STARDAT

DATA ARCHIVING SUITE

Monika Linne, Alexander Mühlbauer, Wolfgang Zenk-Möltgen

DDI - The Basis of Managing the Data Life Cycle, Dec. 8 - 9, 2010, Utrecht, The Netherlands
Overview

- Initial Situation – Different Archiving Tools
  Main Tools – DBKEdit/ DSDM/ CBE
  Intention – Integration of Different Archiving Tools
  Proceeding – Requirements Analysis
  DDI Formats Currently Used
  Requirements Concerning DDI Formats
  Getting Started with STARDAT Development
  Challenges of Going into DDI3
Initial Situation – Different Archiving Tools

Online Publication
- DBK
- ZACAT
- Variable Overview
- DSDM
- DBKEdit

Publication on CD-ROM
- Report
- Report-Tool
- CBE

Longterm-preservation
Overview

Initial Situation – Different Archiving Tools

- Main Tools – DBKEdit/ DSDM/ CBE

Intention – Integration of Different Archiving Tools

Proceeding – Requirements Analysis

DDI Formats Currently Used

Requirements Concerning DDI Formats

Getting Started with STARDAT Development

Challenges of Going into DDI3
Main Tools

1. Data Catalogue Edit-Tool (DBKEdit)

Documentation and management of metadata on study level of all archived studies

A. Inhouse-tool
B. Web-based
C. DDI-supporting

Information provided
For each study/dataset (partly bi-lingual):
- Title & study number
- Primary investigators & fieldwork agencies
- Universe & sampling
- Abstracts
- Related publications
- Access categories
- Questionnaires, etc.
- Notes and comments for internal use

Publication via
- ZACAT (Nesstar)
- CBE (CodebookExplorer)
- DBK (Data Catalogue)
DBKEdit – Management of Metadata on Study-Level
DBK – Research and Information Provided for the User

Data Catalogue

You searched for "(Values and "Title") in Abstract and date of collection: 1999

Found entries: 74

Choice  | Study No.  | Title
---|---|---
1 | 07761 | EVS - European Values Study 1995/2000 (release 2, May 2006) - Spain

Study No. 07761

Title EVS - European Values Study 1995/2000 (release 2, May 2006) - Spain

Current Version Version number (tested automatically; implementation of a uniform versioning policy)

Date of Collection 1999

Principal Investigator/Authoring Entity

[Information]

Data Collector Delta, Madrid

Abstract

Moral, religious, societal, political, economic and social moral concepts of the Europeans.

Topics:

- Secularization with life
- Importance of selected characteristics of occupational work
- Personal employment
- Self-determination (scope)
- Attitude towards Salience of individual goals
- Priority of national over foreign as well as men over women
- Degree of prestige of individual or social reasons for this situation of economic need of individuals
- Freedom of the unemployed to reject a job offer
- Policies, participation
- Attitude to foreign workers in one's country
- Fear of the future
- Assimilation and integration of immigrants.
Main Tools

2. Dataset Documentation Manager (DSDM)

Documentation of metadata on variable-level for special datasets and data collections

A. Inhouse-tool & external projects
B. ACCESS-based desktop-application
C. DDI-supporting

Information provided
For each Variable:
- Exact question and answer texts
- Multilingual documentation
- Documentation of deviations in questionnaires fielded in different waves or countries
- Interviewer instructions
- Forward/backward instructions for filter questions
- Documentation of additional archival notes (e.g. recoding, deviations)
- Generating variable categories

Publication via
- ZACAT (Nesstar)
- CodebookExplorer
DSDM – Management of Metadata on Variable-Level
Main Tools
3. CodebookExplorer (CBE)

Tool on context-level which gives easy retrieval possibilities to end-users for databases of specific topics

A. Tool for end-users or members of external projects
B. Inhouse-tool supporting the publication of context information (e.g. comparable questions, show cards)
C. Not DDI-supporting

Information provided
For each study/data set
- Browsing studies or trends, topics, scales etc. (if the database provides them)
- Structuring of variables according to content or methodical aspects
- Compare question texts
- Make simple analysis like frequencies or crosstabs
- Show questionnaires
- Access to SPSS datasets

Publication via
- CodebookExplorer (CD-ROM/Internet)
- ZACAT
**CBE → Publication and Management via CD-ROM and Internet**

European Values Study

---

**Dataset 2A4009** | Integrated Dataset - EVS 2000
Variable V1: How important is it in your life work (O1A)

- **WE START WITH SOME QUESTIONS ABOUT LIFE IN GENERAL, LEISURE TIME/ACTIVITIES AND WORK**
- **Q1 WORK**
  - 4 other things
  - 4 question not satisfied
  - 3 happy
  - 2 no
  - 1 very important
  - 2 quite important
  - 3 not important
  - 4 not at all important

