WP 9
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (iii): Metadata Models and Terminology

USING ISO/IEC 11179 TO HELP WITH METADATA MANAGEMENT PROBLEMS

Invited paper

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

SUMMARY

The Australian Bureau of Statistics has embarked on a major organisational and methodological change program with respect to economic statistics. For a decade ABS has had an output data warehouse, with associated metadata, from which all dissemination products are produced. In economic statistics, the processing environment has been a series of 'stove-pipes'. The new business statistics environment has a functionally specialised organisation with an input data warehouse as a central feature. Considering metadata management aspects of this new environment has raised many issues for resolution.

This paper firstly provides background to ABS metadata management development. It then describes some of the problems that need to be addressed. The international standard "ISO/IEC 11179: Information technology - Metadata Registries (MDR)", has been used to assist in the resolution of those problems. The paper discusses how we view the 11179 model in the context of existing international statistical frameworks eg SNA, and how it can link to aspects of statistical processing. A number of areas for further work and/or extension are raised, particularly related to definition of data cubes and presentation tables, and handling derived data elements, for example resulting from time series analysis.

[1] Prepared by Graeme Oakley and Alistair Hamilton from Data Management Branch, and Mike McGrath from Economic Standards and Classifications. The views expressed in this paper are those of the authors.

I. INTRODUCTION

1. Planning and implementing the Input Data Warehouse (IDW) for economic statistics has had many implications for metadata management at the conceptual, business and technical levels within the ABS. This paper does not aim to present a comprehensive review of these implications, many of which are still being worked through, but rather provides some illustrative examples. One of these is "strategic", and has resulted in a fundamental rethinking of the way metadata is conceptualised. It includes proposals to bring the ABS into closer alignment with ISO/IEC 11179 (referred to subsequently as 11179). The second is more "tactical"l, and involves interactions between the IDW and the existing Corporate Metadata Repository within the ABS. The third example lies somewhere between the other two, and relates to "practical" metadata required by the IDW which is not yet supported by the CMR.

2. As context, however, the paper first reviews the recent evolution of metadata management within the ABS. This includes summarising the development of the current Corporate Metadata Repository from the outcomes of a consultancy undertaken by Professor Bo Sundgren in 1991. The definition of principles and future directions for metadata management as a result of the development of an ABS End to End Metadata Strategy over the past twelve months is also described. The IDW's approach to metadata management is identified as one of the first embodiments of this strategy.

A. Development of the Corporate Metadata Repository (1991 onwards)

3. The ABS had very early experience with metadata management to support data warehousing. In 1991 the ABS was fortunate enough to have Professor Bo Sundgren undertake a five month review which resulted in the report - "Towards a Unified Data and Metadata System at The Australian Bureau of Statistics". This paper envisaged three components for an "ideal" output database (ABSDB), namely "macrodata", "microdata" and "metadata". Note that even the microdata component in this case was output oriented - being clean and final unit record data.

4. Work commenced first on the "macrodata" and "metadata" components and by 1994 the ABS had established a viable output database. The metadata component included - using the pre 11179 terminology of the time - information about: surveys; value sets; data items; populations; and dataset descriptions ("Elementary Abstract Tables").

5. Professor Sundgren had contrasted metadata, and metadata systems, for "end-user oriented purposes" with those for "production oriented purposes". It quickly became apparent, however, that the quality, coverage and currency of information for end-user oriented purposes could become very inconsistent when done in isolation (and addition) to managing metadata for production oriented purposes.

6. As early as 1994, papers envisaged how various aspects of the "metadata component" could be used to support all aspects of the statistical cycle. In a few cases, in a few areas, this did occur (eg in documenting and approving "data items" to be collected by some surveys).

7. By the late 1990's, the "metadata component" was recognised as being the hub of a Corporate Metadata Repository (CMR) to support "end user" and "production" purposes. This was contrasted with the "macrodata component" and a fledgling clean "microdata component" (eg SuperMARTs described with corporate metadata and managed in accordance with ABS data management principles), which were labelled the ABS

Information Warehouse.

