

**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

USING XML TO STORE DESCRIPTIVE METADATA

Invited Paper

Submitted by the Central Statistics Office, Ireland¹

I. Summary

1. At the last METIS Conference held in Luxembourg 2002, the Central Statistics Office (CSO) Ireland reported the initiation of a project to build a Corporate Data Warehouse, which would fundamentally alter the way that we function as an organisation.
2. This paper outlines one element of this project: the use of Extensible Mark-up Language (XML) to store descriptive metadata. Descriptive metadata could be described as textual or document based metadata.
3. Essentially the development of this XML application would be regarded as being 'proof of concept' and while it has attracted quite a good deal of favourable comment, it has not been officially approved to date.

II. Introduction

4. The most common, but least useful, description of metadata is 'Data about Data'. The problem with this definition is one of perspective i.e. metadata has a different meaning to different people and can comprise of individual elements not necessarily linked to a wider system for encapsulating the sheer depth and complexity of metadata.
5. A truer or more complete definition would be 'metadata means simply data about data, and refers to the definitions, descriptions of procedures, methodologies, system parameters, and operational results which characterise and summarise statistical programs.'² This is a much better or fuller definition but contains one crucial, but acknowledged, weakness i.e. there is a proliferation of

¹ Prepared by Richard Murphy and Rosarie O'Riordan

² Statistical Integration through Metadata Management Michael J Colledge OECD 1999

models, rules, standards etc but there is no definitive 'one best way' to gather all of the elements necessary for an integrated metadata management system.

III. General

6. One of the key problems identified very early in the project to build a Corporate Data Warehouse in the CSO was the fact that data was held in a variety of different formats in different databases and other storage facilities. Similarly, metadata was stored in all the databases, as well as in a variety of formats on PCs, the mainframe, in hard copy as well as in undocumented expert knowledge from subject area specialists. There was no integrated structure for metadata (core or descriptive) storage in the CSO.

IV. Background

7. There are 101 surveys carried out by the CSO, with each one being an island of both data and metadata. One of the first tasks that was carried out was the documentation of each survey detailing the data collected and the methodologies used to produce statistical results under approximately 53 headings and sub headings. This task was called the Business Process Improvement (BPI) Project and was carried out using a survey methodology questionnaire in Lotus Notes. (See Appendix 1)

8. The Business Process Improvement (BPI) survey was a very useful exercise for the CSO, and a very rich source of metadata, with a number of intentional and unintentional results. Our initial aim was to document each survey through the entire life cycle using pre-defined headings, so that all of the data holdings, metadata, methodologies, processes, applications etc were identified. Each step of every survey is now 'known' and documented and the old black box scenario no longer applies. The unintended consequences of the BPI were also very interesting and highlighted a number of issues that have a bearing on the work of the CSO. Items such as the lack of metadata consistency within and between surveys, metadata quality issues, the usage of explained/unexplained abbreviations and terminology, the lack of a Corporate Thesaurus to store explanations were identified among others.

9. The XML metadata repository was developed as a result of the BPI and is now known as the Corporate Business Process Metadata repository, screen shots of which will follow.

V. Project Initiation

10. Our long-term plan is to have an integrated metadata management system in place that will be web enabled and available to both internal and external users of CSO data and metadata. While the initial BPI survey was carried out using a Lotus Notes template, it was our view that XML was a more appropriate mechanism to further this goal. Our next task, therefore, was to build an XML application using the same headings that were in the BPI survey. The reason for using existing headings was that the metadata already supplied could then be transferred in to Sybase, where the data and core metadata will be stored. We then moved the metadata to the new storage mechanism by using a series of SQL commands that extracted the metadata between XML tags into the XML application.

11. A small team was put together and the XML application was built, tested and populated with metadata within 3 months. This included the time set aside for training and project familiarisation.

12. Prior to the beginning of the development, we held a 'brain-storming session', which identified quite a number of issues that would have to be resolved before, during and after the completion of the application. The main issues identified were Infrastructural, Operational, Content, Policy and the use of the Internet/Intranet. This session was useful because the issues identified served to focus the development effort and acted as milestones to measure progress during the development.

