INTEGRATING METADATA WITH SURVEY DEVELOPMENT IN A CAI ENVIRONMENT

Michael Colledge, Fred Wensing, and Eden Brinkley
Australian Bureau of Statistics

ABSTRACT

Metadata management is increasingly recognized as a vital component of the statistical process. This paper discusses the integration of metadata in the survey development process with reference to two major and related re-engineering initiatives at the Australian Bureau of Statistics (ABS).

The first initiative focuses on the management of data and metadata, primarily through the construction and use of a corporate information warehouse. The warehouse provides the metadata basis for the design, development, operation and evaluation of survey processes. It facilitates formulation, standardization, storage and selection of concepts, definitions, and procedures, and it enables analyses and comparisons of datasets. It also provides an output oriented, one stop statistical data shop, with facilities for storage, manipulation, extraction, and dissemination of the ABS's entire range of output data holdings together with the metadata relating to their origin.

The second initiative involves the redevelopment of facilities for collecting and processing household survey data, using computer assisted interviewing (CAI), and taking advantage of the warehouse as a development tool. Survey designers will make use of warehouse metadata to establish concepts and data items, knowing which have been endorsed as "standard", thereby facilitating data compatibility across surveys. The warehouse will also store sufficient metadata in the form of question texts, response categories and routing instructions to support the development of data collection instruments. In the CAI environment, these metadata will be used to generate program statements for the collection instrument and to define the survey output structures into which data are loaded when received from the field.

KEYWORDS

Data Management, Computer Assisted Interviewing, Re-engineering, Metadata, Information Warehouse, Data Warehouse, Statistical Integration, Questionnaire Design

1. Introduction

This paper is concerned with "data management", defined as the management of data and metadata through all stages of the survey life cycle. The survey cycle, the corresponding transformation processes through which the data pass, and the metadata by means of which those processes are monitored and controlled are illustrated in Figure 1.

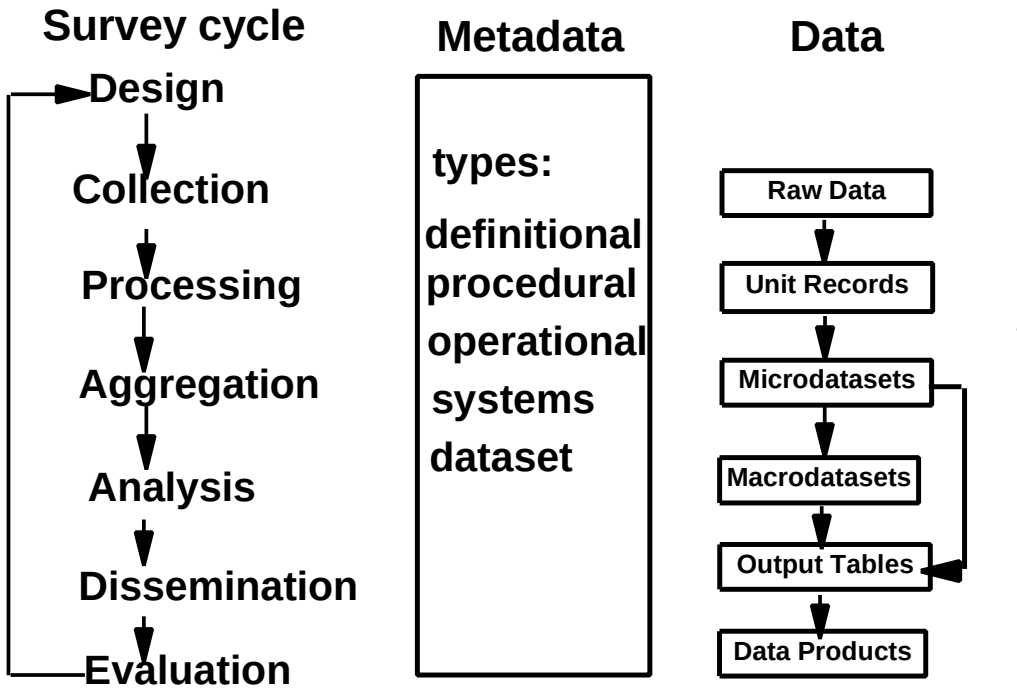

Figure 1: Survey Cycle and Data Transformation

Raw data received from respondents, or administrative or other sources, are converted:

- by derivation, editing and imputing from unit records into clean, complete "micro" datasets;
- by aggregation and estimation (or weighting) into aggregate "macro" datasets; and
- by selection and formatting into output data tables.

Finally, they are combined with explanatory notes and analytical commentary into data products.

For a single survey conducted in isolation from others, data management is conceptually straight forward. For programs of many surveys within national statistical agencies, it is more complex. Figure 2 illustrates the typical situation, indicating basic data flows and storage for four surveys. It reflects the likelihood that the surveys have been developed independently of one another and operate semi-autonomously, with little or no conceptual and physical integration.

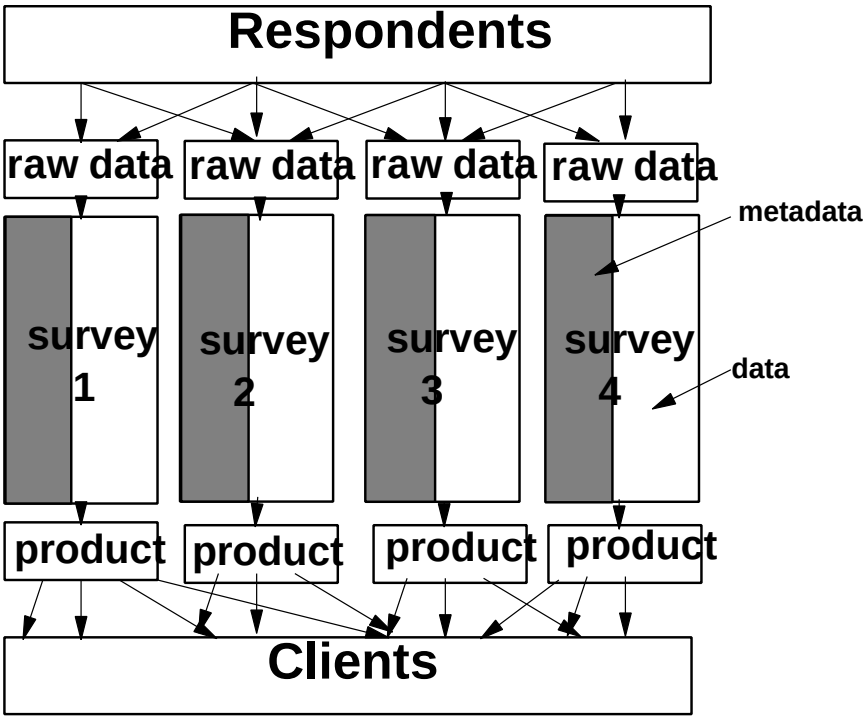

Figure 2: Typical Data Flows and Storage

There are a multitude of data management problems arising from lack of integration of concepts, functionality and data. First, respondents are obliged to reply to survey questionnaires or interviews in which data are collected according to different concepts and sometimes duplicated, thereby increasing respondent burden. Second, there is considerable duplication of data processing and dissemination across the range of surveys, leading to an overall inefficient use of operational and systems maintenance resources. Third, the resulting output datasets can not easily be combined and disseminated jointly due to conceptual differences and physical systems problems. Fourth, clients have to deal with data products which may be mutually inconsistent, and they have to communicate with a range of different output systems and may not know the full extent of the data available. The situation is worse than illustrated in Figure 2 as statistical agencies conduct not four surveys but forty, or even four hundred, during the course of a five year period.

