Session 3: Emerging generic business process models and common frameworks and terminology – A basis for practical cooperation?

Integrating statistical information systems: The Statistical Data and Metadata eXchange Standard (SDMX)

Brian Studman
Australian Bureau of Statistics
brian.studman@abs.gov.au

Session Overview

- SDMX
  - Background
  - History
  - What can it represent
  - SDMX & Other Standards (particularly DDI)
- ABS & SDMX: Integration via IMTP (ABS major business programme)
- Using SDMX in the context of GSBPM
- Examples of Use: ABS main projects that utilise SDMX
- A quick preview of GSIM v 0.1 (if we have time)
SDMX

• Format for statistical returns for aggregated data
• Version 2.1 (April 2011 public comment)
• http://sdmx.org/ (best starting place)

SDMX comes from the international agencies (OECD, IMF, Eurostat, UNSD, World Bank, ECB, BIS)
  – they get aggregate statistical tables from many countries regularly over time
  – they wanted to automate and manage the process
    • they need standard agreed definitions and classifications, standard agreed table structures, standard agreed formats for both data and metadata
  – They commissioned SDMX in 2002
    • started a project, gathered use cases, employed consultants
    • produced a standard and presented it to large numbers of international statistical forums
    • started to use it and to pressure NSOs to use it
  – SDMX is good
    • excellent for managing dissemination of statistical data
      – very good tools for very impressive web sites based on data organised in the SDMX model
    • also some good frameworks for managing evolution of classifications
    • a framework for discussing agreements on concepts and classifications
      – Metadata Common Vocabulary, Cross-Domain Concepts, Domain-specific Concepts
Exchange and Sharing of statistical information

Emphasis on macro-data (aggregated statistics)

Promotes a "data sharing" model

- low-cost
- high-quality of transmitted data
- interoperability between (otherwise) incompatible systems

Structural metadata

Describe data structure

Reference metadata

Describe the contents and the quality of the statistical data

Statistical data

Statistical metadata

Describe data structure

Describe the contents and the quality of the statistical data
**Relationship to other standards**

- ISO 11179 is about what is good “concept” metadata
  - Data Elements and Data Element Concepts
    - and how to represent them (defines XML schemas)
    - no links to data

- ISO 19115 (and ANZLIC) are about geographic classifications
  - essential for geography classifications
  - other data is attached as attributes of a geographical element
    - limited use for statistics

- DDI is about linking metadata to (mainly) unit statistical datasets
  - and the processes that go on around them

- SDMX is about linking metadata to aggregate statistical datasets
  - and organising them for use

- The standards are never in real conflict
  - but there is sometimes a little wriggling and adjustment required
    - they all have their own XML schemas and they do not always line up nicely
    - mapping work is required and is happening
    - more conformance over time is likely as organisations use them together

**Work together with other standards?**

- Business Processes (details)
  - eg BPMN, BPEL

- Semantics
  - eg ISO 11179, RDF, OWL

- Catalogue & Discover
  - eg Dublin Core, ISO 2146

- Geospatial
  - ISO 19nnn (115, 119, 139 etc) [FGDC, INSPIRE, UNGWG, ANZLIC etc]
  - Open Geospatial Consortium (OGC) (GML, KML, WFS etc)

- Sector Specific
  - eg XBRL (business reporting), HL7 (health), SIF (schools)
DDI and SDMX

- Microdata data is an important source of aggregated data
- Crucial overlap and mappings exist between both worlds (but commonly undocumented)
- Interoperability provides users with a full picture of the production process

Why the Difference?

- DDI and SDMX are different because they are designed to do different things:
  - SDMX focuses on the exchange of aggregated statistics
  - DDI focuses on documenting social sciences research data
- There are many similarities and overlaps, but the intended function is different
  - Not all data cleanly fits into one category or the other, however
• SDMX focuses on aggregate data, especially time series
  – It can handle microdata, but is not well optimized for this
• SDMX focuses on collection and dissemination – *exchange* of data and metadata
  – It has an architecture and a good model for management, but it does not have an archival perspective. For archival use, DDI is better.
• SDMX provides support for any set of metadata (including DDI!) but is not optimized for use as a documentation standard for non-exchange activities.