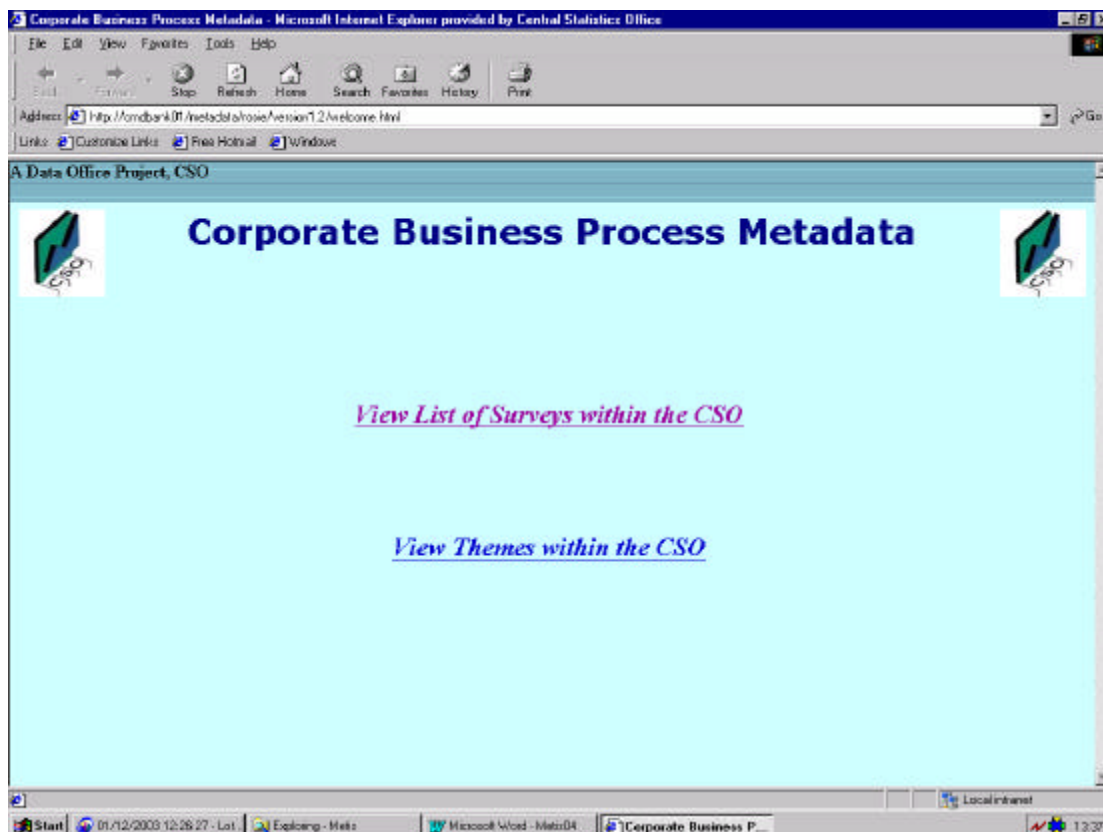
13. We took part in a 2-day introductory training course in XML, which familiarised us with the basic concepts and outlined how the various elements were referenced. We then purchased a number of XML reference textbooks and used both them and the Internet for problem solving during the course of the development.

14. We did not use an XML toolkit for this application. We wrote the code (HTML, XML), style sheets (XSL) and scripts (ASP) in MS Notepad, and used Microsoft Internet Explorer 5 (IE5) to access the application.

Overview of the XML Application

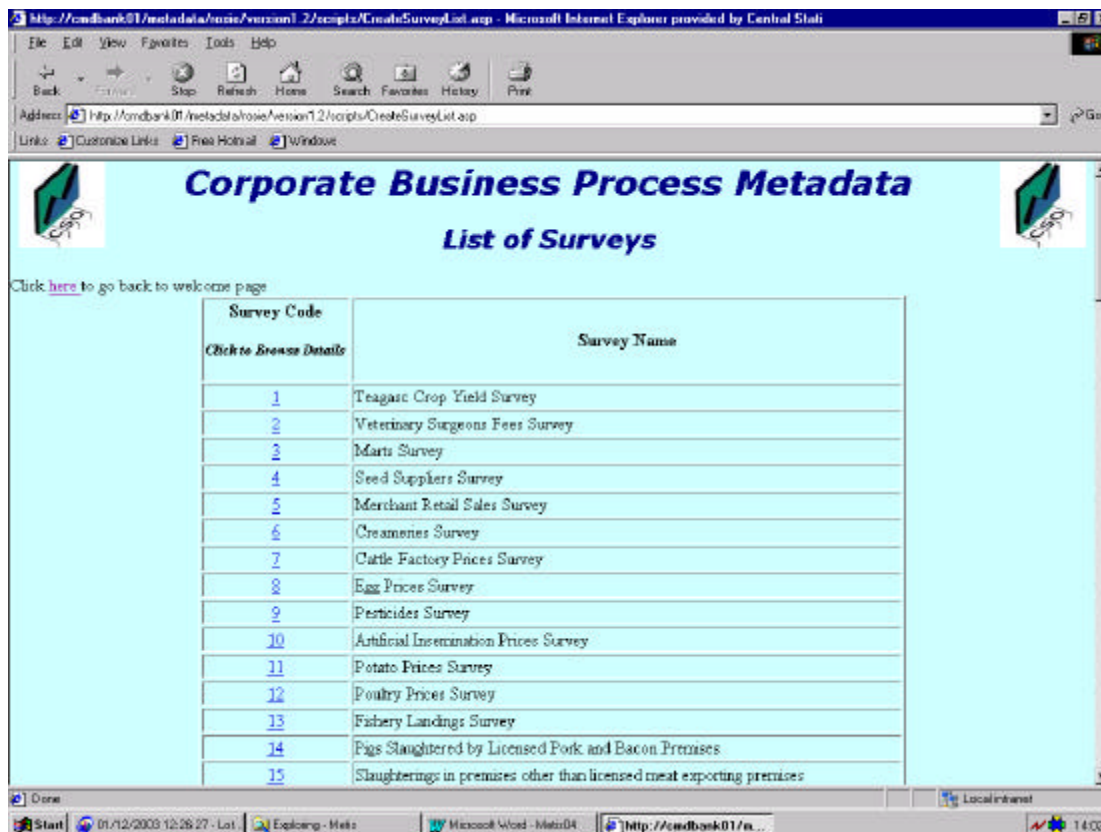
The following screens represent the main features of the application.

Screen 1 is the main menu screen and at the moment is restricted to two options i.e. View List of Surveys within the CSO and View Themes within the CSO. However, this can be expanded when necessary.



Screen 2 View List of Surveys within the CSO

The column on the left-hand side 'Click to Browse Details' is accessible to every user. There is an update facility built into the application. Each individual survey has three staff members with edit or update access. Only one person can edit the survey metadata file at a time. This update facility is currently suppressed but can easily be activated when required. A date stamp and system user id will be used to identify who is updating the file.