Over the last thirty years, the Australian Bureau of Statistics (ABS), along with many other national statistical agencies, have made substantial efforts - with mixed success - to foster integration through re-engineering survey processes and improving data management practices. For example, in 1969 the ABS embarked on the Integrated Census Program for Economic Statistics involving the standardization of statistical units, industrial classification and data concepts and the development of an integrated business register (Australian Bureau of Statistics, 1969). In an attempt to rationalise and standardise the collection of business data, many other agencies also introduced a centralized business register, for example Monsour et al (1976) describe the introduction of the Standard Statistical Establishment List at the U.S. Bureau of the Census. In 1985, Statistics Canada initiated a major project to redesign its business register and business surveys, introducing concepts and shared systems (see Colledge, 1986). There has also been a focus on the development and use of generalised processing systems capable of supporting the needs of a range of surveys. Examples are Statistics Netherlands's Blaise system (see Bethlehem et al, 1989 and 1994) and Statistics Canada's

Generalised Edit and Imputation System (see Kovar et al, 1991). However, integration remains an elusive goal; like continuous improvement it is more a way of life than a project with a completion date. This paper describes two ongoing major and related re-engineering initiatives at the ABS aimed in this direction.

The first initiative, referred to as the "Data Management" Project, now mid way through its third year of full scale activity, involves the design, development and introduction of data management policies, procedures, and systems. The project is breaking new ground in data modelling and taking advantage of the latest computing technology. The underlying thrust is to ensure that corporate data products are readily visible and accessible to clients, and are reliable and mutually coherent. This is being achieved through systematic recording and use of metadata, rationalization of concepts and procedures, and physical integration of output data and metadata in a central repository known as the Information Warehouse, or more briefly, the Warehouse.

The second initiative, known as the "Household Surveys Facilities" Project involves the redevelopment of systems and procedures for developing, collecting and processing household survey data in conjunction with the introduction of computer assisted interviewing (CAI) to all ABS household surveys by the end of 1997. The project will also promote use of the Warehouse as a development tool. Under revised procedures, survey designers will take advantage of Warehouse metadata functionality to define data items, classifications and other concepts, based on information from previous surveys. In particular, designers will have access to all the concepts that have been endorsed as "standard", thereby facilitating data compatibility across surveys. The Warehouse will also store sufficient metadata in the form of question texts, response categories and routing instructions to support the development of data collection instruments. In the CAI environment, these metadata will be used to generate program statements for the collection instrument and to define the survey output structures into which data are loaded when received from the field.

Sections 2 and 3 of the paper describe these two initiatives in more detail, and Section 4 outlines how the resulting Warehouse and Household Survey Facilities will be used in conjunction to re-engineer survey processes and improve metadata management. Section 5 discusses the development of a specific tool to deal with the unique requirements of collection instrument design in a CAI environment.

2. Data Management Project: General Description

2.1 Objectives

In broad terms, the project has twin objectives, first, to improve client service through better catalogued, more visible, and more accessible output data, and, second - the original driving force - to integrate concepts and procedures, and thereby to enhance the information content and ensure the mutual coherence of output data. These objectives are being achieved through the development and implementation of the Information Warehouse which will convert the data and metadata flows and storage indicated in Figure 2 to those shown in Figure 3.

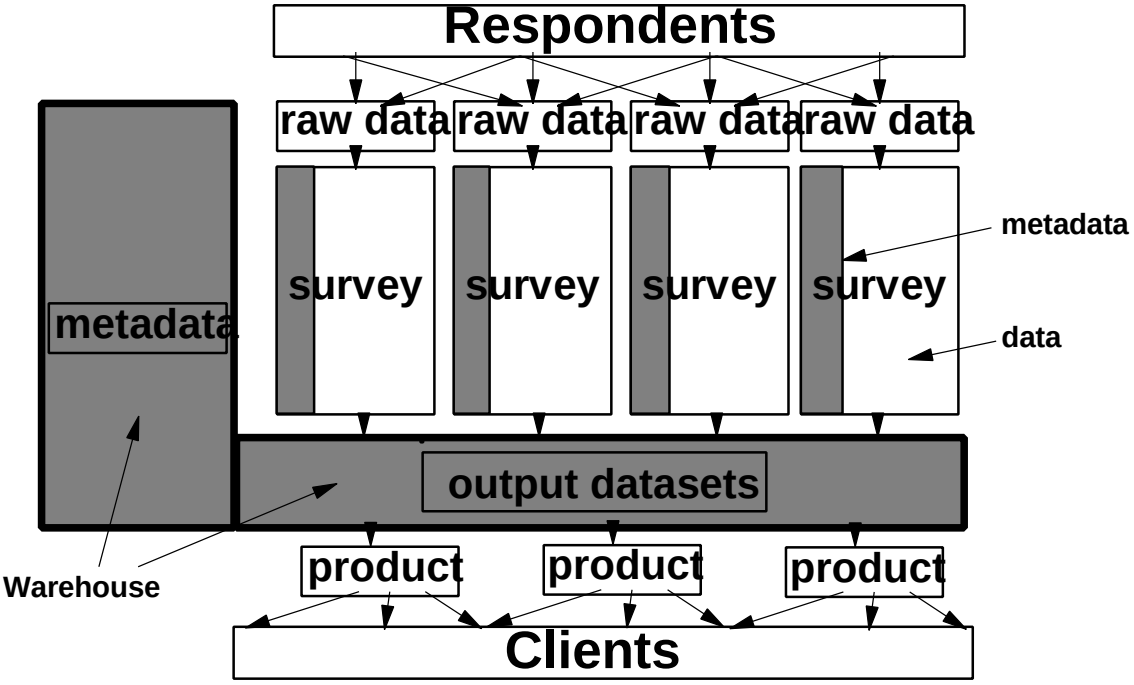

Figure 3: Data Flows and Storage with Warehouse

The Warehouse has two basic functions. First, serving clients within the agency, it provides the metadata basis for the design, development, operation and evaluation of survey processes. It provides a vehicle for formulation, storage and selection of concepts, definitions, and procedures, and it facilitates their standardisation and integration. It supports benchmarking of processes, and enables analyses, combinations and comparisons of datasets. It is the fundamental component of a "statistician's workbench". This role is indicated in Figure 3 by the vertical metadata component of the Warehouse L-shape and the reduced size of the survey specific metadata holdings relative to Figure 2.

Second, the Warehouse provides an output oriented, one stop, statistical data shop, with facilities for storage, manipulation, extraction, and dissemination of the ABS's entire range of output data holdings together with the metadata relating to their definition, collection, and processing. Implementation of the Warehouse will represent a major departure from current practice where data and metadata from individual surveys are separately and independently held and disseminated. Implementation is being accompanied by the introduction of data management policies and standards, and performance measurements, to ensure and support the cultural changes which the new data management approaches require. This second role is indicated in Figure 3 by the horizontal data component of the Warehouse L-shape.

