STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

COMMISSION OF THE EUROPEAN COMMUNITIES

EUROSTAT

Joint UNECE/Eurostat Work Session on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

Working Paper No. 5
English only

Topic (i): Infrastructure issues for statistical metadata

NEW STRATEGIES FOR METADATA AT THE AUSTRALIAN BUREAU OF STATISTICS

Submitted by the Australian Bureau of Statistics, Australia[1]

Invited Paper

SUMMARY

Since the early 1990's the Australian Bureau of Statistics (ABS) has been engaged in the development of its data management strategy. For some background, I have provided a tabular summary of the evolution of data warehousing in the ABS.

This paper addresses a number of the issues identified for the 'Infrastructure issues for statistical metadata' topic by reference to the ABS experience as we address changing business requirements and a new technology environment. In particular, I will address:

- what we are doing with XML;
- organisational changes that are helping us improve the coordination for maintenance and updating of metadata;
- our 'rearchitecture' project which is enabling ABS to embrace a 'services approach' to data management facilities, including reference to the ABS IT Enterprise Architecture;
- the development of a 'topics list' to underpin the Directory of Statistical Sources;
- strategic issues being considered with ABS senior management; and
- directions with metadata and the ABS website.

These issues are addressed from my personal perspective and do not necessarily represent official ABS views. We have more information than can be included in a paper and I would be happy to engage in a dialogue on any of the issues raised.

---

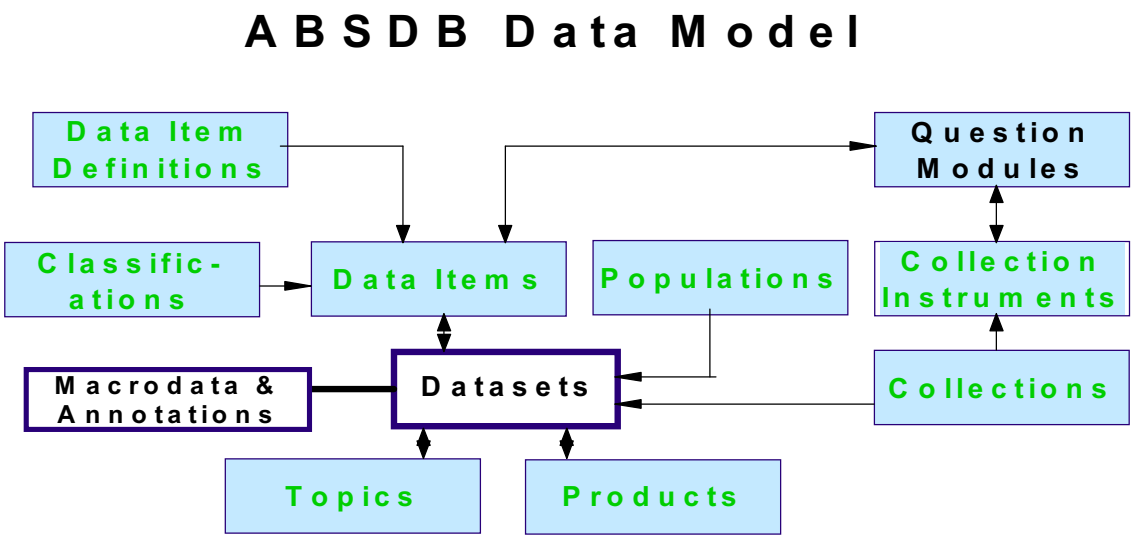
[1] Prepared by Graeme Oakley

I. INTRODUCTION

1. Data management, data item definitions, statistical integration, data dictionaries and user defined processing have been buzz words in the Australian Bureau of Statistics (ABS) for decades. In the early 1990's, terms such as data warehouse and metadata entered the jargon. The most recent cycle of metadata management commenced with a consultancy by Prof. Bo Sundgren and led to the development of the ABS Information Warehouse (ABSIW).

2. What is the ABSIW and what has been achieved in the past decade? The ABSIW (and the Corporate Metadata Repository) is our data and metadata warehouse implementation based on a model for data and metadata that allows the storage of multidimensional cubes, time series and even unit record data, although, in the main, aggregate data is stored in the warehouse. In addition, 'shadow tables' for loaded datasets hold cell level metadata, such as cell annotations, RSE's. The ABS implementation provides access to the seasonal analysis engine, interfaces to a time series toolkit to view and manipulate time series, and interfaces to a variety of vendor products such as FAME and SuperCROSS.

3. The following diagram provides a schematic of the simplified data model. (Not all the classes and linkages have been implemented.)



4. What has been our progress with data management? All subannual collection output (both printed and electronic publications, data cubes and time series) are sourced from the information warehouse and well over 90% of the annual and irregular published output is also sourced from the warehouse. All statistical collection metadata (for some 300+ ABS collections) is defined in the Collection Management System (CMS) and all collection forms have been described to the 'Forms' repository, along with a PDF image of the questionnaire. Over 600 non-ABS data sources have initial entries in the CMS, however the metadata available for these sources is less than for ABS collections. In addition the classification and data element servers hold all standard classifications and data items, along the most of the subject specific metadata objects.

5. Although there are still some concerns about useability of the system and the complexity of dataset design, the ABS Information Warehouse (ABSIW) has become 'part of the furniture' and an integral part of the dissemination stream.

II. EVOLUTION OF DATA WAREHOUSING IN ABS

6. The following table is a summary of how we in the Data Management project see the evolution of the most recent ABS data and metadata management initiative.

