

# Experiences and Plans of the Australian Bureau of Statistics related to Data and Metadata Exchange

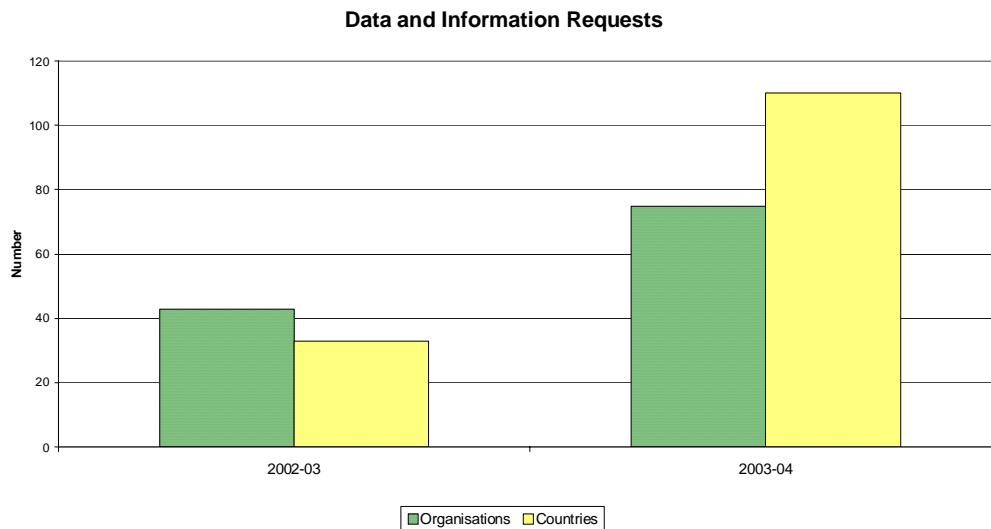
By Graeme Oakley, Alistair Hamilton, Jeremy Michel  
Data Management Branch  
Australian Bureau of Statistics

## A. Introduction

1. In common with other National Statistical Offices, the Australian Bureau of Statistics (ABS) has strategies it is pursuing to make greater use of electronic media for dissemination. The print dissemination environment is declining. The ABS has been positioning itself to use the Internet as the principal channel for dissemination. Our objectives are to: increase the users and uses of statistics for informed decision making; increase user understanding of the content, caveats, contexts and limitations of statistics; and improve cost effectiveness. Among ABS strategic directions are to: improve communication of statistics through a layered approach - simple presentation first through layers to most complex, contextual linking of metadata; improve self help through support for on-line data cubes and manipulation tools and web services; and 'writing once and publishing many times' by developing good data management practices and supporting infrastructure.

2. The ABS provides information to a large number of international organisations and other countries, in a range of formats. In addition, there are domestic customers who received data in electronic formats of many varieties. Data and information requests from both international organisations and specific countries have increased markedly over the last two years. While some requests can be accommodated by pointing to the relevant area of our website this work represents a significant increase of time in managing the requests as well as responding.

## Data and information requests 2002-03 to 2003-04



## Data and information requests from selected international organisations

Organisation	2000-01	2001-02	2002-03	2003-04
International Monetary Fund			7	16
Organisation for Economic Cooperation and Development	2	3		8
United Nations	5	9	11	22
United Nations Economic and Social Commission for Asia and the Pacific	3	6	6	5
United Nations Food and Agriculture Organisation	2	11	7	10
United Nations			2	
United Nations International Labour Organisation	2	4	1	5
UN World Tourism Organisation	3	7	3	2
World Trade Organisation			4	5
Other organisations	3	1	2	2
<b>Total</b>	<b>20</b>	<b>41</b>	<b>43</b>	<b>75</b>

3. This paper describes the experiences of the ABS with our involvement as a pilot agency in the NAWWE project, and our development work to use the SDMX standard. The second part of the paper outlines our future plans with respect to NAWWE, use of SDMX and also an initiative called the National Data Network. There are a number of issues identified that we feel need to be resolved to help our progress. These might form the basis for some group discussion.