Although the Warehouse aims to provide storage and access facilities for a broad range of metadata from survey initiation to dissemination, in order not to create a monstrous "cathedral project" the Warehouse facilities for the storage, manipulation and dissemination of the actual data are strictly limited to output data. Specifically, the Warehouse does not provide data collection, capture or processing capacity. Such functionality is provided by other ABS systems, such as the Household Survey Facilities described in Section 3, with which the Warehouse must interface.

When in full production, the Warehouse will support subject matter staff by:
- providing the metadata starting point for the design of a new survey or the redesign of the next survey cycle;
- providing output data storage and archiving facilities;
- providing data dissemination and publication facilities.

It will service survey and operations staff by:
- enabling the design and development of new surveys, and the redesign and development of existing ones, to be based on a common store of definitional and procedural metadata, including data item definitions and classifications, and sampling, editing, imputing, weighting, and seasonality adjustment procedures;
- providing the functionality for recording survey procedures and operational performance.

It will serve standards and integration staff by:
- promoting integration of concepts and procedures through the provision of comprehensive facilities for the storage and retrieval of definitional and procedural metadata;
- highlighting the definitional inconsistencies between datasets by juxtaposing their metadata.

It will support client service staff by:
- enabling fast and comprehensive searches of output datasets using client oriented terminology;
- enabling fast retrieval of data and their customisation to suit client needs with respect to coverage and data items;
- enabling dissemination of data in the medium to suit the client;
- assessing costs billable to the client prior to retrieval;
- recording search and retrieval statistics, including details of unsatisfied requests.

Finally, it will provide the basis for responding to questions from national and international organisations about the ABS's data collection processes, for example regarding the number and types of surveys which are conducted, the respondent contacts, and the response rates achieved.

2.2 Data Model

As previously noted, in terms of data the Warehouse is restricted to fully processed micro and macro datasets. On the other hand, the Warehouse contains a wide spectrum of metadata, broadly classified into five categories:

- *definitional* metadata - metadata relating to statistical units/populations, classifications, data items, standard questions and question modules for collection instruments, and statistical terminology;
- *procedural* metadata - metadata relating to the procedures by which data are collected and processed;
- *operational* metadata - metadata arising from and summarising the results of implementation of procedures, e.g. response rates, edit failure rates;
- *systems* metadata - file layouts and access paths used by programs;
- *dataset* metadata - the metadata required to describe minimally (excluding procedural and definitional metadata) a particular object dataset, including name, textual description, population, source, topics, data items (classificatory and parametric), and data cell annotations.

A simplified form of the Warehouse data model is shown in Figure 4. At the core of the model is

the dataset, which includes the minimal metadata required to describe a particular dataset together with (optionally) the actual data and annotations. If the data are, in fact, stored elsewhere, i.e., not in the Warehouse, the dataset is referred to as documentary. Dataset metadata include the dataset title, the survey source, the population which the data describe, the broad topics, the actual data items comprising the data, and references to data products generated from the dataset. The data item entities to which datasets refer may be parametric (quantitative) or classificatory (qualitative). The value set for a classificatory data item is defined by a classification, which may be standard, variant, or non-standard. Other key entities in the data model are data item definitions, question modules, collection instruments, and surveys.

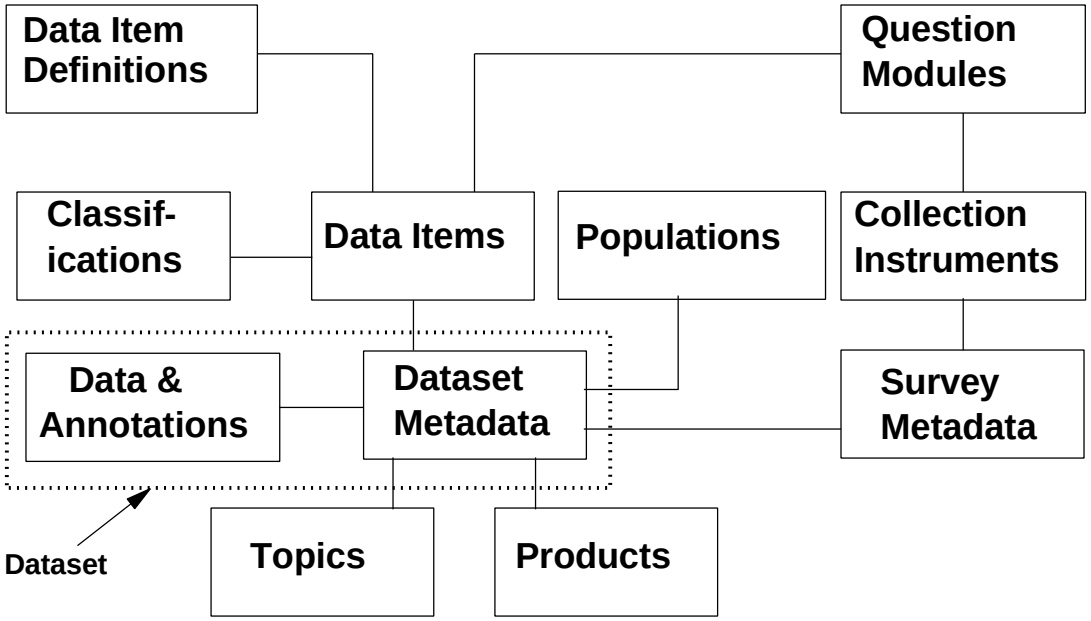

Figure 4: Warehouse Data Model (Simplified)

2.3 Warehouse Functionality

The basic Warehouse functional modules are shown in Figure 5.