Theme	Early to Mid 1990's	Late 90's to 2000/01	Current and going forward
Metadata	• support for basic descriptive metadata • redundant and repetitive because of push to load without workflow and facilities to enable easy reuse • limited linkage between metadata classes eg classifications, data items	• building links to "rich" metadata stores • less redundancy, including limited automated identification of metadata suitable for reuse • better support for workflow and approval processes • improved means of exchanging metadata with systems beyond the ABSIW, but limited support for direct "real time" access to Corporate Metadata Repository	• aiming to further restrict duplication, and reverse it where possible • share metadata across all phases of collection cycle in accordance with the logical workflow • dissemination of metadata in its own right • expand metadata content eg more quality measures
Data	• a few "large payoff" cubes with relatively few statistical issues are defined eg Trade, Motor Vehicles • in most other cases the content of publication tables or legacy output systems were the main basis for dataset design • limited analysis capability - invoke and store seasonal analysis process • during the transition process there were cases of multiple sources for output products from same collection	• more data cubes defined to information warehouse • SuperTABLE views of datasets • "rich" tables (ie integrated data and metadata) provided to Publication system • availability of better guidelines and consultancy support to help subject matter areas design and build cubes • improved ability to load and view datasets in a flexible manner eg as a cube of as a set of time series	• Review early table designs to restructure into data cubes • XML to be used to replace many output formats • more non-ABS data and metadata incorporated into ABSIW • services provided to load and extract data, support greater flexibility in the platforms, applications and interfaces available to end users
Systems	• basic client-server architecture • functionally rich environment, but useability is poor • tool focussed processing, ie users needed to know how to jump between tools to compete processes rather than having a seamless workflow • in-house development • several cases where prototypes "evolved" into production applications without formal design/development • poor integration between	• client-server architecture, very 'fat' client with a few tools having a 'UNIX' version • workflow managers to chain together a set of operations • enhanced tool design and support • continuation of in-house developed tools • improved integration between platforms, allowing relative strengths of Notes, Oracle and UNIX environments to be harnessed	• move to new IT architecture based on API's, more database layer operations • dynamic links within ABSIW and across systems eg PPW, HSF, IDW • buy not build, with emphasis on ability to integrate and to customise where necessary • web enablement for some services • better support for website • services based with the separation of business logic from interfaces maximising flexibility in the

	platforms (eg UNIX, Oracle, Notes) due to limits in third party products; and in some cases the wrong platform was used eg Oracle for text applications and Notes for statistical processing		development of both elements - as a result end users interact with a workflow layer rather than focusing on individual tools ● better adherence to international standards
Management Regime	● system oriented (ie load and utilise according to system requirements; not user workflow and output driven) ● multiple versions of tools - poor version control ● loading of data was by edict and corporate directive, rather than SMAs seeing business benefit ● many exceptions for non-loading ● changes to system designs and new applications tended to be driven by experiences to date (and "squeaky wheel" effect) rather than overall business needs	● better project management including placement of high level stakeholder representatives within project boards so they become accountable as partners in the development ● introduction of improved change management procedures ● better targeting and understanding of clients needs ● tighter control and management of warehouse processes ● users involved in organising and coordinating aspects such as acceptance testing, release sign-off ● Subject matter representatives responsible for data loading practices ● limited data loading exceptions managed and addressed over time	● more collaboration across systems in terms of shared metadata ● improving project management and deployment management ● user driven processes with output and outcome focus ● separate justification and subsequent evaluation of each new development or phase rather than simply managing and reporting on a single multifaceted project ● increased attention to external standards, best practice and benchmarks for metadata and data management ● loading of data and metadata is accepted as just part of the 'way we do things around here'
Objectives and Outcomes	● provision of targeted "high value" cubes to prove concept and generate initial usage ● move areas away from legacy systems and position them to be able to use other standard corporate facilities ● central store of published data ● increase visibility and accessibility of data and metadata ● focus only on 'output' metadata	● central store of all publishable data and metadata ● increasing relatability, reliability and consistency ● expand metadata content and quality	● system integration within ABSIW and other systems in terms of metadata and data flows ● improve integrity of data and metadata ● improve linkage between all metadata repositories ● "future proofing" the investment in data and metadata repositories against obsolescence by ensuring infrastructure can continue evolving

III. USE OF XML

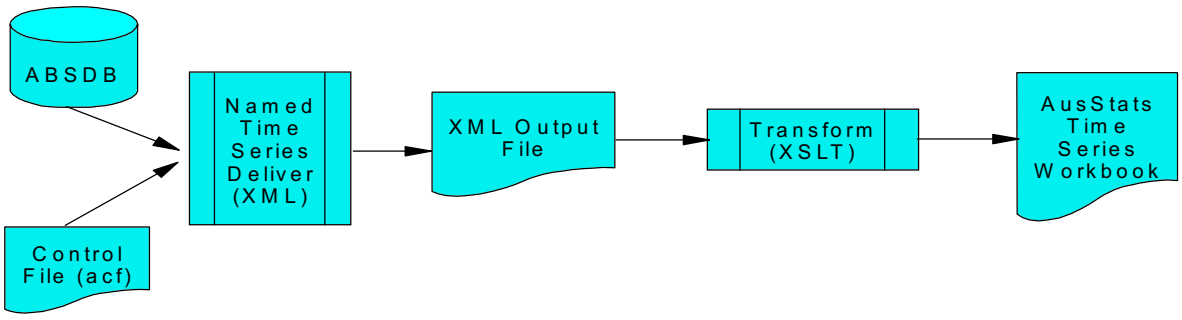
7. For some time, the data management project had been thinking about the use of XML to reduce the large number of internal formats that are used to move data and metadata from the ABSIW to other systems. We jumped into the technology somewhat opportunistically

when a new requirement came along with a very short development lead time.