### B. Experience to date

#### *NAWWE Project (National Accounts World Wide Exchange)*

4. ABS has had an involvement in this project since about mid 2002. Our progress has been slow with higher priority tasks in the National Accounts Branch often necessitating their establishment work taking a lower priority. In the ABS environment, the subject matter area is responsible for defining and loading datasets for dissemination products to our Information Warehouse (ABSIW). The feed of data to OECD for NAWWE is considered to be another product and dissemination channel. In addition to the subject matter work to set-up a dataset with the descriptive classifications required by the OECD, the underpinning transfer mechanism was changed for NAWWE from a predefined spreadsheet with an OECD supplied macro to generate an XML file, to be based on Version 1 of the SDMX standard. ABS supports the change in the 'plumbing' layer.

5. The current status is that the National Accounts area has:
- \* Created a dataset to support the first four tables of NAWWE Annual Excel questionnaire.
  - \* Set up the specification to deliver that dataset as time series to the ABS XML schema format.
6. The development staff in the Data Management Branch have:
- \* Developed a transform SDMX V1.0 from the ABS XML schema format (although some further enhancements are needed)
  - \* Created specifications for the extra content required in ABS XML to satisfy OECD requirements

- \* Set up a mechanism in our environment that enables the subject matter areas to drive the transform process, ie the subject area has control over the generation and approval of the product
- \* Set-up an area on the ABS Website available for NAWWE files -  
<http://www.abs.gov.au/websitedbs/d3110110.nsf/Web+Services/National+Accounts+World+Wide+Exchange>

### ***SDMX V1 (Statistical Data and Metadata Exchange)***

7. For about the last two years, ABS has been involved in the development of SDMX as an interested observer. We have long held an objective to be able to use an internationally agreed data and metadata transfer standard. This would enable us, as a strategic direction for dissemination, to head towards one format for the dissemination of electronic products and so remove the many proprietary formats that have been used. In addition, ABS had been working on the definition of an ABS XML schema to define fully a statistical data set so that we could use this format as the delivery from the ABSIW (Information Warehouse) to various product creation environments, and ideally use XML family tools, such as XSLT, to easily create products in various formats. We have done this with our implementation of time series spreadsheets in EXCEL on the web site.

8. Our work on SDMX is now overlapping with NAWWE work. Therefore, the comments in paragraph 6 above are relevant. In addition, we have commenced discussions with a significant domestic client who currently receives time series data in a proprietary format. We have provided some test files in SDMX V1 format with the view to move this client to that format, once all the relevant ABS time series can be produced in that format by our subject matter areas. Once this client has accepted the approach, there are a number of other clients that would be our next target for a change in format. One of our major selling points is that the old proprietary format does not carry sufficient metadata about quality.

9. During the course of our development with SDMX, a number of matters arose for us to resolve. A walkthrough of our process highlighted some weaknesses. They were in the following areas:

- \* the content of the source schema (ABS XML) was inadequate. We have initiated modifications to that part of our system that 'delivers' time series into the ABS XML in order to provide the information to better populate the SDMX schema.
- \* our understanding of some of the SDMX concepts, and in this case we sent questions to the developers. (see para 11)
- \* limitations in the definition of SDMX XML resulting in the loss of information (see para 12).

10. The issues related to the content of the ABS XML schema were:

- \* Classificatory concepts required better information. Lacking were:
  - \* Classification name to go into the codelist and concept names
  - \* A way to uniquely identify two uses of one classification in a single key family (eg. Birthplace of Mother and Birthplace of Father)
- \* A concept which maps more closely to key family unique identifier was lacking from the schema. (This issue was dealt with by enhancing the 'delivery' system to include the name of the time series map (external name) to use as the key family unique identifier.)

11. Gaps in our understanding were resolved via correspondence through Stuart Feder at the Bank of International Settlements (BIS). Some of the questions asked and the replies were:

- Q. What data exchange format should we be using?  
 A. StructureMessage and GenericDataMessage

Q. What is the distinction between Attributes and Dimensions?

A. Both Attributes and Dimensions describe the data and aid the interpretation of the data. In addition to this, dimensions identify the data ('without knowing the value for a particular dimension, you could not distinguish the data from other, similar data'). 'Attributes do not have an identification function'

12. ABS XML has a rich support for annotations. They are normalised out into a separate section and can be referenced from almost anywhere in the XML. It is not currently clear how this information can be incorporated into the SDMX schema.