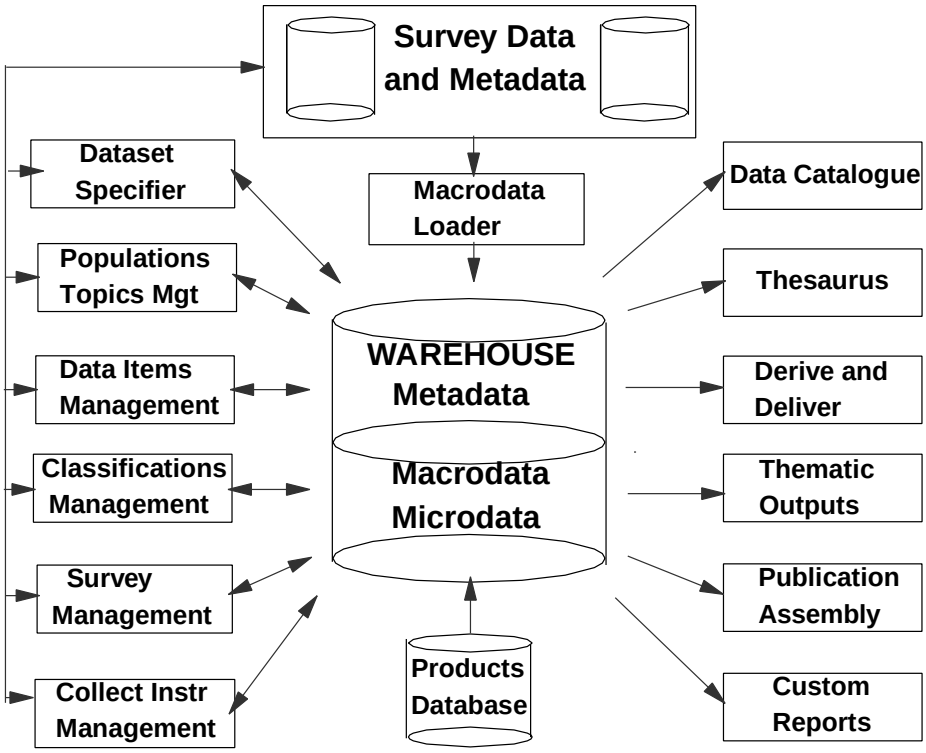

Figure 5: Warehouse Major Components

Metadata Input and Output Functions

There are six basic system components which support metadata loading and access. A common feature of them all is the use of "pick lists" from which users select values. Pick lists not only facilitate data entry but encourage standardisation.

- The Dataset Specifier enables dataset metadata to be entered, viewed and amended.

- The Population and Topics Management System enables updating of the dataset population and topics picklists.

- The Data Item Management System enables creation of and access to detailed data item definitions which exist independently of datasets and to which individual dataset data items may be linked.

- The Classification Management System provides the mechanism whereby users can access, and link together as appropriate, codes, descriptions, levels, and hierarchies of a classification that has been previously loaded. All major standard classification metadata are copied from the ABS Classification System which pre-dates the Warehouse and will eventually be replaced by it. Other (mostly non-standard) classification metadata are loaded from a variety of sources in conjunction with specific datasets.

- The Survey Management System supports the entry, access and update of procedural and operational metadata relating to collections.

- The Collection Instrument Management System enables recording, accessing and updating of metadata relating to paper questionnaires and other collection instruments.

Dataset Input Function

The Macrodata Loader provides the means of loading the Warehouse with data from individual survey processing systems. Microdata may be loaded using the same function.

Dataset Access and Extraction Functions

- The Data Catalogue is the general purpose tool for accessing dataset metadata and other related metadata. It has four basic functions. First, it can generate a list of all the datasets in the Warehouse which satisfy any of a wide range of specified search criteria Lists may be combined across searches and saved or printed for future reference. To improve search speeds, indexes that link dataset data items, topics and populations to the datasets containing them are maintained. Second, it provides access to all the definitional, procedural and operational metadata related to a selected dataset. Third, it can initiate the delivery of the actual data in a specified dataset by invoking the Derive and Deliver function. Fourth, it enables search and interrogation of the Products Database which pre-dates the Warehouse and operates in conjunction with it.

- The Thesaurus expands the search capacity of the Data Catalogue by assisting users to translate their requests into Warehouse "preferred terms" for which title, data item, population and topic based searches are more likely to generate dataset matches. For example, to a user requiring information about "inflation", the Thesaurus would indicate that more effective search terms would be "consumer price index" or "implicit price deflator".

- The Derive and Deliver functions enable the specification of an output table to be generated from a selected dataset; they produce an estimate of the charge for the table based on an ABS pricing algorithm; and they deliver the table in any user specified format, e.g., IMPROV, EXCEL, or comma separated values.

- The Warehouse also has a number of "thematic outputs" designed for the access, derivation and delivery of specific, commonly requested types of data, in particular, international trade, household expenditure, and business register tables. Being purpose built, these thematic outputs perform their operations faster and with more ease, from a user perspective, than the equivalent general functions of the data catalogue and derive and deliver utilities.

- The Publication Assembly System provides the capacity to generate publications directly from the Warehouse, incorporating publication standards.

- Ad hoc reports containing data and metadata that address specific needs of individual clients are generated using the Custom Reporting System.

2.4 Implementation Notes

Although considerable exploratory work had taken place earlier, the starting point of the Data Management project was the report by an information systems consultant from Statistics Sweden (Sundgren, 1991). Significant funding for the project commenced in July 1993 with a total budget in the order of two million dollars per annum. About 25% of this

funding was obtained by regrouping existing activities; the remainder was a new allocation. There are currently about 25 full time project members and many other ABS staff are involved on a part-time basis as the Warehouse has tentacles throughout the organisation.

The development strategy detailed by Richter (1993) involves two phases, the first focusing on facilities for and loading of datasets, the second on statistical integration. As of March 1996, the first phase is essentially complete, a substantial portion of the functionality shown in Figure 5 is in place and the Warehouse contains about 20 gigabytes of data. It is envisaged that the second phase of project will continue through to June 1997 during which time the remaining metadata management facilities will move from prototype into production and corresponding data management policies (Colledge, 1995) will be introduced.

The Warehouse systems are being developed in accordance with ABS policies and standards for databases and computing languages, including use of a client/server relational database environment, ANSI standard C for modules that are portable between Windows and Unix platforms, and SQLWindows for interface work. Use of non-standard features is restricted to security, auditing, and administration facilities. All client systems run against Oracle and SQLBase, the latter being used for standalone demonstrations. Database design is based on a methodology of "minimum commitment", looking beyond the immediate, short term requirements and allowing for addition of new features in accordance with yet to be fully formulated future plans.

As regards development methodology, existing commercially available software - spread sheets, graphics programs, etc - is being used wherever possible. As the initial users are ABS staff, the software is restricted to the existing ABS suite, with only limited investigation of new products, mainly in relation to tabulation. Prototyping is being used for most systems development and all user interfaces. To minimise the development schedule, modules are being built in parallel. As a result, the user interfaces vary in style. They will be rationalised during user review and testing. The underlying philosophy is that any interface system that has cost less than two months to build can be superseded without regret. C modules, which are usually expensive to build, are subject to much more stringent development standards.

3. Household Surveys Facilities: General description

The ABS has recently embarked on a major redevelopment of Household Surveys Facilities (HSF) to support the introduction of computer assisted interviewing (CAI) from late 1997. This redevelopment followed a series of reviews and assessments that looked at requirements for replacement of current household surveys systems which make use of paper forms, optical mark reading (introduced in the late 1980's) and mainframe computer processing.

The introduction of CAI to household surveys has provided an ideal opportunity to completely re-engineer the survey processes and take advantage of advances in technology and other developments that are occurring in the ABS, particularly in the area of metadata management. All the processes ranging from survey development through to office management, data collection and output processing, are now receiving attention for possible re-engineering.

When completed, the HSF will be required to support a substantial household surveys program comprising:

- *Monthly Labour Force survey* - 30,000 households involving an eight month rotating sample using any-responsible-adult methodology, face-to-face interview in the first month and telephone interview from interviewer's home in subsequent months;
- *Supplementary surveys* - generally conducted as an extension to the Monthly Labour Force Survey, covering a range of short topics;
- *Special supplementary surveys* - 15,000 households, carried out once or twice in a year using personal interview methodology, face-to-face, and covering a range of topics in depth;
- *Quarterly omnibus surveys* - 2,000 households covering a range of user-pays items and generally involving any-responsible-adult methodology;
- *Other user-pays surveys* - varying size, content and methodology depending on demand.