Where DDI and SDMX Meet
• Several areas have direct correspondences in SDMX and DDI:
  – IDs and referencing use the same approach (identifiable – versionable - maintainable; URN syntax)
  – Both are organized around schemes
  – Both describe multi-dimensional data
    • A “clean” cube in DDI maps directly to/from SDMX
  – Both have concepts and codelists
  – Both contain mappings (“comparison”) for codes and concepts
Using DDI and SDMX Together

- There are a number of ways in which SDMX and DDI can be used together in the same system, or complement each other in data and metadata exchanges
  - Using DDI metadata as a link to source data for SDMX aggregates
  - SDMX and DDI as complementary formats for processing and dissemination
  - The SDMX Registry as a DDI metadata repository to support the lifecycle

SDMX/DDI Processing Support

- SDMX is easier to use for some tasks:
  - Processing multi-dimensional data for “clean” n-cubes (tabulation, etc.)
  - Representing micro-data sets for dissemination through web services and XML tools
- By using cross-walks, the best XML format for a particular process can be used
- Typically, the DDI and SDMX formats are maintained in parallel for the duration of processing
The SDMX Registry as a DDI Metadata Repository

• Because the SDMX Registry can be used to register, manage, and query DDI metadata instances, it can act as a metadata repository to track metadata versions throughout the DDI lifecycle
• SDMX does not directly address full-text search
  – This becomes part of the implementation
• The SDMX Registry can work as a concept-, question-, or variable bank, or as a metadata resource for processing and dissemination

ABS Experiences with SDMX

(via the Information Management Transformation Program)
Systematically supported production of SDMX V1 outputs from ABS data warehouse.

**History of SDMX in ABS**

- **2002**
  - Participation in OECD NAWWE case study

- **2003**
  - Systematically supported production of SDMX V1 outputs from ABS data warehouse

- **2006**
  - Provided SDMX v1 content to Reserve Bank of Australia

- **2007**
  - RBA feedback very positive overall

**Key solution design learnings:**
- Poor alignment with COGs, DSDs
- Need to duplicate content on ABS data warehouse

**2008**
- SDMX v1 content to supply with SDMX v2
- Cost benefit doubtful, without major reform of production and information processes

**2009**
- SDMX GC 2009: Statistical Information Management (SIM) framework => e2e statistical production processes
- ABS Management recognised maturity of SIM enablers to support coherent e2e within ABS and connecting the world within ABS

**Harnessing existing standards**

ABS Management Meeting in October 2009 recognised:

- Electing to use existing statistical information standards, where fit for purpose, to describe microdata, tabulations & other artefacts associated with the statistical production processes would make it possible to move forward immediately

- The potential for joint use of SDMX & DDI-L (Data Documentation Initiative - Lifecycle) to provide the backbone
  - This was informed by an assessment that SDMX & DDI-L based backbone could aid interoperability with data providers & data consumers who
    - do not use these particular standards but,
    - at least in some instances, use standards capable of interoperating with SDMX & DDI-L (eg geospatial data standards, XRBL, HL7, SIF).
Standards working together: Statistical Processes & SDMX + DDI-L

SDMX needs to be part of supporting integrated end-to-end statistical production processes. Here the possible contribution of SDMX is illustrated impressionistically against the UNECE Generic Statistical Business Process Model (GSBPM)

For NSIs, whose inputs are predominately microdata, SDMX may need to work in with standards such as DDI-L which apply to earlier phases of the production process

<table>
<thead>
<tr>
<th>Concept</th>
<th>Concept</th>
<th>Property</th>
<th>Concept and 11179 Property</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>Code</td>
<td>value</td>
<td>Value domain (permissible value)</td>
<td>Classification code value; 11179 Data element value domain</td>
</tr>
<tr>
<td>Category</td>
<td>Code</td>
<td>value</td>
<td>Value domain (value meaning)</td>
<td>Classification code description; 11179 data element value domain</td>
</tr>
<tr>
<td>Variable</td>
<td>Dimension</td>
<td>Data element 11179 data element; data item</td>
<td>11179 data element; data item</td>
<td>Variable</td>
</tr>
<tr>
<td>Question</td>
<td>Question</td>
<td>Question</td>
<td>Question</td>
<td>Question</td>
</tr>
<tr>
<td>Relationship</td>
<td>Relationship</td>
<td>Relationship</td>
<td>Relationship</td>
<td>Relationship</td>
</tr>
</tbody>
</table>
ABS IMTP

