International Statistical Review (1999), **67**, 1, 79–98, Printed in Mexico

Statistical Integration Through Metadata Management

Michael J. Colledge

Statistics Directorate, Organisation for Economic Cooperation and Development, 2 rue André Pascal, 75775, Paris, France
E-mail: michael.colledge@oecd.org

Summary

Faster and more versatile technology is fuelling user demand for statistical agencies to produce an ever wider range of outputs, and to ensure those outputs are consistent and mutually related to the greatest extent possible. Statistical integration is an approach for enhancing the information content of separate statistical collections conducted by an agency, and is necessary for consistency. It has two aspects—conceptual and physical—the former being a prerequisite for the latter. This paper focuses on methods for achieving statistical integration through better management of metadata. It draws on experiences at the Australian Bureau of Statistics in the development and use of a central repository (the "Information Warehouse") to manage data and metadata. It also makes reference to comparable initiatives at other national statistical agencies.

The main conclusions are as follows. First, a prototyping approach is required in developing new functionality to support statistical integration as it is not clear in advance what tools are needed. Second, metadata from separate collections cannot easily be rationalised until they have been loaded to a central repository and are visible alongside one another so their inconsistencies are evident. Third, to be effective, conceptual integration must be accompanied by physical integration. Fourth, there is great scope for partnerships and exchange of ideas between agencies. Finally, statistical integration must be built into the ongoing collection processes and viewed as a way of life.

Key words: Coordination; Data warehouse; Harmonisation; Information warehouse; Survey management.

1 Introduction

1.1 Background

For historical reasons, the data collection and production operations in national statistical agencies are typically organised into a set of separate processing streams, known as "collections", "surveys", or "enquiries", including some that make use of administrative data sources. An important aspect of quality is the mutual consistency of data output from these different collections. Data that are not mutually consistent cannot be meaningfully combined and used jointly. Even worse, if they refer to the same topic they give rise to confusion and effective loss of information content. What are users to do, for example, if employment statistics from an annual survey of manufacturing industries are inconsistent with those published on the basis of a monthly, all industry employment survey?

On the other hand, if data from different collections can be meaningfully combined, there is likely to be synergistic gain in information content. For example, bringing together data from a survey of production with those from a capital investment survey can produce profit/investment ratios which

could not be generated by either survey alone. Such merging of data requires an appropriate degree of consistency across the collections. In this example it would imply, at a minimum, that the data had been collected from the same or relatable set(s) of business units which were divided into the same or relatable set(s) of industry classes.

One might reasonably assume that a single collection would produce consistent data—consistency being an expected product of good design and editing. In practice, this is not always the case. Inconsistencies can occur at various stages, from respondent misunderstanding of questions through to discrepancies in the data disseminated via different media. For example, the tables in a paper publication may, due to last minute changes, differ from those appearing in an electronic product based on the same collection results. If a collection is repeated, there is further scope for inconsistency between successive cycles. Ensuring consistency across separate collections within an agency is an even more challenging goal as the collections have usually been set up independently to address specific, different needs and they are conducted semi-autonomously. Achieving consistency of data for similar collections across agencies, nationally and internationally, is correspondingly more difficult still.

1.2 Statistical Integration

The basis for consistency is statistical integration. In this paper, "statistical integration" or "integration" for short, refers to the state of having statistical data which are mutually consistent and are related to the greatest extent possible. It also refers to the set of activities which aim to produce data in this state. The qualifier "statistical" simply defines the context within which integration is being considered, i.e., the collection, processing and dissemination of statistical data. Given the context, the qualifier is often omitted, as henceforth in this paper.

Integration embodies the notions of "co-ordination" and "harmonisation". Willeboordse & Ypma (1996) define it as occurring in five stages (elaborated in Section 5.7). This paper considers integration into two principal stages: (1) "conceptual" (or "logical") integration, implying use of common concepts, standards and terminology; which is a prerequisite for (2) "physical" integration, i.e., harmonisation of data inputs, processes, or outputs.

Evidence of integration within a statistical agency is seen, for example, in the use of a business register to provide a common frame for economic surveys, in the use of standard classifications and data item definitions, in the national accounts which combine data from a range of surveys within a unifying framework, and in the centralised, accessible storage of information about collections.

Through sharing of concepts and procedures, integration provides the basis for considerable benefits in addition to consistency of data outputs. These include:

- reduced respondent burden, through easier identification and hence elimination of duplicate requests for data;
- promotion of best practices, through common use of well tried and tested standards and methods;
- reduced risk of introducing variations in what are intended to be common procedures;
- less duplication of data processing functionality, storage and maintenance, through use of common processing tools to service a range of different collections; and
- better understanding by users of the data outputs and their fitness for use.

In summary, integration provides one of the most effective mechanisms for addressing the major challenge faced by statistical agencies, namely that new technology is leading to more sophisticated

users with higher expectations regarding data coverage, quality, and access. Whilst this is not a new situation—statistical agencies having had to adapt to changing circumstances since their inception—it is heightened by the dramatic rate which technology is now advancing coupled with a world wide climate of shrinking government funding. As Keller (1996) notes, "Concerning our output, we see a demand for better access, preferably electronically, and greater user-friendliness. One particular aspect is a demand for an improvement of the coherence of the totality of the information we offer."

Given its many potential benefits, one may wonder why integration has not been at the core of every statistical agency's strategic initiatives. The reason is that integration requires the sacrifice of local (collection) optima for global (agency wide) goals. It requires significant cultural change to convince individual collection managers, who are used to operating semi-autonomously, that it may be more important to respect agency integration objectives than to specialise concepts and procedures in order to optimise local goals.

1.3 Scope and Content of Paper

Integration has many facets. This paper focuses on those aspects of integration:

- that refer to infrastructure and coordination rather than content and harmonisation;
- that can be achieved through better management of metadata, i.e., integration of the metadata describing sample design, data collection, capture, and processing methods, rather than of the procedures themselves; and
- that are being addressed by the Data Management Project at the Australian Bureau of Statistics (ABS).

The paper provides a background to the ABS Data Management Project, outlines the metadata infrastructure being developed in support of integration, describes the corresponding integration activities, ongoing and planned, and outlines very briefly some comparable developments at other statistical agencies.

1.4 Terminology

"Metadata" means simply data about data, and refers to the definitions, descriptions of procedures, system parameters, and operational results which characterise and summarise statistical programs. Metadata may be "passive" ("descriptive"), i.e., in the form of documentation which is used by agency staff, or may be "active" ("prescriptive"), i.e., determining the actions of automated survey processes. Either type of metadata can support integration, but the first case, the integration initiatives must be person driven, whereas in the second case they are machine driven. Priest (1996) refers to metadata used by people as "meta-information", and those used by machines as metadata, but the distinction is not made in this paper in view of the continual blurring of the boundary. Neither was that distinction the basis for the ABS's decision to refer to an "information warehouse" rather than a "data warehouse", which is the more commonly used term in the informatics literature. The former term was preferred because the warehouse contains metadata as well as data, and "information" seems more appropriate to cover both. Other terms used in the paper are explained as they are encountered, and an appendix on terminology is attached for easy reference.