Among the tasks that the Active Server Page script file for this screen performs include

- Creates the survey list.asp
- Opens the welcome.html file
- Uses the browse test.xsl stylesheet
- Controls the Layout of the Page and the Table
- References the various objects, files and directories
- Displays the list of surveys
- Creates the Loop to enable users to scroll through the list
- Lists the variables including the Survey Code and Survey Name
- Closes the loop once all details are written out
- Destroys and de-references various objects within the file

Screen 3 Metadata Details Screen

This next screen is simply a screen shot of the browse access facility that is available to all users to see details of any particular survey and is accessed through the previous screen. A complete list of the fields is included in Appendix 1.

The screenshot shows a Microsoft Internet Explorer browser window displaying the 'Corporate Business Process Metadata' page. The page title is 'Corporate Business Process Metadata' and the URL is 'http://csdbank01/metadata/rosc/version1.2/scripts/ViewUpdateUseDetails.asp?MODE=BROWSE&USER=0'. The page content includes a note: 'Note: Only Data Office Personnel are allowed to edit this information.' Below this is the title 'Teagasc Crop Yield Survey - Details' and a table of metadata fields.

CSODivision	Agriculture Division
CSO Section	Prices Section
Survey Code and Name	1 Teagasc Crop Yield Survey
Purpose of Survey	To compile data on crop yields
Business Specialist No1	CSOCORK/walig
Business Specialist No2	CSOCORK/shuffy
Business Specialist No3	
Any relevant comment to EU Regulation	Council Regulation (EEC) No 2727/75 of 29 October 1975 Council Regulation (EEC) No 837/90 of 26 March 1990 Provides some information required by Agricultural Accounts and Agriculture Prices.
EU Regulation Codes	2727/75,837/90
Is your Survey Linked to CBR	--
Key Relations Comment	Directly related to the June Crops and Livestock Survey. It provides the yield information so that the production of the crops can be estimated.
Any relations to other Surveys	--
Other Administrative Information	--
Any Relevant CARS Codes	--
Any other Documentation	Procedural notes maintained within the sections

Appendix 3 shows a screen shot of a partial XML file for the survey identified above.

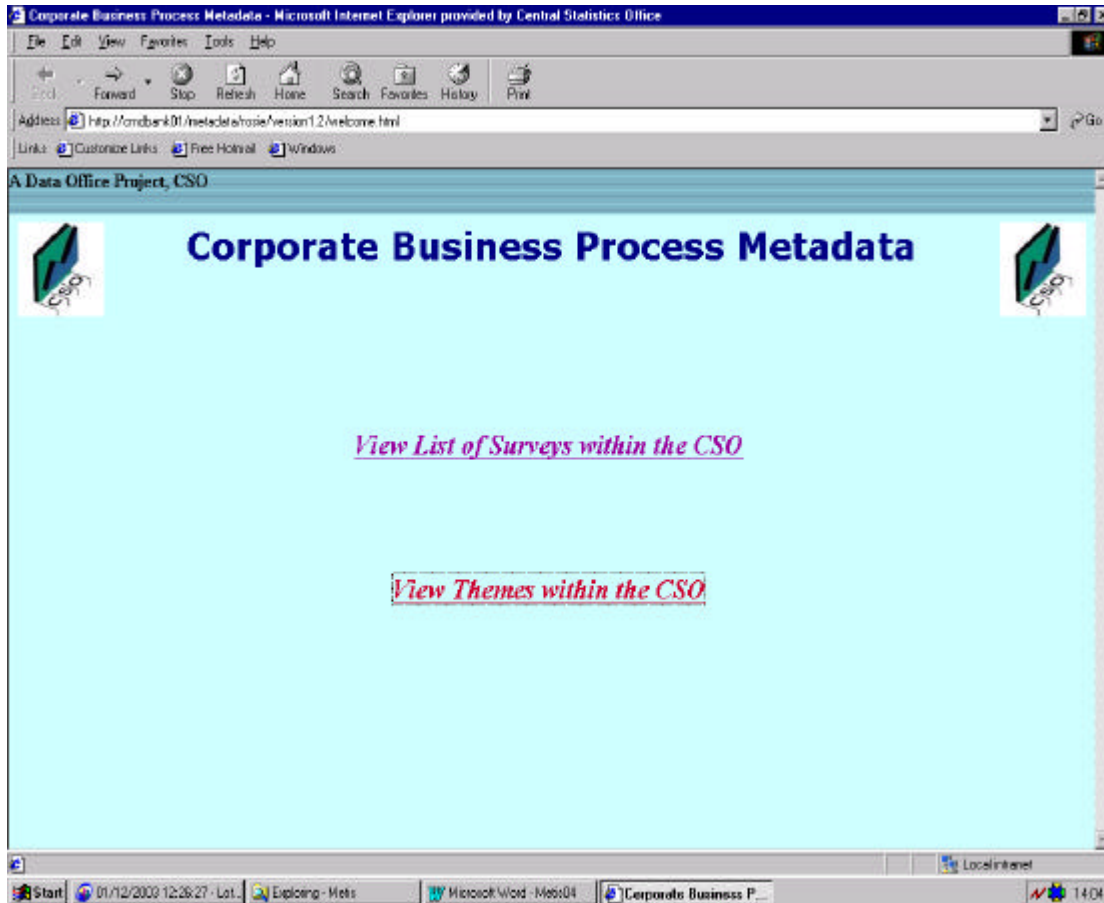
Screen 4 Metadata Details Screen (ctd)

This screen is the same as the previous one except that it gives a better idea how textual metadata appears and is stored in the application.