### ***National Data Network (NDN)***

13. Along with many national statistical organisations, the ABS recognises that other agencies of government have data that would be useful in the form of statistics to policy makers, researchers and the community. However to be really useful, this data has to be able to be 'discovered', and ideally would be based on a framework of common classifications, data definitions, and collection methodologies. For many years the ABS has pursued the initiative of a National Statistical System (NSS). More recently, ABS has identified the need for some infrastructure to enable information clients to find and access the large amount of information held by data custodians. This is called the National Data Network. A demonstration version of the NDN is about to be launched. The 'dNDN' will have limited content initially and be supported by a range of metadata to assist with 'discovery' (based on Dublin Core), to inform clients about quality (based on the OECD quality framework), and to provide definitions about the datasets and data elements (based on ISO/IEC 11179). XML schemas have been defined to convey the metadata.

14. In further versions of the NDN, the ABS plans to include web services that would allow other data custodians or information clients to access statistical services, such as to code business units to the standard industry code, to perform time series analysis, to have a registry to hold data element definitions. The infrastructure is being developed, as much as possible, using open source products. Although not in the initial implementation, we intend advising other data custodians of the SDMX standard and encourage its implementation as one of the formats for provision of data and metadata electronically.

### **C. Plans and Issues to be resolved**

#### ***NAWWE***

15. Advice has been received, following the October 2004 national accounts meeting, of the intention to fully implement NAWWE for the collection of annual national accounts data in 2005, using a new questionnaire and time series identifiers developed by OECD to facilitate this objective. A subsequent communication rescinded the identifier approach and instead a redeveloped key family will be provided by OECD, ECB and Eurostat. These late changes will affect our plans. The ABS National Accounts Branch has indicated they are unlikely to be able to undertake the necessary set-up work to deliver 2004/05 annual data in NAWWE format and would continue with the spreadsheet format.

16. There are a number of issues that might be worthy of discussion:
  - a. Is there value in having series identified with an internationally recognised identifier? Where would these identifiers, if it was agreed that they are useful, be stored in the SDMX schema? ABS intends including its unique series id as an extra attribute associated with each series delivered to NAWWE so that we have a linkage back to our source series. It is noted in communication from the Charles Aspden at OECD that

"although we do not want to use time series identifiers for data interchange of the national accounts, they are needed for analysis. In the OECD's annual and quarterly national accounts databases the time series identifiers are composed of the relevant coordinates plus extra characters to ensure uniqueness." Do NSO's derive these and include in the file or will the OECD generate them?

- b. NAWWE project seems to be moving towards a tightly structured key family to identify a series. This causes ABS problems because the concepts in the key family do not closely align with the concepts used by our subject matter areas in the ABS Information Warehouse to describe the data. For example NAWWE use the classification *Industries (ISIC REV 3 / NACE REV 1)* to describe the industry classification and National Accounts use a classification based on the *Australian and New Zealand Standard Industrial Classification (ANZSIC)*. So for NAWWE the category value for *Agriculture, Hunting and Forestry* is AYA and for *Fishing* is AYB, giving a combined code of AYA+AYB, whereas for ABS National Accounts the value for *Agriculture, Forestry and Fishing* is simply A. [As an aside, ISIC and NACE show the code for Agriculture, Hunting and Forestry as A, not AYA]. Is this an issue for other agencies? What might be possible solutions or work-arounds?
- c. The National Accounts data published domestically does not coincide with OECD requirements. There are differences in classifications and classification levels used to describe the data. This situation will require the ABS to create a mapping of the ABS classification to the specification in the NAWWE classification in order to populate the relevant parts of the key family. As described in b., ABS uses different classifications to describe the data it maintains. So if a key family was to be constructed using ABS classifications / categories it would be different from the key families suggested for the NAWWE interchange. So the key family definition entry for industry in the ABS data would be something like; `<structure:Dimension concept="NA_FLT_7" codelist="NA_FLT_7"/>` vs something like `<structure:Dimension concept="Industries" codelist="Industries"/>`. The use of the industry dimension in the data message would also look differently. Eg. `<generic:Value concept="NA_FLT_7" value="A"/>` vs `<generic:Value concept="Industries" value="AYA+AYB"/>`. [Note: NA\_FLT\_7 is the ABS National Accounts version of an 'industry' classification, with some extensions of a pragmatic nature (eg 'ownership of dwellings') to support a desired output presentation.]
- d. The changes in the NAWWE questionnaire will require some reworking of the NAWWE dissemination process. What is the timetable for the work? When we will be advised of the new key family structure etc?
- e. Our understanding of the SDMX V1.0 and how to implement it is still evolving, and so there could be differing opinions between NAWWE contributors, and also some dead-ends. What can be done to minimise this situation?