8. It is interesting to note that a number of other statistical agencies have questioned the practical viability of a CMR that meets both "end user" and "production" purposes throughout the statistical cycle. As evidenced by the ABS End to End Metadata Strategy outlined below, we believe it is a viable direction (on an incremental basis) - and the use of the CMR by the IDW will be an early test of that.

B. Strategy for End to End Management of ABS Metadata (2002 onwards)

9. In late 2002, ABS senior management commissioned development of a "strategy for end-to-end management of ABS metadata". A group of senior stakeholders from across the organisation participated in the development of the strategy, including experts in: standards and classifications; subject matter; methodology; IT and systems design; and data management. This strategy was developed in four stages:

i. Principles, Definitions and Stakeholders
ii. Where are we now?
iii. Where do we want to go?
iv. How might we get there?

10. Stage 1 of the strategy involved a fundamental reconsideration of concepts and definitions to an extent which had not been experienced within the ABS since Professor Sundgren's work in 1991. It was recognised that external opportunities and drivers had advanced greatly since that time in regard to: standards (eg 11179); implementation of metadata registries by other agencies; and technical enablers (eg XML, Web Services).

11. A key outcome in Stage 1 was the definition, and agreement on, the following key principles for metadata management within the ABS:

- Manage metadata with a life-cycle focus
- All data is well supported by accessible metadata that is of appropriate quality
- Ensure that metadata is readily available and useable in the context of client's information need (whether client is internal or external)
- Single, authoritative source for each metadata element
- Registration process (workflow) associated with each metadata element, so that there is a clear identification of ownership, approval status, date of operation etc.
- Describe metadata flow with the statistical and business processes (alongside the data flow and business logic).
- Reuse metadata where possible for statistical integration as well as efficiency reasons (no new metadata elements are created until the proposer has determined that no appropriate element exists and this fact has been agreed by the 'standards area')
- Capture at source and enter only once, where possible
- Capture derivable metadata automatically, where possible
- Cost/benefit mechanism to ensure that the cost to producers of metadata is justified by the benefit to users of metadata
- Variations from standards are tightly managed/approved, documented and visible.

12. These principles, although not being very new or startling, when drawn together and articulated as a set and stated and supported by senior management, are very powerful. They have been circulated widely within the ABS and have already been used to assess several

current projects, including development of the IDW, and some new project proposals. Within the limits of existing business processes and CMR infrastructure, the design and operation of the IDW was recognised as performing very well in regard to these principles.

13. Stage 4 ("How might we get there?") makes it clear that it is neither practicable nor desirable to initiate one "cathedral project" to completely transform the way the ABS "does metadata". Instead, it is expected that all proposals for new and changed business processes and systems will be reviewed in the light of the ABS Metadata Strategy to determine how best the project can support and advance the strategy. In some cases this may result in the need to incorporate additional information and services within the CMR.

14. The strategy envisages the definition of the CMR being extended greatly beyond its current scope to cover many new elements of metadata, such as technical information required to run business processes, information about outputs produced by the ABS and a greatly extend range of quality measures. Just as is currently the case, the CMR is not envisaged as a single physical information store but as a series of stores that fit together in a coherent and integrated manner and which are subject to a common set of basic business rules. Registration Authorities and responsibility for development and maintenance of content and infrastructure may vary in some cases, but the ABS Data Management Branch will have overall responsibility for ensuring each element of the CMR adheres to ABS metadata management principles and fits appropriate with other elements.

15. The key point is that processes and applications work with the relevant shared metadata in the CMR rather than:

- with purpose specific metadata stores that operate in complete isolation, or
- hard coded information within applications that should have been stored as code independent metadata.

16. A concurrent thrust, based on a design review in 2000 of the ABS Information Warehouse (ABSIW) and CMR, is a move away from applications interacting directly with CMR stores. Instead, a set of "services" are provided that applications can "call" when they need to interact with that store. In this way applications don't need to "know" the physical data model underlying the store, or the low level business rules and constraints which should be applied, but only define what they are seeking to do at a higher business level. The approach has many benefits, including the following:

- improved security and consistency when interacting with CMR
- improved ease of use for developers encourages them to harness CMR appropriately
- underlying structure of CMR can be changed/extended without impacting on applications as long as the services layer is updated appropriately.