ABS Information Management Transformation Program (IMTP) is an enterprise level re-engineering program that aims to support

- A client environment where statistics are readily available, and can be easily integrated with data from other sources
- A statistical production environment that is highly productive, and satisfying for staff to use
- A statistical development environment that is nimble
- A systems environment that is built around standard models and supports shared collaborative development, including internationally.
**Information Management**

- **Principles**
- **Reference Models**
- **Policies, Procedures, Governance**
- **Implementations (MRR, DDD, LISOMX, ...)**

**Statistical Workflow**

- "Granular", high level re-usable business processes/components
- "Atomic", low level reusable services/processes

**Survey design and development**

**Survey Processing**

**Survey Manager**

**Data Analysis**

**Client consultation & service**

**Statistical Business Processes**

**New Apps**

**New User Interfaces**

**Existing Apps**

**Other Corporate Services** (E.g. Security)

**Data Repositories**
**Input Data Store**
- Metadata Registry & Repositories (MRR) (DDI-L, SDMX)

**Output Data Store**
- Process Library
- Output Data Store
- Metadata Registry & Repositories (MRR) (SDMX, SDDI)

**PROCESS METRICS / PARADATA / MI**

**SDMX enabled Dissemination**
- ABS.Stat
- REEM
- GeoSpatial

**Data Collection**
- Direct collection (Unit data)
- Other sources, Government Agencies etc...

**Unit and Summary Data from other sources**

**Denotes example discussed in later slides**

---

**IMTP: Examples**

**Example 1: Metadata Registry**
- Corporate and national statistical data & metadata resources are discoverable, interpretable and (re)usable (readily & reliably) by people and applications

**Example 2: Remote Execution Environment for Microdata (REEM)**
- Researchers are able to meet their own needs for tabulations and analyses from microdata, while preservation of confidentiality is guaranteed

**Example 3: ABS.Stat**
- Direct & flexible standards enabled access to ABS aggregate data, and associated metadata, reduces need for “precanned products” and creates opportunities for third party innovation
Key outcomes/benefits for statistical processes
• Data and metadata are:
  – fully discoverable for use and reuse
  – available to people and applications on a consistent basis, in an
    interpretable and actionable form
• Common services simplify design of business applications by supporting the
  definition, use & management of information
• Standard "information interfaces" mean methods and applications are
  interchangeable within the statistical production process
• Integrity of content is guaranteed over time
• Appropriate content and capabilities from the Registry are externalised
  beyond the ABS to support federated resource discovery
• Store information once - represent in as many formats as appropriate (SDMX,
  DDI-L, RDF etc)

Features
• Registry service interface aligned with SDMX
  – registers additional objects, eg
    • from DDI and detailed processes, workflows, business rules
  – supports some additional services/methods

Progress to date and roadmap :
• High level design approved December 2010
• Proof of Concept to be delivered June 2011
• Core functionality to be delivered 2011/12
• Migrate content to Registry, applications to use Registry Services by 2015/16

Learnings
• Need to be able to work (input/output) with some content using multiple standards eg
  SDMX Codelists, DDI-L Category & Code schemes
• Some gaps across SDMX & DDI-L currently, eg
  • description of formal, standard, statistical classifications as per,
    eg Neuchâtel Terminology Model
  • detailed machine actionable documentation of derivations, validations, "edits", business
    rules in a generic standard form

Opportunities
• Overlaps & gaps lead to idea of a "canonical" format, with appropriate mappings/transforms
  to relevant standards.
• This also helps manage version changes in international standards
• "Bottom up" input to information model
Example 2: REEM (Remote Execution Environment for Microdata)