2 Integration Framework at the ABS

2.1 Introductory Remarks

Figure 1 illustrates the sort of data collection, processing and dissemination environment that, in the absence of specific interventions, tends to evolve over time at the ABS (or any statistical agency) as new collections are added to the bank of existing ones in response to specific data demands. Collections independently obtain data from respondents, transform them into statistical aggregates and disseminate them to clients. The transformation processes are under the control of, and described by, collection specific metadata held in local databases. In this model, there is a conspicuous absence of exchange of information between the collections, the main crossflow (unfortunately) being via respondents and clients.

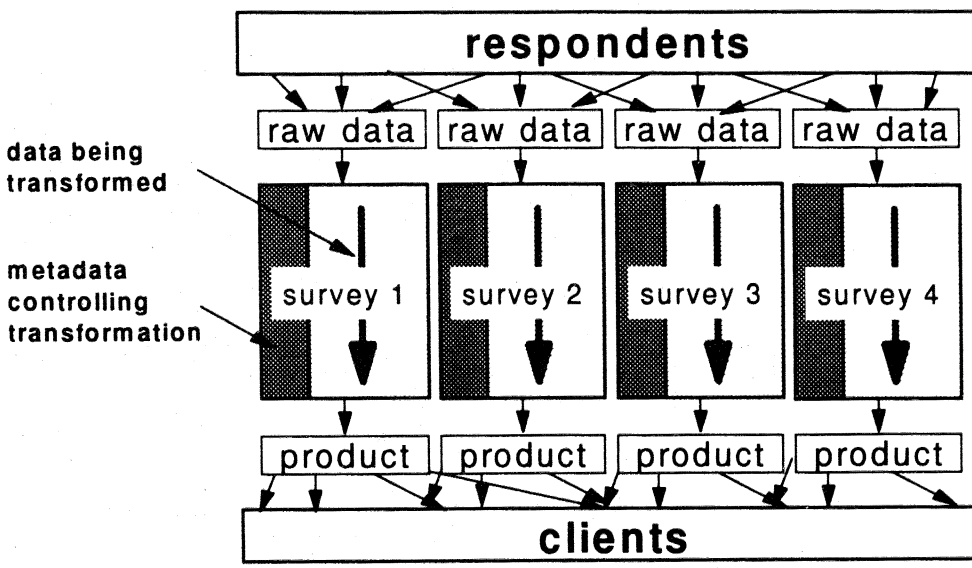

Figure 1. *Typical processing environment*

There are numerous problems with this arrangement. Respondents are asked for information in a confusing variety of formats, and even for the same information by different collections. Clients see the products of different collections or groups of collections, but do not get a complete and unambiguous picture of the agency's entire data holdings. Concepts are duplicated or contradictory across collections, the same term being used to mean different things, or different terms the same thing. There is little sharing of procedures and systems with consequent duplication of effort in their creation and maintenance.

The ABS has been aware of these problems and trying to deal with them for many years, through a variety of organisational arrangements and initiatives. The integrated economic censuses project (ABS, 1970) was an early example. The current approaches are described below. In this respect, the Data Management Project is just another attempt to deal with an old problem.

2.2 *Integration Related Functions and Initiatives*

Like quality management, of which it is an aspect, integration is a broad goal with many different facets. Typically, it is not managed as a single activity but through a range of initiatives built into the mandates of functional areas and the objectives of specific projects. (Here, "functional area" implies ongoing operations and budget, in contrast to a "project" which has a resource allocation for a limited period, for a development purpose.)

Functional area involvement in integration may be briefly summarised as follows.

- Methods and standards areas are responsible for developing and promoting the use of standard statistical units and populations, classifications (nomenclatures), data item (variable) definitions, content and phrasing of collection instruments, and other statistical terminology.
- Statistical service areas are responsible for the identification, development and use of survey design best practices, including collection instrument design, survey frame construction, sampling, editing, imputation, and analysis.
- Technology support areas are responsible for identification, development and implementation of an informatics infrastructure which promotes relatability, interconnectivity, and use of a common tool kit. Opportunities for integration occur as new systems are developed and old ones decommissioned. The gradual phase out of the main frame architecture, the introduction of Notes based workflow applications, and addressing Year 2000 problems are currently providing particular focal points for this sort of activity. In addition, these areas are pursuing an "object management" strategy aimed at organising the ABS's unstructured, non-statistical information. This has implications for data management as the boundaries between statistical metadata and other information becomes increasingly fuzzy.
- The business register area provides both definitions for and lists of business units from which the frames for individual surveys can be constructed and samples drawn. This is an example of conceptual and physical integration.
- The national accounts area brings together data from a wide range of economic surveys and integrates them within the international System of National Accounts framework, also embodying both conceptual and physical integration.

In addition, as of July 1997, there is the newly constituted statistical coordination area which is introducing and administering a "Statistical Clearing House" on behalf of the Australian Government. The primary purpose of the Clearing House is to ensure that statistical collections impose minimum load on business respondents and that the data resulting from these collections are fit for their intended uses. Collections affecting 50 or more businesses conducted by or on behalf of Australian Government departments or agencies, including the ABS, are subject to a clearance process. Those that are not approved are not permitted to start or continue. The Clearing House review of business collections not only aims to reduce respondent burden, promote good design, enforce documentation, and facilitate data sharing, it also provides an excellent opportunity to promote integration not only within the ABS but across the federal government as a whole. Metadata for all the collections falling within Clearing House scope are recorded in a Register of Surveys, published on the Internet (www.abs.sch.gov.au). More details are provided by Colledge (1998).

Complementing the ongoing integration activities of functional areas, there are several substantial ongoing projects in which integration plays a significant role. They include redevelopment of the ABS Business Register, introduction of a common "Integrated Processing System" for collection and capture of data, use of a standard "Survey Processing Environment for Economic Surveys", redevelopment of general purpose "Household Survey Facilities" for collection, capture and processing of social data, and the "Data Management Project". These initiatives are coordinated with

one another and with ongoing functional activities through exchange of information between the project teams and commonality of membership of the project steering committees. Lee (1996) gives a more complete picture.

2.3 *Data Management Project*

Of the above initiatives, the Data Management Project has the broadest scope in terms of support for statistical integration. If is focused on the management of data and metadata and has twin goals: improved client service through better catalogued, more visible, and more accessible output data; and integration of concepts and procedures to enhance the information content and mutual coherence of data and to reduce systems maintenance costs. These goals are being approached through the development, loading and use of a corporate repository—the "Information Warehouse", or "Warehouse" for short—from which most, if not all, ABS data products will ultimately be generated. The Warehouse provides an output oriented, one-stop, statistical data shop, with facilities for storage, manipulation, extraction, and dissemination of the ABS's output data holdings together with the metadata relating to their origin. It also facilitates the formulation, standardisation, storage and selection of concepts, definitions, and procedures. Development of the Warehouse is being accompanied by the introduction of policies and procedures for better management of data and metadata.

