WP. 21
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, 3-5 April 2006)

Topic (iii): Metadata and Statistical Cycle

QUALITY INFRASTRUCTURE SYSTEM - A CASE STUDY OF AN E2E APPLICATION AT THE ABS

Supporting Paper

Submitted by Graeme Oakley, Australian Bureau of Statistics
[graeme.oakley@abs.gov.au]

I. INTRODUCTION

1. The Australian Bureau of Statistics (ABS) has the objective, as part of its electronic publishing of metadata vision, to produce information about the quality of a statistical dataset alongside the actual data. The information about the quality of the data would follow a framework similar to that used by the OECD eg relevance, accuracy, coherence etc. It would have a descriptive component (text, relatively static) and metrics such as response rates of the survey, standard error of main variables (numeric, changing with each production cycle).

2. This case study explores the architecture of the Quality Infrastructure System (QIS) to capture the data that contributes to the quality metadata metrics from various processes through the statistical cycle. This data is associated with descriptive metadata from the Collection Management System to create a Quality Declaration which is associated with statistical products published to the web site. This development is still a work in progress,

however, several components are in production and deployment of back-end processes into surveys has commenced.

II. OBJECTIVES OF THE QUALITY INFRASTRUCTURE SYSTEM

4. The QIS has its origins in a project called 'Making Quality Visible' which aimed to help ABS progress its mission to "provide a high, quality, objective and responsive national statistical service". The 'Making Quality Visible' focus has four main objectives. Firstly, the ABS is committed to improving dissemination by increasing the information available externally about the quality of ABS data. The second aim is to educate users about the different dimensions of quality and how to use information on data quality to make informed use of data. The third objective is to publish and promote guidelines and framework about quality. The fourth objective is to use information about data quality internally to manage and improve statistical processes.

5. The Quality Infrastructure System (QIS) is primarily focused on the first and fourth objectives, by providing an integrated architecture to define, collect, store, access and analyse quality measures. The objectives are summarised as being to:

- understand the performance of statistical processing functions
- understand the quality of statistical outputs
- enable methodologist and subject matter specialists to drill down to understand why performance, or quality, is what it is
- manage quality and statistical processes more efficiently
- deliver to users (either internal or external) reports on performance and quality that *Make Quality Visible.*

6. It is envisaged that these objectives will be met by creating a repository of measures integrated with the Corporate Metadata Strategy. The system will:

- automatically capture quality measures
- store quality measures
- enable desktop viewing of quality measures
- enable dissemination of quality measures both within and outside the ABS.

7. The implementation of the repository will be achieved by:

- defining the set of interactions that must occur with a repository of quality and performance measures
- defining and building Application Program Interfaces (APIs) to support those interactions
- populating it with performance and quality measures (for both within cycle and end of cycle measures) that have been defined by one authority
- defining metadata for quality measures. It is envisaged that a basic set of measures, aimed at clearance documentation and quality statements, will be defined by Methodology Division but that others may be added by users for their own purposes.

8. The information stored in the repository will:

- be equally applicable to business surveys, household surveys, the population census, and indeed any other collection operations

- have useable interfaces for a range of purposes, especially exploring, viewing, presenting, and analysing measures
- be seamlessly interfaced with existing or forthcoming IT tools, as well as other corporate metadata repositories
- be supported by definitional metadata unambiguously defining quality data items and active metadata.

III. SOME DETAILS OF THE SYSTEM

9. The QIS project has a strong metadata focus. While quality measures are a class of metadata in themselves, they require their own definitional and operational metadata. The Methodology Division acts as the registration authority for standard quality measures. This aids in consistency and comparability across collections. Subject Matter areas are still able to define their own custom measures.

10. The key principles of the ABS Metadata Management framework that are invoked in this project include:

- requirement for a 'registration authority' for each metadata element, which is the single authoritative source;
- reuse metadata; and
- capture metadata at source.

11. Metadata related to quality measures that needs to be stored:

- definitions of quality measures
- operational metadata enabling derivation of derived quality measures
- quality measures required to be collected
- levels at which quality measures should be collected
- frequency at which quality measures should be collected
- content of customised reports (including what quality measures to include, when to report, who to report to)
- structure and content of clearance documentation
- registration authority details.

