Metadata Management and Production System for surveys in Empirical Socio-economic Research

A Project funded by EU under the 5th Framework Programme
Key Action: Improving the Socio-economic Knowledge Base
- Data infrastructure – HPSE-CT-2002-00122
Project period 01.01.2003 – 31.12.2005

MetaDater: Towards Standards and Tools to document comparative surveys

Ekkehard Mochmann (Project Co-ordinator)
Uwe Jensen (Project Manager)
GESIS / ZA University of Cologne

Edinburgh, May 27, 2005

Internet: www.metadater.org
Objectives & Products

- **Standards for Metadata**
  - Data model for metadata covering the life-cycle of complex comparative surveys on space & time axis; no panel

- **Efficient User system**
  - for Data Collector and Provider to produce, manage & exchange metadata according to new standards
Project Status 2003 – 05.2005

1. User analysis
   Data archive / Researcher & field work institute

2. Data modelling
   - Design of a unique data model based on analysis of events & processes of the life-cycle of a comparative studies
   - Conceptual Data models > Relational Data model (2nd edition)
   - Provision & Evaluation of the Data model by DDI / CSDI

3. System Architecture & Functional Core
   - Design of 3 tier system (GUI, Application Server, RDBMS)
   - Programming in Java

4. Application development & programming
   - Modular system for Data provider & collector

5. Usability
   - Preparation of a test database & tests

6. Dissemination activities
1. User Analysis - Results

Archive – Analysis on work sequences & systems

Common shared aspects
• Acquisition > Entry control & Study documentation > Archiving > Dissemination

Differences in detail (Archives; Studies)
• Complexity: Cross-section studies > comparative Studies
• Technical & functional differences in used systems

Researcher & Research institutes

Capture of Metadata at first occurrence
• General Interest vs. missing concepts or tools

Specific Interests given
• Capture of Metadata on Study design & Field work
• Support on Questionnaire development / Question DB
• Easy exchange of metadata between researcher & archives
2. Data modelling of metadata

1. Conceptual Data Model
   Major Life-cycle phases of Studies
   1. Study design and documentation
   2. Questionnaire development & testing
   3. Sample selection & field work
   5. Study extension, harmonization & standardization
   6. Indexing & Classification
   7. Data and Metadata publication & distribution
   8. Analysis, correction and update of Metadata Documentation

Recent discussions & evaluation in the DDI & CSDI network
   • Study Life-cycle is adopted in DDI data modelling; DDI version 3 shall be compatible
   • After evaluation provision to interested users

2. Relational Data Model
   • Relational schema of Metadata & metadata management by RDBMS
   • Compatible with DDI v2.0 > extension for comparative Studies
3. System Architecture

3-tier system
- User interface for the modules
- Application server
- Relational database management system
4. User Modules & functional components

Overview:

M1  Study description
M2  Q/V editor – define standard as a reference
M3  Process single data set (cross-section; comparative study)
M4  Process extension: harmonize & standardise comparative data
M5  Import & Export of XML files or SPSS metadata
Module M1 Study Description (cross-section dataset)

Study description with metadata sections on
Project > Study > Data set > Design - Questionnaire – Data - Context – Classification

This right panel illustrates the view of the "citation - title statement folder."
Module M2 Q/V editor

Basic task – define the standard for a study or a study collection

- Capture or edit **questions** from the (basic) questionnaire
- Define the depending **variables**
  - add variables to the data file structure (standard setup)
- Organise, route & filter, index & classify questions - variables

Principle options to work on questions with the Q/V editor:

- **Import** (basic) questionnaire or questions to DB
  - Xml compatible & provided by a project; from XML codebook
- **Copy** from database
  - complete basic questionnaire (e.g. replicate ISSP)
  - single questions from previous wave (e.g. EB trend questions)
- **Capture** Questions from scratch with the Q/V editor

Q/V screens to capture of different question types:

- **Question type 1**: Single response on simple Question
- **Question type 2**: Single response on Question with Items
- **Question type 3**: Multiple response by Answer Instances (1st. item; 2nd. Item)
- **Question type 4**: Multiple response by Answer categories (Mentioned / not mentioned)
- **Question type 5**: Multiple response on single Items of a question (Grid)
Module M2 Q/V editor / example

Question type 4: Multiple response on Question with Items

**Step 1:**
- select Question type
- enter 1.st / last Item No. …
- CREATE activates Question / Variable pane
- offers respective no. of table lines for
  - Items 1 to 17
  - depending variables