The HSF being developed consists of six major subsystems as follows:

- *Sample frame creation and maintenance* - facilities to assist in the establishment and maintenance of the lists of dwellings in the base sample and to provide for selections of units to be made for individual surveys;
- *Survey design and construction* - facilities to assist in the development and testing of collection instruments;
- *Survey data collection* - facilities to manage the distribution of workloads to interviewers, to enable the collection of data by the interviewer and the transfer/transmission of that data to the office;
- *Survey processing and analysis* - facilities to further process the collected data, adding structure, deriving new items and applying imputation and estimation techniques;
- *Survey output* - facilities to extract final data for tables, graphs, publication and special requests;
- *Survey management* - facilities to assist in the overall management of the survey processes.

Some of the facilities will be in the office, others in the field. The systems architecture for office based facilities will primarily make use of the standard ABS client server computing environment comprising Windows on the desk top and the Oracle relational data base management system on a Unix midrange platform. Facilities in the field will be DOS based notebook computers using Blaise version III software (developed by the Netherlands Central Bureau of Statistics) for data collection.

For each subsystem, the processes are being redefined to take into account the potential of the systems architecture and the links needed between the various system elements. These links are being managed through the establishment of tables of metadata which record the characteristics (both system and descriptive) of the data items and unit data stores for each survey and cycle. A fundamental characteristic of the design is the establishment of standard table definitions (for unit data storage) which will be used by all surveys.

Central to the system will be the links to other corporate facilities and, in particular, to the Warehouse. As previously noted, when the data management policy is fully implemented all ABS surveys will be required to routinely deliver data (and metadata) to the Warehouse before data can be released. The HSF will include processes to enable that delivery to be carried out with relative ease for household surveys. The data items in the HSF will be reconciled with data items in the Warehouse through a concordance table which maps one to the other. It will be ensured that items can only be defined once in the concordance table, and subsequent uses of the item will continue to map to the same Warehouse item. This close relationship between HSF and the Warehouse is illustrated in Figure 6.

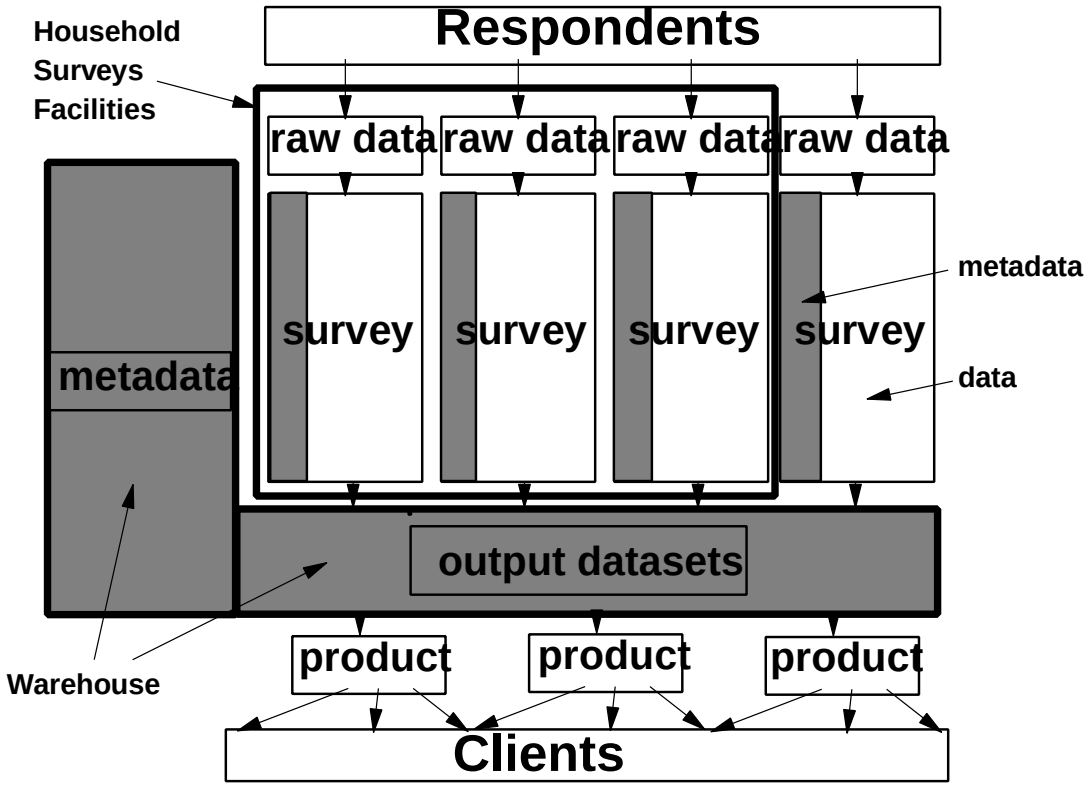

Figure 6: Warehouse and Survey Facilities

The potential exists, therefore, to make use of the metadata in the Warehouse to drive or assist in the survey development process, particularly within the CAI context where the question texts, categories and routing information are to be converted to program statements rather than to a paper form. How this is to be done is described in the next two sections.

4. Use of the Warehouse and Household Surveys Facilities in survey development and operation

4.1 Introductory Remarks

The envisaged use of the Warehouse and Household Surveys Facilities in the development and implementation of a new ABS survey is illustrated in Figure 7. There are six more or less sequential groups of activities which are outlined in the following paragraphs.