Sampling Procedures	For economic and operational reasons, the sample was first stratified to distinguish between town areas (i.e. 1,000plus inhabitants) and country areas (i.e. less than 1,000 inhabitants). A two stage sample design was used for the 1999/2000 Survey. This comprised of a first stage sample of 2,600 blocks (or survey areas) randomly selected at county level to proportionately represent the following eight strata: 1. County Boroughs 2. Suburbs of County Boroughs 3. Environs of County Boroughs 4. Towns 10,000 plus 5. Towns 5,001 to 10,000 6. Towns 1,000 to 5,000 7. Mixed Urban/Rural Areas 8. Rural Areas. Each block was selected to contain, on average, 75 households. In the rural areas, each block was further divided into 4 sub-blocks, each containing approximately 18 households. The rural survey areas were randomly selected from within these sub-blocks. The second sampling stage involved the random selection of two independent samples of 4 original households and 4 substitute households for each survey area. The number of original sample households constituted the quota to be realised in each survey area and the field interviewer systematically approached as many substitute households as was necessary to realise this quota. In this fashion, variations in response by region and town size were controlled. A feature of Household Budget Surveys since 1987 has been the integration of the Teagasc National Farm Survey (NFS) sample of farm households. These were accommodated by identifying and eliminating any non Teagasc NFS farm households in the original and substitute samples of households in each survey area. As the NFS households were distributed throughout the country (the sample was not clustered geographically), they were linked to the survey area in which they were located or to which they were most adjacent.
Measure to reduce non-response	Each member aged 15 years or older of a fully co-operating household was issued a pounds 10 gratuity payment and free entry to one of two prize draws with 1st, 2nd and 3rd prizes of pounds 5,000, pounds 3,000 and pounds 2,000 respectively for each prize draw.
Scrutiny of returns	
Data coding	All returns are coded by the Processing Section. This involves the keying of up to 480 descriptive codes and 1000 different expenditure and income codes.
Data capture procedures	Returned household files are keyed into ASCII files using the Viking System. See attached document for further information (SEE DOC LDNK).
Data editing/Quality control-correction	Phase 1 of the HBS Process consists of a number of options, which carry out consistency edits, validation and range checks. Included also are options which allow for the correction of invalid entries and the combination of valid batches of records to the master file. In addition, further checks are carried out after phase 2 and the file is updated with any corrections.
Other relevant information for data collection and preparation	--

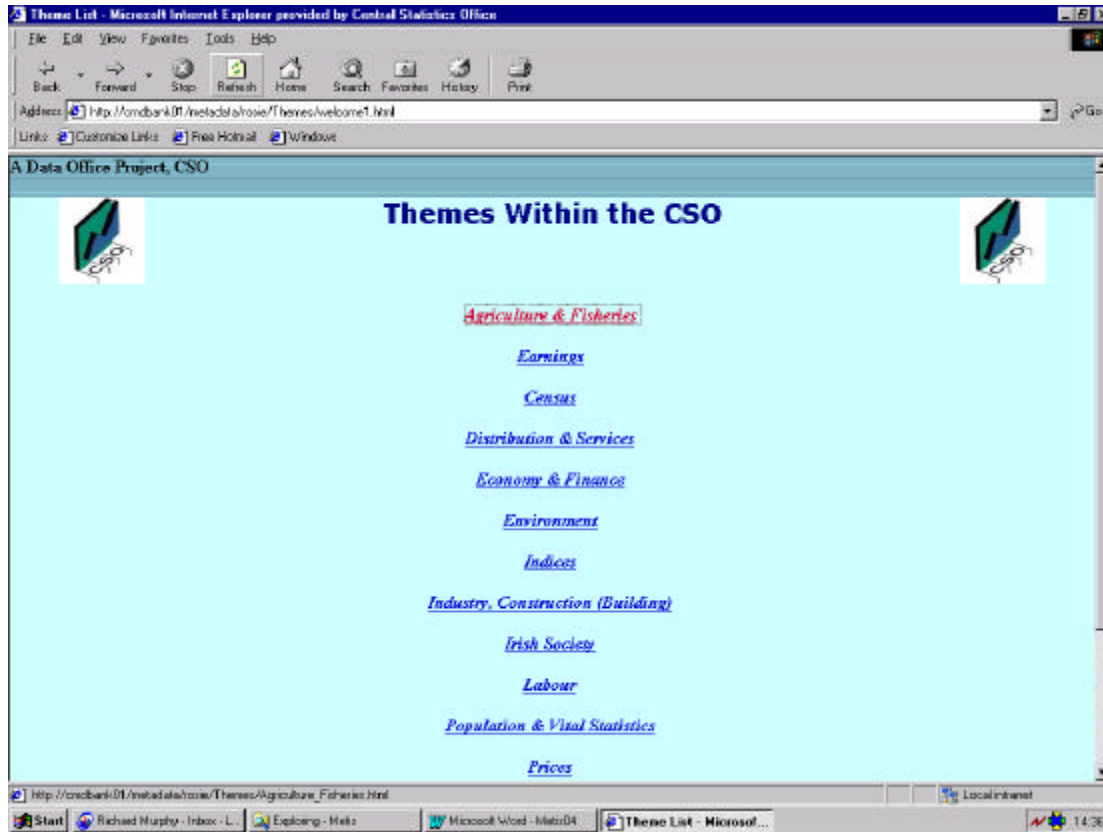
Screen 5 Main Menu

Screen 5 is the main menu screen again



Screen 6 View Themes within the CSO

This is the second option on the main menu and outlines a potential corporate theme list for the CSO. This theme list could function as an independent module and act as a portal into the metadata repository if necessary.

