Creating a Model for a Central Metadata Repository

Klas Blomqvist

EDDI 14, London 2014-12-02
Background

- Statistics Sweden - NSI
- > 200 “statistical products”
- Traditional stove-pipes production system
  - “Product” oriented
- Decentralised organisation
  - Relatively independent departments
  - 50% financed by commissions – “solution for each customer”
- Strategies (Decision by the Director General)
  - Register system must be coordinated
  - Data storage changed from accessing individual products to accessing the entire data warehouse
Content

- Register and data warehousing vision and strategy
- A central metadata repository
- Standards based
  - GSBPM
  - GSIM
- Testing in production systems
GSBPM - Generic Statistical Business Process Model

Quality Management / Metadata Management

1 Specify Needs
   1.1 Determine needs for information
   1.2 Consult & confirm needs
   1.3 Establish output objectives
   1.4 Identify concepts
   1.5 Check data availability
   1.6 Prepare business case

2 Design
   2.1 Design outputs
   2.2 Design variable descriptions
   2.3 Design data collection methodology
   2.4 Design frame & sample methodology
   2.5 Design statistical processing methodology

3 Build
   3.1 Build data collection instrument
   3.2 Build or enhance process components
   3.3 Configure workflows
   3.4 Test production system

4 Collect
   4.1 Select sample
   4.2 Set up collection
   4.3 Run collection
   4.4 Finalize collection

5 Process
   5.1 Integrate data
   5.2 Classify & code
   5.3 Review, Validate & edit
   5.4 Impute
   5.5 Derive new variables & statistical units
   5.6 Calculate weights
   5.7 Calculate aggregates

6 Analyse
   6.1 Prepare draft outputs
   6.2 Validate outputs
   6.3 Scrutinize & explain
   6.4 Apply disclosure control
   6.5 Finalize outputs

7 Disseminate
   7.1 Update output systems
   7.2 Produce dissemination products
   7.3 Manage release of dissemination products
   7.4 Promote dissemination products
   7.5 Manage user support

8 Archive
   8.1 Define archive rules
   8.2 Manage archive repository
   8.3 Preserve data & associated metadata
   8.4 Dispose of data & associated metadata

9 Evaluate
   9.1 Gather evaluation inputs
   9.2 Conduct evaluation
   9.3 Agree action plan
Conceptual model

Implementation standards

Other relevant standards

DDI

SDMX

Geospatial standards
Strategy

- Support process oriented production
- Active Communication with suppliers
- Use of registers when possible
- Standardization
SCB’s strategy on coordinated statistics production

- The centre of a sphere of interest
- “Owns” the population
- Continuously updated
- Independent of other base registers
- Has links to the other base registers
- Contains mainly stock variables
- Time stamped data

- Process data “automatically” provided
- Variable values checked individually or by comparisons
- Longitudinal editing and analysis supported
- Physical data transportation minimized
- No “private” versions
A Central Metadata Repository

Metadata Store
- SCBDOK
- DoS
- Business Processes
- Variables
- Populations
- Unit Types
- Value Domains
- Question Bank
- Thesaurus
- Process Methods
- Process Steps
- Rules
- Business Services Catalogue
- Tech info

Support and infrastructure
- Operations Support
- SCB.SE

The Platform
- Specify needs
- Design and plan
- Build and test
- Collect
- Process
- Analyse
- Disseminate and communicate

Data Store
- Process Metrics
- Statistical Data
- Production Data

Evaluate and feedback
- Specifying needs
- Designing and planning
- Building and testing
- Collecting data
- Processing data
- Analyzing data
- Disseminating and communicating

Support and infrastructure
- SCB
- Statistics Sweden

Business Processes
- Variables
- Value Domains
- Populations
Current Metadata Situation at SCB

- **Input**
  - Raw data
  - Local systems

- **Throughput**
  - Micro data
  - SCBDOK
  - MetaPlus

- **Output**
  - Macro data
  - BaS
  - SSD

- **Scenarios**
  - Local systems
  - SCBDOK
  - MetaPlus
  - BaS
  - SSD
Desired Metadata Situation

Metadata Repository

- SCBDOK
- BaS
- MetaPlus
- MacroMeta

Input
- Raw data

Throughput
- Micro data

Output
- Macro data

Common vocabulary
- MetaCollect
- Question Bank
- ETL rules

Process metadata
- Aggr. rules

MacroMeta
- MetaPlus
- SCBDOK
- BaS

Question Bank
- MetaCollect
- SCBDOK
- BaS

MetaPlus
- SCBDOK
- BaS
- MetaPlus
- ETL rules

ETL rules
- MetaCollect
- MetaPlus
- ETL rules

Common vocabulary
- MetaCollect
- MetaPlus
- ETL rules

MetaCollect
- MetaPlus
- ETL rules

MetaPlus
- ETL rules
- Aggr. rules

Aggr. rules
- MetaPlus
- MacroMeta

Process metadata
- MacroMeta
Metadata Expectations

- More active metadata
  - Documentation during the design phase
  - Metadata driven production
- Enable tracing
  - What was the question that led to this published figure?
  - Ideally a two-stage tracing for NA
- More and better process data
  - Who did what when and why?
- More comprehensive metadata
  - Improved search mechanisms
- Unified terminology
  - Thesaurus
Runs and feeds the processes

- Input
- Process
- Output/input
- Process
- Output
Macro editing with drill-down

Input
Raw data

Throughput
Micro data

Output
Macro data

Macro edit

Display source micro data

Insert new generation

Deemed as being incorrect

Display edit results

Click suspect item
The Current Metadata Situation

Input
- Raw data

Throughput
- Micro data

Output
- Macro data

Local systems
- SCBDOOK
- DoS
- MetaPlus
- SSD

Support and infrastructure
- Operations Support
- SCB.SE

Evaluate and feedback
- Specify needs
- Design and plan
- Build and test
- Collect
- Process
- Analyse
- Disseminate and communicate

Metadata

Data
MetaPlus Structure

- Register
  - Register variant
  - Register version
- Object class
  - Population
  - Context Population
- Variable
  - Object variable
  - Context variable
  - Value
    - Conceptual value domain
    - Value domain
- Data base table
  - Column
Central Metadata Store for efficient statistics production

**Business**

**Exchange**
Statistics product, Supplier, User, Contract, Agreement

**Structures**
Dataset, Reference metadata, Data Structure, Dimension, Record

**Concepts**
Unit, Population, Variable, Value Domain, Classification
Common and specific metadata

Common to the entire statistics production process
- Process step
- Service catalogue
- Variable
- Classification

Individual product implementation
- Process step within a round
- Service within a round
- Value domain
- Instance variable

Frame unit data
Sample unit data
Report unit data
Observation unit data

Statistical data collection
Process data

Data Store

Common metadata definitions
Definitions of occurrences specific to a production round
Actual occurrences
SCB’s strategy on coordinated statistics production
Ongoing

- Prototype - Access database
- Coordinating economic short term statistics (5 products)
- Testing in Structural business statistics and National accounts (Excessive Deficit Procedure - EDP)
- Case on sampling filling the blue part
- Version 1.1
- Next step - Incorporating in/with the platform
The end…