WP.4
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 (i): Functions of metadata in statistical production

MODERNISING STATISTICAL SYSTEMS TO IMPROVE QUALITY THE EXPERIENCES OF THE OFFICE FOR NATIONAL STATISTICS PAPER FOR THE METANET CONFERENCE - GENEVA - FEB 2004

Invited Paper

Submitted by the Office for National Statistics, United Kingdom[1]

1. Introduction

The UK's Office for National Statistics (ONS) has embarked on an ambitious modernisation programme to deliver a standard technical infrastructure; standard agreed methodologies and statistical tools, over a five-year period. The Programme was set up by Len Cook - National Statistician, in 2001.

The ONS was created in 1996 out of the former Central Statistical Office, the Office for Population Censuses and Surveys (which includes Civil Registration of births, marriages and deaths), and the statistical division of the Department for Employment. This brought together into ONS a wide range of statistical activities spanning economic, social, population and labour market statistics. The legacy of the three organisations was a diverse set of IT platforms with software and statistical methodologies that required central support. The Statistical Infrastructure Development Programme (SIDP) and the Information Management Programme (IMP) were created to develop and capture the benefits of an integrated statistical system for ONS' varied statistical outputs. This has now been combined with a structural reorganisation, with the office moving away from a traditional social/economic split to one based on data sources/analysis directorates. Customers for ONS' data should benefit from the provision of a more responsive and high quality analysis service, as resources are switched from maintaining systems, to the production of value added statistical analyses. The ONS' stated aim is to become a provider of "World Class" statistics, and to implement systems that are responsive to new developments in technology and statistical methods.

[1] Prepared by John Kinder, Margaret Lane, Debby Osman, and Emma Heath.

Undertaking a major programme of modernisation has to be adequately funded if the benefits of standardisation are to be realised. During the summer of 2002 ONS submitted a business case for modernisation to HM Treasury, and a sum of £75million (approx. Euro115m or $120m) was allocated to the programme for spending between 2003 and 2006. The systems currently used for the production of ONS' outputs have been put on a 'care and maintenance' basis in order to free up staff resources for developing and testing the new tools and software being acquired.

2. ONS' modernisation strategy

In simple terms ONS' strategy is to converge on an agreed set of statistical methods and tools that function efficiently within a technical environment which is sufficiently robust to meet current business requirements and future demands. The Information Management Programme was charged with examining the future IT requirements of the office, and to design a technical architecture that would support the methodology, statistical tools and software identified by SIDP as the standard for ONS. The Methodology Group (MG) within ONS is a partner in the Modernisation Programme, and is responsible for developing standard methodology, assessing the suitability of statistical tools and is working with SIDP on the implementation of the standard tools across the office. The SIDP team comprises of a mix of professional statisticians, project managers and general administrators. This team only 'Project Manages' the programme, it is not involved in the actual assessment and development of methodological work. The modernisation programme has set a "buy, not build" strategy for most of its standardisation, however the Metadata System is being developed within ONS. The ONS has looked to National Statistical Institutes across the world for work on statistical tools in a similar technological environment so that such tools may be imported and implemented.

SIDP has also set up a Communications Strategy in partnership with IMP and the Re-engineering Programme, and has appointed a member of staff with communications and training skills to steer the strategy. Investment in keeping the whole of the office fully informed about progress is seen as a critical part of the strategy, as the areas within ONS which are conducting 'business as usual' are an important factor in the success of a modernisation programme. Among other initiatives the Modernisation Programme has provided the facility to cascade information about progress of redevelopment on the ONS Intranet. Open sessions are also regularly held, where prototypes of new tools are demonstrated.

3. Progress so far

SIDPs first task, when it was set up in 2001, was to conduct a review of the ONS' statistical business and to identify any urgent re-engineering projects. This was the first time that a comprehensive review had been undertaken since the merger of the three organisations into the ONS in the 1990s. The SIDP reviewed the work (processes and practices) of the ONS during the summer of 2001. The review found that the office used a collection of many diverse systems, methods and software in the business areas. There was a heavy reliance on the use of complex spreadsheet based systems for the production of outputs, and the line between data collection and analysis varied across business areas. The review also painted a picture of the office's statistical activity using the concept of a Statistical Value Chain (SVC) which describes statistical components from survey design to data dissemination. This brought together the office's work in terms of the activities undertaken and the statistical tools that support them in a data lifecycle approach. The SVC is now being cascaded for use

throughout ONS and is the basis for setting up and defining the new infrastructure - see Annex A for a visual representation of the SVC.