Key outcomes/benefits for statistical processes
- REEM Survey Table Builder (STB) allows researchers to specify & populate tabulations “on demand” (“just in time”) from microdata
  - 994,010,994,000 different four dimensional tabulations are possible from 1000 variables on a microdata file
- NSI cannot pre-generate every combination “just in case” it’s needed
- REEM Analysis Service Modules (ASM) will support statistical analysis of microdata while preserving confidentiality and ensuring consistency
- ABS is able to support the needs of researchers more efficiently and effectively
  - Statistics of relevance to specific research will be made available which once would have been prohibitively expensive and time consuming to obtain

Features
- Aggregate tabulations described using SDMX
  - Machine to machine access via SDMX web services (design anticipates SDMX 2.1 specification) in addition to a Graphical User Interface (GUI)
- Original microdata described using DDI-L
- Assured protection of confidentiality on an automated basis

Progress to date and roadmap:
- REEM STB Release 1 made available to selected users December 2010
- Content for selected other (historical) surveys to be made available June 2011
- Analysis Service Modules released June 2012
- Enhanced capabilities (eg analysis for linked and longitudinal datasets, improved geospatial discovery & presentation services) delivered June 2014

Learnings
- Initial experience with REEM highlights challenges for business units producing statistics
  - Finding time for adoption of new approaches along with existing work
  - Mixed quality & availability of existing metadata to migrate to new standards

Opportunities
- Microdata access is an area of interest for collaboration for statistical agencies (eg via proposed OECD Expert Group) & others (eg Data Archives)
- Partnering with medium size niche market vendor with an interest in supporting open standards
  - More on REEM’s approach to SDMX (& DDI-L) is available in a vendor presentation.
Key outcomes/benefits for statistical processes

- Provides users (eg econometric modellers) with direct & flexible access to aggregate ABS data
  - efficient & quality assured “machine to machine” data feeds to the databases/models these analysts use become possible
  - users can save (and share) the specific query of interest to them and have the results updated automatically once new data becomes available
- Provides standards aligned output data from which standards enabled GUIs and analytical & presentation components can source content (data & metadata)
  - components/tools can be developed collaboratively by statistical agencies, or by innovative third parties, and then shared
  - standards enabled GUIs and tools benefit all users of ABS aggregate data, not just “power users”
- ABS can make available shared queries corresponding to existing packaged products (eg published tables, data cubes)
  - saves resources, improves navigation, improves quality, guarantees consistency

Features

- ABS.Stat is the ABS implementation of the OECD.Stat platform shared also with IMF, Statistics New Zealand and ISTAT (Italy)
- DotStat support for SDMX is a key consideration for ABS
- The platform supports machine to machine access via SDMX web services & saving of plugins, widgets, visualisations) on top of data feeds
- ABS can make available shared queries corresponding to existing packaged products (eg published tables, data cubes)
  - saves resources, improves navigation, improves quality, guarantees consistency

Progress to date and roadmap:

- MoU (Memorandum of Understanding) signed with OECD and pilot data service provided to key client (Treasury) June 2010
- ABS.Stat Basic made available for external users to evaluate “in beta” June 2011
- ABS.Stat Enhanced made available to external users on a “production” basis and integrated with ABS statistical production processes June 2012
- Ongoing enhancement including increased visualisation capabilities and integration with the Metadata Registry June 2013
- Keen to promote third party “innovation” (eg development of plugins, widgets, visualisations) on top of data services
- Keen to see improved support for SDMX, including SDMX 2.1 (including SDMX 2.1 services specification), from DotStat
- Keen to provide to key client (Treasury) 2010
- ABS.Stat Basic: made available for external users to evaluate “in beta” June 2011
- ABS.Stat Enhanced made available to external users on a “production” basis and integrated with ABS statistical production processes June 2012
- Ongoing enhancement including increased visualisation capabilities and integration with the Metadata Registry June 2013