### **SDMX**

17. ABS plans to continue its work to use SDMX V1.0 to underpin the delivery of time series in electronic format to clients. As this work progresses there have been a number of issues that we refer to the development team (via Stuart Feder at BIS) for advice. Some of the work program elements mentioned earlier in the paper will be addressed during this calendar year. Our intention is to finalise our transform to create compliant SDMX V1 outputs. A quick read of the documentation submitted to ISO on the SDMX standard raised a number of questions for us. These follow and might be worth talking about at the meeting.

- a. Compliance document. The standard requires that an application using SDMX must have a compliance statement if the authors want to claim that they comply with the standard. ABS has drafted such a statement for its transform application - see Attachment A.
  - b. 'Structural Definitions Maintenance Agency (SDMA)' [see line 247 of the SDMX Implementers Guide] concept is described as "an institution that devises key families". Does this concept apply to a national statistical officer? In what circumstances? Who are these 'agencies' for other statistical domains? [It is assumed that OECD is the 'key family maintenance agency' for national accounts, although recent documentation might suggest that Eurostat has that role.]
18. There are some further issues where we need guidance. They are:
- a. structure of Annotations - How are annotations represented in SDMX?
  - b. infrastructure for the creation and maintenance of Key Families; Has someone already created a key family repository? How to integrate this with existing delivery mechanisms?
  - c. What would ABS like to see in SDMX V2.0? See Attachment B for some discussion of the issues as we see them. The meeting might want to talk about some of the matters raised there.

## **Attachment A**

### **Conformance Statement**

#### Application

ABS Information Warehouse Extraction to SDMX XML. This application extracts time series data from the ABS Information Warehouse in an ABS specific (ABS XML) XML format. It then transforms it to SDMX V1.0 via a standard XML transform, then some extra functionality which gets rid of duplicates from the codelists, etc.

#### Message Types Supported:

Structure Message Types;  
SDMX-ML Key Family  
SDMX-ML Concept  
SDMX-ML Codelist  
SDMX-ML Agency

Data Message Types  
SDMX-ML Generic Data

#### Functionality Supported

Application supports write functionality only (no read or delete functionality is supported)

#### Structural Dependencies

All structural dependencies (codelists, concepts, and agencies) must be included inline in the structure message to be supported by the application.

#### Message Format

The Structure Message and Generic Data Message are valid XML and have a root element of StructureMessage and GenericDataMessage respectively.

## Attachment B

### Priorities for SDMX V2.0

1. ABS priorities in regard to new features that could be supported in SDMX V2.0 align closely with the list documented in Part VIII of *Project Team Summary Review of Issues for Version 1.0 SDMX Standards*.
2. Support of hierarchical code lists is an important issue to the ABS. We would favour flexibility in the definition of code lists. For example, ABS modelling allows for cases where the same code is valid at multiple levels of a hierarchy. This saves data redundancy and neatly identifies cases where concepts are identical. In these cases, and in other cases, the ABS also avoids the assumption that meaning is built into the structure of the codes themselves. In other words, while there may be a relationship between the code assigned to a parent and the code assigned to a child (eg '14' has a child of '141') there may not be (eg '14' has a child of '76').
3. If SDMX support for hierarchical code lists is not as flexible then the ABS will need to add layers of mapping and recoding, introducing additional complexity and risk, between our own data and metadata repositories and the production and processing of SDMX files.
4. Support for data cubes is also a priority for the ABS. If the format proved suitable, there is the potential the ABS would seek to describe large and sparse data cubes in SDMX (eg detailed Trade data). One issue at the moment is that this would lead us to significantly exceed the 35 character limit recommended for time series keys. Detailed Trade data does have time as a dimension, but it is not necessarily pre-eminent for analytical purposes. While the key family concept is seen as continuing to apply in the case of such data cubes, its value and relevance as a "time series key" seems to diminish. It will be a challenge for SDMX V2.0 to remain a coherent standard while moving beyond its time series heritage to be able to also efficiently and effectively represent data cubes.
5. The ABS is re-engineering the internal metadata repositories and processes to align fully with ISO/IEC 11179. In addition to international support for the standard, several other agencies within Australia have either undertaken similar work or are planning to do so. We are therefore keen to see maximum alignment between SDMX V2.0 and 11179. The seeking of alignment should form part of the work program for developing SDMX V2.0, rather than mapping occurring only after V2.0 has been developed independently. As noted by the Project Team, this is especially the case because it is planned that SDMX V2.0 address "reference metadata" - a priority which is strongly supported by the ABS.
6. Both standards will continue to evolve over time and it is important the alignment process is not delayed indefinitely waiting for "a perfect opportunity". An update process was completed last year to ensure the other five parts of the standard are now consistent with Version 2.0 of the 11179 metamodel. Unless there is unexpectedly fast progress in regard to defining the next generation of the 11179 metamodel, and/or unexpectedly slow progress in developing SDMX V2.0, it is recommended alignment be with Version 2.0 of the 11179 metamodel. Maximising the alignment between these two standards should assist in regard to the efficiency, consistency and quality of the generation and interpretation of SDMX files.
7. Part VII of the summary from the Project Team notes their decision to make a more complete separation between the technical standards they were developing for submission to ISO and content standards. The reasons for the Project Team making this separation are understood. Across the SDMX initiative more broadly, however, this raises a risk that the agreed technical standard will start being