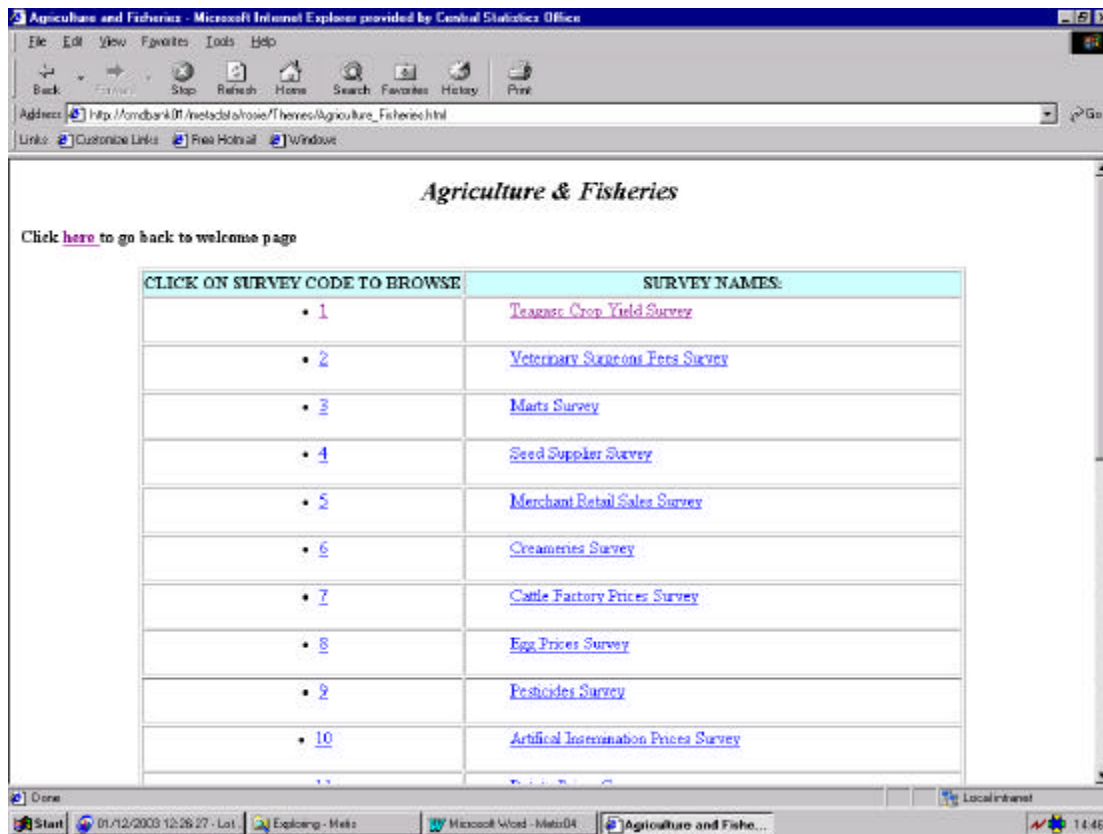


Among the tasks performed on this screen include

- Opening the welcome.html file
- Listing the 14 HTML files that make up the themes

Screen 7 View Themes within the CSO

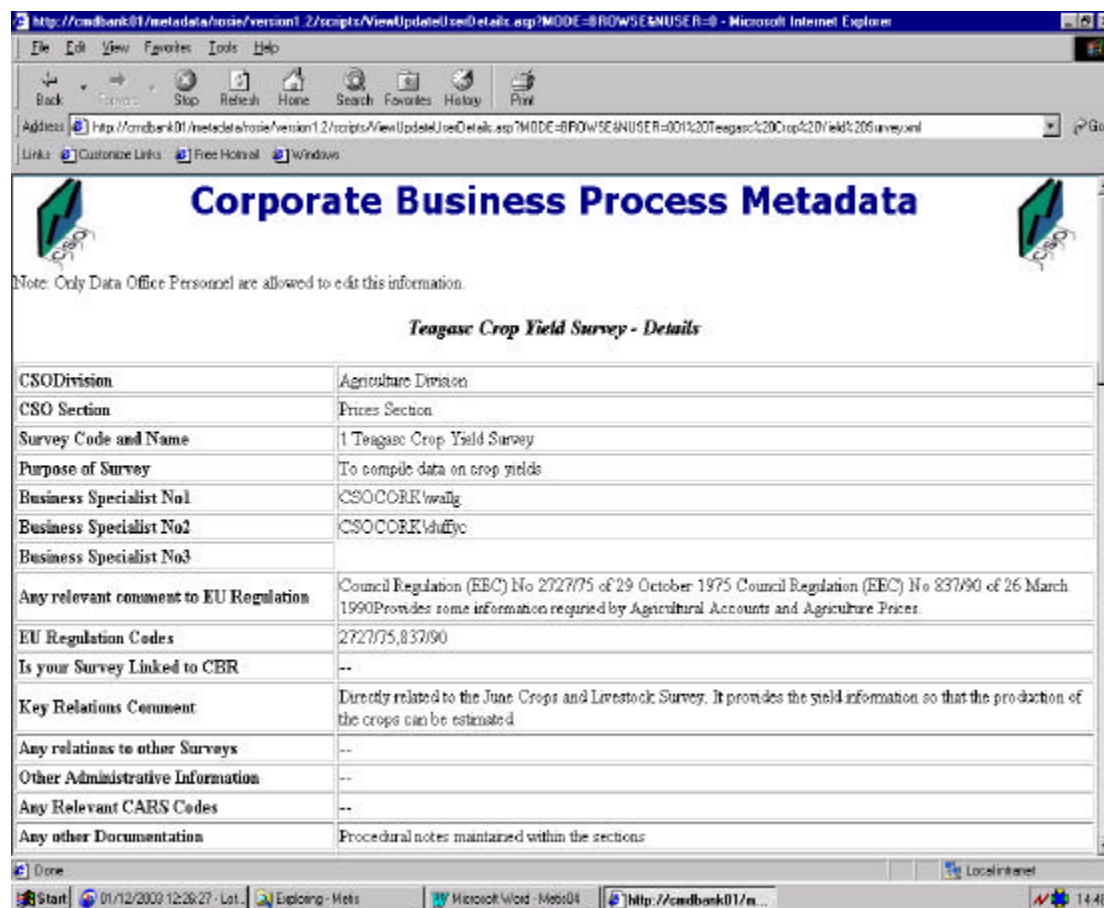
This screen shows the surveys available when the Agriculture & Fisheries option is selected on the previous screen. Both the survey number and the survey name are enabled, so users can access the next level by clicking on either option.