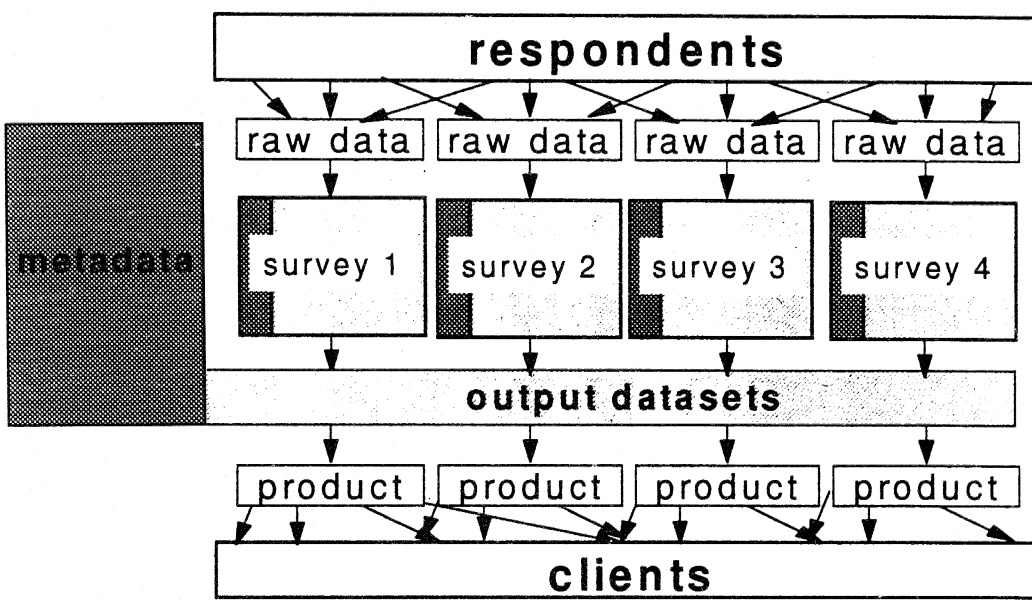

Figure 2. *Processing with Warehouse*

Although considerable exploratory work had taken place earlier, the starting point of the project was the report by Sundgren (1991). Significant funding for the project commenced in July 1993 with a total budget in the order of two million dollars per annum. As of August 1997 there are some 25 full time project members. As the Warehouse affects a wide range of activities throughout the ABS, many other staff are also involved, including those in subject matter, standards, methodology, and client service areas. The original project definition document (Richter, 1993) outlined the development and implementation strategy in terms of two project phases, the first being the construction and loading of facilities to store and retrieve output datasets, the second being the expansion of these facilities and loading of metadata to support statistical integration.

Figure 2 illustrates the affect the Warehouse will have on the processing environment depicted in Figure 1. The Warehouse is represented by the L-shaped block. The horizontal part of the block indicates the intended result of the first phase development, i.e., the storage of output datasets and their dissemination from the Warehouse. The vertical part illustrates the result of the second phase, i.e., the storage of collection metadata so that they are visible and accessible for use throughout the agency rather than being locked up in local collection databases.

Colledge & Richter (1994) and Richter (1996) provide more comprehensive descriptions of the general aims and development framework of the Data Management Project, and Colledge, Wensing & Brinkley (1995) indicate how data management may be applied in the context of computer assisted interviewing. The next two sections elaborate those aspects of the initiative which enhance the metadata infrastructure in support of integration, and the consequent integration activities which are taking place or planned.

3 Warehouse Metadata Infrastructure

3.1 Warehouse Metadata Model

The high level metadata entities envisaged for the Warehouse in full production are shown in Figure 3. At the core of the data model, reflecting its output orientation, is the dataset entity. A dataset identifies a table of data together with corresponding row, column and cell annotations. It is linked to the population and data items described by the data, and to the products which draw on the data. It is also associated with one or more topics drawn from a prescribed list. Topics are used by search routines to facilitate identification of required datasets. The search routines will eventually access synonyms and related terms recorded in a thesaurus in order to extend search power. For example, a search for datasets involving "cars" or "automobiles" will also include "motor vehicles" which is the name actually used in the data set titles.

The dataset is the core entity. It is the principal object of data searches, being preferred to the data item entity, because data items never appear singly, they always appear in a context, and that context is embodied in a dataset.

A dataset is generated from one or more cycles of a collection. Associated with each collection cycle are descriptions of both the procedures it embodies and the operational metadata resulting from the cycle, for example response rates and cost. Also associated with a collection cycle is the collection instrument (or instruments), comprising question modules, which generate data items. Each data item is linked to a definition and other information relevant to its conception and use. In addition, a qualitative data item is linked to the classification which defines its value set. A glossary of statistical terms, linked to the thesaurus, is available to support descriptions, not only of datasets and collection cycles as indicated in Figure 3, but of all other entities.