**Step 2:** enter Question text … & Variable definitions

**Step 3:** next screen
Module M2 Q/V editor / example

Question type 4: Multiple response on Question with Items

**Step 3** at Tab Value/ Missing > select valid & missing values
Application design & programming

**M3 – Process single data set (cross-section; comparative study) with reference to the Q/V standard**
- Document deviations in question wording or demographic items
- Process metadata of the single datasets load SPSS data and adjust them to the standard
- Technical standardisation of space or time bound variables
- Adjust the variables to the standard (recode, rename, filtering etc)

**M4 - Process Extension: harmonize & standardise comparative data**
- integrated dataset (space compound datasets) > e.g. ISSP
- cumulated dataset (time compound datasets) > e.g. ALLBUS
- cumulated-integrated dataset > e.g. EB Trend file
- Generalized dataset harmonization

Index & classify on Study and variable level

**M5 - Import & Export of XML or SPSS based metadata**
- Metadata exchange & publications (XML) > SD, CB, Quest. – SPSS metadata: portable / syntax / setup
- Access via WWW front-ends (MADIERA; local systems;)

Preparation on Usability testing & user workshops

Other modules to come 2005
The end of the presentation
2. Data modelling of metadata

1. Conceptual Data Model
   • Life-cycle of Studies from the Concept > Dissemination
   • Relational schema of Metadata
   • Compatible with DDI v2.0 > extension for comparative Studies
   • After evaluation provision to interested users

2. Relational Data Model
   • Model implemented at the RDBMS; DD & ERD
   • Recent evaluation in the DDI network
     • Study Life-cycle is adopted in DDI data modelling
     • DDI version 3 shall be compatible
     • Detailed analysis is under work of DDI expert groups
2.1 Conceptual Model II. Behavioural model

Main process groups along the study life-cycle

1. Study design and documentation
   - Concept, Questionnaire, Field work, Data / Study description
2. Contact and Data acquisition
3. Data & document. control – Process Cross-section
4. Study extension, harmonization & standardization
5. Indexing & Classification
6. Data and Metadata publication & distribution
7. Correction and Update of Metadata Documentation
Model for simple & replicated survey

- **Simple cross-section:**
  Project > Study with > one design > one set of data > a set of Variables > Values

- **Repeated over space and time:**
  Basic object: dataset requires proper definition to be managed by the system
  - single dataset = one sample at one time
  - compound dataset = collection of datasets > many sets of data
    - collected in different time instances (time compound)
    - collected in different spaces (space compound)
  - processed dataset = integration of single data sets to a new one
    - cumulative (covering different time instances)
    - integrated (covering different space instances)
    - cumulated-integrated (covering time and space instances)
Conceptual model Concepts II

Types of Metadata (app 7)

- **Definitional**
  - Data definition, methodology, measurement

- **Operational**
  - Researchers, Use of data

- **Contextual**
  - Social background; party information … Documents

- **Classification**
  - Study, variable, question level by variety of vocabularies

- **Administrative**
  - Embargo, access categories, responsibility to change data …
Integration & Cumulation processes (app 8)

- Cumulation (z. B. ISSP)
- Time series (z. B. Politbarometer; ALLBUS)
- Trends (Eurobarometer)
- Harmonisation of different Data sets
  - involved datasets already in the systems
  - Harmonisation over space or time
  - Procedures: ex post overall – or adding new dataset
- Integration of collections of studies into the system
  - Import already processed metadata to the system
Conceptual model Concepts IV

Questionnaire documentation & Question DB (app9)
  ● The Questionnaire
    ● Questionnaire ‘editor’
    ● Import & export questionnaires
    ● Indexing of questions & questionnaire
  ● The questions – to be classified
    ● Question types
    ● Questions forms
  ● Relation between dataset and questionnaire & questions
Conceptual model Concepts V

- Documentation object
  (DOID = Document object id (~ ISBN) to identity uniquely every object in system
  Assign administrational attributes to DO (not to each basic entity)

- Attributes like: who, when, release, permission level
  or subtypes of document objects
  - Study
  - Data set
  - Publication
  - Classification
  - Object of observation
  - Attribute
2.2 Relational Model E-R-D Schema & DD
3-tier system

- **Client tier**
  - Swing-Client
  - Web-Client

- **Middle tier**
  - Application server (running EJB) "JBoss"

- **Database tier**
  - Database Oracle
  - Database MySQL
  - ...

3. System Architecture
Question type 1: Single response on simple Question

Step 1: select Question type and define ... Open empty fields > step 2

Step 2: enter Question and Variable information

Step 3: search and select Answer scale / missing values
Question type 2: Single response on Question with items

Step 1: select Question type and define ...
Open empty fields → step 2

Step 2: enter Question and Variable information

Step 3:
- Select Tab for the Answer text / missing value
- search and select Answer scale / missing values

Result > next screen
4.1 Modules – M3 Q/V editor / example 2

Question type 2: Single response on Question with items

Step 3 Tab with select Answer scale / missing values