SDMX and the DDI: Using the Right Tool for the Job

Arofan Gregory
Executive Manager, Open Data Foundation
A Choice of Tools

For any given task, we have a choice of tools:

(For screwing things in…)

(For wrenching things…)

(For hammering things…)

(For getting hammered…)

IT is a Bag of Hammers

• A good tool box has a variety of tools

• But not everybody understands that (especially in IT!)

“The hammer in your hand is the best one for the job…” NOT!
SDMX and DDI

• Overview of SDMX major features
  – Comparison with DDI
• Selecting the right standard
• Direct mappings between them
• Using SDMX and DDI together
SDMX Background

• SDMX is the “Statistical Data and Metadata Exchange” initiative, formed in 2001
• Now ISO/TS 17369 (version 1.0)
• Produced by 7 large supra-national organizations
  – BIS, IMF, OECD, World Bank, UN, Eurostat, European Central Bank
• Adoption doubled in the past year
  – More than 40 organizations are using it (or starting to)
  – UN Statistical Commission declared it the preferred standard for statistical exchanges this year
What is SDMX?

• The problem space:
  – Statistical collection, processing, and exchange is time-consuming and resource-intensive
  – Various international and national organisations have individual approaches for their constituencies
  – Uncertainties about how to proceed with new technologies (XML, web services …)
Major Products of SDMX

• Technical standards for formatting aggregate data (versions 1.0 and 2.0)
  – Supports XML and EDIFACT formats
• Technical standards for formatting metadata (version 2.0)
  – XML format only
• Information model for managing statistical collection, processing, and dissemination (version 2.0)
• Registry-based architecture, based on web services/SOA (version 2.0)
• Content-oriented guidelines (now in draft):
  – Classification of all statistical activities (high level)
  – “Common Metadata Vocabulary” provides definition of terms and concepts
  – Cross-domain metadata concepts – common concepts for structuring data and metadata sets
Data Formats

• Message for describing multi-dimensional data structures (XML, EDIFACT)

• Message for transmitting multi-dimensional data (4 equivalent flavors of XML, EDIFACT)
  – The different XML flavors support different use cases
  – They are identical in content, and can be transformed back and forth

• Data concepts are configured by the user, and can describe *any* multi-dimensional data
Metadata Formats

• Metadata structures are described in an XML message
• Metadata reports have equivalent 2 flavors of XML (for different use cases)
• Metadata reports can be configured to contain any metadata
  – This includes DDI, Dublin Core, etc.
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
The Old JEDH (Joint External Debt Hub) Site

- BIS
- IMF
- OECD
- World Bank

(Various Formats)  (3-month production cycle)
JEDH with SDMX

Retrieves data from sites

SDMX-ML

SDMX-ML

SDMX-ML

SDMX-ML

Debtor database

SDMX-ML

Agent

Discover data and URLs

Data provided in real time to site

JEDH Site

BIS

IMF

OECD

World Bank
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
DDI and SDMX

DDI
- Microdata
  - Low level observations
  - Single time period
  - Single geography
  - Controlled access
  - Expert Audience

SDMX
- Aggregated data
  - Indicators, Time Series
  - Across time
  - Across geography
  - Open Access
  - Easy to use

• Microdata data is a important source of aggregated data
• Crucial overlap and mappings exists 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
A Practical Difference: Tools Support

• SDMX is older than DDI 3.0, but younger than DDI 1.*/2.*
  – Not surprisingly, DDI 1.*/2.* has the best tools support, but SDMX has a growing set of tools which nearly match them
  – DDI 3.0 has a small but growing set of tools, but is not as well supported as SDMX today
Which is the one to use…
when you’re using only one?

• 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
A Better Toolbox: 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
Linking Source Data and Aggregates

• DDI provides a wealth of information about the micro-data which serves as an input to SDMX aggregates
  – It is possible to capture these links in SDMX, at the cell level or higher, to provide automated access to source data
  – An SDMX registry can be used to provide easy access to these links
  – The user/collector of aggregate data can access the rich DDI metadata, and possibly the data
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
The Full Toolbox DDI, SDMX, and More

- DDI and SDMX were both created with an awareness of other useful standards
  - ISO/IEC 11179 and related standards
  - METS
  - OAIS (PREMIS)
  - Web-Services and XML Standards
  - ISO 19115
  - Dublin Core

- All of these standards can work together to provide a more complete set of standards-based functionality
- Standard mappings are being defined by people from many different organizations (see presentation from METIS 2008 in Luxembourg)
High-Level Vision – Standards Mappings

Federated Registries (Based on SDMX, ebXML, web services)

Semantic definitions

ISO 11179

Aggregated Data/Metadata (SDMX)

Organized using

References to source data

registered

METS/PREMIS

Standard classifications

Registered

Used in

ISO 19115 Geographies

Dublin Core Citations

DDI Microdata Sets

XBRL Business Reports
Summary

• It is not as simple as DDI-or-SDMX
• The two standards are designed to perform different functions, but also to be complementary
• SDMX (especially the registry) can be used as a platform to support DDI-driven systems