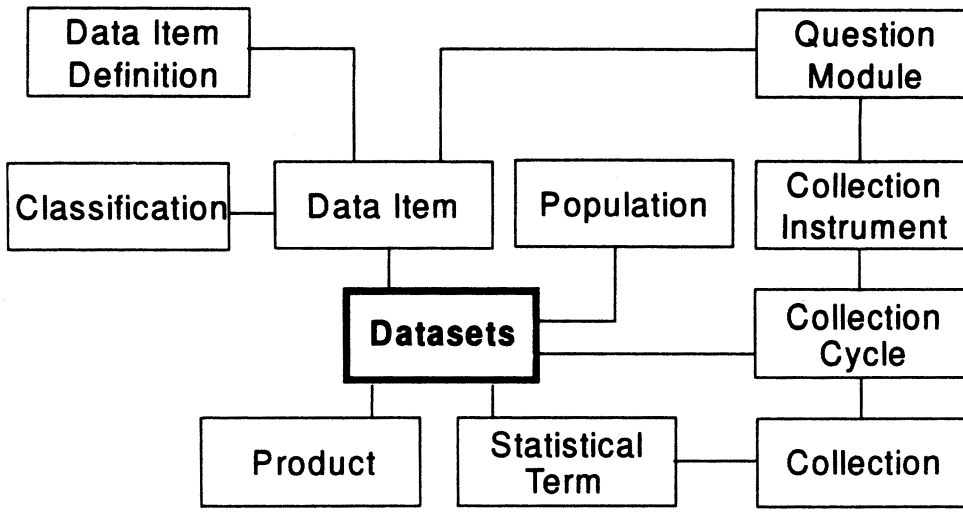

Figure 3. *Warehouse high level data model*

3.2 Metadata Naming Conventions, Standards and Policies

Naming conventions, templates, policies, standards and recommended practices have been and are being developed for defining, classifying, and categorising the Warehouse metadata entities to assist in their organisation, maintenance, and retrieval. When the Warehouse development began, there were virtually no relevant international or accepted metadata standards. Now there is a variety of standards at various stages of production, including the international draft standard for data elements (ISO/IEC 11179, 1994–96), standards developed in the US for spatial data transfer and for cultural and demographic metadata, the proposed "Dublin Core" metadata standards for unstructured data, the US Government Information Locator Services Standard, and the generic statistical message standard (GESMES) for exchange of array data developed by Eurostat (1995). Furthermore, there are numerous ongoing initiatives aimed at creating more comprehensive standards, for example, the Joint Workshop on Metadata Registries, University of California, July 1997. Development of metadata standards is an area ripe for international co-operation, not only to facilitate sharing new ideas, but also to underpin consistency of data at the international level.

Warehouse metadata are managed in terms of five broad categories.

Definitional. Definitional metadata relate to statistical units and populations, topics, classifications, data item definitions, questions and question modules, and statistical terms in the glossary and thesaurus. There are no specific naming conventions for any of these metadata entities. Warehouse users may define new data items, classifications, questions and statistical terms as they wish, but creation of populations and topics is controlled. The attributes of data item definitions which were determined before the advent of an international standard will be extended to make them ISO 11179 compatible.

Procedural. Procedural metadata relate to the procedures by which data are collected and processed. There is a prescribed template for recording metadata. It contains 100 or so fields, organised into 15 groups. Some fields require selection from prescribed picklists, for example, to indicate the type, frequency and reference period of a collection. Other fields allow free text entry and are used

to record a summary description of each aspect of the collection process, for example, follow-up procedures, non-response adjustments. There is also provision for recording links to more detailed documentation contained elsewhere. In summary, each record contains a structured synopsis of the entire collection process and an index to selected, in depth references.

Operational. Operational metadata arise from and summarise the results of implementing the procedures. They include measures of respondent burden, response rates, edit failure rates, costs, and other quality and performance indicators. To date, templates have been defined only for response and respondent burden measurements.

Systems. Systems metadata are active metadata used to drive automated operations, including, for example, file layouts and access paths.

Dataset. Dataset metadata comprise the minimal systems metadata required to describe, access and update datasets. They are categorised separately from other metadata in view of their significance for the first phase of the Warehouse development. There is a naming convention for dataset titles, namely: *population: data item by (up to five classifications), (periodicity or time); comments*, generating, for example a title like: "Deaths: Counts by Major Cause, by Age Group, by Statistical District (1991)".

To accompany the introduction of the Warehouse, a data management policy was formulated and ratified by senior management (Australian Bureau of Statistics, 1996). There was no difficulty or delay in getting broad support for the policy statements as they are essentially common sense. They stipulate, for example (1) that data cannot be collected from respondents until the collection objectives and principal outputs have been documented and recorded in the Warehouse, (2) that data cannot be disseminated until the collection procedures are documented and the output datasets are recorded in the Warehouse, and (3) that any concepts, sources, and methods descriptions associated with a publication or planned dissemination activity must be drawn from the Warehouse. These statements certainly underpin the drive for integration. Other statements (reproduced below) are even more explicit in their support.

"1. Statistical integration provides the basis for relatable data. It may be viewed in terms of three related components:
   - rationalisation, standardisation and integration of concepts (statistical units, data item definitions, classifications), and data inputs (questions, forms);
   - rationalisation, standardisation and integration of collection procedures (sample design, collection, processing, estimation, tabulation and dissemination);
   - confrontation, analysis, rationalisation and physical integration of data outputs.
2. The objective of this policy is to ensure that statistical integration initiatives occur, fully exploiting the definitional, procedural, operational and dataset metadata contained in corporate repositories.
3. It is ABS policy to advance statistical integration by conducting an ongoing program of reviews of statistical units, data item definitions, and classifications, resulting in recommendations and directives for standardisation.
4. It is ABS policy to conduct an ongoing program of reviews of sample design, collection, processing and dissemination procedures, resulting in recommendations and directives for adoption of best practices.
5. It is ABS policy to conduct an ongoing program of reviews of data outputs, resulting in recommendations and directives for improving data relatability."

Notwithstanding widespread support in principle for the policy, implementation may well take several years.

3.3 *Warehouse Metadata Functions*

Figure 4 (reproduced from Colledge & Richter, 1994) illustrates the principal Warehouse metadata functions as of August 1997. The functions on the left hand side are those used to get metadata in and out of the Warehouse; those on the right hand side are for viewing, retrieval and dissemination of data. In ballpark terms, as of August 1997, about 90% of the systems development for the first (dataset output) phase has been completed. One significant gap in functionality is the ability to merge datasets having different structures. The conditions under which merging makes sense have yet to be worked out. As regards the second (metadata based integration) phase, the systems development is perhaps 50% complete.

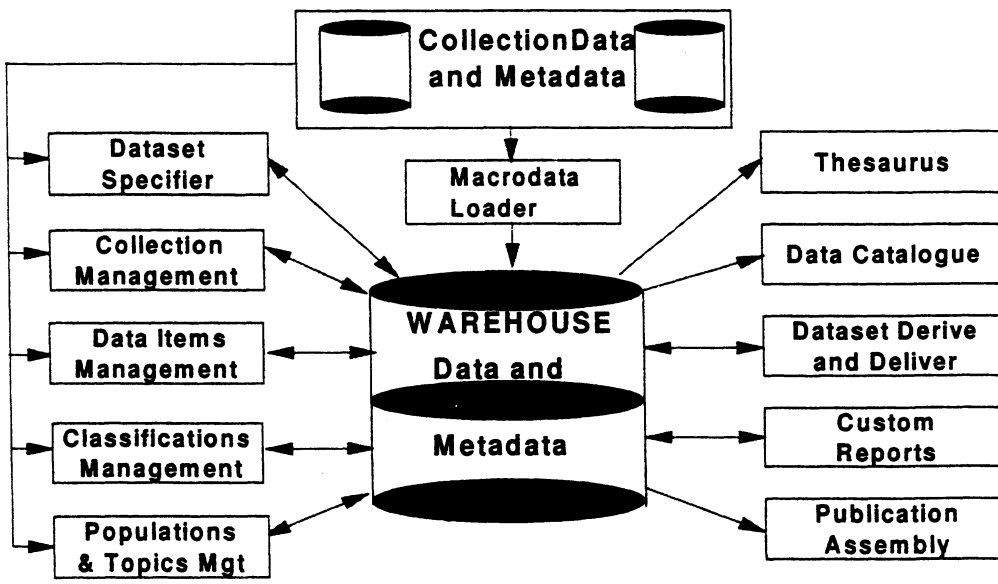

Figure 4. *Principal warehouse functions*

Because the Warehouse was a novel development when it began in 1993, the functional requirements were not easy to specify in advance. Sometimes there was not even an obvious client to service and to be responsible for specification and evaluation. Thus, a prototyping approach was used, and, on the whole, it has proved very successful. For example, a stand-alone prototype classification comparison tool was developed, based on very sketchy requirements. It was given to standards area staff for experimentation in detecting and comparing classifications with different names but similar or even identical structures and meanings. Subsequently, based on much more solidly defined user requirements, the tool was rebuilt and incorporated in the Warehouse suite. Usability concerns were and are being addressed using Hiser Group methodology which entails systematic recording and analysis of client actions as they use the facilities for specified tasks.