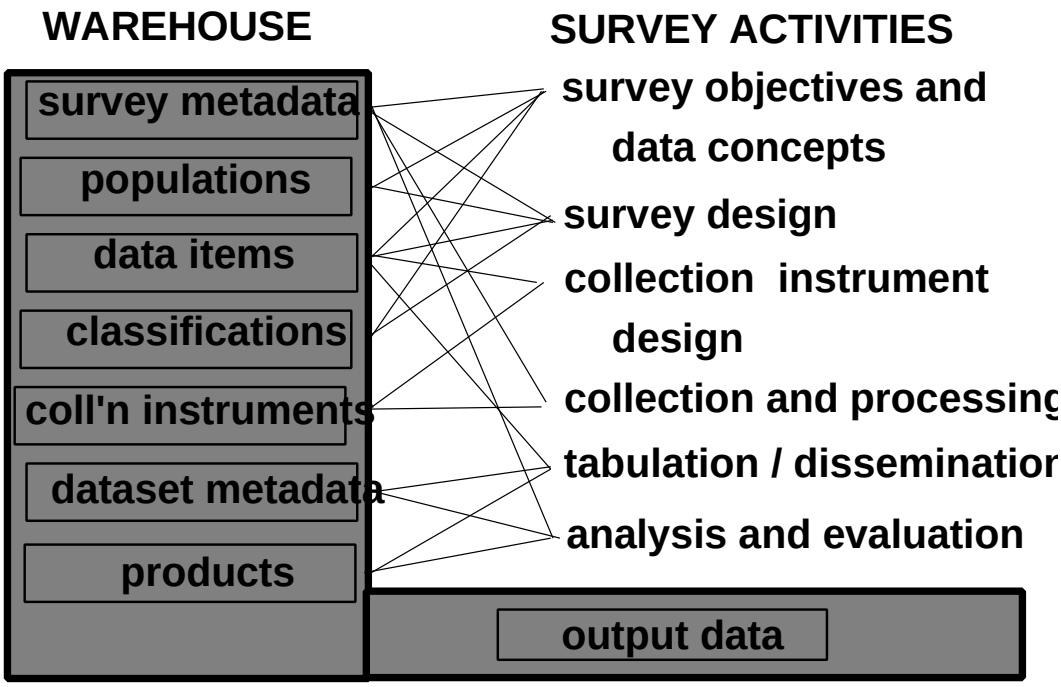

Figure 7: Use of Warehouse and Survey Facilities

4.2 Survey Objectives, Scope, Content and Organisation

Senior ABS staff and the potential survey clients jointly identify the issues to be examined, outline the corresponding data requirements, and establish the survey objectives. ABS staff secure funding for the survey, set the budget, define the survey development process, establish a steering committee, assemble a project team, and discuss the arrangements with clients. Responsibility for the survey development is assigned to a project team typically comprising staff from the subject matter, standards, survey design, operations, systems, and client service areas.

The subject matter staff define the data requirements more precisely. They identify and document the underlying concepts, target (sub)populations, classifications, data items, and the basic content of the collection, and they outline the principal output tables. In doing so, they refer to the procedural and definitional metadata associated with any other surveys which might be relevant.

Warehouse Activities

- Subject matter staff open a Warehouse record for the new survey and enter the proposed survey name, reference period, frequency, geographical coverage, collection type, data sources, and an overview description of the survey including purpose and content. They record the historical background to the survey, the intended users and uses in so far as they are known, and the names of any other surveys to which the survey is related in the sense of acquiring similar data, using similar methodology, or being part of an integrated group. They enter the organizational unit with primary responsibility for survey implementation, the principal contact persons, the date the survey is to be officially presented in parliament, and estimates of the survey cycle costs.

- Subject matter staff also enter the conceptual framework (if any) underpinning the survey and other relevant concepts, the target population, the provisional lists of input and output data items, the classifications used, the proposed release dates, and any terminology particular to the survey. In doing so, they access, review, and use as appropriate, metadata for related surveys recorded in the Warehouse. They select a definition for each data item from amongst the Warehouse standard and variant definitions. Where necessary they create proposed new data items and definitions.

- Standards staff review the proposed classifications and lists of data items and definitions for conformity to relevant standards and guidelines, and feedback comments to the subject matter area.

4.3 Survey Design

Survey design staff jointly define the collection methodology, the operational design, the sampling frame and specify the statistical units for selection. In the case of a sample survey, survey design staff define the stratification and sample selection procedures, and select the sample. They define in broad terms the estimation methods, outlier treatment, non response adjustment, and variance calculation procedures.

Warehouse Activities

- Survey design staff record in the Warehouse summary descriptions of the sampled population, the frame construction procedures, the resulting types and numbers of sampling units, the frame updating procedures (for repeated collections), the known limitations of the frame and the procedures to overcome them. Survey design staff also enter the results of studies and pilot tests, and the measures to be used in improving response and minimising respondent burden.

- Survey design staff record summary descriptions of the sample stratification, allocation and selection procedures and the resulting sample size, also broad level details of the procedures for weighting, treatment of outliers, non response adjustment, and variance calculation.

4.4 Collection Instrument Design

Survey design staff jointly formulate the questions and the interviewer reporting instructions, and select and generate the appropriate collection instrument. They define the acquisition, capture, follow-up and first stage editing procedures.

Warehouse Activities

- Survey design staff access the provisional list of input data items, together with their definitions

and the corresponding standard questions, response categories and explanatory material recorded in the Warehouse. They assemble existing questions, develop new questions and specify response categories, reporting instructions, and first stage editing rules, and they transfer the results to the selected collection instrument medium (ie. paper or CAI program). In doing so they automatically refine the input data items associated with the survey. (These activities are elaborated in Section 5 for CAI surveys.)

4.5 Data Collection, Capture, Editing, Imputation and Estimation

Survey design staff finalise the weighting, outlier treatment, non response adjustment and variance calculation procedures. Operations staff collect the data and transfer them to the collection input database. In the case of CAI, first stage editing rules are built into the collection instrument, and most if not all first stage editing is done at the time of data entry. Operations staff follow up non responses, resolve first stage edit failures, and arrange for second stage ("statistical") edits to be applied. Missing data items are imputed; adjustments are made for outliers and for non response; the principal estimates are calculated; and the micro and/or macro output datasets are generated.

Warehouse Activities

- Survey design staff complete and install the procedures and systems for capture, coding, editing, weighting, treatment of outliers, non response adjustment, and variance calculation, and summarise them in the Warehouse.

- ***Control Point****. Before data collection can commence, all relevant procedural and definitional metadata must be recorded in the Warehouse.*

- Operations staff enter response rates, edit failure rates, imputation rates, and other quality/ performance indicators into the Warehouse, also summaries of written directions issued to or legal proceedings initiated (if any) against nonrespondents.

- Subject matter staff finalise the metadata descriptions of the principal output datasets, define any other output datasets and record them in the Warehouse. In so doing, they automatically identify the survey output data items. Operations staff load the corresponding data are into the Warehouse, or other data store, as agreed.

4.6 Tabulation, Dissemination, and Archiving

Subject matter staff specify and produce output tables for pre-planned publications and products from the output datasets. They check that all table cells satisfy confidentiality constraints, and confidentialize the tables by cell suppression or adjustment where necessary. They check that all table cells satisfy the quality conditions specified for the survey, and suppress those which do not. They sign off the tables for release to the public, but keep the data embargoed until the prespecified release date and time, at which point they release the tables in a variety of different formats and media.

Following release of the data, client service and subject matter staff respond to requests for additional information. The former handle requests which are relatively straightforward and can be satisfied through data which have been released; the latter deal with more complex requests or involving data which have not been released. Subject matter staff subsequently archive any actual data that have been in the Warehouse for more than a prescribed period without being called upon to satisfy a single request.

Warehouse Activities

- Subject matter staff summarise in the Warehouse the procedures required to confidentialize datasets prior to release into the public domain, also any quality based release criteria or other

release conditions and procedures.

- ***Control Point.*** *All pre-planned outputs must be specified in terms of metadata stored in the Warehouse, and the corresponding object data must be drawn from the Warehouse, or if stored elsewhere, must be extracted based on specifications generated through the Warehouse. The explanatory notes in publications must also be derivable from metadata stored in the Warehouse. This implies that all the corresponding definitional, operational and procedural metadata must be recorded in the Warehouse activities before any dissemination can take place.*

- Subject matter staff specify standard (and other) products to be generated from the survey data, they create Warehouse links to these products, and they specify the publication contents and layout in general terms. For a repeated survey, publication staff create a publication template in accordance with corporate standards, with the aid of which subject matter staff generate the publication, drawing data and metadata from the Warehouse.