implemented for many exchange agreements, with each exchange agreement arriving at independent decisions about matters of content that could easily have been based on standards for SDMX Core Statistical Concepts had these been available at the time. Once such agreements have become operational, there is likely to be more significantly most cost, risk and resistance associated with any move to change content to reflect SDMX Core Statistical Concepts.

8. The ABS sees the opportunity to standardise at least a small number of core concepts as a key potential benefit from the SDMX initiative. Conversely, if the same core concept needs to be represented in different ways in different SDMX exchanges this increases cost, complexity and risk. It would be good to see a workplan which aims to have a "core" of SDMX Core Statistical Concepts defined, at the latest, in time for them to be referenced by the implementation guidelines for SDMX V2.0. Recommendations on consistent representation of units of measure, for example, would seem to be a high priority and to address a universal requirement.

9. A related element of work which should maximise the benefits realised from implementation of SDMX V2.0 (and V1.0) would be progressing the definition of a registry framework and, potentially, standard infrastructure based on it. It should be noted Part 6 of ISO/IEC 11179 is broadly applicable to the registration process, including those for metadata items such as "key families" which are not part of the current 11179 "metamodel". Operationalising this framework would, once again, bring major benefits in terms of efficiency, consistency and quality for partners participating in SDMX exchanges. It would be good to see a workplan which aims to have the registry framework defined and supported, at the latest, in time for the implementation of SDMX V2.0.

10. Another issue, which is not considered such a high priority by the ABS but might warrant further consideration during the development of SDMX V2.0, is support for optional association of user defined "time series identifiers" with time series keys. Possible uses and benefits include:

- \* eases the transition process to producing/consuming SDMX messages for organisations that are used to exchanging data as named time series
- \* may improve validation processes for organisations that continue to structure data internally as named time series
- \* selectively naming a few series from a large cross sectional exchange could be used to highlight that component of the data which is likely to be of most interest from a time series analysis perspective
- \* provides an optional ability to allocate short and simple synonyms for series whose time series key is complex and lengthy.

11. It might be possible to identify multiple "sets" of identifiers (eg identifiers used by the ABS vs identifiers used by a different organisation) where each time series key could be associated with 0..1 identifiers from each set. This should support maximum flexibility in use and reuse of a single SDMX message.

12. Finally, anything that can be done in the way of the design of SDMX V2.0, or tools to support it, which assists in the structural validation of messages produced and received would be highly valued.

13. An important aspect in the delivery of data electronically, is that information about the quality of the data source be included in the package. Through the use of annotations associated with individual cells, the data provider should be able to attach annotations related to RSEs, or that the data cell has been 'revised'. It is assumed that this can be handled in V1. At the data source level, many statistical organisations have a quality framework that covers aspects such as Relevance, Accuracy, Interpretability,

Coherence, Accessibility, Timeliness. ABS would like to include in a data delivery to a client, as a minimum, a link to the relevant Quality Declaration on its website by embedding a URL. There could be some discussion about whether or not the SDMX V2 schema carries more information about quality.