In any project with a four year time span there is danger of development being overtaken by new technology. In 1993, the ABS was using Microsoft desktop products and the Internet access was very restricted. Now ABS is a Lotus Notes shop with Internet access through Notes Domino. This has had two major consequences for Warehouse functionality.

First, the original plan was to house the entire Warehouse metadata suite in an Oracle database with a SQLWindows interface. However, in view of the Notes desktop which is ubiquitous throughout the agency, there has since been a switch to the use of Notes databases for the lightly structured and unstructured data, such as paragraphs of descriptive metadata. The interfaces between these Notes and Oracle databases and software are not yet so seamless as to make this an easy transition.

Second, the initial focus was on creating a Warehouse to service staff through the ABS internal network. The new goal of making Warehouse data and metadata accessible outside the agency using the Internet is an, as yet not fully defined, third phase.

4 Warehouse Based Integration Activities

4.1 Framework

Given that conceptual integration is a prerequisite for physical integration, one might think that conceptual integration alone would produce most of the gains associated with integration. In other words, that physical integration of metadata with a central repository would not be necessary so long as these metadata were maintained by the collection areas in accordance with the appropriate standards. However, experience has shown that, unless conceptual integration is put to the test in terms of physical integration, it may not actually occur. This has been illustrated time and again during Warehouse loading of data and the metadata that are supposed to describe them. The loading process has revealed many discrepancies in the metadata that would otherwise have continued unnoticed, at least until detected by users outside the agency! Thus, the Data Management Project has adopted an approach of conceptual *and* physical integration.

Given that data and metadata are to be physically integrated by loading them into the Warehouse, an approach involving detection and correction of inconsistencies prior to or during the loading process was considered. This seemed intuitively preferable to loading inconsistent metadata and then trying to clean up the mess. However, past experience in attempting to resolve inconsistencies before or while loading suggests that this sort of process almost invariably gets bogged down. First, it is difficult to view or anticipate and thus plan for the inconsistencies which will be encountered. Second, loading the metadata to a central repository usually implies additional work for the collection staff, at least in the first instance, and any additional barriers, such the requirement for removal of inconsistencies can bring the loading process to a near standstill. Thus, the Data Management approach is to "load, confront, and integrate". This implies, first, loading metadata within a framework that ensures whatever consistency is possible without involving significant extra work, next, reviewing all the inconsistencies, and finally, taking steps to resolve them.

For convenience of description, integration related activities are divided into four groups:

1) specification, development and introduction of Warehouse systems and procedures to support integration—as outlined in the previous section;
2) loading metadata into the Warehouse—not an integration activity *per se*, rather a precursor to integration;
3) "clean up" activities—dealing with the inconsistencies that are revealed following loading;
4) "ongoing integration" activities—activities that ensure that integration, once achieved, is maintained.

The distinction between clean up and ongoing integration activities may be explained in terms of a polluted lake (representing the Warehouse with inconsistent metadata), being supplied by a

polluted stream (representing an incoming flow of inconsistent metadata from collection areas). The aim of clean up activities is to clean up the lake (remove inconsistencies in the Warehouse). The aim of ongoing integration is to clean up the stream (ensure collection areas create no further inconsistencies).

4.2 Loading the Warehouse

Warehouse loading began in 1994. As of August 1997, in ballpark terms, perhaps 50% of the ultimate stock of datasets has been loaded. It is a slow and painstaking process. In the first place, until the Warehouse replaces rather than parallels existing data dissemination systems, its loading implies extra work for the collection areas. Second, even within the framework imposed by the Warehouse data model, there are a variety of alternative ways in which the clean, edited microdata generated by a collection can be structured into datasets. Best practices for defining datasets, and the accompanying annotations, have been slow to evolve. There are many tradeoffs to consider, most of which have a bearing on integration. For example, it is important to find the appropriate balance between making datasets as large and disaggregated as possible (to preserve the maximum information content) and dividing the data into separate, smaller datasets (which are easier to load, faster to access, meet many users needs, but contain less information in total).

Population, classification and data item label metadata have been loaded in conjunction with datasets which reference them. As regards other definitional metadata, the Warehouse contains a fairly comprehensive bank of economic data item definitions, and a prototype glossary and a thesaurus. There are procedural metadata of variable quality and completeness, describing one cycle of every collection conducted over the last six years, but almost no operational metadata as yet. In summary, the loading of definitional, procedural and operational metadata is perhaps 40% complete.

Having metadata in the Warehouse rather than in separate collection specific databases allows them to be viewed together. The set of collections has been fully enumerated and can be analysed ensemble for the first time. For example, it is now known that, over a 6 year period to June 1997, the ABS has conducted, or is still conducting, 300 collections, of which 15 are in current development, 29 have a monthly reference period, 33 are by-products of administrative processes, etc. There are some interesting challenges from the integration perspective. For example, at the latest count there were 27 different versions of the data item "sex".

4.3 Clean up Integration Activities

Clean up activities are required to deal with inconsistencies due to previous lack of integration. They include review and rationalisation of all the principal entities in the Warehouse—collections, data items, classifications, populations, topics, statistical terms. This is painstaking work which few people really enjoy, particularly as it often has to be accomplished in addition to regular activities. Data Management staff do not have the resources of skills to undertake these clean up activities themselves, and even if they had, they would be the wrong people as they do not own any of the Warehouse metadata. The Data Management role is facilitator, expediter, provider of tools, and exploiter of opportunities, while the collection, standards and statistical service areas have to do the actual integration work.

4.4 Ongoing Integration Activities

Quality management involves building quality practices into processes so that they become the normal way of doing things, not extra tasks that are liable to be dropped when resources are tight or difficulties encountered. Integration is an aspect of quality, and the same approach is appropriate. Thus, the Data Management long term aim is to incorporate integration activities into the core collection processes. With this in mind, generic descriptions have been constructed of statistical processes such as the development of a new collection, the development of a new collection cycle, and the conduct of a collection cycle. They will form the basis of future workflow applications which will reengineer current processes, not only assisting ABS staff to perform their functions more easily, but ensuring that optimum use is made of Warehouse facilities to promote and maintain integration, including control points to check compliance with data management policy.