- In response to requests from clients which can be satisfied from released data, client service staff access the Warehouse *output* database, extract the data and supply them in a format to suit the clients' preferences.

- For requests which cannot be satisfied from the *Warehouse output* database, subject matter staff extract the appropriate data from the Warehouse *internal database* (or other database), check them for confidentiality and quality, confidentialize and suppress cells as required and supply the resulting tables to the clients.

- ***Control Point****. All ad hoc outputs in response to client requests must be specified in terms of metadata stored in the Warehouse, and the corresponding data must be drawn from the Warehouse, or if stored elsewhere, must be extracted based on specifications generated through the Warehouse.*

- Through the *Warehouse register of confidentially sensitive datasets*, subject matter staff record details of all datasets (and the clients requesting them) that may lead to future restrictions in release of related data due to confidentiality constraints. They also check dataset access details and use Warehouse facilities to archive any data which have not been accessed for more than the prescribed period, leaving the metadata in the Warehouse.

4.7 Analysis, Quality Assurance and Evaluation

Subject matter staff analyse the output data and compare them with data from related surveys. Jointly with design and operations staff, they analyse the quality and performance measures, and assess the impacts of the various sources of error. Jointly with client service staff, they analyse the output dataset access patterns and match the actual users and uses against those envisaged. They prepare a survey evaluation, and, if the survey is to be repeated, recommend appropriate changes and enhancements.

Warehouse Activities

- Staff use the Warehouse to retrieve processing and performance measures and dataset access patterns and to summarise sources of error and measures to deal with them, evaluation results and recommended changes and enhancements.

5. Use of the Warehouse in Development of CAI Surveys

As the repository for concepts and definitions associated with data items, the Warehouse provides a wealth of metadata that can be used to drive the development of questionnaires. In particular, the Warehouse has the functionality to record the questions that may be used on a paper questionnaire together with the associated explanatory notes and to maintain links between these questions and the corresponding data item definitions. This facility is already proving useful in the ABS to help standardise the questions on business survey questionnaires.

The CAI environment poses more challenges, however, because:

- the underlying texts of questions and responses are imbedded within the program code for CAI instruments;
- the questions themselves can include program elements to provide alternative wording for different situations - this metadata information should also be found in the Warehouse;
- CAI enables more complex questionnaires to be designed because a program can be used to define:
  - the route which is taken and the conditions that define when a question is asked;
  - the application of edits to be resolved in the field; and
  - derivation of new items from response information.

All of the instructions to the CAI instrument to cover the features mentioned above need to be stored within the metadata of that instrument. This metadata basis gives CAI instruments the potential to be more directly "linked" to the Warehouse (through their metadata elements) than paper forms. Those links can be used not only to supply metadata to the Warehouse but also for the development of new questionnaires. It is this development aspect that is the focus of the remainder of this paper.

The solution being developed for CAI in the ABS is to provide a special "Collection Instrument Development Tool" which links with the Warehouse and accommodates the specific needs of CAI (including generation of the questionnaire program) while providing a framework within which the survey developers can operate and share their work. The Warehouse provides the initial ingredients for the survey development and is also the repository for the metadata representing the final outcome. The Warehouse role as it is being developed within HSF is shown in a simplified form in Figure 8.

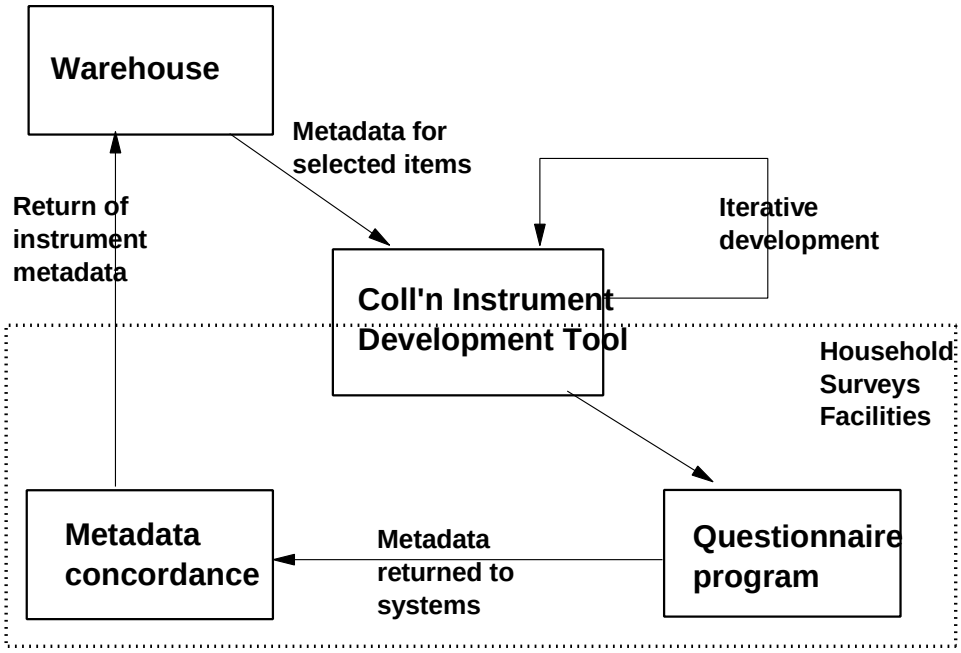

Figure 8: Role of the Warehouse in development of CAI instruments

The Warehouse provides the repository of the data items and their definitions in descriptive terms. The Collection Instrument Development Tool provides the facilities which enable the questionnaire elements to be developed within a structured format. Once developed, the elements can then be used to generate the computer program which forms the survey instrument which controls the interview and collates the respondent data. The metadata concordance tables provide the linkage of the elements of the instrument back to the definitions of data items in the Warehouse.

During survey development, sample specification, target populations, data items, questions, edits and derivations often undergo many iterations. Elements are added or taken away, elements are modified, ordered, and so on until the questionnaire gradually takes shape. The Collection Instrument Development Tool is expected to provide this kind of functionality. A description of the tool follows.

The basic building unit of a questionnaire under this tool is a module of related questions. Such a module may measure a single concept or group of related concepts and will generally be applicable to a defined population in the survey. Earlier prototypes using a question as the building unit gave too many elements and were found to be difficult to manage. For convenience a module can be seen as containing up to 10 questions although provision of larger modules may also be necessary. A module has the advantage of having closely related questions kept together and could be small enough to treat as a single unit that can be stored and reused for other surveys (without the likelihood of breaking apart).

When presented on the screen, a module can be likened to a "form" where the material is presented in a structured way. Appendix 1 shows an example of such a module as it is has been developed and tested in the ABS context.

Each of the fields on the form are saved into a database where they can be accessed for reporting or translation into CAI program statements. The module has provision for recording edits and derivation specifications to create the data items for the survey. Some additional fields are included on the form to ensure that the program for the CAI instrument, or most of it, can be generated from the module.

The module is expected to be self-contained and targets a defined subgroup, or population, of the survey. It is expected that the module population will be defined through data items defined in preceding modules. The routing instructions recorded within a module would be expected to remain within the module. Routing outside of a module is handled by derived items or population definition. These restrictions are necessary to ensure the relative independence of modules and will make them easier to be reused for other surveys.