8. The business problem to be addressed related to the website. The ABS places spreadsheets on its website in the cost recovery part called AusStats. These spreadsheets have been created in .wks format. The source for the data and metadata is the ABS Information Warehouse. Clients have had a number of difficulties with time series included in these spreadsheets, namely limitations on the number of observations (because the time dimension was placed across the columns), difficulty in programming macros to automatically access regularly retrieved spreadsheets (because series could only be identified by absolute position), and annotations metadata on cells are not supported. In addition, most clients had a preference for spreadsheets in EXCEL format.

9. To address the client needs, ABS decided to create the spreadsheets in EXCEL format, move the time dimension to rows, provide additional metadata for each series, including a time series identifier and cell annotations. In addition, we could include a copyright notice and provide links to other metadata on the website. The clients would be in a better position to program macros because of a series name. It was decided to use XML to describe the internal transfer protocol for data being disseminated from the warehouse to applications that create product for clients.

The Process

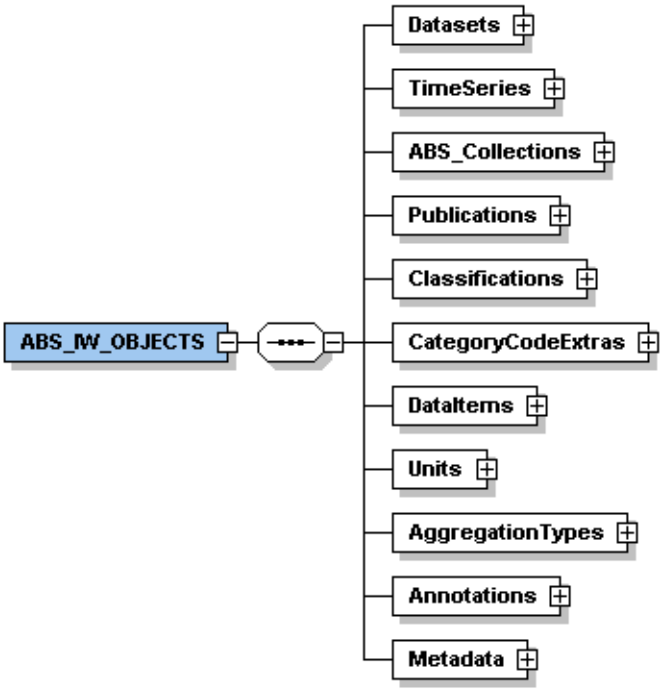


10. The information warehouse exports a number of other formats eg a format that we call a 'Rich Table Object (RTO)' which is used by the ABS tool called PPW (Publication Production Workbench). PPW is used to prepare paper publications and some elements for website publishing. The schema used for the time series spreadsheet exercise has all the elements required for the RTO. In due course I expect that we will remove all other formats and just use XML internally. In addition, some of the current proprietary formats are used to create products that are exported to clients of ABS services and so we have started negotiations with those clients about taking an XML feed.

11. We have concentrated on developing and implementing those bits of the schema required for our first application, namely time series. The other elements are less developed and some entries have been put in the schema only as placeholders. In due course we will develop the schema content to support the multidimensional dataset and metadata that is input into PPW. The internal schema is quite long and complex. I expect that we will need to create specific and simpler sub-schemas for clients eg for time series only, and there could be benefits for all statistical agencies and our clients if a common definition is used.

12. The following diagram contains the first version of the top level diagram of the XML schema that ABS has created for the time series work. [We have used the product XMLSPY

to define, amend, display and control versions of our schema.] A package of material about our work has been placed at: http://www.epros.ed.ac.uk/metanet/resources/resources.htm



13. In structuring the XML, objects that have the potential to be used in multiple places in the definition have been extracted. For example, annotations can be referred to in many places, so these have been extracted into an annotation section. This distillation of objects reduces redundancy of storage, and also provides scope for particular objects that are frequently used to be extracted once, and used for subsequent retrievals from the ABS Information Warehouse. For example, there are a small and fixed number of units of quantity, so it would be possible to build the XML for these and store it in a central place. Then, whenever an object was retrieved from the ABSIW, this units of quantity object could be incorporated.

IV. ORGANISING AND MANAGING METADATA

14. For the best part of the last decade the ABS data and metadata focus has been on the output end of the statistical processing cycle with the major emphasis on developing our internal systems, policies and procedures, and loading and disseminating data. This strong emphasis on 'output' use only for metadata when it comes to 'populating' the repository, and the tight coupling of the metadata repositories to the Oracle rdbms implementation of our warehouse data and metadata model, has created a 'stove-pipe' in thinking about metadata. The notion of a metadata system providing services for the entire statistical processing cycle is not widely held nor understood and to an even lesser extent do people see the existing metadata facilities as being applicable to other parts of the cycle. The data management project started about eighteen months to two years ago to deliberately change that perception.

15. What did we do? The ABS is a Lotus Notes shop and staff are very familiar with that product for email, document creation and collaborative work. The project had already begun to create Lotus Notes applications as 'front-ends' to the metadata stored within the Oracle data and metadata model. In most cases the metadata attributes were extended and some approval

workflow was incorporated. We started referring to this collection of metadata repositories as "The Corporate Metadata Repository" (CMR). The CMR has separate databases containing metadata about collections, classifications, data items, forms and topics. The model specifies linkages between these objects although only a few have been implemented.

16. In addition, we started talking about the ABS Information Warehouse (ABSIW). This concept is broader than just the Oracle based ABSDB because we have thrown the net over other storage mechanisms in order to bring a larger number of ABS data holdings under the data management umbrella. A slogan that we are adopting is to refer to the ABSIW as the 'single version of the truth'. The aim is to have all publishable ABS information in the ABSIW, with that information quality assured, signed-off for release, and all relevant metadata available. Then all output products should be created from the ABSIW. So far we have included Space Time Research SuperMarts in the ABSIW (with the description of each SuperMart fully defined in the CMR). SuperMarts hold much of our household survey unit record data which is often the level required to satisfy client requests because of the large number of potential classificatory variables. In the future, we will consider including FAME databases (they hold time series) as part of the ABSIW.