II. EXAMPLE OF PROBLEMS TO ADDRESS USING A CASE STUDY OF IDW

A. Input Data Warehouse (IDW) (2000 onwards)

17. The importance of sound metadata management was recognised as fundamental to the success of the IDW. The proposal to move the IDW from "Proof of Concept" to "Production Pilot" in 2001 noted that "*Consideration of **the data management and methodology issues will be** as **critical to the success of the IDW** as the technical and subject matter issues. Without them our ability to compare across collections and make statistically valid use of IDW data will be reduced*" and also "*Desirable outcomes, such as the use of shared metadata sourced from the Corporate Metadata Repository should be supported by the IDW.*

Outcomes in Phase 1, for administrative data within the IDW, 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; ...*
- *improved understanding IDW data management roles within a broader data management context."*

18. One crucial aspect was the "gatekeeper" role accepted by the Economic Standards and Classifications (ESC) section in regard to approving the data element concepts and classification schemes that could be associated with data held on the IDW. In the case of the ABSIW, while the relevant standards area had defined a number of data element concepts and classification schemes to the CMR, users still had the option (and often found it easier for various reasons) to define their own metadata and associate it with their data. This led to a proliferation of non standard metadata and hindered attempts at data confrontation and statistical integration in regard to the content of the ABSIW.

B. Strategic Example: Framework of Concepts, Properties, Classifications, Facts

19. This represents a fundamental body of work in regard to high level metadata modelling which has been "triggered" by the IDW development. One driver was an early expectation that ESC may need to separately define and approve each individual Data Element, or even Data Item, to be recognised within the IDW. For example, separately approving each of the following (and more) for "Bananas", then for "Pawpaws" and a similar list for every other agricultural commodity.

- Production Bananas (t)
- Quantity purchased Bananas (t)
- Area (not yet bearing) Bananas (ha)
- Area of bearing age Bananas (ha)
- Yield Bananas (t/ha)
- Purchase cost Bananas ($)
- Gross value Bananas ($m)
- Gross unit value Bananas ($/kg)
- Gross unit value Bananas ($'000/t)
- Production Pawpaws (kg)
- Quantity purchased Pawpaws (t)
- etc

20. This quickly led to the conclusion that ESC needed to manage and approve metadata related to classification schemes, object classes, properties and representations to be used in the IDW rather than managing every individual data element which might be formed.

21. It was recognised that the concepts and terminology associated with 11179 could be of great value in managing this situation but these were not widely recognised and understood within the ABS. It took several iterations, worked examples and consultations with 11179 experts within, and outside, the ABS before the high level framework - and its application to statistical activities within the ABS - could be understood and agreed.

22. The development of this framework will be a big step forward. While some details of workflows and systems supporting the framework may vary to some extent, at the highest level the framework should prove applicable to the statistical cycle as a whole, across

economic and social subject matter areas and across both "input" data and "output" data.

23. While it will require a significant cultural change process within the ABS (with updates to many documents and interfaces) even the adoption of terminology from 11179 seems set to bring long term benefits. For example, the term "data item" is in current use by different areas within the ABS to refer to "Data Items", "Data Elements" or "Data Element Concepts" as defined in 11179. Such differences in meanings and perspectives within the ABS have been a significant barrier to achieving common approaches to metadata management in the past.

24. Activities currently underway in regard to the framework include:

- confirming the high level applicability of the framework in an "end to end" context
- confirming the definition, and metadata requirements, of statistical processes associated with the framework
- particular focus on processes which it is expected will be interacting with the IDW
- identifying which aspects of the requirements can be met by existing CMR information stores and services and what extra structures, databases and services may be required.

25. In relation to the last point, the absence of a suitable CMR facility for managing question wording and question modules is, for example, recognised as a significant deficiency. This then means that it is not possible to trace at a detailed level from a question to a collection instrument and back to the underlying data element concept (or vice versa). Some infrastructure being used by the Population Statistics Group within the ABS is likely to be considered in this regard, along with the approach taken by the US Bureau of the Census in linking their Generalized Instrument Design System with their Data Element Register.