**Comparability:**

**Card 1**

- 1: Very important
- 2: Quite important
- 3: Not important
- 4: Not at all important

---

**Dataset 2A5001** | Integrated Dataset - EVS 1995-2000
Variable V1: How important is it in your leisure time (O1B)

- Q1 Leisure time
  - 4 other things
  - 4 question not satisfied
  - 3 happy
  - 2 no
  - 1 very important
  - 2 quite important
  - 3 not important
  - 4 not at all important

**Comparability:**
Overview

Initial Situation – Different Archiving Tools

Main Tools – DBKEdit/ DSDM/ CBE

Intention – Integration of Different Archiving Tools

Proceeding – Requirements Analysis

DDI Formats Currently Used

Requirements Concerning DDI Formats

Getting Started with STARDAT Development

Challenges of Going into DDI3
Intention – Integration of Different Archiving Tools

- Integrated management system for metadata
- Transfer of the features of DBKEdit, DSDM, CBE and further tools
- Interoperability with standards like DDI 3, ISO 20252
- Multi-language documentation on study and variable-level
- Web based modul for structured metadata capture, management and dissemination (Web Based Data Ingest)
- Controlled vocabularys (Thesauri)
- Related publications, continuity guides, scales, trends and additional metadata
- Longterm-preservation with DDI
- Export in different portals like ZACAT, Cessda Data Portal, Sowiport
Overview

Initial Situation – Different Archiving Tools

Main Tools – DBKEdit/ DSDM/ CBE

Intention – Integration of Different Archiving Tools

Proceeding – Requirements Analysis

DDI Formats Currently Used

Requirements Concerning DDI Formats

Getting Started with STARDAT Development

Challenges of Going into DDI3
Requirements Analysis

- Structured interviews with the main stakeholders:
  - Investigation of usage of the tools and workflows
  - Identification of problems
  - Feedback from stakeholders about necessary/desired changes
  - Result: Description of all functionalities, also structured in diagrams

- Requirements specification document
  - Detailed description of requirements
  - Transfer into BPMN-Diagrams
  - Coverage of not yet existing functionalities

- Next steps:
  - Building software-modules
  - Evaluation of the new tool by users
Overview

Initial Situation – Different Archiving Tools

Main Tools – DBKEdit/ DSDM/ CBE

Intention – Integration of Different Archiving Tools

Proceeding – Requirements Analysis

DDI Formats Currently Used

Requirements Concerning DDI Formats

Getting Started with STARDAT Development

Challenges of Going into DDI3
DDI Formats Currently Used

**DBK**
- Export to DDI 2.0
  - for publication on ZACAT (Nesstar) server
  - for long-term archiving
  - for data exchange with portals like da|ra, sowiport

**DSDM**
- Export to DDI 2.0
  - for publication on ZACAT (Nesstar) server
  - for long-term archiving
- Export to DDI 3.1
  - for Enhanced Publication Editor (linking publications to datasets)

**CBE**
- No DDI currently used
DDI Formats Currently Used

Example: Study Description from DBK to DDI 2.0 contains

- Study number and title
- Principal investigators
- Abstract and classifications
- Research area and data collection information
- Further documents and bibliographic references
- Dataset information
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE codeBook SYSTEM 'http://info1.gesis.org/DDI/Version2-0.dtd'>
<codeBook>
  <docDscr>
    <citation>
      <titlStmt>
        <titl>Dataset Title</titl>
        <subTitl/>
        <altTitl/>
        <parTitl/>
      </titlStmt>
    </citation>
  </docDscr>
  <stdyDscr>
    <citation>
      <titlStmt>
        <titl>Dataset Title</titl>
        <subTitl/>
        <altTitl/>
        <parTitl/>
        <IDNo>Study Number</IDNo>
      </titlStmt>
      <rspStmt>
        <AuthEnty>Principal Investigator, Affiliation: Institute</AuthEnty>*
      </rspStmt>
      <prodStmt>
        <producer></producer>
        <fundAg></fundAg>
      </prodStmt>
      <distStmt>
        <distrbtr abbr='GESIS' URI='http://www.gesis.org/'>
          GESIS - Leibniz-Institut für Sozialwissenschaften
        </distrbtr>
        <depositor/>
        <depDate/>
      </distStmt>
    </citation>
  </stdyDscr>
  <serStmt><serName/></serStmt>
  <verStmt><version/></verStmt>
</codeBook>
DDI Formats Currently Used

