STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE/Eurostat Work Session
on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

COMMISSION OF THE
EUROPEAN COMMUNITIES

EUROSTAT

Working Paper No. 7
English only

Topic (i): Infrastructure issues for statistical metadata

PLANS FOR METADATA MANAGEMENT

Submitted by Central Statistics Office, Ireland[1]

Contributed paper

SUMMARY

Over the last five years, the Central Statistics Office (CSO) has engaged a number of consultants to carry out an assessment of the way that the Office conducts its business and to draw up an IT Strategy document for the future. On the basis of the recommendations contained in the strategy document the CSO has embarked on a major IT infrastructure upgrade, which will radically alter the way in which the Office functions. The move to a client server architecture, the introduction of Relational Database technology, and the deployment of Lotus Notes will all challenge the way that the Office will we conduct our its business. A project to build a Corporate Data Warehouse has been initiated which will contain all of the data as it passes through the data production cycle. All of the initiatives underway within the CSO will mean that data and metadata as currently produced and stored within the CSO will have to be revised, reformatted and updated. This brief paper focuses on experiences to date and the way the Office will evolve over the next couple of years.

I. INTRODUCTION

1. Until relatively recently the CSO used a mainframe computer for all of the data processing. Of late a 'client server' architecture has been installed and upgraded and an increasing portion of data storage and business processing is now carried out on this network. By operating on a number of different platforms, a fairly disparate set of tools, applications and languages has been accumulated for data capture, editing, analysis, processing, table production and dissemination. This tool set includes:

[1] Prepared by Richard Murphy and Dave Jennings.

Function	Toolset	*Function*	Toolset
Operating Systems	Win NT4.0, Open VMS 7.2	*Graphics*	SAS, PC-Axis and Excel
Data Capture	Blaise, Viking, Power Builder, Centura	*Analysis*	SAS, S-plus, X-11-ARIMA etc.
Database Design	Power Designer	*Text Handling*	MS-Word
Databases	Sybase, Oracle, MS-Access, Excel, SAS	*Publications*	Ventura, Adobe (PageMaker, Framemaker + SGML, Illustrator and PhotoShop)
System Development & Data Processing	Blaise, Btrieve, Centura, Cobol, Corvision, DCL, Dec Forms, Power Builder, Precision, SAS SQL, SSA Names3, Teleforms, TPL, Visual Basic	*Web Servers*	MS Internet Information Server
Classification Server	CARS and Business Register	*Web Browsers*	Internet Explorer, Netscape Navigator
Spreadsheet	MS-Excel, Lotus 1-2-3	*GroupWare*	Lotus Notes
Tables	SAS, PC-Axis, Ivation 20/20	*File Formats*	PC Troll, XML, HTML, GESMES, ASCII etc.

2. The challenge for the CSO is to develop the Data Management Strategy in ways that best meet the needs of the statistical divisions of the Office, that maximise efficiencies in collecting, processing and disseminating statistical information, and that provide added value to statistical products through the use of leading edge analytical tools. As part of this, it is intended to design and implement a Corporate Metadata database and management system.

3. From an operational point of view the CSO has some 70 survey/data collection processes which process around 8,000,000 data records annually (excluding the quincentennial Census of Population). On the electronic dissemination side the electronic databank contains 200 datasets with 37,000 series.

II. BACKGROUND

4. There have been a number of initiatives that affect both the business and IT areas including:

- the deployment of Lotus Notes as a GroupWare product. This has been very successfully used for email, as a corporate discussion tool, as a communications medium and as a document management and retrieval mechanism;
- sybase for data and some metadata storage;
- sybase/Centura and SAS for applications development;
- the purchase of a Classification application CARS (Classifications and Related Standards) from Statistics New Zealand;
- the purchase of the Central Business Register System from Statistics New Zealand.

5. These are the corner stones of the Data Management Strategy and will be crucial for the development of a corporate data warehouse as the principal storage mechanism for data and metadata holdings.

6. Essentially this structure includes four databases (input, output, aggregate and disseminate), a classifications server, a central business register and a metadata management system. See Figure 1.

7. This diagram is an interpretation of the flow of information to the end user taken from the CSO Data Management/Warehouse Strategy document. The idea is that all information will be stored once centrally in these databases from where they can be disseminated to the public or other institutions whether by paper or electronic means. The data model to be used in the Aggregate/Disseminate databases will be based on the data cube. The basis for the data cube or 'hypercube' model is Bo Sundgren's (a, b, g, t) model. This was accepted as a standard at the Eurostat Statistical Metadata workshop in Luxembourg in February 2000. The dimensions of the cube would be fed by a classification server (the CARS system, developed by Statistics New Zealand, will be the classification server of choice).