There is no controlling Active Server Page script file for this screen as the user simply clicked on an HTML file to access it.

Screen 8 Metadata Details Screen

This screen is a screen shot of the browse access facility that is available to all users. It is the same as screen 3 and references the same file, however, it is accessed from a different vantage point.



VI. Lessons Learned

15. We put a team together to undertake the development but we did not have an expert in any of the web programming languages on the team. Our lack of experience in XML meant that we literally started with the basics. Occasionally we had difficulty describing or defining the problems that we were encountering before we could look for possible solutions. This slowed down progress at times, but was a very good learning experience.

16. We wrote all of the code from the beginning using MS Notepad. We did not have access to an XML Toolkit, which would have made the task much easier. Access to an XML Toolkit would have allowed us to write code quicker, to trouble shoot and problem solve in a more coherent and timely manner.

17. We initially kept the functionality of the development to a minimum. The screens were kept reasonably clutter free, with a simple layout that was easy to manage and navigate. We used style sheets to fill in the design detail such as colour, font and to reference the content. The same colour and font detail was used in every screen, where possible, for consistency in the 'look and feel' of the application. Corporate GUI standards are being drawn up for all CSO applications and will be uniformly applied.

18. Learning to walk before we could run. While the development went reasonably smoothly (if slowly at times), there was a natural inclination to try and build in extra functionality as the development progressed, instead of proving the concept and subsequently refining the application when time, resources and feedback allowed.

VII. Strike a Balance

19. While we have identified other development modules that could be integrated into the XML application, we are also mindful of the need to strike a balance between strategic and operational pressures, between increased complexity and user friendliness. We must maintain the user friendliness, ease of navigation and simplicity of the application and be aware of the demands made on data production staff to maintain and update a large amount of metadata as part of the survey process. Demands for too much added functionality and linkages to other applications will make it less user friendly, will reduce response times and make it more complex to maintain both in terms of software programming and metadata content.

VIII. Future Plans

20. This application has been developed but not yet approved for availability to our internal users. All of our future plans are predicated on this application going live in the short-medium term.

21. Our starting point for this application was simply to prove that we could build something that was functional, user friendly and simple to use. We did not pay particular attention to the proliferation of XML standards for web services although there are some standards used within the application. Also, the application is rather basic in terms of screen layout, use of icons, search facilities, hotspots, colour etc and does not conform to the CSO's corporate standards for our web site. Therefore, one of the first tasks will be to re-design the front end of the application to give it the corporate look and feel.

22. Having proved the concept, we would also need to pay closer attention to the XML standards for web services and to incorporate them where necessary.

23. There is an 'updating' function in the application, which is currently suppressed. This piece of functionality will allow three designated survey specialists to update the metadata on an ongoing basis.

24. The way in which the application was developed ensured that it was scalable to incorporate any changes or additional fields subsequently requested. During the development phase we identified a number of existing headings that could be expanded and a range of other new headings under which we could collect metadata (See Appendix 2). During the testing phase we included these extra fields (55 in number) in the application without any difficulty, or without any deterioration in the performance of the application. This can be repeated if necessary.

25. We recognise that we must integrate this XML application with the core data and metadata, which will be stored in Sybase. By doing this we will be able to seamlessly associate the descriptive metadata held in XML with the data and core metadata held in Sybase to give the complete picture.

26. We are considering introducing a version control and archiving facility to the application. This would allow us to rename XML files with a version number and date field. The older versions could then be archived off in a structured manner to another area of the server. Archived files could be made available for browsing to all interested users.

27. We have also developed an XML prototype of a Statistical Release Calendar (see Appendix 4). By using the same survey numbers from the metadata storage application, we can access the textual metadata that gives the background details of each survey. Essentially we are trying to make the metadata readily available from many different access points within and outside the production cycle, all of which will refer back to the same single source i.e. store once but reference many.

28. We have identified a number of other independent development modules capable of being integrated into the application, but which could also function on a 'stand alone' basis. These include a Corporate Theme List (already shown) to act as a portal to the descriptive metadata, a Publication Catalogue, a Module for Frequently Asked Questions and a Glossary of Terms/Thesaurus.

IX. CONCLUSION

29. Initially the major objective of this application was simply to prove the concept. The application was built quickly, and tailored to our own needs. It is a relatively simple application, it is scaleable for further development, and can be put on both the CSO's Corporate Intranet and Internet reasonably quickly when required. Solutions for problems do not have to be highly technical, complex and expensive. The best solutions are often simple, inexpensive and fill an identifiable need.