12. Central to the development of QIS was the production a repository for the storage of quality measures. The repository is built around a “star schema“ data warehouse. This model allows for a variety of different types and levels of quality measures to be stored for both economic and household collections. A star schema is composed of a central “fact table”, a database table that contains the observed values of all quality measures and series of “keys” that link each quality observation or “fact” to a number of different “dimensions”. These dimensions are themselves database tables, containing the keys and classificatory details. Dimensions used in the QIS repository include: quality measure (e.g. response rate, standard error), geography (e.g. state, part of state) and industry. Wherever possible standard classifications are used, for example the industry dimension uses the Australian and New Zealand Standard Industry Classification (ANZSIC). The repository is built using the Oracle rdbms.

13. A 'services' architecture has been employed. Two different service types have been developed to store quality measures. The PutQM service enables existing systems (such

as provider management system or estimation system) to send quality measures directly to the QIS Repository. Existing donor systems undergo minor modification to enable them to send quality measure mail messages to the ABS business process management infrastructure. A "subscription service" monitors this process management infrastructure and automatically loads quality measures to the repository as they arrive.

14. The LoadQM services are the services that source quality measures from existing systems and load them to the QIS Repository. LoadQM services, independently from existing systems, access data stores these existing systems create and load these measures to the QIS Repository. These services do not require any re-engineering of donor systems. The PutQM services are preferred as it is based on a single integrated system.

15. On-line analytical products, such as Oracle Discoverer or SAS, can access this repository. Generic services have also been generated to allow standardised access to the repository. The StoreQM service allows quality measures to automatically copied from QIS to other parts of the ABS Corporate Metadata Repository. For example, the Collection Management System is a centralised store that records details on the concepts, sources, methods, timing, collection procedures and output of each ABS collection. The StoreQM service allows for the relevant quality measures to be included in this documentation.

16. The GetQM service is a standard service that allows analytical systems to specify the quality measures required and retrieves these from the QIS repository. These analytical systems allow the user to explore the quality measures, drilling down to finer levels of details or making comparisons across data collections and collection cycles. These services again use the ABS business process management infrastructure to access the quality measure repository.

17. Diagram 1 illustrates the system.

18. How does the dissemination end of the statistical cycle work? Diagram 2 shows the production of the Quality Declaration (QD). The reader should note that ABS does not currently produce and publish a QD - that is work in progress. We still have to finalise the content of the QD in terms of the static text (eg related to relevance, coherence) and the metrics that vary from collection cycle to cycle. Once that is completed then the development of a QD template and product production process will follow a model that we already use for our Directory of Statistical Sources.