17. This strategy is beginning to have an impact. The processing system used in the ABS for household surveys, known as the Household Surveys Facilities (HSF), provides the functionality required to handle input and estimation/tabulation processing for many social surveys. HSF acquires and stores data item and classification metadata associated with the input processing cycle by linking to the Corporate Metadata Repository. The Population Statistics Group (PSG) Classifications and Data Standards Section uses the facilities to record standard definitions etc that should be used across all collections that require data related to a particular concept. This approach is intended to assist with statistical integration.

18. From the Data Management project viewpoint, collaborating with the HSF developers is intended to reduce the number of metadata repositories, to capture metadata as close to the point of creation as possible and then utilise it actively in other processes, and to work with the relevant standards area to ensure that an approved definition is available and is used in order to control the proliferation of data element definitions. We still have a long way to go.

19. Also, the ABS is embarking on the development of an Input Data Warehouse (IDW) and the project owners (the ABS Economic Statistics Group) have accepted that the metadata repository for this work will be the CMR. The data management group is part of the advisory group to the project team. The IDW is a managed data store serving analysis, collection activities (including editing), research and management needs between initial data capture until movement of data to the managed output/dissemination store (ABS Information Warehouse). In Phase 1 of the project, the team will begin to address the data management principles that will be needed for an effective longer term IDW. This is not easy. It will cut across the governance, workflow, control and communication processes around the responsibility for content in various data and metadata repositories. Clear and accepted roles for creation, approval, maintenance of all types of metadata through the statistical life-cycle of data and metadata will be needed.

20. The IDW team recognise that outcomes, such as the use of shared metadata sourced from the Corporate Metadata Repository, should be supported. Phase 1 outcomes to be pursued include:

- experience in using common shared metadata facilities (between ABSIW and the

IDW) supported by the Corporate Metadata Repository (CMR);
- identification of IDW metadata requirements from its use of classification metadata sourced from the CMR, to new IDW metadata not in the current CMR;
- understanding the possibilities of using common tools, including a common OLAP tool between ABSIW and IDW;
- understanding how data could be migrated to the ABSIW;
- improved understanding IDW data management roles within a broader data management context.

21. Until two years ago, the work of data management was mainly carried by the Data Management Branch. This organisational unit had, at various times, been located in our IT Division and in our Information Services Division (dissemination of statistical products). For a large part of the time it was an IT systems development project and also had responsibility for deployment of the facilities and assisting subject matter areas to load data and metadata. Crucial elements of the deployment activities include coordination of user acceptance testing, training staff and advice about using the facilities. On the system development front, the project team members also had to work with subject matter areas to identify their requirements, and prepare detailed specifications in priority order.

22. Data Management groups have been set up with the ABS economic statistics area (about 24 months ago) and population/social statistics group (about 6 months ago). These DM groups were given the specific responsibility of assisting their subject matter colleagues achieve the stated corporate data management outcome of loading all publishable data and metadata into the corporate warehouse. In recent times this role has been extended to assist SMAs in adopting the new ABS dissemination directions which are aimed at more dissemination in electronic form and away from paper based publications. In working to achieve the stated outcome, these groups have assisted the DMB by coordinating specification of user requirements and setting priorities, by ensuring a good spread of subject areas are involved in user acceptance testing, by participating in preparation and sign-off of software for release (ie change management processes), by assisting in the delivery of training, and providing advice about the data management facilities.

23. The move to locally based data management groups has certainly increased the accountability in subject matter areas for achieving data management goals and improved communication between the DMB and its clients.

24. The ABS has regional offices in each State and Territory capital city (8 in total) and most of these regional offices contain statistical units that use the corporate data management facilities and have responsibility for publishing information in their field of statistics. The three largest offices have created 'user groups' to share their knowledge and experience in the use of data management facilities. This collaborative knowledge sharing activity is supported by the central office data management groups. One of the main outcomes is the increased capability in the regions in terms of trained people with respect to data management and this is proving useful when the office becomes involved in work about data and metadata issues with State/Territory government agencies.

25. The Methodology Division in ABS has a particular focus on metadata about the quality of a statistical output. Some of the indicators can be acquired from the earliest stages of the statistical cycle eg response rate, through to creation of population estimates and measures of quality such as the relative standard error. Alliance with MD provides a client

and strong champion for the data management project's ambitions about storing metadata about quality in the ABSIW/CMR and making that information available with statistical products. Our current progress is best described as being at a prototyping stage for the systems, although a lot of analysis work has been done on the quality framework and relevant indicators. Some of this has been implemented and populated in the CMR but that information is not yet being released with statistical products.

26. Another example of an alliance to further promote metadata development has been with the standards areas in our economic and social statistical groups. One function already in place is a workflow application to ensure that statistical collection questionnaires are approved as using the appropriate data element definitions and standard classifications. This workflow is embedded as part of the 'forms repository' in the CMR and a full history of the development of a statistical collection form is captured, including the electronic sign-off by the relevant senior executive officer.

V. ABSIW 'REARCHITECTURE' PROJECT

A. The Project

27. About two years ago the ABS employed consultants to review the implementation of our information warehouse - the ABSDB Future Design Review Project. That process concluded that our data model was sound, however the team recommended that the architecture and tools used in the system implementation should be upgraded to ensure that the ABS investment in data management was preserved, to deal with some potential data integrity issues and to provide a solid foundation for improving the useability of the tools. The technical recommendations included the development of a set of Application Program Interfaces (API's) to the Information Warehouse data and metadata services, investigation of IT tools such as OLAP to possibly replace ABS developed software and/or provide options for improving useability, adoption of XML and alignment with industry standards.