The report concluded that there were areas in ONS that needed to be re-engineered onto a modern infrastructure making use of standard statistical tools and methodologies. In early 2002 Phase 1 of the re-engineering programme began, and four projects were set up. These were:

a) Production of National Accounts

b) Labour Market Statistics

c) Population Estimates & Projections

d) Production of Price Indices.

The first phase of the re-engineering projects has been set up and the programme of work, which will take place over the next three years, is in its development and scoping stage for three of the projects. Labour Market Statistics is due to go live in the summer 2004. The re-engineering projects were selected by assessing their potential impact on a series of Performance Indicators set up for monitoring the Modernisation Programme. They are expected to score highly in increasing the value of ONS' outputs, improving integration and harmonisation, reducing staff and analysis costs and reducing time to publication.

The SIDP team then proceeded to identify the statistical tools that would support the SVC and the requirements of the re-engineering projects. SIDP and MG are working together on Statistical Infrastructure Projects (SIPs), set up to deliver standard statistical tools to support both the re-engineering projects and the technical infrastructure being designed for the ONS.

These have been designated as Phase 1 SIPs, and are all due for implementation during 2004. The Phase 1 SIPs are to produce tools for -

a) Time Series Analysis - the selection of X12 ARIMA as the office standard has been approved by the Statistical Policy Committee

b) Tabulation - SuperCROSS has been selected as the corporate standard that is already being rolled out and older packages have been removed

c) Disclosure and Confidentiality

d) Index Number Construction

e) Weighting, Estimation and Standard Error Calculation Tools

f) Quality Measures

g) Coding Tools

h) Editing and Imputation.

Recently, the SIDP team has considered a further set of tools. These are Geo-statistical tools and geographic referencing instruments.

All the tools being acquired and developed must be capable of operating within the technical infrastructure being set up for ONS. Oracle has been selected as the relational database, a Content Management System has been bought, and the Time series tools project has been extended to identify and acquire the tools that will be needed for managing and analysing time series via the web.

These tools are specific examples of the function of metadata in statistical production. This is because there will be a need to record metadata about the processing the tool executes. A

set of Metadata will be captured at input, during processing and after the tool has completed its processing. This metadata will be captured in the ASDMS subsystem, which is explained in section 4.

An example of metadata captured during statistical production is detailed below using the SIP tool Time series:

Metadata to be captured prior to time series tool processing is:

Input

- Non Seasonal Adjustment of time series
- Title of time series
- Title of time series group
- Run title

Parameters

- Periodicity
- Time Series/Time Series Group
- Extent of Processing
- Handling partial failures

Processing metadata to be captured

- Status of processing for each time series
- Reason for failure for each time series
- Warnings displayed to the user (e.g. quick in place of Methodology Group derived)
- Outliers detected by time series and location within time series
- Specification used for each time series
- Input set as provided to X12ARIMA
- Output set as generated by X12ARIMA (including: Analytical output, Graphical Output)
- Command line for each Time Series

Metadata to be captured after the time series tool processing has completed:

Output

- Seasonality (Seasonal Adjustment/Non Seasonal Adjustment) of each Time Series

Quality Measures

- Outliers detected in each time series.
- Metadata indicating which points have been identified as outliers.
- Periodicity
- Time Series/Time Series Group
- Extent of Processing
- Handling partial failures.
- Seasonality (Seasonal Adjustment/Non Seasonal Adjustment)of each Time Series
- Specification used for Time Series.
- Relevant diagnostics from output set as generated by X12ARIMA (including: Analytical output, Graphical Output,) (reference to file set)