C. Tactical Example: Existing linkages between IDW and CMR

26. Several elements of the "star schema" associated with the IDW are currently being populated by various means with metadata from the CMR. This includes:

- basic metadata about data sources (eg collections/surveys and their cycles)
- definitions of key classifications
- definitions of "data items" (ie content of the data item dimension in the schema).

27. For performance reasons, in most cases basic information is read into the star schema in a "snapshot" manner using services provided by the CMR rather than being looked up dynamically. Flags indicate which metadata has been approved by ESC for use on the IDW. Unapproved metadata cannot be read into the star schema. Links to more detailed definitional information in the CMR are retained. The CMR also remains as the authoritative source of this metadata as no independent updates can be performed to the information which has been read into the star schema. If updates are required they must be authorised and then applied to the CMR. Processes check for changes to relevant metadata in the CMR and notify of updates since it was last "snapshotted" by the IDW.

28. Examples of classifications and value sets obtained from the CMR include standard time codes, Australian and New Zealand Industry Classification (ANZSIC), Standard Institutional Sector Classification of Australia (SISCA), Type of Legal Organisation (TOLO), Processing Stage Indicator (PSI) and Unit Status Indicator (USI).

29. Additional services are to be provided to the IDW in 2004 related to the:

- inclusion in the XML returned by CMR classification services of information about how classification scheme items relate to each other in a hierarchical manner.
- provision of CMR services about Collection information sourced from the Collection Management System.
- provision of CMR services about Form information sourced from the Repository of ABS Survey Forms.
- provision of more sophisticated searching mechanisms.
- development of a mechanism to allow data and associated metadata to be transferred from IDW to the CMR / ABS Information Warehouse.

D. "Useful Metadata Structures" not yet supported

30. In loading and managing data on the IDW to date, and envisaging the operationalisation of the IDW for survey data, the design team have identified the need for several metadata structures which are likely to prove useful with, and beyond, the IDW. The following are examples:

i. *Concordances*. The IDW project and ESC have identified a high priority need around improved support for concordance metadata and associated services by the CMR. They have identified that similar structures may be appropriate for describing relationships between a value domain used on the IDW and an underlying "standard" classification scheme. (Several items from the underlying classification scheme may, for example, have been collapsed together in a describable manner to form a subtotal while others are represented independently in the value domain.) There will be particularly heavy demand for concordance metadata when the 2006 version of the Australian and New Zealand Industry Classification replaces the 1993 version, but there will also be a need for updated concordances to other industry classifications (eg NAICS) used throughout the world.

ii. *Definition of Legal, Illegal and Questionable Combinations.* It has been recognised in some cases it is either conceptually impossible or very unlikely in practice to have certain combinations reported. For example, in terms of agricultural production, it is unlikely to have commercial production of tropical fruit in a very cold climate. (Note that the combination of tropical fruit and a cold climate may be quite sensible when recording consumption, as opposed to production.) For data validation and editing purposes, the IDW team are keen to be able to model these "combinations". It is recognised that a similar concept of combinations can be very useful metadata for output and dissemination purposes. For example, in a large cube of data there may be "holes" (absences of data) into which users who do not have a good knowledge of the subject matter can fall. Such "holes" can exist because the combination requested from the cube is conceptually impossible or because the data was simply not collected or not releasable.

iii. *Specialised transformations*. IDW has a requirement to be able to describe various types of transforms, such as those associated with ETL (Extract - Transform - Load) processes used to populate data within the IDW in accordance with controlled value domains that have been agreed for use on the IDW, and those associated with deriving new data elements conditional on the value of other data elements. Being able to describe such transforms through standard metadata, making them fully repeatable and reusable by multiple systems, potentially offers many benefits, such as guaranteed consistency, easier to change software choices as transforms don't need to be rewritten, easier to maintain transform as logical metadata rather than as software specific code.