Diagram 1. The Quality Infrastructure System

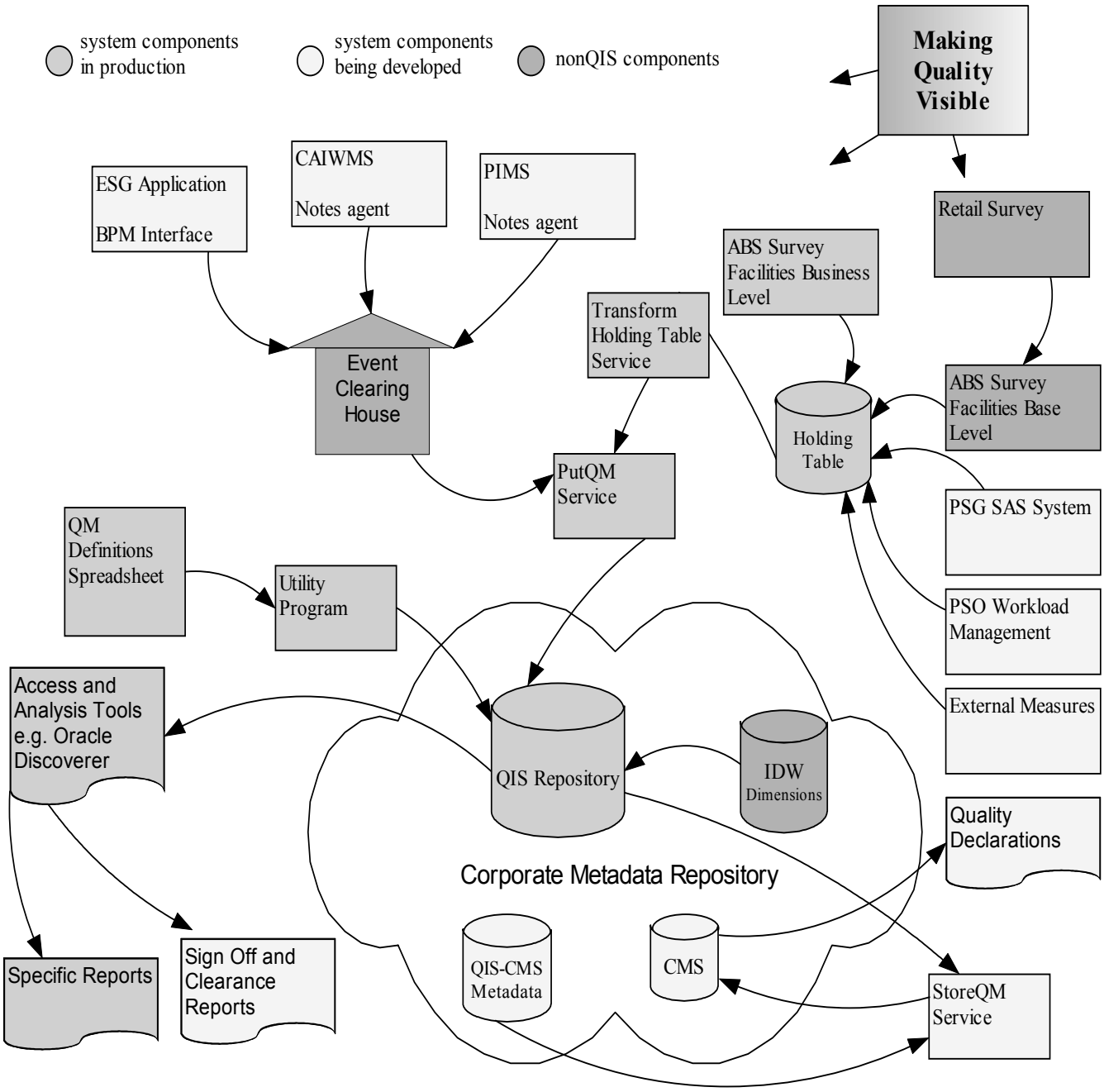
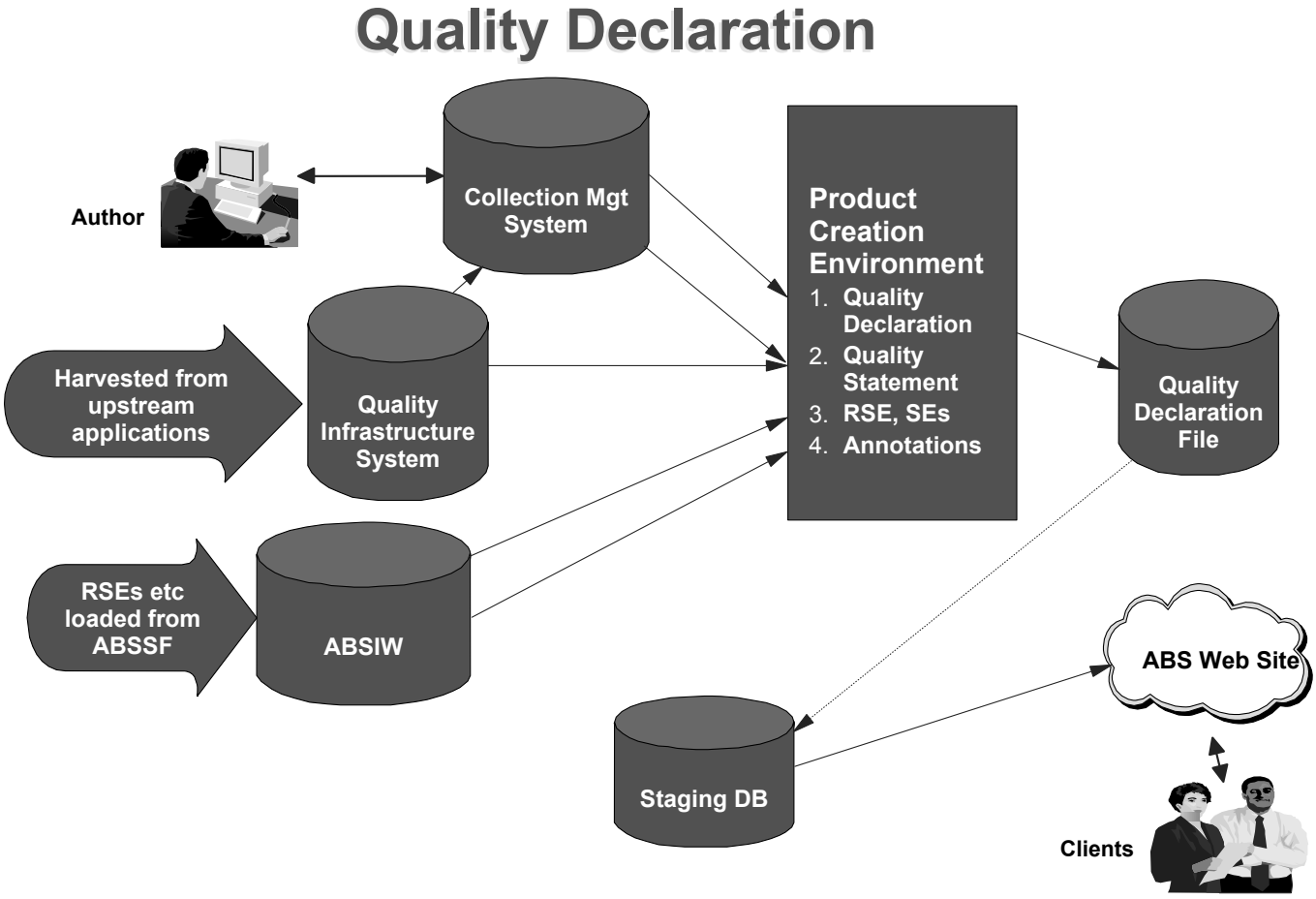