ABS.stat Roadmap

<table>
<thead>
<tr>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
<th>Phase 4</th>
<th>Phase 5</th>
<th>Phase 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Completed</td>
<td>Completed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Research of software options for an output data store & dissemination solution
- Selected OECD Stat
- Establish opportunities for collaboration
- Develop business & technical understanding of Stat
- Form international agreement with OECD (MOU)
- Pilot of Data Services with a small number of datasets for key client(s)
- Install live version of .Stat without enhancements
- Subject Matter Experts (SME) engagement on data model
- Upload main economic indicators available from existing data warehouse
- Collaborate with .Stat community
- Establish internal .Stat user community
- New concepts website
- Evaluate user feedback on phase 3
- Build Automatic Loader
- Integrate .Stat with existing dissemination systems & processes
- Develop & implement enhancements to .Stat
- Define corporate approach to structuring data in .Stat aligned with national & international requirements
- Integrate data within SMAs where possible for upload into .Stat
- Establish external .Stat community
- Work with Data Management & METP on options for data integration
- Continue to install enhancements to .Stat
- Customised browser interface including data visualisation
- Continue integrated data migration
- Integrate with Metadata Registry
- Repository to support integrated discovery of content across ABS.Stat & other sources
- Complete integrated data migration
- Evaluation & recommendations

Learnings

- Pilot data service and other research confirms increasing
  - user demand for direct & flexible access to data
  - key client demand and capability for harnessing machine to machine web services
- Practical experience of differences (+/-) between collaborative development and vendor provided services

Opportunities

Example 3 : ABS.Stat
Integration: What you need to consider

- Business processes
  - yours
  - as well as upstream and downstream processes
- Metadata
  - what is required e.g. classifications, questions, variables
  - how it is going to be searched
  - where it is stored now
- Systems – data/metadata stores, applications, services
  - what you interact with and how
  - what you are going to provide
  - what you need from other projects
- Flows
  - process, data, metadata
  - what metadata to accumulate through e2e life cycle
- Timeframe – critical dates for your project(s)
- Implementation in stages –
  - e.g. Metadata only first, live connection later
  - alpha, beta, test, user acceptance, production
- …..

Challenges and Opportunities

- Buy-In from Subject Matter Experts
- Building Capability
- Common Practice
- Achieving Semantic Interoperability
  - within Statistical Subject Matter Domains
  - within Statistical Information Management
- SDMX and DDI working together
Summary

- After "dabbling" for 7 years, in October 2009 ABS "decided" (took a decisive, irrevocable step) in regard to SDMX
  - For ABS, SDMX represents one key enabler for much broader business transformation (not an end in itself)
    - needs to work with other enablers, including other standards
  - Practical work on initial implementation projects that use SDMX is
    - demonstrating feasibility & value
    - generating learnings
    - confirming known opportunities & identifying additional ones
  - While substantial, we believe remaining challenges can be overcome
    - resolving these, to the benefit of all, is much more likely through collaboration and community than through efforts in isolation
  - No single standard meets all requirements, more work needed!
  - Initial discussions between the SDMX community and the DDI community have been initiated & other related work is ongoing: GSIM, ESSnet CORA/CORE, etc.

Acknowledgements:

Statistical Investment in the Future – Program, 14-15 September 2009
SDMX Implementation in the statistical practice
Mimmi Phillips
Executive, Statistical Information Technologies Unit

SDMX and the DDI: Using the Right Tool for the Job, 2010
and
Exploring the Relationship between DDI, SDMX and the Generic Statistical Business Process Model, 2011
Austin Gregory
Open Data Foundation

SDMX and DDI February 2010
Overview for ADB Executive
Bryan Fitzgibbon
Rapassa Consulting Limited

IMTF – OCIMF Project Update 28 May 2011
OCIMF (Operational Common Metadata / Information Management Framework)
ADB OCIMF Reference and Working Groups

Planning Individually Strength Implementation of SDMX
A case study from the Australian Bureau of Statistics (ABS)
SDMX Global Conference 2011
Albert Hamilton
ABS Chief Statistical Information Architect
Information Management Transformation Program
QUESTIONS?

http://sdmx.org/

Brian Studman
Australian Bureau of Statistics
brian.studman@abs.gov.au

ANOTHER ABS SDMX PROJECT EXAMPLE
IMTP Example: GaSI
(Geospatial and Statistical Integration)
• Still in applied research phase (unlike ABS.Stat & REEM)