30. One of the consequences of carrying out this project was that it highlighted the inconsistent use of terminology. People use common terms but with different meanings. This implies that what is readily understood by one user can be easily misunderstood by another. This issue was raised by Dan Gillman³ in a presentation that he gave at AMRADS in Ljubljana in March 2003. XML can facilitate the resolution of some of the problems identified here by simply acting as a repository for all of the metadata to be gathered into one place and making it accessible and visible to all users. The lack of coherence or consistency within the metadata will become obvious and can be improved upon as problems are identified.

31. There is a proliferation of models, rules and standards that are widely available within the metadata area and all of which serve a useful, but specific, purpose. However, to repeat something mentioned in the introduction to this paper, there is no definitive 'one best way', which leads me to question whether there is a need to have a definitive 'one best way'? My background is as a Systems Analyst⁴ and my training inclines me to document everything and to apply rules and standards where possible. More recently I have been working on a large-scale migration project (data and process) and my perspective has changed somewhat. The rush to try and standardise can, (especially where standards are being rigidly applied in unsuitable circumstances), lead to further problems or complications at a later stage. Rules and standards must be 'sympathetic' or suitable with the circumstances to which they apply.

32. When we undertook this development our goal was to have something useful and user friendly. The application was developed in accordance with our own corporate needs rather than adhering to a rigid set of guidelines that might have steered the development in an unintended direction.

33. Perhaps the time has come to start shaping metadata rules and standards to fit corporate needs rather than bending application needs to fit metadata standards.

³ Terminology and Metadata AMRADS Training Workshop on Metadata March 2003

⁴ The views expressed here are personal and do not purport to reflect official CSO Corporate policy

Appendix 1

BUSINESS PROCESS IMPROVEMENT PROJECT Description of statistical survey process

I. Administrative Information

CSO Division
 CSO Section
 Business Specialist(s)
 Purpose of the survey
 Form in which survey conducted
 EU/national regulations or requirements
 Links to Business Register
 Links to CARS
 Key dependencies on, and relationships with, other surveys
 Documentation
 Other relevant information

II. Survey Inputs and Outputs

Periodicity of survey
 Size of survey
 Target population
 Principal variables collected
 Survey outputs
 Users
 Other relevant information

III. Data Collection and Preparation

Sampling frame
 Sampling procedures
 Measures to reduce non-response
 Scrutiny of returns
 Data coding
 Data capture procedures
 Data editing or other quality control/correction
 Other relevant information

IV. Survey Data

Number of numeric variables
 (Micro) Dataset codes
 Other relevant information

V. Statistical Processing and Presentation

Non-response methodology
 Description of aggregation (grossing, imputation, estimation etc)
 Macro edits or consistency checks
 Seasonal adjustment
 Index numbers
 Treatment of statistical confidentiality
 Presentation and dissemination procedure
 (Macro) Dataset codes for aggregated results or statistics ready for dissemination
 Time lag between the end of the survey period and dissemination
 Other relevant information

VI. IT Systems and Processing Software**Systems****Software**

Sampling frame maintenance/manipulation

Sampling

Data entry

Statistical processing

Tabulation, presentation and dissemination

Other (e.g. for ad hoc reports, query handling,
electronic data transfer)

Other relevant information

VII. Remarks(e.g. anticipated changes to survey design,
planned new surveys, any other remarks)

Appendix 2

Key: Black – New Fields

*Red (italics)– Expand existing Fields*Blue (underlined) – New Fields Modular Development**Other Information fields for XML Application****1. At the introduction after Survey Code and Survey Name**

- Status of Survey

2. Administrative Information

- List of Tables/Datasets from each survey
- Survey Span (first instance/last instance)
- Pre-Release Arrangements
- Copyright Declaration

Expand Business Specialist to include

- *CSO Contact Details*
 - *Section Name*
 - *Address*
 - *Data Custodian*
 - *Telephone*
 - *Fax*
 - *Email (group email)*

Modular Development (stand alone)

- Release Calendar (weekly/4 monthly)
- Publication Catalogue

3. Survey Inputs + Outputs

- Sources (primary + secondary)
- Frequency of Collection (including reference period)
- Geographical Coverage
 - Coverage not a classification
 - Specific Exclusions (areas)
- References (internal and external)
- Disclosure Control (link to confidentiality issues)
- Access Control (Relevant when the Disseminate DB is available to Public)

No metadata currently available on Survey/Questionnaire design

- Questionnaire Design
- Questionnaire Type
- List of questions (Maybe link to PDF version on web)
- Instructions to Respondents/Interviewers
- Comments by Respondents/Interviewers
- Trialling of Questionnaire

Expand Target Population to include

- *Specific Exclusions (groups)*

4. Data Collection and Preparation

- Measurement Unit

Expand Sampling Procedures to include

- *Sample Design*
- *Sampling Method*
- *Sampling Unit*

5. Statistical Processing and Presentation

- Other Adjustments (listed after Seasonal Adjustment)
- Weightings Used
- Revisions
- Rebasing
- Discontinuities
 - Code Change
 - Change in Data Structure
 - Change in Methodology
- Sampling Errors
 - Non-Sampling Errors

Expand Data Editing to include

- *Editing Rules*
- *Data Entry*
- *Data Checking/Validation*
- *Consistency Checking*