The forms that describe the modules of questions are stored in a database. Selected identifying information about them is available to be presented to the user through a series of views. The views can also be printed as reports.

In addition to the basic questionnaire module there are other modules to define broad survey parameters (e.g., sample/ subsample specification, dependent interviewing requirements, and global substitutions) to assist in the questionnaire design.

The following characteristics are considered to be necessary in such a tool:

- *appropriateness* - corresponds to natural thought processes for questionnaire design;
- *accessibility* - the facility needs to be available to all the players in surveys: subject matter specialists; survey designers; managers and programmers;
- *ease of use* - it should be relatively easy to use and understand, and efficient to enter and change data;
- *structure* - the facility needs to show the developing survey metadata within a structured format that presents both the broad flow of the questions and the survey content, and assist in validation;
- *reporting* - the facility should provide a range of reports on the survey content;
- *functional* - the facility needs to provide a range of functions associated with the development process, including fetching metadata from the warehouse and delivering program metadata for the CAI instrument;
- *integration* - the facility should provide for the integration of questionnaire elements, in particular the links between data items and question modules.

The linkage of data items from the Warehouse to the Collection Instrument Development Tool will be achieved by an export process which is carried out after the developer has used Warehouse tools to selected prospective data items for the survey. The Collection Instrument Development Tool will eventually become an integral part of the Warehouse.

The linkage of the Collection Instrument Development Tool to the CAI instrument is carried out by exporting the fields in each module to a special program which generates the program code for those questions. Each module defined in the Collection Instrument Development Tool database is thereby converted to a separate CAI instrument module. These instrument modules can then be put together to form the CAI questionnaire.

The intention in generating program code is to reduce the redundant transcription aspects of instrument preparation. It will still be necessary, and desireable, for some programmer involvement in the further preparation of the CAI instrument, to ensure that the instrument is fully functional and that the various features of CAI software are well utilised.

Once the CAI instrument is tested and finalised then both the instrument modules and the associated metadata will be returned to the Warehouse as a record of the final outcome. The return links between the CAI instrument and the Warehouse will make use of utilities that can extract the metadata from compiled survey instruments.

6. Concluding Remarks

Most of the conceptual thinking which underpins the Warehouse development has been completed, the basic functionality for loading, manipulating and accessing datsets is operational, and 20 gigabytes of data have been loaded. During the next fifteen months or so, the definitional, procedural, and operational metadata manipulation systems will move from prototype into production, and statistical integration - that elusive target - will begin. The Household Surveys Facilities are at an earlier stage of development, though prototype systems such as the CAI Collection Instrument Development Tool are operational and in active use.

Both projects represent a major commitment by the ABS. Senior management is providing sustained support to the project teams, and the developments are being viewed by the potential users within the ABS with considerable interest if not actual enthusiasm. (Who likes change!) At completion the Warehouse and household surveys facilities will provide the tools for better control of metadata and data as the latter are transformed during the survey cycle than has ever been available to date.

The presence of the Warehouse and its support by management has brought about a change of corporate culture in the ABS. For developments such as the Household Surveys Facilities, it is now a foregone conclusion that the Warehouse will play a significant role. The Collection Instrument Design Tool described in this paper is an example of how this is being put into practice.

Acknowledgements

This paper is based on the work of all the ABS Data Management and Household Surveys Facilities Project team members past and present. Special reference must be made to Warren Richter (Data Management team manager), Rob Edmondson (Warehouse chief architect), Peter Damcevski (Warehouse loading), and Ken Smith (Household Surveys Facilities).

The authors would also like to thank John Hodgson, Ann Bromwich, and Glenn Cocking for reviewing the document.

References

Australian Bureau of Statistics (1969), "Integrated Censuses", Year Book, Chapter 31, Catalogue No 1300.0, Australian Bureau of Statistics, Belconnen, ACT 2616, Australia.

Bethlehem J.G., Hunderpool A.J., Schuerhoff M.H. and Vermeulen L.F.M. (1989), "Blaise 2.0 / An Introduction" Central Bureau of Statistics, Voorburg, Netherlands.

Bethlehem J.G., Hofman L., Schuerhoff M.H. (1994), "Blaise III / Overview" Central Bureau of Statistics, Voorburg, Netherlands.

Colledge M.J. (1986), "Business Survey Redesign Project: Implementation of a New Strategy at Statistics Canada",

Proceedings, Third Annual Research Conference, U.S. Bureau of the Census, Washington DC, U.S.A.

Colledge M.J. (1995) "Proposed Data Management Policy", Working Paper, Australian Bureau of Statistics, Belconnen, ACT 2616, Australia.

Kovar J.G., MacMillan, J.H., Whitridge P. (1991), "Overview and Strategy for the Generalised Edit and Imputation System", Working Paper, Statistics Canada, Ottawa, K1A 0T6, Canada.

Richter W., (1992), "Project Definition Document", Working Paper, Australian Bureau of Statistics, Belconnen, ACT 2616, Australia.

Sundgren B., (1991), "Towards a Unified Data and Metadata System at The Australian Bureau of Statistics", Working Paper, Australian Bureau of Statistics, Belconnen, ACT 2616, Australia

Appendix 1. Sample question module

Question Module

Collection:	Survey of Income and Housing costs		
Module Name:	Main Job - Wage and Salary Details		
Module Seq:	2a		
Status:	Qns: Draft/Final Edits: Draft/Final Derivs: Draft/Final		
Population:	W&S earner		
Populations in	P2/0		Path
Questions:	Q1	I WOULD NOW LIKE TO ASK YOU ABOUT YOUR PAY FROM YOUR (MAIN) JOB. WHAT WAS THE TOTAL AMOUNT OF YOUR MOST RECENT PAY *BEFORE* TAX OR ANYTHING ELSE WAS TAKEN OUT? *0 *1..99999	 Q3 Q2
	Q2	IS THAT YOUR USUAL PAY? 1 = Yes 2 = No	 Q4 Q3
	Q3	HOW MUCH DO YOU USUALLY RECEIVE EACH PAY? Interviewer: Enter amount in dollars *0..99999	 Q4
	Q4	WHAT PERIOD DOES THAT PAY COVER? Interviewer: Code to 'Weeks' if response is in weeks Code to 'Months' if response is in months 1 = Weeks 2 = Months	 Q5 Q6
	Q5	(WHAT PERIOD DOES THAT PAY COVER?) Interviewer: Enter number of weeks/months *1..52	 End
	---		

	Q10		

Population Derivations

Population Label	Specification

Output Derivations

Dataitem Label	Specification	In/post field
WKLYWS	IF Q2=Yes THEN DO IF Q4=Weeks THEN WKLYWS=Q1/Q5 ELSE WKLYWS=Q1/Q5 * 12/52 END ELSE DO IF Q4=Weeks THEN WKLYWS=Q3/Q5 ELSE WKLYWS=Q3/Q5 * 12/52 END	In

Edits

Edit Label	Specification	Interviewer prompt	Fatal/Query	In/post field

ED1	WKLYWS<2000	Weekly wages are generally less than $2,000	Q	In