In the shorter term, prior to the impact of major workflow reengineering initiatives, Data Management will likely use a mixture of "stick" and "carrot" to maintain metadata in support of integration. An example of the stick approach will be monitoring performance against data management policy and building appropriate targets into workplace agreements. The carrot approach involves the provision of new facilities that support integration and will sell themselves by assisting agency staff to carry out their tasks more easily and effectively. It is believed that the latter approach will be more effective, certainly in the longer term, reflecting the "tools not rules" strategy recommended by Willeboordse & Ypma (1996).

5 Integration Through Metadata Management at Other Agencies

5.1 Introductory Remarks

Many other statistical agencies are engaged in initiatives to promote statistical integration through metadata management. The following paragraphs contain very brief summaries of developments at seven other agencies—two in Australia, five outside—based on published and working papers, and correspondence and comments of the staff responsible. This does not constitute a detailed review, which would require a paper in itself. Nor is there an implication that these agencies are the only ones deeply involved in such developments. Rather, this section is intended simply to validate the general direction in which the ABS is moving, to give a flavour of the similarities and differences in approach at other agencies, and to emphasise the opportunities which exist for exchange of ideas, standards, and even software.

5.2 Australian Institute of Health and Welfare

The Australian Institute of Health and Welfare has developed a "Knowledgebase" with the general aim of promoting the sharing of health related metadata. The Knowledgebase can be dynamically interrogated through the Internet and contains the national dictionary of health related definitions and concepts mapped to an enterprise level health information model, together with agreements to collect health data and relevant work programs. The definitions are uniquely identified and recorded in accordance with ISO/IEC (1994–96) 11197 standard for data elements. The Knowledgebase provides links but not automated access to the data to which it refers. Although designed for Australian health statistics, the Knowledgebase would be equally applicable to statistics for any other geographical or subject matter area. More documentation is available directly from the agency's Web site (www.aihw.gov.au).

5.3 *(Australian) Bureau of Resource Sciences*

On behalf of the Australia New Zealand Land Information Council (ANZLIC) and the Commonwealth Spatial Data Committee, the Bureau of Resource Sciences is building and maintaining a directory of spatial datasets. A "spatial dataset" is defined to be any dataset which has a spatial component. Interpreted broadly this includes not only maps, photographs, etc., but statistical data with a regional component. The directory does not contain the actual data, only metadata. The object of the directory is to ensure that spatial data produced by federal and state governments are readily visible. Metadata guidelines (ANZLIC, 1996) have been developed to support the dataset registration and recording process. The guidelines define the core ("page 0") metadata elements that each submitting agency is expected to provide. The metadata are presently stored in the National Directory of Australian Resources and accessible through the Internet (www.nric.gov.au).

5.4 *Statistics Canada*

Statistics Canada has initiatives which parallel the two phases of the ABS data management project. The CANSIM2 redevelopment described by Bassett & Stoyka (1996) aims to create a database which will contain all the agency's aggregate output data together with the corresponding metadata, and to provide access to these data via the World Wide Web. The corporate meta-information base and collection template described by Priest (1996a) consolidates metadata from the collection areas into a single, authoritative, publicly browsable repository. A pilot corporate meta-information database has been created and is now in active use as the major reference tool in the ongoing harmonisation activity, proposed by Priest (1996b). The initial focus of the initiative is a core set of data items (variables) that are likely to be included in any social survey and/or the next census, belonging to areas for which experienced, subject specialised expertise is available. For each data item, the existing range of concepts, definitions and (sometimes) processing specifications are analysed, and a "best practice" proposal is developed as the basis for discussion, negotiation and finally adoption by the relevant program areas. As of June 1997, best practice standards for about 170 data items had been prepared. Consultation is also taking place with internal and external subject matter committees, and is currently being extended to include other statistical agencies with a view to moving closer to international practices to the extent possible.

5.5 *US Bureau of the Census (USBC)*

The USBC is also engaged in developments closely paralleling the ABS Warehouse. As noted by Gillman *et al.* (1996), the agency has enthusiastically endorsed the same approach to data management as was recommended by Sundgren (1991) for the ABS. It is developing metadata standards and a central metadata repository which will be publicly accessible (Laplant *et al.*, 1996). A comprehensive metadata model has been developed and adopted by the major data dissemination projects in the agency. Work to build a corporate metadata repository based on the model has started. The repository will support the metadata needs of the data dissemination projects, and hopefully takes into account the broader based ambitions for a US National Statistical Information Infrastructure described by Dippo & Tupek (1997).

5.6 *Statistics Sweden*

Statistics Sweden's plans and most recent developments in the area of output databases containing data and metadata are also aligned with those of the ABS, which is not at all surprising in view of the architectural role which Professor Bo Sundgren has played in both cases, and the ongoing exchange of ideas between the two agencies. On 1st January, 1997, Statistics Sweden's "Statistical Databases"

came into operation. They contain data structured in a similar fashion to the ABS Warehouse's datasets. Likewise they contain structured data item and classification metadata, textual descriptions of collection procedures structured in accordance with a prescribed template, and links between the data and metadata elements which allow traversal in any direction. The facilities are in advance of the ABS Warehouse in that they enable the information to be publicly viewed and accessed through the Internet (www.scb.se). Statistics Sweden has made a commitment that all official statistics for which the agency is responsible will be available via the Statistical Databases by the year 2000.

Sundgren (1997), from which this paragraph is drawn, provides a comprehensive description of the Statistical Databases, outlining the basic and value added services and pricing policy which accompany their introduction, and stressing the importance of the the open systems architecture which allows additional databases and software to be readily bolted on. It is evident that the development embodies a suite of tools which will provide great support for integration activities.

5.7 Statistics Netherlands

Statistics Netherlands is also involved in similar initiatives. Keller (1996) outlined the development of the the agency's STATLINE database which parallels the ABS Warehouse output database. A STATLINE prototype is now available for dynamic interrogation via the Web (statline.cbs.nl). In November 1996 the agency hosted the Conference on Output Databases, which provided an excellent forum for exchange of ideas. Willeboordse & Ypma (1996) defined "coherence" (meaning integration in the terminology of this paper) in terms of five levels, namely, well defined concepts, unequivocal terminology, reduction in the number of concepts, coordination of concepts, and coordination/tuning/integration of processes. They noted that STATLINE did not in itself integrate data and they outlined the agency's route towards coherence in terms of short, medium and long term objectives. In essence, the short term objectives are more standardisation and use of user oriented terminology in tables and explanatory notes; the medium term objectives involve systematic tracing and treatment of major inconsistencies and redundancies; and long term objectives include modification of STATLINE to enable merging of data from different collections and creation of a supporting metadata base containing, in the first instance, classifications and data item definitions, relations, keywords, synonyms, thematic clusters, representations and value sets. Kent & Schueroff (1996) provide more systems details.

