Controlled vocabularies for DDI3
a work in progress

1st Annual European DDI Users Group Meeting 3-4 December 2009

Taina.Jaaskelainen@uta.fi
Meinhard.Moschner@gesis.org

DDI CVG

- Develop/propose controlled vocabularies on DDI3 elements or attributes
- Co-ordination with Technical Implementation Committee (TIC)
- Will become part of the official DDI3 documentation
- Resource for all DDI users

- Multinational working group within in DDI Alliance context
- Regular videoconferences (usually every two weeks)
- Members:
  - Atle Alvheim, NSD
  - Sanda Ionescu, ICPSR
  - Taina Jääskeläinen (chair), FSD
  - Chryssa Kappi, EKKE
  - Fredy Kuhn, FORS
  - Ken Miller, UK-DA (recently retired)
  - Meinhard Moschner, GESIS

[Note:]

NSD: Norwegian Social Science Data Services (Bergen)
ICPSR: Inter-university Consortium for Political and Social Research (Ann Arbor, MI)
FSD: Finnish Social Science Data Archive (Tampere)
EKKE: National Center for Social Research / Greek Social Data Bank (Athens)
FORS: Swiss Foundation for Research in Social Sciences (Lausanne), formerly SIDOS
UK-DA: UK Data Archive (Essex)
GESIS – Leibniz Institute for the Social Sciences (Data archive for the Social Sciences, Cologne)

gesis
Leibniz Institute for the Social Sciences

DDI CVG

- **Challenges:**
  - Limited experience with (knowledge of) DDI3
  - Not-so-clear use of DDI3 elements or attributes
  - Overview of DDI3 elements in the different modules
  - Potential heterogeneity of DDI user community
  - Lack of existing controlled vocabularies (e.g. for subsetting)
  - Lack of standard use of terms
  - Optimizing both indexing and searching (as large and complex as necessary, but as small and simple as possible)

www.fsd.uta.fi

[Note:]
Issues which are not covered by the CVG:

Technical aspects: CV publication format (recommended to publish the machine-actionable part of the CV in Genericode); application requirements for handling hierarchies, if present in a CV, for indexing and searching; separation of index terms and structural description in a CV etc.

Business practices: Selection of a platform for publishing CVs (currently planned to become part of DDI3 documentation) ; decision about term structure in CVs; versioning and cross-referencing of terms in CVs; policies governing CV maintenance etc.

gesis
Leibniz Institute for the Social Sciences

Controlled vocabularies

- Organized list of subject terms for indexing and retrieval
- (Ideally) **exhaustive** list of terms
- **Mutual exclusive** terms (no overlapping)
- **Clearly defined** subject terms
- The only choices for usage in a specific context
- Scope notes to avoid misunderstanding if needed
- From a short (flat) list to a complex (hierarchical) thesaurus including relationships between terms like ELSST ...

www.fsd.uta.fi

[Note:]

ELSST: European Language Social Science Thesaurus (not in the scope of the CVG)

gesis
Leibniz Institute for the Social Sciences

Importance of CVs

- Metadata formats:
  - machine readable (structured or semi-structured text)
  - machine interpretable (DDI2)
  - **machine actionable (DDI3)**

- **Consistency and efficiency** in the production of metadata
- Semantic/technical **interoperability between organizations**
- Semantic/technical **interoperability between systems**
- **Precision** in searching the metadata

- CVs usually do **not replace textual description**

gesis
Leibniz Institute for the Social Sciences

Some types of vocabularies

- ... as illustrative examples from selected DDI3 modules

  - STUDY UNIT
  - DATA COLLECTION
  - PHYSICAL DATA PRODUCT
  - PHYSICAL INSTANCE
  - COMPARATIVE MODULE
  - REUSABLE MODULE

[Note:]