III. HOW ARE WE USING ISO/IEC 11179 (PLUS EXTENSIONS)

31. The previous section described a particular problem with the ABS IDW project concerning the definition of data elements. Having agreed that data element definitions would be registered before any data can be loaded to the IDW, the question arises as to what exactly needs to be registered. We have determined that to manage data elements at the level resulting from combining the concepts of properties, object classes, classifications and qualifiers would be unworkable because there are too many potential data elements. For some areas (eg International Investment, Population Census, and Manufacturing) this would involve inordinate effort specifying, documenting and loading many thousands of separate entries onto the data items database. Secondly, once loaded, any change to any one of the classifications used would need to be reflected in any and all affected entries.

32. It also makes sense conceptually to define at a higher level because the inclusion of either the object class or the qualifier generally does not add to the property, i.e. the core definition of what is being measured remains the same regardless of the level at which it is being disaggregated, with only the name of the classification level being replaced each time. The concept remains the same. How well this principle can be applied in practice to all cases, as opposed to most cases, is currently being tested by applying the framework to a variety of standard and challenging (but realistic) examples.

33. The framework presented in this paper now fully incorporates 11179 but extends it firstly to take in the overarching framework at the beginning, and secondly, to incorporate consideration of collection instruments and instrumentation arrangements at the end of the process, so that it ends with the question wording used to actually collect data. There are a number of issues that still need to be worked through to ensure the framework performs the role required of it, and to review the infrastructure and functionality needed to support it.

A. Describing the framework components

34. From a practical and theoretical point of view, it makes sense to describe a collection (or the data to be gathered in a collection) from a conceptual framework perspective and to establish and document links to other frameworks (e.g collecting R&D information under the Frascatti Manual, elements consistent with - but perhaps extensions to - SNA93/BPM5 are ..., elements inconsistent with SNA93/BPM5 are ..., etc}. This not only ensures that good integration principles are being followed (avoiding unnecessary duplication), but also ensures that the platform is laid for the resolution of problems that arise during the survey cycle.

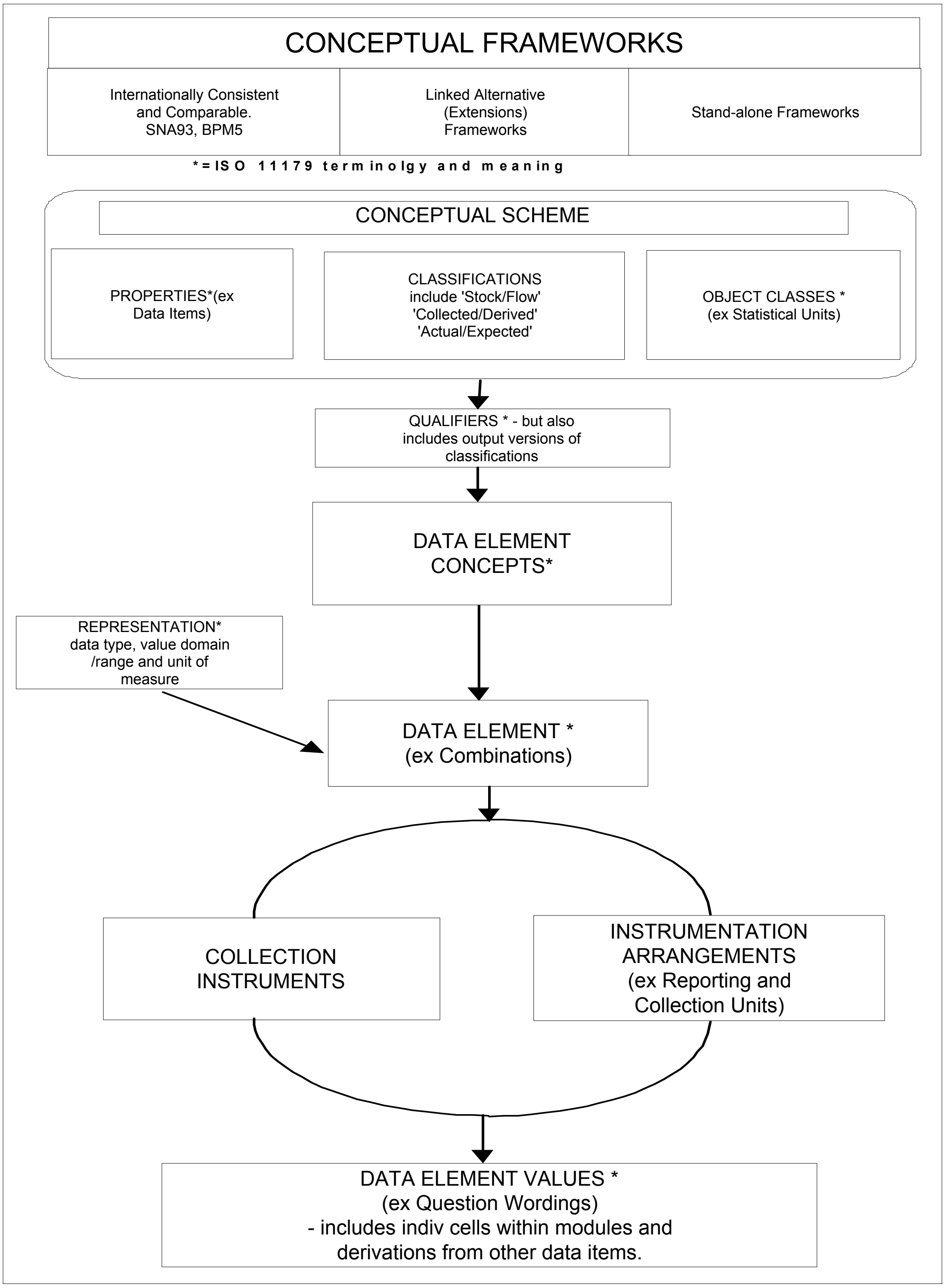
35. The conceptual framework step lays the foundation for the whole collection and starts the process of establishing:

- what questions are we trying to answer; and
- what outputs should be produced to help answer those questions.

36. Having established these answers, the next steps involve determining:

- what information is needed (11179 'properties', ABS 'data items') to create the outputs;
- about what (11179 'object classes', ABS 'statistical units') is the information required;
- what level of detail is required to help understand the aggregate level results -i.e disaggregations of properties and object classes (11179 'classifications' and 'qualifiers', but also ABS 'output-versions-of-classifications'); and
- the nature of the data itself (11179 'representation', ABS 'instance values').

The Framework



37. The result of combining the levels of detail required in the first three areas in the list above generates what are referred to in 11179 as data element concepts. The result of combining data element concepts with representations generates what are referred to in 11179 as data elements (ex 'combinations'). Having now established the data elements required, the final steps of the framework involve considerations of both:

- the nature of the collection's instruments (e.g form, CATI, internet); and
- the implications of any instrumentation arrangements (i.e the 'from' units rather than

the 'about' units).

38. We then arrive at the data element values. At its simplest this means we arrive at the question wording used to collect a particular data element, from a particular set of entities using a particular type of collection instrument. In the case of derived data elements, however, we arrive at the question wordings used to collect the base data, coupled with the rules for derivation (which were created at the time the derived property was defined).

39. The aim in all of this is that the metadata stored against each question wording will be the sum of that created at each stage above.

- Having created and stored the necessary descriptive and definitional metadata about each concept being used, we reuse this to create the metadata for the data element concepts.
- This is then combined with metadata created and stored for 'value meanings' for the codes or values of allowable value domains, to create the metadata for data elements.
- Combining this with metadata about the collection instrument(s) and instrumentation arrangements creates the metadata for data element values.

40. Representation: In 11179 (and therefore in the ABS framework) the 'Representation' component of the model is limited in both breadth and depth to embody three parts:

- A *data type* - a format used for the collection of letters, digits or symbols to depict the values of a data element, determined by the operation that may be performed on the data element. Examples are integer, alphabetic, and alphanumeric.
- A *value domain* - a set of permissible values for a data element and can be fully enumerated or expressed as a set of rules. Examples are M,F; 0,1; and >=0.
- A *unit of measure* - accompanies numeric value domains. Examples are units, '000s, dollars, and tonnes.