4. Metadata and its contribution to standardisation

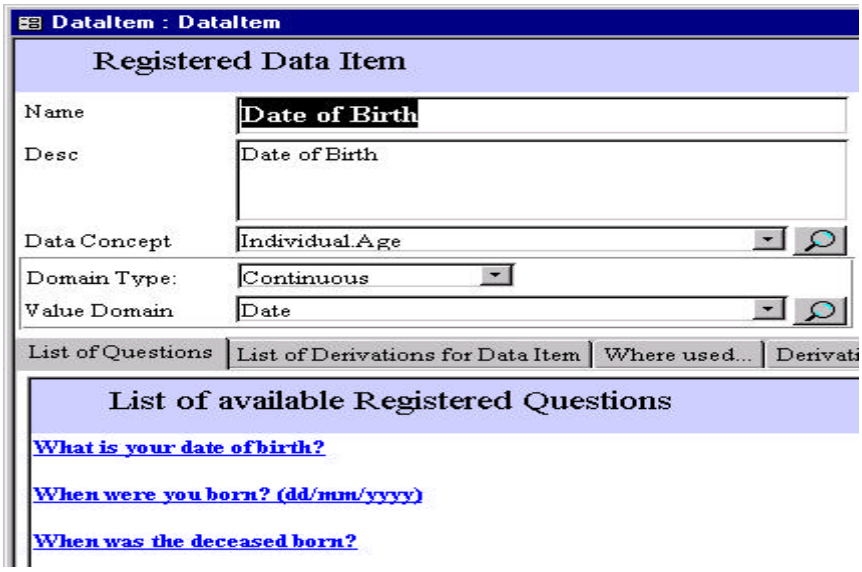
In addition to the tools listed above a Metadata Programme has been set up to deliver a corporate metadata system. A comprehensive set of metadata is seen as fundamental to the modernisation strategy. The aim of the strategy is ensure that a set of metadata is created as data are produced at each stage of the statistical value chain. When data are then eventually presented to the dissemination database and accessed via ONS' website, all outputs will be metadata referenced. The Metadata Programme has brought together several projects designed to provide standard metadata across the office's outputs. These include:

- *Discovery metadata standards* - which is defining a definition of standards for the discovery and control metadata needed for each of the types of 'content' which may be prepared for dissemination via the website. This will enable users to readily find and item of information. The project will set standards for ONS and work to influence standards set by other organisations, to ensure that those standards meet National Statistics needs.
- *Controlled vocabularies* - which is identifying all the classifications and other entities for which a controlled vocabulary needs to be established; establishment of responsibilities and procedures for agreeing or updating such vocabularies; management of the initial round of obtaining an agreed list for each entity.

To help promote the need for a Corporate Metadata system a Metadata Demonstrator was developed and presented around the ONS. This system illustrated to staff what a corporate metadata system might look like, helped stakeholders to understand what metadata is and it's benefits across the organisation, and encouraged feedback from business areas on the potential impact of a corporate system.

The system successfully demonstrated the importance of integrating standard metadata into a modern statistical infrastructure that will deliver outputs with a consistent set of explanatory data across the organisation.

Shown below are two of example screens of the Demonstrator:



Work is now well underway to develop this Corporate Metadata system. This system will provide:

- repositories to hold metadata;
- tools to create, transform and disseminate metadata;
- common vocabularies for metadata entities, to enable effective storage and retrieval;
- agreed procedures and responsibilities to support the capture, use and management of metadata.

Data interchange

The new systems will operate within ONS, but account will be taken of the need to acquire and transfer metadata to/from the National Statistics data providers and users. The advantages to other Government Departments of this metadata system have been identified as:
Improving quality for users by giving a consistent approach and by having all metadata in one place.

Improving the metadata that people are supplying to use and reducing supplier burden by means of a simpler process to load data on the National Statistics website. This will not have immediate impact but will evolve over two or three years.

The ONS corporate Metadata system will be made up by two subsystems. These subsystems are:

1. **Classifications**
2. **Administrative and Survey Data Management System (ASDMS)**

The **Classifications** subsystem is due for delivery in April 2004. This will hold all of the versions of Classification used by ONS with an indication of the current approved version. This subsystem will also include the tools to access and retrieve classification metadata.

The **Administrative and Survey Data Management System** will hold all the other metadata that relates to survey and administrative data collected by the ONS and other organisations, including the tools to access and retrieve the metadata. There are four types of metadata and these are:

1. *Statistical metadata* - This is information about the content of statistical data, to support understanding and interpretation of the data eg. Statistical methods, data definitions, classifications, quality measures. It also informs users about how the data was collected, sample sizes , collection date, degree of confidence in results, geographic coding and classifications, etc.
2. *Discovery metadata* - This enables users to identify and find appropriate content or information, using eg. title, author, keywords, publication date, variable names, geography areas. These metadata provide a way of finding relevant data and content, using search engines and indexes.
3. *Control metadata* - (also known outside ONS as Administrative Metadata). This is information needed to automate workflow systems and manage content, eg. Creation date, data supplier, editor, approver, release date, archive date. These metadata are for internal operations, to dissemination. These metadata are concerned with the processes affecting data rather than the data and their content.
4. *Technical metadata* - This is information about the location and format of the data, used by the systems for data interchange and manipulation eg. Types of file, size of dataset, record length

The ASDMS is to be delivered in phases. Phase 1 is due to for delivery in the summer of 2004, and contains around 130 metadata items for surveys.

