following the analysis phase of the IMIM project, it is proposed to discuss an interface to the Common Application Platform, to provide for compatible metadata exchange with the NewCronos system.

The exploitation of these project results by EUROSTAT can be explored according to three criteria, namely:

- Standardisation of Metadata Modelling
- Metadata Capture and Re-Use
- Metadata Flows from National to Supra-National Administrations

Standardisation of Metadata Modelling

The IMIM and IDARESA projects will both explore and develop models for metadata use, with IMIM focusing on the statistical production process and IDARESA on the dissemination characteristics. The objective is to achieve models that lend themselves to exchange and re-use, while being detailed enough to support a range of activities. The models therefore represent various transformations, with each stage of the statistical production and use cycle requiring different “views”.

Metadata Capture and Re-Use

The focus of metadata management in the production process is based on capture at source, and refinement throughout the statistical process. This needs to be supported by appropriate informatics tools. From the IMIM and IDARESA project results, EUROSTAT will be able to study the national level models and determine what is of interest for statistical harmonisation and distribution. The re-use of the metadata objects will allow comparisons of data over time, and between countries. It will also significantly reduce the statistical burden of metadata development.

Metadata Flows from National to Supra-National Administrations

The statistical metadata content requirements are different for national statistical institutes, and the supra-national organisations such as EUROSTAT. Basically, a great deal of metadata is lost in the flows from national to supra-national level, particularly metadata concerned with conceptual changes and historical factors of data.

These flows were identified as a problem area for the statistical production units within EUROSTAT during the course of the SUPCOM project. If, for example, the responsables

within EUROSTAT have access to a framework such as CAP for data sources, it will be able to select the metadata required to add value to the data from the national level.

Conclusions

The production of the prototype system in the course of the project may be useful in determining the use of metadata in information flows between the various statistical units of EUROSTAT. The underlying problem is, however, that EUROSTAT needs to formalise its understanding of metadata both as to structures and function. The most promising vehicles in this area are currently:

- The current and (future) extended NewCronos environment
- The UN/EDIFACT work on statistical message

During the analysis of the metadata available within the production environments and EUROSTAT Reference Environment generally, it became clear that it is necessary to identify the metadata resource types which occur, or are likely to be available. This is particularly important for assessing the longer-term requirements with respect to the EUROSTAT Reference Environment and the eventual flow of information to dissemination.

Metadata in the production environment are the working documents and resources which describe the data in terms of structural, methodological and historical derivation. Examples of these classes of resources are:

- Dictionaries and Glossaries describing concepts
- Classifications, Nomenclatures and Correspondence Tables
- Data Exchange Standards
- Questionnaires and Response Frameworks
- Sectorial Calendars (i.e. Business Statistics, Education Statistics)
- Formulae and Processing Rules
- Historical Notes on data changes and interpretation
- Quality Notes
- Estimation Methods

In practical terms, the above list would have several other possible elements. The information will, in almost all cases, be available only in text documents or in support files and databases. Any typologies applied to structure the information will be either developed for sectorial reasons, or for information retrieval support.

Metadata in the reference environment are the objects necessary for the storage, retrieval and dissemination of the underlying data. This is more highly structured than the production information, but the level of metadata content is more conceptual than methodological. Examples of these classes of resources are:

- Dictionary files and data structure documents
- Methodological Notes
- Flags and Footnotes
- Change logs on the data
- Generated reference documents (i.e. HTML files)

The difference between the metadata groups is shown by the utilisation of the metadata. In the production units, the metadata is used to support the day-to-day tasks of the statisticians, and is normally compiled by the statisticians themselves. In the reference framework, the metadata needs and resources have been mainly specified and developed from the requirements to represent metadata and data from a wide array of statistical sources.

From a technical point of view, the use of the HTML notation for the representation of data and metadata in distribution is an issue which will need to be addressed in the near future. At present, HTML is a useful interim format for information in electronic form, particularly as the notation used is simple and intuitive. In the past year, a number of extensions to word-processing packages and other commercial desktop tools has made the transition to the HTML format automatic.

While HTML is adequate for distribution purposes, it is not currently suitable for the full exchange of data and metadata foreseen by EUROSTAT. This is due to the need to represent the information as objects which can be processed by third-party applications, and at present there is no means of defining detailed structures within HTML.

Within the UN/EDIFACT projects, the message implementations currently use the EDIFACT syntax and the SGML (Standard Generalised Mark-up Language) notation as a means of representing the necessary structures. SGML has many possibilities as a representation notation within EUROSTAT because it allows the definition of standard document structures which can be used as statistical data object formats. Another advantage is that HTML is a subset of SGML, and the transition between the two formats is well defined.

Emerging standards in the interchange of objects between application systems such as the Common Object Broking Architecture (CORBA) and the Common Object Model (COM) will need to be considered in the longer term, as future versions of commercial applications are developed to use the models.

CORBA and COM are standards for the specification of formal interfaces between information systems, to establish protocols for the exchange of object specifications and data. In principle, a system with a CORBA interface can process a data object from an external source, according to a specification of the objects's information structure and processing interfaces. The advantage from the point of view of metadata is that a statistical system can access and relevant information from another system without needing to have code specifically for processing the objects themselves.

From EUROSTAT's perspective, the consideration of these technologies must come after the full establishment of requirements for metadata representation and exchange. The IMIM and IDARESA projects are both expected to show some results in this area.