8. The task for the Office is how best to devise a metadata management system that covers the four stages of the production cycle, that is easy to use, easy to maintain, and that will enable a first class service to be provided to users.

III. METADATA MANAGEMENT

9. There are three basic descriptions for Metadata storage – Structured Metadata, Semi-Structured Metadata and Structure Free Metadata.

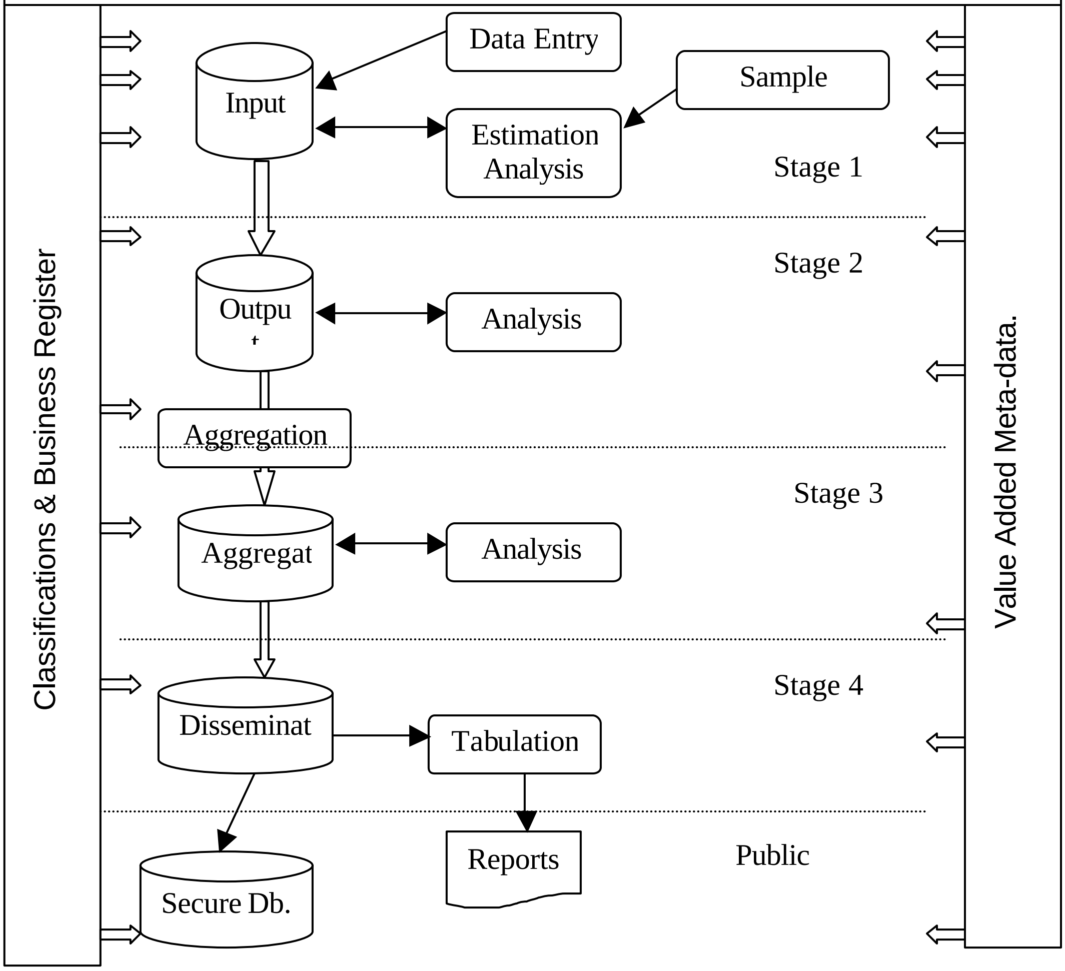
10. Metadata is used for 3 functions. It is used to find, understand and evaluate data. The primary goal of the metadata management system is therefore, to provide a repository for survey and output information gathered through the survey cycle. The system must include items such as:

- information relating to surveys / outputs that does not normally change with each instance of the survey. For example, contact names, survey design, questionnaire image, output variables, classifications, glossary of terms, operating instructions, processing systems;
- information which can and does change with each instance of the survey. For example, hand-over reports, analysis and error reports, key occurrence details, data quality, output variables, milestones planned and achieved, progress logs, publication details;
- the dataset description that contains information relating to the definitive dataset for each instance of the survey.