Work is currently in progress to establish the content of phase 2, although no delivery timetable has been detailed as yet.

Phase 2 will contain further metadata items, giving the users of the system more detailed information regarding the statistical data they are using. It will contain further functionality that includes:

- Time Series
- Question Library - which will hold metadata related to standard questions held in a library, to enable consistent collection of data by ONS and other organisations, including the tools to access and retrieve these metadata
- A Glossary

Population of the Metadata Systems

- **Migration of existing metadata**

In order to populate both the Classifications System and the ASDMS and other sub-systems, each item of existing metadata will need to be scrutinised to see how well it fits the new models. It is hoped that a proportion of existing metadata will be able to be migrated automatically (but the process for doing this needs to be created). Where automated migration is possible, resource from business areas will be needed purely for the quality / proof-reading process.

Manual migration will be done by a migration team within SIDP who will be responsible for looking at existing metadata for each system or business area, one at a time, and inputting it into the new system. The staff within the migration team will be trained in interpreting existing metadata, producing metadata where none exists and rewriting where necessary. As many drop-down boxes as possible will be utilised in the new systems. Business areas will quality assure the work.

- **Input of new metadata**

Business areas will be responsible for the population and utilisation of the metadata repositories i.e. they must ensure that their business processes create, edit, and use metadata at appropriate points. Where existing metadata have been incorporated into the metadata system, business areas will be responsible for adding to and fully populating the repositories.

Similarly, business areas will need to make use of metadata when developing their products, by identifying relevant metadata and including the links when specifying web pages etc.

5. Programme for 2003 - 2006

The next steps in the Programme are already being taken. Following the office's restructuring into Sources and Analysis Directorates, the creation of the Sources Directorate will produce the most significant culture change for the office. Collection and first stage analysis of data from business and household surveys will be integrated. The transition into a Sources Directorate will create vertical streams of expertise in such areas as editing, estimation and survey design. To achieve this transition, a Business Transformation Office (BTO) has been set up to assist Sources Directorate in identifying the programme of work it needs to achieve its goals, and to manage the workstreams identified alongside the Re-engineering projects and the SIPs.

The BTO will now manage, in addition to the SID Programme, a range of re-engineering projects and statistical tools, and steer the transition programme drawn up by Sources Directorate. Survey and data sources integration will be a major plank in the transition planning. An integrated social survey is being developed to bring together the continuous social surveys conducted by the office. A cross economy monthly business survey is in the pipeline, whereby data derived from Value Added Tax (sales tax) will support survey data to give estimates of the monthly movements in production and turnover. Developments in methods of capturing data are evolving and will provide a wide range of alternatives from the traditional use of mail, to Telephone Data Entry (TDE), and eventually online electronic questionnaires.

The BTO is also sponsoring a project that examines ways of integrating the wealth of administrative data collected and available to ONS, into the data held by Sources Directorate. From its outset the project plans to bring all datasets into one area where they may be held and validated in a central repository. The IM Programme is designing a prototype Central ONS Repository for Data (CORD) that will hold all of ONS' outputs in a standard format with agreed metadata. This is one of the foundation tools for the new infrastructure that will hold all data produced and used by the office, and interface directly with web based dissemination tools to provide a more integrated and harmonised set of outputs for customers.

6. Conclusion

This paper has described the ONS' approach to modernising and standardising its statistical and information management systems. Substantial benefits are expected from the programme. Some benefits will be realised in terms of greater operational efficiency, which may provide finance to be ploughed back into the business. Other benefits will be seen in the Office's ability to produce an increased range of products at a higher quality. Finally customers will see an improvement in a range of aspects of ONS' performance from increased speed of access to data, to more robust systems, and also an office better equipped to develop the new statistics required in the 21st century. Demonstrating the improvements in ONS' performance against an agreed set of indicators is a key part in convincing customers of the long-term benefits. This is an ambitious programme designed to deliver modern systems built in modules, so that future developments in methods and tools may be incorporated into ONS' Corporate Toolbox, whilst maintaining continuity in the business of the office.

Annex A:

The Statistical Value Chain