List descriptions of aggregation separately

- *Aggregation*
- *Grossing*
- *Imputation*
- *Estimation*

6. Remarks

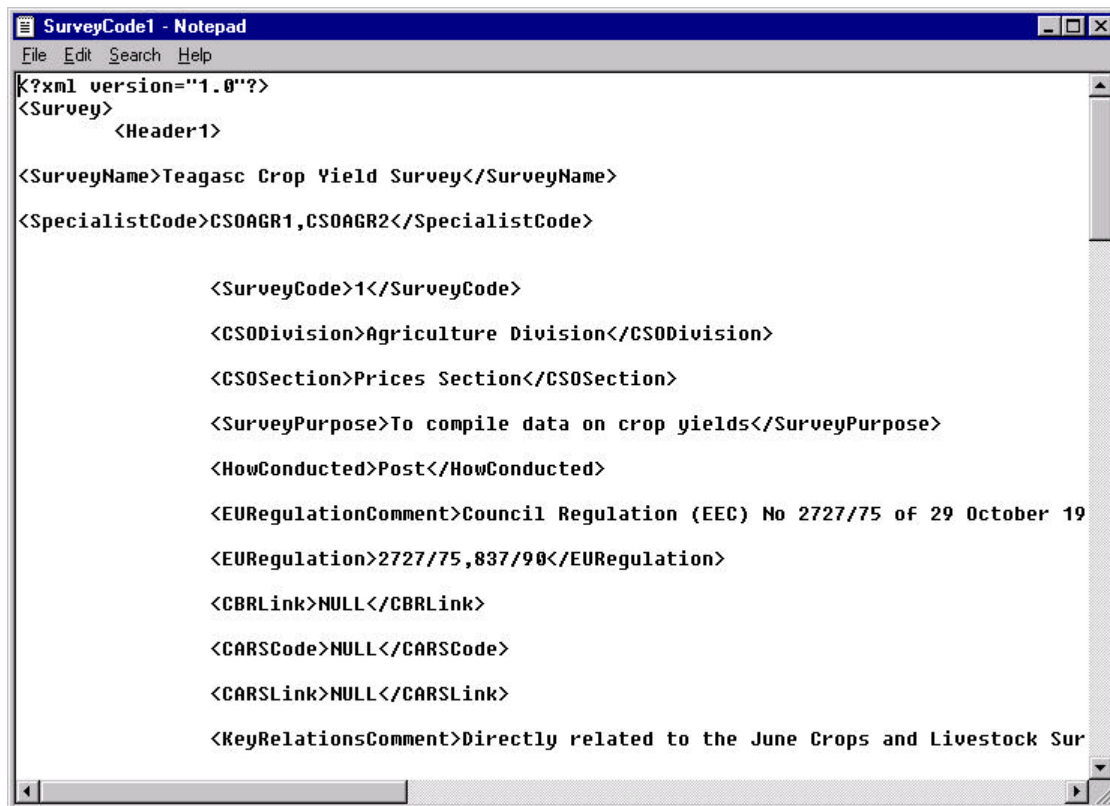
- Event Log

Modular Development (stand alone)

- Frequently Asked Questions
- Glossary of Terms/Thesaurus
 - Abbreviations
 - Symbols
 - Acronyms

Appendix 3

Example of an XML file.



```
SurveyCode1 - Notepad
File Edit Search Help
<?xml version="1.0"?>
<Survey>
  <Header1>
<SurveyName>Teagasc Crop Yield Survey</SurveyName>
<SpecialistCode>CSOAGR1,CSOAGR2</SpecialistCode>

    <SurveyCode>1</SurveyCode>
    <CSODivision>Agriculture Division</CSODivision>
    <CSOSection>Prices Section</CSOSection>
    <SurveyPurpose>To compile data on crop yields</SurveyPurpose>
    <HowConducted>Post</HowConducted>
    <EURegulationComment>Council Regulation (EEC) No 2727/75 of 29 October 19
    <EURegulation>2727/75,837/90</EURegulation>
    <CBRLink>NULL</CBRLink>
    <CARSCode>NULL</CARSCode>
    <CARSLink>NULL</CARSLink>
    <KeyRelationsComment>Directly related to the June Crops and Livestock Sur
```

Appendix 4 Release Calendar

The screenshot shows a web browser window with the following content:

LIST OF RELEASES	LIST OF ASSOCIATED SURVEYS:
Transport and Tourism <ul style="list-style-type: none"> • Vehicles Licensed for the First Time (Monthly) • Vehicles Licensed for the First Time (Annual) 	
Transport and Tourism <ul style="list-style-type: none"> • Statistics of Port Traffic 	92 Statistics of Port Traffic Survey
Transport and Tourism <ul style="list-style-type: none"> • Tourism and Travel (Quarterly) 	96 Passenger Card Inquiry Survey 97 Country of Residence Survey
Transport and Tourism <ul style="list-style-type: none"> • Tourism and Travel (Annual) 	96 Passenger Card Inquiry Survey 97 Country of Residence Survey
Transport and Tourism <ul style="list-style-type: none"> • Household Travel Survey (Quarterly) 	98 Household Travel Survey

The browser window title is "H:\Rosie\Advance Calende\Calendar.html - Microsoft Internet Explorer provided by Central Statistics Office". The address bar shows "H:\Rosie\Advance Calende\Calendar.html". The taskbar at the bottom shows several open applications: "Start", "Richard Murphy - Inbox - L...", "Exploring - Advance Calen...", "Microsoft Word - Main04", and "H:\Rosie\Advance C...". The system clock shows "16:50".

Bibliography

Statistical Integration through Metadata Management Michael J Colledge OECD 1999

AMRADS (EU 5TH Framework Project)

Accompanying **M**ea**s**ure to **R**esearch and **D**evelopment in Official **S**tatistics)