• Medium term aim
  – SDMX to WFS (Web Feature Service) for data
    • together with appropriate geospatial metadata

• Ultimately want full interoperability between “dimensional” (SDMX) & geospatial standards based views of statistics for purposes of
  – discovery & selection of data of interest
  – presentation of data of interest
  – analysis of data of interest

SDMX MODELLING CONCEPTS IN MORE DETAIL
• SDMX has
  – Organisations organised into schemes
    • Organisations own and manage artefacts, and provide or receive things
  – Concepts organised into schemes
  – Codelists, including classifications
    • a Codelist combines DDI Categories and Codes
  – Data Structure Definitions (Key Families)
    • a DSD describes a conceptual multi-dimensional cube used in a Data Flow and referenced in Datasets

• SDMX has
  – Data Flows
    • described by a DSD, linked to registered data sets, and categorised
  – Categories organised into schemes
    • not the same as a DDI Category
    • provide a basis for indexing and searching data
  – Hierarchical Codelists
    • a misnomer – maps relationships amongst inter-related classifications
    • explicit, actionable representations of relationships
  – Process metadata
    • a Process has steps with descriptions, transition rules, computation information, inputs, outputs
    • all actionable, linked to other SDMX artefacts or to external sources
• SDMX has
  – Structure Sets
    • additional linking of related DSD and Flows
  – Reporting Taxonomies
    • information about assembling reports or publications
  – Reference Metadata, Metadata Structure Definitions, and Metadata Flows
    • additional, probably useful, options for attaching metadata to data
  – Annotations almost everywhere
    • good options for managed, actionable extensions

SDMX Registry

• Standard interfaces are provided for implementing a web-services-based SDMX Registry
• A registry classifies and indexes data and metadata sets, but the data and metadata sets can be held in any repository or web server
• A registry functions for distributed systems like a card catalog functions for a traditional print library
A Note about the SDMX Registry

- SDMX was intentionally designed to work with other standards
  - DDI (and other standard XML formats) can be registered in an SDMX registry using the simple, user-configured metadata format
  - This makes DDI accessible as a resource in an SDMX system
- The DDI lifecycle model can be represented as an SDMX "Process"
  - This can help with tracking DDI metadata through the lifecycle

SDMX SEMANTICS – ISSUES IN PRACTICAL IMPLEMENTATION
Semantic Interoperability: Statistical Subject Matter Domains

- Achieving consistent, extensive, practical (eg potentially machine actionable) semantic interoperability is a challenge for ABS
  - In the case of EIF also: Achieving semantic interoperability in the EU context is a relatively new undertaking, not achieved before on this scale.

- Areas of emphasis include
  - controlled vocabularies (common use and/or relatable)
    - SDMX cross domain concept/codelist approach seen as adding value here
  - well designed, well defined, relatable DSDs
    - DSDs to meet national and local needs sometimes vary from internationally agreed DSDs
      - eg to incorporate national standards for classifying industry, occupation
      - structures should be relatable (eg mappable)

Semantic Interoperability: Statistical Information Management

- As a common reference model, Generic Statistical Business Process Model (GSBPM) has a vital enabling role in supporting statistical business process management/reuse

- Similarly it is envisaged a Generic Statistical Information Model (GSIM) may have a vital enabling role in supporting statistical information management/reuse
  - propose work through the METIS Group, convened by the UNECE, to develop & agree GSIM
    - the same approach was followed for GSBPM

- Once agreed as a common reference model, GSIM could be operationalised through
  - mapping it to information models (semantic), and ultimately XML schemas (technical), associated with existing standards such as SDMX and DDI-L
  - addressing (eg through proposed extensions to existing standards) any gaps identified during the mapping process

- Similarly to GSBPM (but even more so) GSIM (& its operationalisation) would be intended to facilitate collaboration & sharing across NSIs
Semantic Interoperability (for SIM)

Some Challenges

- SDMX Metadata Common Vocabulary (MCV) not yet harmonised & machine referenceable
  - aware of promising MCV Ontology work associated with SDMX ESSnet
  - anticipate connections with GSIM in future