In an entirely separate development described by Keuning (1997), the basic principles of national accounts are being applied to integrate a much wider range of data. The System of Economic and Social Accounting Matrices and Extensions (SESAME) is a single unifying information system from which a core set of economic, environmental and social indicators is derived.

5.8 Eurostat

In January 1996, the Integrated Meta-Information (IMIM) project was initiated at EUROSTAT. It is a three year project of which the primary objective is to facilitate the metadata management processes within statistical agencies by providing methods and tools for integrating the many (complete and partial) metadata sources and flows that usually exist. More details can be obtained from the IMIM Web Site (www.imim.scb.se).

6 Concluding Remarks

Integration is a component of "quality" defined in the broad sense of fitness for use, and, like quality management, integration is a multifaceted way of life, not a concrete deliverable which will, one day, be achieved.

There are (at least) two aspects of integration: conceptual and physical. In principle, conceptual integration alone should produce most of the benefits, but practice shows that it must be supported by physical integration, otherwise, for all sorts of reasons, it may not actually occur.

A prototyping approach is ideal for development of new systems in previously uncharted areas, such as metadata management, as users cannot initially specify exactly what they want. Prototypes may bring some immediate gains as well as fleshing out requirements.

An infrastructure for recording and accessing metadata is a prerequisite for integration. Given an infrastructure, there are essentially two types of integration activities: "cleaning up" the results of previous lack of integration, and "ongoing integration".

An approach analogous to prototyping is suitable for clean up activities, i.e., in very broad terms, load first and rationalise afterwards. The reasons for this approach are, first, it is not obvious what rationalisation is possible until the metadata have been brought together, and, second, loading is time consuming and resource intensive, often with no immediate benefits to the loaders, so it is important not to put any more barriers in the way.

As regards ongoing integration, the key is to build integrating activities into core processes by ensuring that integrating metadata play an active rather than passive role.

Until very recently, agency dissemination practices have focused almost entirely on data and scarcely at all on metadata. This may give rise to an assumption that data and metadata should be handled in a similar way as regards dissemination. In fact, almost the opposite is the case. Whereas the confidentiality of individual data must be protected under an agency's statistics act, the act does not apply to metadata which are created by the agency at tax payer's expense and to which the public typically have a right under a freedom of information act. Transparency of government operations is an important principle. The US and New Zealand governments, and probably others too, have explicit commitments to making metadata widely available through the Internet. Users will demand metadata—not only concepts and definitions, but procedures and operational statistics. It will become increasingly difficult to conceal dubious statistical procedures, in particular, lack of integration, behind a bogus cloak of confidentiality.

An increased focus on metadata accessibility will have profound and positive implications for integration across agencies and countries. Just as librarians have defined and use common cataloguing standards to share catalogue entries instead of creating them individually, so statistical agencies need common metadata frameworks, as a basis for sharing data item definitions, classifications, procedures, etc. As of August 1997, there are few internationally or even nationally recognised standards such as ISO 11179 for data elements. Standards are required for the statistical metadata model and its principal elements, for metadata release and archiving practices, and so on. There is little virtue in each statistical agency continuing to invent its own metadata standards elements in isolation.

An essential feature of any metadata model and its practical implementation is that it should be possible to traverse the entities in any direction which makes sense.

The Internet is a major enabling mechanism. It supports integration by enabling metadata exchange between agencies and dissemination to users. Possible stages in developing its use are (1) dissemination by static (preformed) Web pages, (2) dissemination by dynamic Web pages, i.e., created in response to user interrogation, (3) two way flow of metadata involving dissemination by dynamic Web pages and receipt by e-mail, (4) two way flow with dissemination by dynamic Web pages and dynamic update; (5) multimedia extensions—pictorial representations of hierarchical classification structures, for example, of the relationships between units and data items. An agency's internal security will typically be preserved along the lines outlined by Sundgren (1997), by use of an internal network separated by a firewall from an external network with is open to the Internet.

Comparison of ABS integration activities with those of other agencies suggests that the ABS is going in the right direction, or, at the very least, in a commonly followed direction!

Acknowledgements

This paper is largely based on the work of ABS Data Management Project team members, past and present, including, in particular, Warren Richter (project manager), and Rob Edmondson, Craig Watson, and Brian Studman (Warehouse architects). The author would also like to thank Peter White, Nigel Mercer, and Joe Christienson at the Australian Institute of Health and Welfare, Gordon Priest at Statistics Canada, Dan Gillman and Marty Appel at the US Bureau of the Census, Jean-Paul Kent, Martin Schuerhoff, and Ad Willeboordse at Statistics Netherlands, and Bo Sundgren at Statistics Sweden for information about integration related activities at their agencies, and Susan Linacre for comments on content and style.

References

Australia New Zealand Information Council (1996). Meta Guidelines, Version, 1, July 1996, Australia New Zealand Land Information Council, Belconnen, ACT, Australia.

Australian Bureau of Statistics (1970). Australian Integrated Economic Censuses, 1968–69, Chapter 31, *Official Yearbook of the Commonwealth of Australia*, **56**, Belconnen, ACT, Australia.

Australian Bureau of Statistics (1996). Statistical Data Management. *Policy and Legislation Manual*. Australian Bureau of Statistics, Belconnen, ACT, Australia.

Bassett P. & Stoyka, A. (1996). Statistics Canada's Aggregate Output Database–CANSIM II. In *Proceedings of the Conference on Output Databases*, Voorburg, November, Statistics Netherlands, Voorburg, The Netherlands.

Colledge, M.J. & Richter, W. (1994). Data Management and the Information Warehouse: Infrastructure for Redevelopment. In *Proceedings, Statistics Canada Symposium 94*, Statistics Canada, Ottawa, Canada.

Colledge, M.J., Wensing, F. & Brinkley, E. (1996). Integrating Metadata with Survey Development in a CAI Environment. In *Proceedings, Annual Research Conference 1996*, US Bureau of the Census, Washington, USA.

Colledge, M.J. (1998). Commonwealth Government Statistical Clearinghouse. *Small Enterprise Research*, Volume **6**, No. 1, University of Newcastle, Callaghan, NSW, Australia.

Dippo, C.S. & Tupek, A.R. (1997). Creating a National Statistical Information Infrastructure. In *Bulletin of the International Statistical Institute, Proceedings of the 51st Session*, Istanbul; Voorburg, The Netherlands: International Statistical Institute.

Eurostat (1995). *GESMES The International Standard for Exchange of Array Data*. Statistical Office of the European Communities, Luxembourg.

Gillman, D.W., Appel, M.V., Laplant, W.P. Jr. & Sundgren, B. (1996). Towards a United Data and Metadata System at the Census Bureau. In *Proceedings of the Annual Research Conference*, March 1996, US Bureau of the Census, Washington, USA.