41. Qualifiers: While the 11179 definition of 'qualifier' is rather limited (in that it is unstructured in the standard and doesn't have much status) the model proposed above advocates a stronger role and is really more 'classification' than 'qualifier'. In the model being proposed, in cases where the registered classifications aren't able to be employed to provide the sub-setting necessary, there is a separate qualifier created and registered for each different concept embodied in a fact. As well as providing a way to disaggregate or subset broad measures, e.g breaking 'sales' into 'sales-by-IOCCproduct' - an output-version of ANZSPC, registering qualifiers also provides a way to simplify the task of managing the metadata embodied in a single fact or data item (instead of operating at the level of permutations and combinations, which is too onerous from both creation and maintenance perspectives).

42. The critical point is that each and every data element value will have more than one classification and qualifier embodied in it. The classifications and qualifiers will apply to the object class (e.g classifying them perhaps by ANZSIC, geographic region, size groupings, marital status etc) and to the property being measured. Note that while many classifications and qualifiers only apply to properties, some (e.g geographic region) can apply to both properties and to object classes. There is a worked example in the appendix.

IV. FURTHER WORK

43. ABS senior management has broadly endorsed the proposals relating to the ABS metadata management strategy. That work will now move forward into an implementation phase involving a series of parallel projects. The framework described above has also been

endorsed by the relevant committee to move into a more detailed design and implementation phase. Effort will be required to ensure that the framework can be used by all business and household based collections. Effort will also be required to analyse the separate concepts embodied in the explanatory information carried in processing systems (e.g. conceptual qualifiers, scope qualifiers and subsetting qualifiers).

44. Another early task is the establishment of exactly what metadata is required within each of the framework components. To a degree this depends on decisions regarding how the different metadata stores need to interact, both within the framework but also with other stores within the broader business process (e.g in estimation, imputation and analysis). Having decided on the desirable content and functionality of the component metadata stores, the next step will be to establish the compatibility of existing infrastructure (such as the Data Items Database, Classifications Databases, and the Forms Repository Database). From this work there will be a series of recommended actions in regard to component metadata stores and associated services. These are likely to include:

- establishment of new stores for metadata which is not currently managed within the existing CMR
  - the preferred option here, wherever feasible, will be "buy" and integrate rather than "build from scratch"
    - we will aim to co-opt existing models and processes associated with the particular type of metadata (whether currently used in specific parts of the ABS or by other agencies) and integrate them effectively within the CMR.

- establishment of replacement stores and services for metadata which are currently managed within the CMR in a manner which does not best support the framework
  - for example, the Data Item Database currently has aspects of:
    - defining and managing properties
    - defining (via a very simplistic textural picklist) associated of object classes
    - defining qualifiers (a mediocre ability to assign subsetting qualifiers and a primitive ability to handle conceptual and scope qualifiers)
    - defining additional information data element concepts.

  - this is likely to be broken down into a number of stores and processes which better support the framework
    - there may be business value in then porting at least some of the existing content of the existing Data Items Database to the new structure

- extension of existing CMR stores and services where tuning rather than redevelopment is required .

45. In November 2003, we had the benefit of a number of international visitors to the ABS - people who are very familiar with metadata issues such as ISO 11179 standard and the MetaNet reference model. As a result of those discussions there are a number of further investigations. The following sub-paragraphs provide further details and could be of interest to statistical agencies that are using 11179, and might be worth some discussion.

i. Definition of a table or matrix, ie the result of the 'aggregation' process. The 11179 standard has been set up primarily with unit record data descriptions in mind. In a statistical organisation, we know that data element values from data providers are subject to processes such as aggregation across unit records or derivation within unit records or at the aggregate level, that create new data element values that are linked to the unit record description. For example, the *property* - sales; *object class* - businesses, manufacturing, Australia, provides a data element concept for input processing of a collection. After the aggregation process, there might be the *property* - sales; object class - Manufacturing (businesses), Australia, providing a new but related data element concept. Other properties associated with this new object class could be mean, RSE, count. These aggregate cells are part of a table, or data cube or matrix. Our interest is how to define these 'derived' data element concepts in 11179 and maintain the link to contributing data element concepts. Also, how to define the broader 'container' of the table or cube.