- Need to clarify semantic relationship between SDMX & DDI-L information models
  - also seek agreed common practice for complementary implementation of the two standards

SDMX & DDI-L Working Together

More…

- The conceptual, & the leading practice, working relationship between SDMX & DDI-L is a key area of interest for more than just ABS, eg
  - In June 2010 the inaugural meeting of Informal CSTAT workgroup on stronger collaboration on Statistical Information Management Systems identified SDMX and DDI-L as two statistical information standards upon which “information interoperability” for collaborative developments could be founded.
  - The most recent meeting of the International Working Group on Microdata Access agreed to trial use of DDI-L and SDMX
  - In December 2010 an ongoing informal dialogue was established that engages the two standards bodies (eg representatives of the SDMX Secretariat & the DDI Alliance) as well as third party stakeholders who have an interest in the two standards & how they work together.

- Having resolved to apply SDMX Cross Domain Concepts & Data Structure Definitions to support coherence of output, there is a need to establish leading practice in efficiently & effectively working toward this outcome during the (often microdata/DDI-L oriented) design, collect & process phases of a coherently designed & managed statistical production process
DDI MODELLING CONCEPTS
IN MORE DETAIL

- DDI comes from the data archive organisations across many countries
  - trying to capture and store survey data for future use
    - and to document it so future users can understand it and make sense of it
    - mostly social science collections from researchers
    - funding organisations are requiring such data to be preserved for further use
  - mostly they had to grab data and try to salvage metadata after the event
    - but DDI now aims to capture all metadata “at source”
  - early versions were narrowly focused on an individual data set
    - grew out of their documentation processes
  - latest version (DDI V3) is much more extensive, better organised
    - common analysis/designer support with SDMX
    - an end-to-end model compatible with the Generic Statistical Business Process Model (GSBPM)
      - the international version of the BSIP “caterpillar”
• The DDI package is a “Resource Package” or a “Study Unit”
  - but these is a high-level containers
    - they can contain everything or nothing
    - we probably want to have methods that deliver Resource Packages and possibly Study Units
    - but mostly we want to focus on lower-level artefacts
      - specifically when specifying Resource Packages
      - but also for general retrieval
  - the DDI artefacts are not totally clearly defined
    - they are not necessarily split the way we would like
      - and they have a fair amount of “noise”

• SDMX metadata is more clearly defined and explicit
  - but it does not cover all types that we will want
    - more focussed at the dissemination end
  - the SDMX package is a Structure document
    - but the artefacts are well defined

• DDI has
  - Survey-level metadata
    - Citation, Abstract, Purpose, Coverage, Analysis Unit, Embargo, …
  - Data Collection Metadata
    - Methodology, Sampling, Collection strategy
    - Questions, Control constructs, and Interviewer Instructions organised into schemes
  - Processing metadata
    - Coding, Editing, Derivation, Weighting
  - Conceptual metadata
    - Concepts organised into schemes
      - Including 11179 links
    - Universes organised into schemes
    - Geography structures and locations organised into schemes
• DDI has (cont)
  – Logical metadata
    • Categories organised into schemes
      – (categories are labels and descriptions for question responses, eg, Male, Unemployed, Plumber, Queensland, …)
    • Codes organised into schemes and linked to Categories
      – Codes are representations for Categories, eg “M” for Male, “Qld” for Queensland)
    • Variables organised into schemes
      – Variables are the places where we hold the codes that correspond to a response to a question
    • Data relationship metadata
      – eg, how Persons are linked to Households and Dwellings
    • NCube schemes
      – descriptions for tables

• DDI has (cont)
  – Physical metadata
    • record structures and layouts
  – File instance metadata
    • specific data files linked to their record structures
  – Archive metadata
    • archival formats, locations, retention times, etc
  – Places for other stuff not elsewhere described
    • Notes, Other Material
  – References to “Agencies” which own artefacts but no explicit stricture to describe them
  – Inheritance and links embedded in most schemes
    • but need to be ferreted out, not necessarily easily usable