ISO/IEC 11179 (1994–96). *Specification and Standardization of Data Elements*. International Standards Organization, Geneva, Switzerland. Parts 3–6 available via Internet (http://www.iso.ch/cate/cat.html).

Kent, J.-P. & Schuerhoff, M. (1996). *Some Thoughts about a Metadata Management System*, paper presented to InterCASIC, November 1996; Voorburg, The Netherlands: Statistics Netherlands.

Keller, W.J. (1996). EDI: Electronic Data Interchange for Statistical Data Collection and Dissemination. In *Proceedings of the Annual Research Conference*, March 1996, US Bureau of the Census, Washington, USA.

Keuning, S.J. (1997). SESAME: an Integrated Economic and Social Accounting System. *International Statistical Review*, **65**, 111–121.

Laplant W.P. Jr., Appel, M.V., Gillman, D.W. & Lestina G., Jr. (1996). Proposal for a Statistical Metadata Standard. In *Proceedings of the Annual Research Conference*, March 1996, US Bureau of the Census, Washington, USA.

Lee, G. (1996). *Synergy between Survey Computing, Data Quality and Methodological Improvement.* Belconnen, ACT, Australia: Australian Bureau of Statistics.

Priest, G. (1996a). *A Corporate Metadata Information System.* Internal Working Paper, Statistics Canada, Ottawa, Canada.

Priest, G. (1996b). *The Issue of Harmonization of Data from Diverse Sources.* Prepared for Eurostat Workshop, Statistics Canada, Ottawa, Canada.

Richter, W. (1993). *Project Definition Document.* Working Paper, Australian Bureau of Statistics, Belconnen, ACT 2616, Australia.

Richter, W. (1996). The ABS Information Warehouse—Present and Future. In *Proceedings of the Conference on Output Databases*, Voorburg, November 1996, Statistics Netherlands, Voorburg, Netherlands.

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.

Sundgren, B. (1997). *Sweden's Statistical Databases: an infrastructure for flexible dissemination of statistics.* Report to the UN/ECE Conference of European Statisticians, June 1997; Statistics Sweden, Stockholm, Sweden.

Willeboordse, A. & Ypma, W. (1996). From Rules to Tools—New Opportunities to Establish Coherence among Statistics. In *Proceedings of the Conference on Output Databases*, Voorburg, November 1996, Statistics Netherlands, Voorburg, Netherlands.

Résumé

La vitesse et la versatilité de la nouvelle technologie génèrent une demande des utilisateurs pour avoir une plus vaste gamme de produits des agences statistiques qui doivent s'assurer que ces produits sont plus cohérents et reliés entre eux lorsque possible. L'intégration statistique est un façon d'améliorer le contenu des enquêtes indépendantes menées par un bureau de statistiques. Elle est nécessaire pour assurer la cohérence des données. Elle a deux aspects—conceptuel et physique—le premier aspect étant une condition pour le deuxième. Cet article s'intéresse aux méthodes permettant d'assurer l'intégration statistique par une meilleure gestion des données. L'article inclue des expériences du Bureau australien de statistiques au sujet du développement et de l'usage d'un répertoire central (le "Information Warehouse") pour les données et les métadonnées. Il fait aussi référence aux activités similaires chez quelques autres bureaux nationaux de statistiques.

Les conclusions principales sont les suivantes. Premièrement, quand on développe des fonctions nouvelles pour appuyer l'intégration statistique, il faut utiliser une approche prototype parce que on ne connaît pas à l'avance les outils nécessaires. Deuxièmement, on ne peut pas rationaliser facilement les métadonnées des enquêtes indépendantes tant quelles ne sont pas dans le répertoire et visibles les unes à côté des autres pour que leurs incohérences soient manifestes. Troisièmement, il faut appuyer l'intégration conceptuelle avec l'intégration physique. Quatrièmement, il y a plusieurs possibilités pour des associations et des échanges d'idées entre les bureaux nationaux de statistiques. Finalement, on doit inclure l'intégration statistique à l'intérieur des enquêtes et la considérer comme un mode de vie.

Appendix: Terminology as Used or Implied in the Paper

active metadata	metadata determining the actions of automated survey processes; also referred to as "prescriptive metadata".
best practice	a practice based on experience and promoted by an agency as being consistent with its objectives; not obligatory, from which deviations are allowed with justification.
classification	a set of mutually exclusive and exhaustive categories into which members of a population can be divided.
coherence	applied to data; means the same as consistency (qv).
collection	by implication a "statistical collection" (qv).
consistency	mutual compatibility, coherence; is applied to datasets rather than concepts; coordination of concepts is a necessary but not sufficient condition for consistent data.
coordination	associated with concepts rather than datasets; implies concepts are understood and aligned without necessarily being identical, i.e., implies the use of the units, classifications and data items that have the same definitions, *or* definitions with known and stated differences; can be the result of an agreed compromise (c.f. harmonisation).
data item	an atribute of a population member or of the population as a whole; synonymous with "variable".
data management	management of data and associated metadata throughout the life cycle of a statistical collection.
dataset	the metadata describing a table of data referring in essence to a single population and a specific set of data items, together with, optionally, the actual data.
guideline	recommended practice (qv).
harmonisation	user of the same unit, classification and data item definitions, typically but not necessarily those recognised as being standard; a state of harmony, not just a compromise—a stronger requirement than coordination.
integration	the state of having statistical data which are mutually consistent and are related to the greatest extent possible; the set of activities which aim to produce data in this state; includes both "logical" ("conceptual") integration and "physical" integration; includes "coordination" and "harmonisation"; applies to procedures and systems as well as concepts and definitions.
metadata	data about data; refers to the definitions, descriptions of procedures, systems parameters, and operational results which characterise and summarise statistical programs.

metadata management	management of metadata throughout the life cycle of a statistical collection.
metainformation	(in this paper) metadata; sometimes (but not here) metadata used by humans rather than machines.
nomenclature	synonym for classifcation (qv).
norm	informal standard (qv).
passive metadata	metadata in the form of documentation; also referred to as "descriptive" metadata; in contrast to "active metadata".
prescriptive metadata	metadata determining the actions of automated survey processes; also referred to as "active metadata".
policy	a directive from agency management, from which deviation requires strong justification.
recommended practice	a practice promoted by an agency as being consistent with its objectives; not obligatory, from which deviations are allowed without justification.
standard	an approved or designated version.
standardisation	the use of standards; a tool for harmonisation.
statistical collection	the activity of collecting data primarily for the purpose of producing summarised information or scientific inferences; excludes the collection of data for the purposes of monitoring, controlling or otherwise focusing on the activities of individuals.
statistical integration	(in this paper) synonym for integration (qv).
survey	(in this paper) statistical collection (qv).
variable	synonymous with data item (qv).

[*Received February 1998, accepted August 1998*]