ii. When to use 'qualifiers' and when to use a hierarchy of 'object classes' to achieve subsets of populations. Statistics Canada did a lot of work in this area during 2002. While there are some overall differences of scope and emphasis within the ABS framework compared to the Canadian framework (which was also designed within the 11179 paradigm), it is likely that the well developed Canadian thinking on managing the challenges of the Object Class dimension will be very useful for the ABS.

iii. Expression of the 11179 model in an XML schema. Directions in ABS IT Enterprise Architecture are towards use of XML schemas to define data structures, particularly for the exchange of data and metadata. Therefore we are particularly interested in the representation of the ISO 11179 model in XML. It is understood that some work has been done in this area by members of the committee responsible for ISO 11179 but that work is incomplete and no longer active. Any other thinking, or work being undertaken, in this area would be of great interest to the ABS.

iv. ABS would like to define value domains (containing the permissible values) and 'rules' that would enable metadata to drive 'editing engines' rather than having 'editing rules' which include permissible values and tolerances implemented in program code. We know that this has been done in specific implementations but what ABS is interested in exploring is how this requirement might be described using the 11179 model for value domains, and what extensions would be needed. Where an extension is required, then it would be preferable for that extension to be one agreed by the statistical community.

It should be possible to create a 'permissible values' set from a 'derivation rule', possibly involving cross product of value domains in order to validate data element values during an IDW ETL process. There can be cases for a unit record where one data element value can make another data element value "non permissible". For example, if a business reports it has a total of 10 employees then it cannot report (on a consistent definitional basis) that it has 20 part time male employees. "10" and "20" are likely to be permissible values for the respective data elements but they are not permissible in combination. The latter "conditional permissibility" does not appear to be well defined within 11179.

We would like to apply similar concepts to drive editing processes eg query if value > y. In a sense this introduces the concept of a "questionable value", which is technically permissible if determined to be accurate - but whose accuracy should be questioned in the first instance. This concept is not defined within 11179 at all, but we should be able to be design a metadata extension to address it while the core definition of data elements remains consistent with 11179.

APPENDIX: Worked Example.

1. Level estimates of Private New Capital Expenditure: - collection development perspective

Conceptual framework = Australian Accounting Standards

Property concept = Expenditure

Conceptual Qualifiers:
- Classification concept = Accounting basis (cash / **accrual**)
- Classification concept = Series type (**value** / price / volume)
- Classification concept = Periodicity (daily, weekly, monthly, **qrtly**, annual)
- Classification concept = Observation perspective (**backward** / forward)
- Classification concept = Valuation basis (basic / **purchasers'** / producers')
- Classification concept = Type of derivation (**as-collected** / additive / multiplicative / index)
- Classification concept = Type of time series adjustment (**original** / seasonally adjusted / trend)

Scope Qualifiers:
- Classification concept = Type of expenditure (**capital** / current)
- Classification concept = Type of capital expenditure (**fixed** / financial)
- Classification concept = Type of fixed-capital (**new** / used)

Subsetting Qualifiers:
- Classification concept = Type of asset (**ANZSPC- Asset output variant, level 'x'**)
- Classification concept = Reference period (..., **ended 31 Mar 03**, ...)

Object Class = Business (**Type of Activity Unit** / Enterprise / EntGrp)

Scope Qualifiers:
- Classification concept = Institutional sectors (SISCA categories:
  1. Non-financial corporations;
  2. Financial corporations;
  3 General Government;
  4 Households;
  5. Nonprofit institutions serving households; and
  6 Rest of the world).
- Classification concept = Public/Private classification (Public / **Private**)
- Classification concept = Industry / Activity (ANZSIC categories -mix of 1&2 digit levels with exclusions)
- Classification concept = Employing unit (**yes** / no)

Subsetting Qualifiers:
- Classification concept = Industry (**ANZSIC, level '1' - and '2' in Mfg**)
- Classification concept = Location (**State / Territory**)

Representation:

data type = Integer
value domain = >=0
unit of measure = AUD 000's