Classification

Abstract

Research area

Collection Dates

Data Collectors

Sample Information

Data Collection Mode

Access Class

Related Materials

Related Materials
DDI Formats Currently Used

<relMat>
  <citation>
    <titlStmt>
      <titl><![CDATA[ Document Name and Type ]]]></titl>
      <holdings URI='Document URL'/>
    </titlStmt>
  </citation>
</relMat>

<relStdy/>
<relPubl>
  <citation>
    <titlStmt>
      <titl>Related Publications (only for first relPubl-Element)</titl>
    </titlStmt>
  </citation><![CDATA[ Bibliographic Reference ]>
</relPubl> *
<othRefs>
  <citation>
    <titlStmt>
      <titl>Further Remarks</titl>
    </titlStmt>
  </citation><![CDATA[ Further Remarks ]>
</othRefs>

</othrStdyMat>
<fileDscr>
  <fileTxt>
    <dimensns>
      <caseQnty>Number of Cases</caseQnty>
      <varQnty>Number of Variables</varQnty>
      <recPrCas>Records per Case</recPrCas>
    </dimensns>
    <fileType>File Format</fileType>
  </fileTxt>
</fileDscr>
</codeBook>
DDI Formats Currently Used

Example: Variable Descriptions from DSDM to DDI 2.0
(for long-term archiving)
contains
- Study number and specific list (countries, waves, panels)
- SPSS variable names, labels, values and value labels
- Variable groups
- Question numbers, texts, interviewer instructions, filter instructions
- Sub-question numbers and texts
- Answer codes and categories, answer groups
- Several variable notes, e.g. for derivation
DDI Formats Currently Used

<?xml version="1.0" encoding="UTF-16" ?>
<!DOCTYPE codeBook SYSTEM "http://info1.gesis.org/dfd/DSM/DSM-LTA-200.dtd">
<!-- Selection of DDI 2.0 DTD Elements and Attributes used by DSDM for long-term archiving -->
<!-- The original DDI 2.0 DTD can be found at http://www.icpsr.umich.edu/DDI/Version2-0.dtd -->
<codebook ID="Archive Study ID">
<docDscr>
<citation>
<titlStmt>
<titl>Langfristsicherungsdokumentation der Studie 'Archive Study ID'</titl>
</titlStmt>
<rspStmt>
<AuthEnty>GESIS</AuthEnty></rspStmt>
<prodStmt>
<producer>GESIS</producer>
<copyright>© 2007-2009 GESIS</copyright>
<prodDate>Date (DD.MM.YYYY)</prodDate>
<software>DSDM DDI Long-Term Archiving 2.0.0</software></prodStmt>
</citation>
</docDscr>
<stdyDscr>
<citation>
<titlStmt>
<titl></titl>
</titlStmt>
<stdyInfo>
<sumDscr>
<nation* ID="country identifier" abbr="short name of country">long name of country</nation>
</sumDscr>
</stdyInfo>
</stdyDscr>
<dataDscr>
<varGrp* ID="identifier of variable group" var="list of variable references, coded under <var>">
<labl>name of variable group</labl>
</varGrp>
<var* ID="variable identifier" name="SPSS variable name">
  <labl>SPSS variable label</labl>

  <qstn* ID="identifier of question"
    seqNo="Questionnaire No"
    sdatrefs="list of country references, coded under <nation>">
    <preQTxt?>text before question</preQTxt>
    <qstnLit*>literal question</qstnLit>
    <forward*>forward instruction</forward>
    <backward*>backward instruction</backward>
    <ivulInstr*>interviewer instruction</ivulInstr>
    <postQTxt?>text after question</postQTxt>
  </qstn>
  Or (if only reference):
  <qstn* qstn="question reference, coded under <qstn>"/>

  <txt* ID="identifier of sub question"
    Level?):"sub question no"
    sdatrefs="list of country references, coded under <nation>" sub question text</txt>
  Or (if only reference):
  <txt* sdatrefs="sub question reference, coded under <var><txt>"/>

  <catgryGrp* ID="identifier of answer value group"
    catGrp="list of answer value references, coded under <catgry>"
    <labl>name of answer value group</labl>
  </catgryGrp>
  Or (if only reference):
  <catgryGrp* catGrp="answer value group reference, coded under <catgryGrp>"/>
<catgry*> ID="identifier of answer value" only exists if tag is AnswerValue in DSDM
Missing?="SPSS value missing" as Y/N only exists if tag is VariableValue in DSDM
<catValu>answer value code and SPSS value only numeric! </catValu>
<labl?>SPSS value label</labl> only exists if tag is VariableValue in DSDM
<txt* ID="identifier of specific answer text">country specific value label</txt>
</catgry>