28. The 'rearchitecture' project is now underway. After the review team report was considered, we undertook a 'proof of concept' phase to show that API's could be developed with our data model, that an OLAP tool could be interfaced to the data model and that the business benefits were worth the investment. Following detailed senior management consideration, the project was approved and budget earmarked for six phases of six months duration. Phase 1 was completed at the end of December 2001 and Phase 2 has commenced.

29. The main business benefits that we are pursuing with the project are to:

- Improve the useability and usefulness of the ABS Information Warehouse (ABSIW) and Corporate Metadata Repository (CMR) facilities by dealing with client area's concerns about the user interface and operation of the facilities; and by changing the underlying architecture to be based on APIs so that development programmers can more easily integrate the facilities with SMA application systems and to ensure reuse of ABSIW code (and hence improve maintainability).
- Enhance the integrity of the data and metadata and security of the data model. Currently programmers and some SMA staff can directly access the underlying ABSIW database tables through direct programming (ie embedding 'model knowledge' in their application programs) or the use of 'scripts' that usually manipulate specific parts of the data model in isolation from other parts. (The use of 'scripts' arose in the early days of the ABSDB when functionality was limited and the

use of 'scripts' has persisted.)

30. What is our vision for the ABSIW/CMR at the end of the final phase? This project is not primarily intended to add new functionality, although there will be some. For example, the number and range of delivery mechanisms will be augmented by inclusion of XML and this will lead to the planned reduction of some formats, such as X-files which is used for time series delivery. There will be major improvements in the way users and programmers work with the product. Applications will have the opportunity to link to ABSIW processes at the program level and so replace the current need to call a variety of warehouse tools eg data and metadata loading. The API layer will be in place fully protecting the data model. The APIs will be deployed into applications systems that currently use program code to directly interact with the warehouse model and most of the 'scripts' will have been replaced. In addition, ABSIW tools will also use these APIs. At the end of the project, there will be a reduction in programmer support costs by one third because the system will reuse program code and be easier-to-maintain. The APIs will be used by applications to interact with the warehouse and metadata repository. The most notable opportunity on the horizon is the Input Data Warehouse (IDW) development which will use metadata APIs to create and retrieve metadata from the CMR, thereby enhancing reuse and statistical integration.

31. The client interface to the information warehouse will be redeveloped to deal with the identified useability issues. This will occur in a number of possible ways: - vendor tools such as OLAP could replace existing ABSIW tools that are considered to have useability problems; where vendor tools are not available the project team will use human-centred design techniques to redesign user interfaces; and the 'process manager' facility will be modified to use the API layer and adopt ABS work flow approaches. Overall the interface will be more visual and intuitive to use.

B. ABS IT Enterprise Architecture

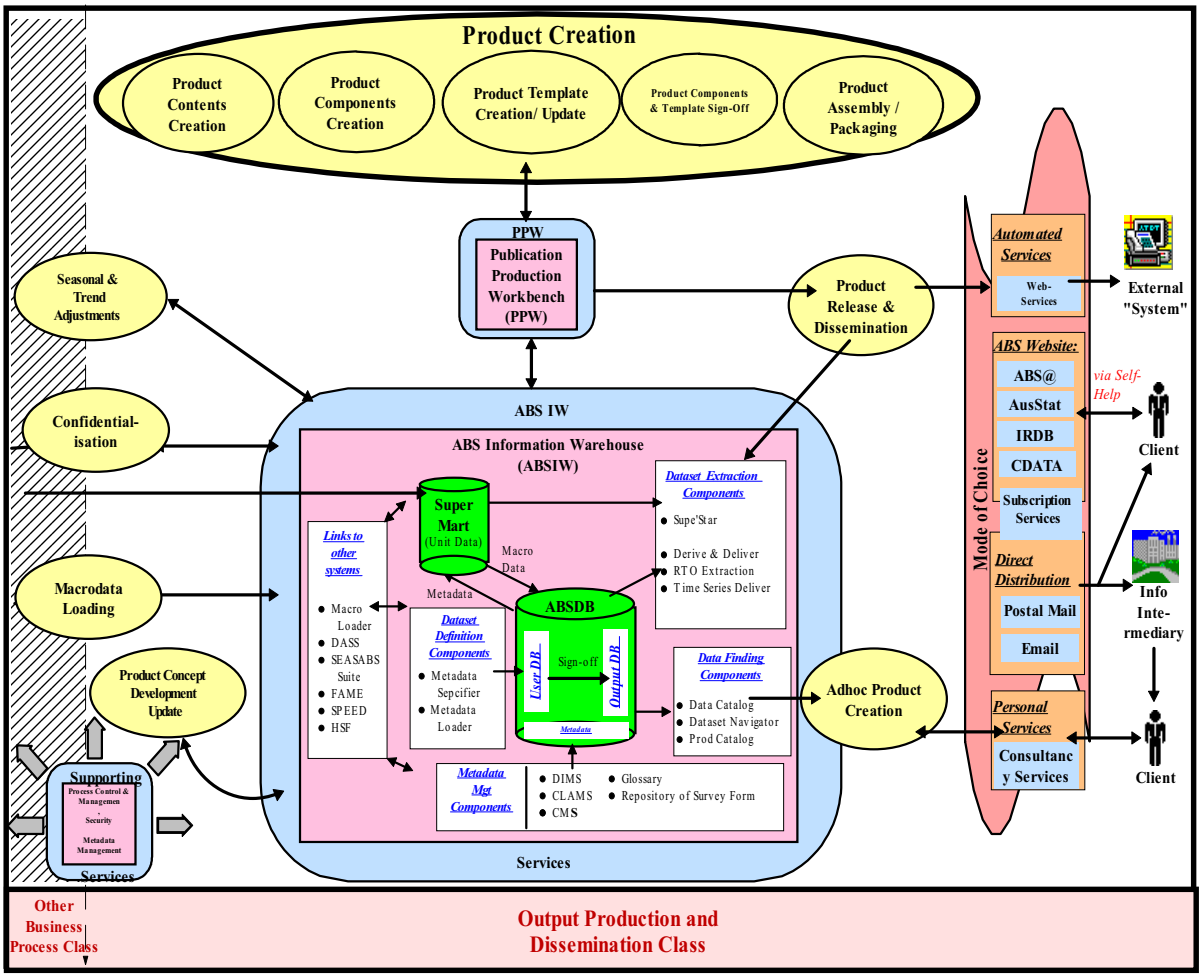
32. During 2001, the ABS Technology Services Division commenced a project to define the ABS IT Enterprise Architecture. This architecture has been used to guide the ABS Information Warehouse Rearchitecture Project mentioned above. The ABS Enterprise Architecture has the following objectives and principles:

- Underpin the ABS Business, Corporate Plan strategies, and ABS efforts to improve structures, processes and productivity, and to position the ABS for eBusiness future and externalising of services.
- Provide a framework for more integrated business systems, emphasising common systems for common processes, keeping unique systems to a minimum, and, ultimately, leading to a significant reduction in the number of systems.
- Promote and support a culture of "assemble and integrate" rather than "build from scratch" for business systems.
- Identify and exploit commercial applications and toolset solutions wherever possible and avoid 'rolling our own' particularly in areas that are maturing and changing and in which home-grown solutions are likely to be obsolete before they can be completed.
- Follow mainstream standards and technologies that fit well into our IT environment.
- Exploit and extract value from business and IT infrastructure, updating key components where necessary to ensure they fit into this framework.
- Provide a toolset that fits well into this environment, is suitable for general use, and encourages the development of low-complexity, maintainable solutions.

- Collaborate and partner with the IT industry where effective to do so.

33. Within the Enterprise Architecture, there has been defined a Business Process Taxonomy and one of those classes is the Output Production & Dissemination Class. Other classes refer to the data collecting and processing activities of different types of surveys or collections. Output Production & Dissemination Class describes the activities involved in the final stage of a statistical survey life cycle to generate ABS products for multi-modal dissemination. This class applies confidentialisation and seasonal and trend adjustments (if required) to the clean aggregated data. The confidentialised and adjusted data are then delivered to the Product Creation process to produce the final products that are ready for release and dissemination via different output media.

34. This diagram is the Output Production and Dissemination class.



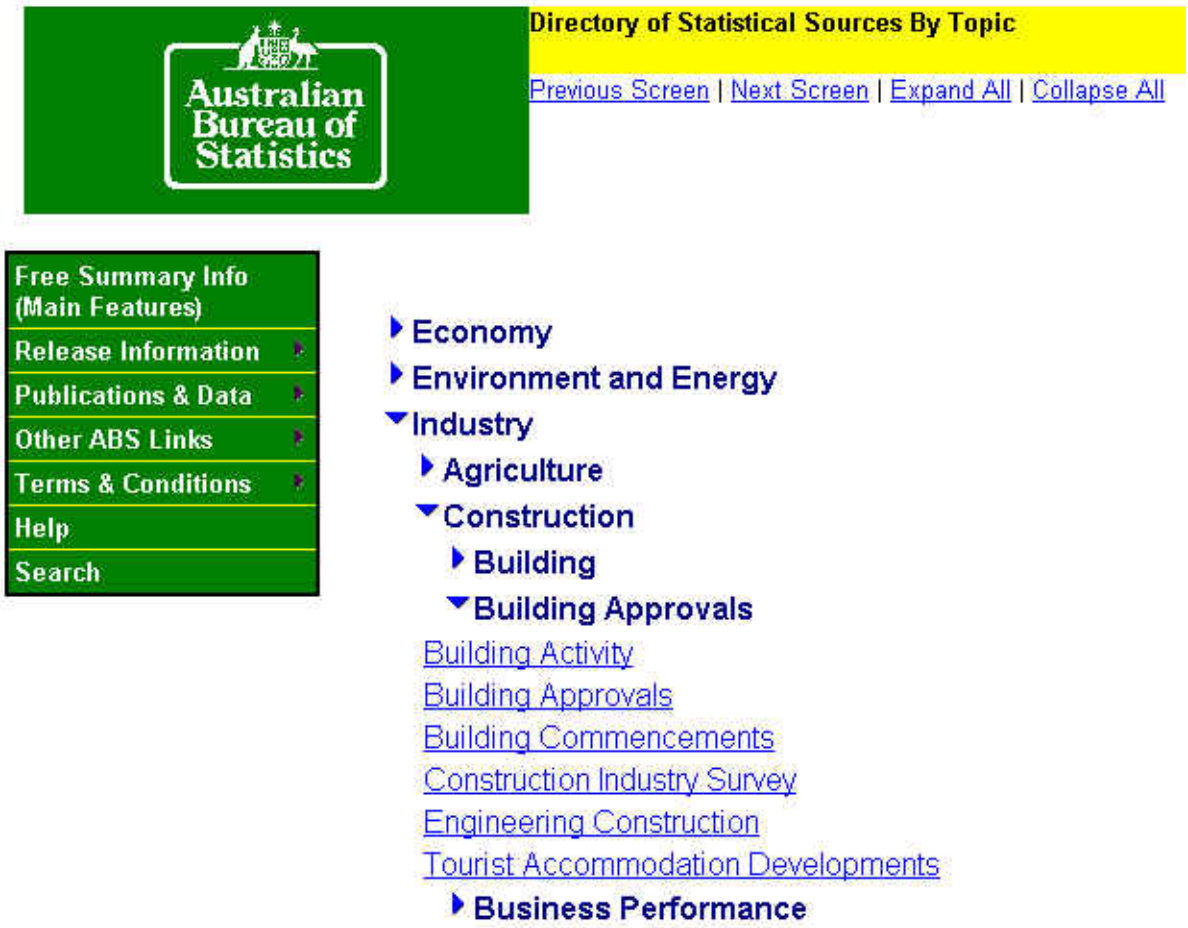
35. The following technologies have been identified as the 'preferred' technologies for ABS development projects:

- We prefer to keep fairly close to the mainstream of products used in environments like ours.
- We will build around well-established powerful technologies such as SQL, PKI, Directories, and Web/Internet and identify emerging well-supported technologies.
- XML and its specific schemas (eg XBRL) are of major interest.
- We will emphasise component architectures and interfaces that fit well into our

applications environment eg SOAP, COM+, Object-Oriented, Unified Modelling Language (UML). There is an interest in exploring the usefulness of the Microsoft .Net architecture.

VI. DEVELOPMENT OF 'TOPICS' LIST

36. Recently, the data management project was involved in an extensive collaborative project to determine a 'topics list' that would be used as part of a search mechanism on the website. This exercise took longer than planned and the resulting list is not much different from similar lists available on the websites of other statistical offices. However, there was value in the exercise because every statistical subject matter area was involved and so it has raised awareness of metadata on the web, has delivered a measure of 'buy-in' for the work and created recognition of the need for quality metadata that may not have been achieved if a list had been imposed. The following image of the top level screen from the prototype facility gives an impression of the hierarchy of the list.



37. As part of the development, we undertook useability testing of the 'topics' list. Although not an extensive test, it still involved 17 external users who had a selection from some 30 tasks to perform. The tasks were related to finding statistics based on searching through the topic list. Each participant performed about 7 to 10 tasks in their test period, so we had 160 attempts to review. As a result of the testing, adjustments were made to the topic hierarchy and some of the entries.

VII. STRATEGIC ISSUES BEING DISCUSSED WITH ABS SENIOR MANAGEMENT

38. In the ABS context, the senior executive management team comprises the Australian Statistician, Deputy Australian Statisticians and Division Heads - some 9 people in total. This group provides leadership, and approves directions, policy and budget for ABS projects. A crucial success factor for all successful data warehouse and metadata projects (in fact any major project) is that there is senior management 'buy-in' and support. This has been the case in the ABS over the years with respect to data management and means that at least three Australian Statisticians and many other senior executives have seen the business value of data and metadata management activities. It has not been smooth sailing and achievement of our objectives has taken longer than expected. Along the way there has been considerable learning about project management, change management, transition and deployment planning, impact analysis and integration of corporate infrastructure.

39. In order to have divisional management continue to support corporate metadata policies in a context of other competing demands, requires that these people be persuaded that corporate benefits are being achieved and that the investment is worthwhile. One main avenue followed by the ABS data management project is to involve the senior management group by having them consider data and metadata issues. This has been done through the annual Branch report to senior management. During this meeting, the senior management team is engaged in a dialogue about the objectives for the program and also by considering a range of issues relevant to progressing specific aspects of data management. This latter discussion often leads to the senior management team commissioning further papers that they discuss during the year.

40. At the most recent meeting, a number of papers were commissioned which will see the data management topic before the senior executive group possibly 3 or 4 times in the coming twelve months. The topics to be covered include:

- Linking Data Management and Dissemination Strategies. This discussion will explore the concept of database publishing and focus on the metadata packages that should be provided with particular data releases; explore the quality assurance, approval and sign-off processes relevant to an electronic, web based environment - ABS history is a paper paradigm and we need to move our processes and culture to electronic; consider a framework for subject area dissemination strategies that considers products such as data cubes; metadata to describe context and quality; and search metadata.

- Strategy for Metadata on the Web. The ABS has accumulated a considerable amount of metadata in the Corporate Metadata Repository (CMR), however very little of this material is available for external consumption. For example, the 'Forms' repository has a complete set of ABS questionnaires with relevant information about each, and the 'Collection Management System' repository contains information about all ABS collections, such as methods used, outputs, quality measures, contact information, data elements and classifications used. In fact, with respect to metadata in the CMS we have identified with the SMAs those items that would be useful to clients and that should be published - we flag them with a 'book' symbol.

- Policy for non-ABS data and metadata loading. The discussion of this policy will bring senior management attention to metadata issues related to non-ABS data sources. The National Statistical Service objective in the ABS corporate plan seeks a shared

commitment at the Commonwealth, State and Local Government levels to the effective delivery of the statistics required by key users, no matter what their source. With respect to the metadata aspects of the draft policy on non-ABS data, the proposals cover the data custodian role to ensure that concepts, sources, methods and quality information are document in the ABSIW; responsibilities for loading and maintenance of data and metadata; and sources of non-ABS data that are included in ABS dissemination products.

41. A further aspect of our new strategy for metadata is to ensure that data and metadata management aspects of new developments are understood. I have mentioned earlier the work on Data Base publishing - a concept linking the back-end Information Warehouse to our dissemination channels. In pursuing this activity, we have engaged with seven subject matter areas in case studies to explore the issues that SMAs need to consider in their dissemination strategy and how the DB publishing concept might work with their particular outputs. In particular, we are exploring the quality assurance and approval workflow. The ultimate objective is to identify processes and infrastructure needed to break the paper paradigm, ie to have those responsible for approving product releases not necessarily continuing to rely on the hardcopy publication.

