Mapping Metadata between SDMX and DDI

Bryan Fitzpatrick
Rapanea Consulting Limited
March 2010
What is the Issue?

• SDMX and DDI are two standards for managing statistical metadata and data

• Statistical offices are interested in improving the management and automation of its statistical processes

• Key to managing the statistical processes is managing and using metadata of all sorts

• DDI and SDMX (along with other standards such as ISO11179 and the geographical standard, ISO 19115 or country variants) a relevant, useful, “fit-for-purpose” standards for use in this program

• SDMX and DDI both map metadata to data
The Use Case

• End-To-End Management of the Statistical Process
  – From conception of requirement to dissemination of the statistics
    • Typically multiple, repeating, ongoing cycles
      – Monthly, quarterly annually
    • Planning, data gathering, edit/impute/derive, weight and aggregate, analyse, disseminate
    • Integration, coherency, relatability goals
      – Metadata consistency, relatability of aggregate statistics
    • Automation of process and capture of knowledge
    • Metadata and data sharing
      – National Statistical Offices and International Organisations
  • All sorts of statistics – social, economic, agriculture, trade, transport…
  • Commonality across statistics areas and organisations/countries
SDMX and DDI

• SDMX
  – Focuses on the aggregate data dissemination end
    • Defining and describing multi-dimensional cubes
      – Often conceptual, spanning cycles and countries
      – Value added with and quality usage information
        » “Reference metadata”
  • Sharing the metadata and the data
    – Service interfaces for registry/repository
  • Origin in international agencies and “official statistics”
    – OECD, Eurostat, IMF, ECB, …
  • Emphasis on Concepts and Classifications
    – Classification is a major entity in its own right
  • Superb for managing aggregate statistical web sites
    – Useful models for evolution of classifications
SDMX and DDI

• 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)
    • no Registry/Repository
DDI Metadata

- 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 Metadata

• 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 Metadata

• 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 structure to describe them
  – Inheritance and links embedded in most schemes
    • but need to be ferreted out, not necessarily easily usable
SDMX Metadata

• 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 Metadata

• 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
    • 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 Metadata

• 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
The objective

• Map and convert DDI and SDMX metadata
  – DDI
    • Concepts and Concept Schemes, Category Schemes and Code Schemes, Variables
      – Perhaps NCubes

  – SDMX
    • Concepts, Codelists (Classifications), Data Structure Definitions
    • Hierarchical codelists for mapping evolution of classifications

• Aim is round-trip migration of metadata with no loss
  – Probably integrated into an external metadata store
The Work

• Started in November 2009
  – at DDI Expert Group Meeting in Dagstuhl, Germany
  – Arofan Gregory and Bryan Fitzpatrick
  – not yet completed
    • aim to complete it over next few months
The Work

• ConceptScheme -> ConceptScheme
• Concept -> Concept
• CodeScheme -> Codelist
  – With information from CategoryScheme
• Code -> Code
  – With information from Category
• If multiple Code Schemes for one Category Scheme, information is duplicated
• Must hold information from DDIInstance and ResourcePackage somewhere
  – I used annotations on SDMX Comments element (not much use for anything else)
  – Really need annotations at SDMX outer element (Structure)
The Work

• Not all information from one standard has a natural correspondent in the other
  – aim to hold this information in “generic” places
    • Annotations in SDMX
    • OtherMaterial in DDI

• Some information has a direct correspondent but they are not compatible
  – IDs and URNs
The Future

• We will finish the mapping
  – and present it in DDI and SDMX forums

• We will use the mapping in practice
  – in process management work in National Statistical Offices

• We will look for more consistency and correspondence between SDMX and DDI
  – work to define a consistent Registry/Repository is underway
  – I will be looking for more cooperation and integration on metadata artefacts
Questions

- BryanM Fitzpatrick@Yahoo.CO.UK