Or (if reference):
<catgry* sdatrefs="specific answer text reference" >
<labl?>SPSS value label</labl> only exists if tag is VariableValue in DSDM
</catgry>

<derivation>
<drvdscr>description of derivation</drvdscr>
</derivation>

<notes? type="NoteNote">userdefined notes to variable: Note</notes>
<notes? type="Text">userdefined notes to variable: Text</notes>
<notes? type="Note">userdefined notes to variable: Archive Note</notes>
<notes? type="Problem">userdefined notes to variable: Problem</notes>
</var>
</dataDscr>
</codeBook>
DDI Formats Currently Used

Example: Variable Descriptions from DSDM to DDI 3.1
(for Enhanced Publications)
contains
- Study number
- SPSS variable names, labels, and URNs
- Variable groups
- Complete question, answer texts and notes for each variable
- SPSS values and value labels
<?xml version="1.0" encoding="UTF-16" standalone="no"?>
<ddi:DDIInstance
  id="gesis_Archive Study ID"
  agency="de.gesis"
  version="Version (Format Major.Minor.Revision default 1.0.0)"
  versionDate="Date (ISO Format YYYY-MM-DD)"
  xmlns:ddi="ddi:instance:3_1"
  xmlns:s="ddi:studyunit:3_1"
  xmlns:pd="ddi:physicaldataproperty:3_1"
  xmlns:pi="ddi:physicalinstance:3_1"
  xmlns:c="ddi:conceptualcomponent:3_1"
  xmlns:l="ddi:logicalproduct:3_1"
  xmlns:r="ddi:reusable:3_1"
  xmlns:dc="ddi:datacollection:3_1"
  xmlns:a="ddi:archive:3_1"
  xmlns:xhtml="http://www.w3.org/1999/xhtml"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="ddi:instance:3_1 http://info1.gesis.org/ddi/3_1/instance.xsd">
  <s:StudyUnit id="Archive Study ID_SU">
    <r:UserID type="DBK Study Number">
      Archive Study ID
    </r:UserID>
    <r:Citation>
      <r:Title>
        Archive Study ID
      </r:Title>
    </r:Citation>
    <s:Abstract id="Archive Study ID_A">
      <r:Content>Study description not available.</r:Content>
    </s:Abstract>
    <r:UniverseReference>
      <r:ID>UNIVERSE_REF</r:ID>
    </r:UniverseReference>
    <s:Purpose id="Archive Study ID_P">
      <r:Content>
        Archive Study ID DSDM to EPE DDI 3.1 Export (1.4.0)
      </r:Content>
    </s:Purpose>
    <dc:DataCollection id="Archive Study ID_DatCol">
      <dc:QuestionScheme id="Archive Study ID_QueSch">
        <dc:QuestionItem id="Archive Study ID_QSPSS Variable Name"/>
      </dc:QuestionScheme>
    </dc:DataCollection>
  </s:StudyUnit>
</ddi:DDIInstance>
DDI Formats Currently Used

```xml
<l:LogicalProduct id="Archive Study ID_LogPrd">
  <l:DataRelationship id="DataRel">
    <l:LogicalRecord id="LogRec">
      <l:VariablesInRecord allVariablesInLogicalProduct="true">
        <l:VariableSchemeReference>
          <r:ID>Archive Study ID_VarSch</r:ID>
        </l:VariableSchemeReference>
      </l:VariablesInRecord>
    </l:LogicalRecord>
  </l:DataRelationship>

  <l:CategoryScheme id="Archive Study ID_QSPSS Variable Name_CatSch">
    <l:Category id="Archive Study ID_QSPSS Variable Name_CatValue">
      <r:Label>SPSS Value Label</r:Label>
    </l:Category>
  </l:CategoryScheme>

  <l:CodeScheme id="Archive Study ID_QSPSS Variable Name_CodSch">
    <l:Code>
      <l:CategoryReference>
        <r:ID>Archive Study ID_QSPSS Variable Name_CatValue</r:ID>
      </l:CategoryReference>
      <l:Value>SPSS Value</l:Value>
    </l:Code>
  </l:CodeScheme>

  <l:VariableScheme id="Archive Study ID_VarSch">
    <l:Variable id="SPSS Variable Name">
      <r:UserID type="SPSS Variable Name">SPSS Variable Name</r:UserID>
      <r:Label maxLength="255" type="label">SPSS Variable Label</r:Label>
      <l:QuestionReference>
        <r:ID>Archive Study ID_QSPSS Variable Name</r:ID>
      </l:QuestionReference>
    </l:Variable>
  </l:VariableScheme>
</l:LogicalProduct>
```
Overview