Diagram 2 Production of Dissemination Products



IV. CONCLUSION

19. What are some of the issues that are being addressed by this project? They are probably a microcosm of the metadata issues faced by any project that spans a large number of different collection domains and types, being both cultural, people and technical. Some examples of these issues are:

- agreeing on definitions for quality metrics across a large number of subject matter survey areas. Historically, collection areas have developed their own management information system and the definition of data elements is not always consistent, although the same data element name may be used. ABS Methodology Division is attempting to implement standard definitions and terminology, backed up by a 'registration process'. This process is a lengthy negotiation to establish agreement and buy-in, and has very little to do with the tools and systems that are provided. This is a common metadata management issue that requires time, persistence, senior management endorsement for the objectives of the project and good change management processes.
- deployment and timetabling issues. Deployment of the quality infrastructure system has to fit in with the 'business as usual' processing cycle ie not interfere with the core work of the collection. Again a good change management process is needed because the collection area may not perceive any direct benefit to them. The benefits accrue to others eg methodologists who will eventually only

have to work with one system to access quality information from many surveys. ABS is expecting a lengthy process for deployment in order to take into account survey processing schedules - could be up to 18 months to establish the bulk of the collections into the system.

- differences between internal and external clients with respect to quality metadata. The quality metadata required for association with disseminated products is a subset (and possibly derivation) of the basic observation data collected into the quality metadata repository through the statistical processing cycle. Internal users such as methodologists and survey managers are most interested in the detail and being able to drill down. This leads to tensions in terms of focus on the development of facilities (effort devoted to internal requirements ahead of dissemination) and the preparation of descriptive and analytical text (comes second to internal processes).
- efficiency and size of the database, throughput. The infrastructure could produce many data observations during a collection cycle when all the end-to-end processes are instrumented. The database could become large and so careful management by the database administrator will be necessary to ensure that updating efficiency and access efficiency are given appropriate weight. Also, the architecture based on services and an 'event clearing house' could lead to significant network traffic which means attention to throughput will be needed.
- security. Probably the major issue to be considered because the quality information being collected will include early estimates with standard errors, ahead of the official embargo time. Therefore, only a limited number of officers should have access to the quality data. The quality infrastructure system repository needs a robust security model which ensures that the survey manager controls who can have access to sensitive quality measures, and BI tools are set-up to prevent access. [Note that only ABS staff could possibly access the repository because it is on the internal network, but our enterprise risk management strategy requires a 'need to know' approach to access to embargoed statistical data.
- appropriate allocation of resources to areas that have to instrument systems. Availability of resources is always an issue. In this case, many processes in the statistical cycle could require instrumentation, although the increasing use in the ABS of generalised processing engines means that instrumenting one processing engine can pick up many surveys. Still resources are required to do this work and then the survey area will need to specify some 'process metadata' to 'drive' the particular process for their survey eg specify details for some core quality metrics and decide if they want to define collection specific metrics.

20. The Quality Infrastructure System has a number of components in production, the core (phase 0) quality metrics have been agreed and registered, and instrumentation is starting in a few of the processes that service a number of surveys, such as the ABS Survey Facilities (providing estimation and imputation service), and the Economic Surveys Data Centre (providing measures related to population and sample sizes, responses etc). Work has commenced in defining requirements for internal reports, such as 'clearance documents', and the external dissemination products.