42. Where is this taking the ABS and what is the expected impact on subject matter areas? There are many changes already underway or being anticipated that are going to influence the ABS dissemination business. Some of these include:

- more electronic products such as spreadsheets, data cubes;
- provision of contextual and quality metadata with this range of new products;
- decline in the use of printed publications for dissemination;
- new services such as self-help, e-commerce, print on demand;
- policy framework for Community Service Obligation, non-ABS data, pricing.

43. We are working on a dissemination framework to guide SMAs through consideration of the dissemination strategy for their area of statistics and take into account the changes mentioned above. The following table is an extract from the framework with the particular focus on issues related to metadata.

Issue	**Matters to Consider**
Non-ABS data [include link to ABS policy about non-ABS data]	• Determine if any non-ABS data is relevant to the dissemination strategy. If so you will need to acquire relevant metadata to assess the quality of the information, as well as obtaining the data. • If non-ABS data is used, you will need to allocate the role of data custodian. This might be a new role to you. If another area is undertaking the custodian role, you will need to make them aware of your use of the non-ABS data. • Are the metadata, quality indicators etc complete for the non-ABS collections? • Ensure that the non-ABS data sources appear in the Directory of Statistical Sources. • Decide if the non-ABS datasets should be held in the ABS information warehouse.
Metadata	• Is relevant 'search' metadata defined and stored eg topics list? • Is a thematic directory relevant to this field of statistics? • Are all relevant metadata used by the underpinning collections for this statistical field 'signed-off' as approved and complete - covers CMS entries, approved classifications, data items? • Are standard classifications and data items being used where

	appropriate ie datasets are linked to classifications and data items marked as 'standard' on the metadata repositories? ● Are data element metadata being created and used in all phases of the statistical processing cycle?
Quality [reference to set of subject matter collection quality measures defined by Methodology Division]	● Have the appropriate quality measures been determined? ● Have quality statements been prepared and circulated with all output products? ● Is there a process to ensure that all the data and metadata content components are validated, signed-off and locked
Data retention [refer to ABS policy about data retention]	● Are the data retention statements for all relevant datasets complete and stored in the CMS? [This covers retention requirements for collection instruments, final unit record files, aggregate data in the ABS Information Warehouse.] ● Is there a data custodian (ie an officer with the responsibility for managing the data holding) for this collection or set of statistics?

VIII. DIRECTIONS CONCERNING METADATA AND ABS WEBSITE

44. After a good start with our data management initiative and with a well populated ABS website, we find that we are in 'catch up mode' with respect to publishing metadata. The website contains all publications since 1998 in .pdf format together with some additional data in the form of spreadsheets and data cubes. In addition, there is a lot of information in the form of main features; media releases; subject matter theme pages; the annual Australian Yearbook completely on-line; concepts, sources and methods in the Statistical Concept Library; and information papers.

45. However, we are lacking in a few areas. They include: metadata in the form of directories to enhance information location; provision of keywords linked to specific datasets so that the powerful website search facility can help clients locate specific data; a time series service that enables clients to find and extract specific time series; metadata about collection methods, collection quality parameters etc; and metadata in machine readable form so that metadata could be incorporated in the processing systems of other agencies that might contribute to the national statistical service.

46. We note that our peer agencies have tackled a large number of these matters and provided facilities to their clients. So ABS is looking to strategies to 'catch up' in the provision of these services. The Directory of Statistical Sources will be deployed soon - the software infrastructure is in place, and most of the metadata is available in the Collection Management System. Currently SMA staff are adding 'topic' metadata to each collection and reviewing the fields to be published to ensure they are good quality. Part of our work with XML is the development of that part of the schema that defines classifications. The intention is to extract standard classifications in XML form and deploy both the XML form and probably another well-known form eg .txt, on the web so that clients could download the classification for use in their systems. A next step is for the ABS to provide a 'coder service' by placing the coding application on the website (possibly using the SOAP technology) linked to standard classifications and enable users to invoke the service remotely.

47. Again using XML technology, a further step with time series after the EXCEL spreadsheets are deployed, is to replace the special data services that we provide in

proprietary formats with a time series feed to those clients based on XML. The follow on is then to provide a full time series service on the website to make all our time series accessible to clients for searching, retrieval, and downloading, with payment via e-commerce. [Note the ABS operates a cost recovery regime for certain data as required by our Federal Government policy since 1988.] A significant issue being addressed now in preparation for such a service is the structure and content of a corporate time series name that can be automatically generated from existing stored metadata and that will become meaningful to regular time series users. At present we have many local time series naming conventions.

48. Another area for catch up is to increase the number of data cubes and the breadth of their subject matter coverage, and to provide web based services that enable clients to 'slice and dice' those cubes before downloading. Currently, the client only has the option of downloading an entire cube (paying for the entire cube) and then performing any extraction on their machine. In conjunction with this service, we need to ensure that all relevant metadata is linked to the cube and only one click away.

49. The final area for catch up by the ABS is in our knowledge and understanding of developments that have been undertaken by groups such as MetaNet, METIS and the developments under various Eurostat programs. It seems to me that there are a number of potentially useful initiatives with metadata models, use of UML and XML and some software modules. ABS contacts with this work and hence our ability to know what is going on and potentially to contribute ideas and comments has been limited. I feel that we are missing out on some interesting exchanges.

IX. CONCLUSION

50. This paper has covered a lot of territory with only sufficient detail to enable the reader to assess if they have an interest in pursuing a one-to-one dialogue with the ABS about any of these matters. My staff and I would be interested in any comments on our work or information about similar work that you are undertaking.

51. For the work session, I think that we could have a profitable discussion about the process we might follow to achieve a level of standardisation in the XML descriptions of common statistical objects. It seems that enough agencies are pursuing solutions with XML for there to be a critical mass to discuss this issue.