Initial Situation – Different Archiving Tools

Main Tools – DBKEdit/ DSDM/ CBE

Intention – Integration of Different Archiving Tools

Proceeding – Requirements Analysis

DDI Formats Currently Used

Requirements Concerning DDI Formats

Getting Started with STARDAT Development

Challenges of Going into DDI3
Requirements Concerning DDI Formats

- Export to DDI 2.1 still needed
  - for publication on ZACAT (Nesstar) server
  - for data exchange with portals like da|ra, sowiport
- Export to DDI 3.1 needed
  - for long-term archiving
  - for Enhanced Publication Editor (linking publications to datasets)
- Import from DDI all versions needed
  - for data exchange with primary researchers/projects
- Future DDI versions need to be supported
- Usage of Resource Packages needed for re-using elements
  - for own elements and elements from other institutions
- Concept for long-term archiving of re-used elements needed
- Migration issues need to be solved
Overview

Initial Situation – Different Archiving Tools

Main Tools – DBKEdit/ DSDM/ CBE

Intention – Integration of Different Archiving Tools

Proceeding – Requirements Analysis

DDI Formats Currently Used

Requirements Concerning DDI Formats

Getting Started with STARDAT Development

Challenges of Going into DDI3
Getting Started with STARDAT Development

- Architectural requirements
  - Browser-based web application
  - Transaction-based multi-user access
  - Distributed undo-/redo-mechanism
  - Multi-threading
  - Automation of metadata management tasks
    e. g. long running tasks with progress reporting and cancelation support
  - Extensible framework for further needs of metadata documentation
  - Keeping in mind: REST and SOA

- Experimental prototyping
  - First milestone until May 2011
  - Implementation of only a small set of functional requirements
  - Check of feasibility of architectural and technical ideas
  - Iterative process to improve development and implementation
Getting Started with STARDAT Development
First Draft of Architecture
Overview

Initial Situation – Different Archiving Tools

Main Tools – DBKEdit/ DSDM/ CBE

Intention – Integration of Different Archiving Tools

Proceeding – Requirements Analysis

DDI Formats Currently Used

Requirements Concerning DDI Formats

Getting Started with STARDAT Development

 Challenges of Going into DDI3
Challenges of Going into DDI 3
Really Internalize Lifecycle Orientation

- Managing documentation process of complex social science data
  - Apply adequate grouping approach
  - Identify a strategy to establish resource packages

- Migration issues
  - Find equivalent elements
  - Identify additional elements needed
  - Identify re-usable elements
  - Re-structure the elements

- Building software
  - Existing software tools are *static*
    - Only their combination supports lifecycle management
  - New software tool (!) shall be *dynamic*
    - Lifecycle management is inherently contained
Challenges of Going into DDI 3
To Be ... or Not to Be

What does it mean when we talk about DDI 3 usage?

- supporting DDI 3
  - proprietary domain model
  - proprietary storage
  → I/O module with some squeezing mapping

- compatible with DDI 3
  - DDI 3 domain model, perhaps some proprietary extensions
  - storage in relational database
  → mapping between XML and relational database

- based on DDI 3
  - DDI 3 domain model, no proprietary extensions
  - storage in flat XML files or native XML databases
  → no mapping, full interoperability
Challenges of Going into DDI 3

What DDI 3.x? How to manage DDI version evolution?

- Extensions
  - no problem

- Modifications
  - not trivial
  - but assumed to be manageable

- Downgrade mismatch
  - unconditionally avoid nasty lock-in effect
  - but how?
Challenges of Going into DDI 3

Performance Issues

- The performance topic has been addressed

- Solutions are already discussed by the community
  - DDI light
  - XML – RDB transformation
  - We need to understand and adapt to our needs

- We need always keep in mind performance because of its extensive side effects on architecture
  - What usage of DDI is possible?
  - What storage technique do we use?
  - Where are the bottle necks?
Challenges of Going into DDI 3
Not at least Reuse, Reuse, Reuse of Software!

- More generalized thinking
- Stringent component orientation in architecture design
- Advanced documentation needs
- We would like to offer our results to the DDI3 community
- We are interested in exchange.
  ➔ One first step done: We already started an exchange with DDA, we hope to reuse DDI3 Editor code.