11. The system will enable the following to be done:

- learn about the surveys and outputs of the CSO: Purpose, design, contacts, output variables, operating instructions, glossary of terms and acronyms, significant events, etc.
- provide information to the publication process, e.g. technical notes;
- provide information to external users;
- compare work practices across surveys / outputs and learn from previous practices;
- monitor survey performance – planned versus achieved;
- disaster recovery – work practices are documented.

Figure 1. Data Flow through the Statistical Production Cycle



IV. EXPERIENCE WITH METADATA

12. Fundamentally the system must be visible, accessible, relatable, reliable, understandable and media independent. Therefore, the system must allow for both 'Free Text' and structured metadata to be stored centrally, that is easily accessible and user friendly.

13. It is worth bearing in mind that metadata is crucial in the dissemination process. Therefore, improving the quality, quantity and consistency of metadata provides value added to both printed and electronic products.

14. The metadata that is currently used within the Office is at two levels. The Variable (data item) level metadata includes information on:

Field	Description	*Field*	Description
Code	7 or 8 digit alphanumeric code for the variable	***Pcode***	The periodicity of the time series
File	The dataset in which the variable appears	***SACode***	Seasonal adjustment indicator
Title	The full title of the variable	***Dtlu***	Date the series was last updated
Label	Short hand title for the variable	***Tmlu***	The time the series was last updated
Start	The starting period for the variable	***Diska***	Technical item
End	The last period for which data is entered	***Diskoth***	Technical item
Units	The measurement units used for the variable	***Filea***	Technical item
NDec	The number of decimal places the data will be reported to	***Fileoth***	Technical item
SACode	Seasonal adjustment indicator		

15. The file or dataset level metadata includes:

Field	Description
Title	The name of the dataset
Periodicity	The periodicity of the dataset
File Description	Free text giving a description of dataset contents.
Revisions	States the revisions policy followed for the data items in the dataset
Updating	The usual publication lag for the data items in the dataset.
Technical Notes	This field can contain anything from the relevant EU legislation to sample size or technical definitions.
Rebasing	If the dataset contains index number series, this gives the date and period to which the series is rebased.
Discontinuities	Includes code number changes or data structure changes.
Seasonal Adjustment	Method of seasonal adjustment if relevant
References	Publications giving detailed breakdown of methodologies or in which the data appears.
Contact	Contact details of person who oversees the production of the data
Source	Source survey or administrative source data originates from.

These will serve as a starting point for the population of the new integrated corporate metadata database.

V. PLANNING

16. In the new metadata system the metadata relating to the data tables in the Input and Output databases (see Figure 1) will be stored at table level and then at column level. Where free text metadata is required the relevant metadata table column will contain a reference (and link) to the free text source. Examples are given in the following tables.

Tables

TableName	Section	Data_ Custodian	Description	Documentation	Created	Updated
I-BC-A-032006	Building	A N Other	B & C employment survey returns	LN/Industry & Building/ Building	30/03/06 17:19.027	21/05/06 10:05.213
…						

Columns

TableName	Column	Classifi_ cation	StoreFormat	Description
I-BC-A-032006	SurveyMonth	T	yyyyMmm	Reference month. Returns cover a specified week.
I-BC-A-032006	RefNo	N	(9)x	Building Section reference number.
I-BC-A-032006	SizeCategory	Y	x	Based on average employment. Updated every January.
I-BC-A-032006	PermStaffA	N	(6)n	Managerial, clerical etc.
I-BC-A-032006	PermStaffB	N	(6)n	Manual staff.
I-BC-A-032006	LabOnly	N	(6)n	Labour-only subcontractors.
I-BC-A-032006	TotalPE	N	(6)n	All persons engaged.
…				

17. For the Aggregate and Disseminate databases the metadata tables will also be mainly at the data table and column level and will reflect the fact that the ‘data cube’ model is being used. It is intended that one of the main dissemination tools that will be used in the new system will be PC-AXIS and so the metadata system is being designed to interact easily with PC-AXIS. Examples of metadata tables are as follows:

Tables

TableName	Text	SText	PubCat	Database	Timescale
D-BC-A	Building and Construction Employment Indices (Monthly)	B & C Emp Ind	O	DISS	month
D-HB-GEN-2007	Household Budget Survey (HBS) Main Tables	HBS Main Tables	O	DISS	HB-TS
D-HB-HLD3-2007	Household Budget Survey Household Data	HBS Household Data	P	DISS	HB-TS
...					