STUDY UNIT: corresponds to the "study" in a traditional codebook, i.e. nature and scope of the data collection
DATA COLLECTION: metadata about the data collection process (data sources, sampling, measurement instrument etc.)
PHYSICAL DATA PRODUCT: describing the physical layout of the data
PHYSICAL INSTANCE: information about the physical instance of an actual data file
COMPARATIVE MODULE: relations between study units in terms of universe, concepts, questions, variables, categories and code schemes
REUSABLE MODULE: common features for all modules

gesis
Leibniz Institute for the Social Sciences

DDI3 STUDY UNIT

- **KindOfDataType**

  Kind of data documented in the logical product of a study unit (i.e. kind of distributed data, not the source of data)

  - Administrative data
    - Clinical records
    - Sales records
    - Court proceedings
    - School records
    - Police records
    - Local authorities records
    - National government records.
  - Assessment data
    - Examination results
    - Psychological/Intelligence tests
    - Evaluation/Accreditation results
  - Audio data
    - Music
    - Radio
    - Speech
  - Demographic data
    - Birth rates
    - Death rates
    - Causes of Death (?)
    - Divorce rates
    - Fertility data
    - Marriage rates
    - Census/Enumeration data
  - Measurement data
    - Physical measurement
    - Biological measurement
    - Economic measurement
    - Environmental data
    - Statistics
    - Ratings
    - Web logs
  - Image data
    - Photographs
    - Film
    - TV programs
    - Maps
  - Personal communication data
    - Oral accounts
    - Written narratives
    - Diaries
    - Letters/E-mails
  - Public communication data
    - Company accounts
    - Catalogues
    - Printed publications
    - Online publications
    - Web Sites
  - Survey data
    - Government surveys
    - Market research surveys
    - Opinion polls
    - Independent survey
  - Voting data
    - Election returns
    - Exit polls
    - Parliament votes
  - Event/Transaction data
  - Other

gesis
Leibniz Institute for the Social Sciences

DDI3 DATA COLLECTION

- **TimeMethod**

  Describes how time fits into the data collection methodology

  - Longitudinal
    - Cohort/Event-based
    - Trend/Repeated cross-section
    - Panel
      - Continuous
      - Interval
  - Time Series
    - Continuous
    - Discrete
  - Cross-sectional
  - Cross-sectional ad-hoc follow-up

- **ModeOfDataCollection**
  - Interview
    - Face-to-face
    - Telephone
    - E-mail
    - CATI
    - CAPI
  - Self-completed questionnaire
    - Paper/pencil
    - Web-based
    - CASI
    - ACASI
  - Coding
  - Transcription
  - Compilation
  - Synthesis
  - Recording
  - Simulation
  - Observation
    - Field
    - Laboratory
    - Participant
  - Experiments
  - Focus Group
  - Other

www.fsd.uta.fi

gesis
Leibniz Institute for the Social Sciences

DDI3 PHYSICAL DATA PRODUCT

- **CharacterSet**
  used in the data file

- ASCII
- ISO-8859-1
- ISO-8859-2
- ISO-8859-3
- ISO-8859-4
- ISO-8859-5
- ISO-8859-6
- ISO-8859-7
- ISO-8859-8
- ISO-8859-9
- ISO-8859-10
- ISO-8859-11
- ISO-8859-13
- ISO-8859-14
- ISO-8859-15
- ISO-8859-16
- Mac OS Roman
- Unicode 5.1
- UTF-8
- UTF-16

- **SoftwareName**

  physicaldataproduct:
  proprietary

- AcaStat
- ADaMSoft
- Analyse-it
- Auguri
- BioStat
- BMDP
- BrightStat
- Dataplot
- EasyReg
- Epi Info
- EViews
- GAUSS
- Golden Helix
- GraphPad Prism
- gretl
- JMP
- Limdep
- MacAnova

gesis
Leibniz Institute for the Social Sciences

DDI3 PHYSICAL INSTANCE

- **CategoryStatisticsCodedType**
  (category level statistics)

- Absolute Frequency
- Percent of N
- Valid Percent
- Percent of total sum
- Cumulative Frequency
- Cumulative Percent
- Percentile Rank - lower or equal
- Percentile Rank - lower
- Standard Error (SE)
- Confidence Interval - level of confidence: 90%
- Confidence Interval - level of confidence: 95%
- Confidence Interval - level of confidence: 99%
- Other

- **SummaryStatisticsCodedType**

- Arithmetic Mean (X)
- Geometric Mean
- Harmonic Mean
- Trimmed Mean
- Standard Error of the Mean
- Mode (Mo)
- Median (Mdn)
- Valid Cases
- Invalid Cases
- Minimum
- Maximum
- Range
- Sum
- Variance (s2)
- Standard Deviation (s)
- Coefficient of variation (CV)
- Average absolute deviation (AAD)
- Median Absolute Deviation (MAD)
- First Quartile
- Second Quartile
- Third Quartile
- Interquartile range

gesis
Leibniz Institute for the Social Sciences

DDI3 COMPARATIVE

- **CommunalityTypeCoded**
  UniverseMap, ConceptMap, QuestionMap, VariableMap, Categorymap
- Initially suggested values in DDI3:
  - Identical
  - High
  - Medium
  - Low
- CVG proposal after consulting researchers about usability:
  - Identical
  - Some
  - None

gesis
Leibniz Institute for the Social Sciences

DDI3 REUSABLE

- **ContributorType:** Role
  - Data Collector
  - Data Producer
  - Depositor
  - Metadata Producer
  - Research Instigator
  - Other
- **Publisher:** Role
  (if added by TIC)
  - Publisher
  - Distributor

- **LifeCycleEventType**
  - Study Proposal
  - Study Design
  - Instrument Design
  - Funding
  - Interviewer training
  - Ethics Review
  - Sampling
  - Instrument pre-testing
  - Pilot study
  - Questionnaire translation
  - Documentation translation
  - Data collection
  - Data collection reports
  - Post-collection processing
    - Data production
    - Initial data quality checks
    - Metadata production
    - Original release
  - Deposit
  - Post-production processing
    - Data quality checks
    - Data editing

gesis
Leibniz Institute for the Social Sciences

Why CESSDA wants to use CVs?

- Improve and harmonize archive CVs
- Multilingual access and documentation
- Temporal, spatial and topical comparability
  - retrieval
  - standardization in the context of question data bases
  - data harmonisation routines (→ CHARMCATS)
- Authentication and authorisation procedures

www.fsd.uta.fi

[Note:]

CESSDA (Council of European Social Science Data Archives)

CHARMCATS (Cessda HARMonisation of CATegories and Scales) about the use of DDI3 for publishing harmonisation routines. See EDDI09 presentation by Martin Friedrichs (GESIS/CESSDA-PPP) in session A2.

gesis
Leibniz Institute for the Social Sciences

Recommendations for CESSDA-ERIC

- Need for an agreed metadata/DDI template
- Use of CVs **obligarory for agreed DDI elements**
- Use of CVs recommended also for other elements
- Use of CVs as the first step towards DDI3
- Most CVs already applicable in DDI2
- Translation of CESSDA CVs into local languages

www.fsd.uta.fi

[Note:]

Recommendations from the EU funded CESSDA Preparatory Phase Project (CESDDA-PPP) for the **CESSDA-ERIC** (European Research Infrastructure Consortium) starting in 2010.

gesis
Leibniz Institute for the Social Sciences

Work in progress

- Systematic review of current CVs
  - application for different data holdings
  - missing terms
  - unclear definitions (trying translation)
  - reviewers beyond the data archive community are welcome
  - deadline: January 15, 2010
- Publication and maintenance within the DDI3 documentation

gesis
Leibniz Institute for the Social Sciences

Contact and Resources

- Review material:
  http://www.fsd.uta.fi/jemma/eddi-cv-review/
  User Name: eddi-cv-review
  Password: vC6lBmLc
- CVG Contact:
  taina.jaaskelainen@uta.fi

www.fsd.uta.fi