Columns

TableName	Column	Text	SACode	Subject Code	Copyright
D-BC-A	Index	B & C Empl Index (1995=100)	0	IB	1
D-BC-A	SAIndex	SA B & C Empl Index (1995=100)	1	IB	1
D-HB-GEN-2007	Value	Estimate	0	PH	1
D-HB-HLD3-2007	Value	Household data	0	PH	1
...					

18. While the metadata model is being developed to accompany the data model, it must also be able to stand apart from and function independently of the data model. As work proceeds with the models and in order to put proper foundations in place decisions on some basic issues must be made such as:

– user interfaces;
– updating procedures;
– rules and standards for content;
– migrating data and metadata.

19. User Interfaces: User Interfaces, which retrieve data by using SQL statements, will be developed. These interfaces will ensure that the link between the different metadata repositories is seamless and will not hinder metadata retrieval.

20. Updating procedures: As the quality of metadata is the responsibility of the producer sections, the system will have in-built controls to ensure the integrity of the metadata is maintained.

21. Rules and standards for content: Experience to date with metadata is that if rules and standards for the metadata content are not provided, then the quality varies significantly across producer sections. Capturing metadata will be an integral part of the production cycle and not a procedure that is carried out after the data production cycle is complete. We are intending to have four metadata databases to correspond with the databases. This increases the need for quality management, consistency and coherence of the metadata.

22. Migrating Data and Metadata: With the move to the new architecture there will need to be a process of grooming the data and metadata. This process of migration will highlight inconsistencies, often within published data. Intuitively it would be preferable to clean data and metadata as it is migrated to the new environment. However, this may not be a realistic option as it could lead to the migration process being slowed down dramatically while the inconsistencies are resolved. To avoid the slowing of the migration, we can copy the metadata to a temporary area for cleaning before loading it into the new storage environment. The use of the new central data management system for data and metadata will highlight inconsistencies, which can be rectified immediately. The metadata management system currently being used for the dissemination of data on the mainframe has been re-developed in Lotus Notes for testing and demonstration purposes.

VI. WHO USES OUR DATA/METADATA?

23. The new system will have to be able to meet the needs of all users who will use it to search, analyse and interpret the data. Typically users want data or information that is easily understood, accessible and coherent. Essentially, there are three types of data users:

- casual users who have no in-depth knowledge of the subject matter and who want quick answers;

- normal users who have some knowledge and who are prepared to put in some amount of effort in interpreting/analysing data. They will want some more background information;

- expert users who have an in-depth knowledge of the content, who know how to analyse data and will be able to make their own decisions in interpretation. They will require a good deal of background information.

VII. FUTURE DEVELOPMENTS

24. The system now being introduced will cater for the both data producer section and data users and will be flexible enough not to hinder future development in any area of the office. Future initiatives will include E-Government, Increased use of web technologies, the development of a knowledge management system, value for money audit and quality analysis.

VIII. CONCLUSION

25. This paper is necessarily brief and does not give a complete outline of all the work that is currently being undertaken within the Office. The Data Management Strategy is still at an early stage of development where the core elements/concepts have been identified. The methodology employed to bring the Strategy to a successful completion remains open. There is still room for refinements to the implementation plan. The target completion date is 2005, which suggests that there is adequate time for all of the tasks. However, as anyone who has been involved in the project management of large-scale projects knows, there are always unforeseen elements that put pressure on resources and deadlines. Therefore, ideas, suggestions, discussions or initiatives undertaken elsewhere will be evaluated with a view to incorporating them into the Strategy in order to smooth the path to a successful completion.

BIBLIOGRAPHY

List of Metadata Items for OECD’s Main Economic Indicators Gerald Petit, Pierre Bezis and Rob van Eck OECD Statistics Directorate October 1996

MISSION (EU 5TH Framework Project)
Multi-agent Integration of Shared Statistical Information Over the interNet.

FASTER (EU 5TH Framework Project)
Flexible Access to Statistical Tables and Electronic Resources

ADDSIA (EU 4th Framework Project)
Access to Distributed Databases and Statistical Information Analysis
Classifying Metadata Work Package 04.02.01e 1997