Title	**Data and Metadata Extensions of the CESSDA RI (D8.3)**
Work Package	WP8 Enhancement of data and metadata infrastructures for the CESSDA RI
Authors	Uwe Jensen (GESIS)
Source	WP8
Dissemination Level	PU (Public)

Summary/abstract

The report provides an overview on major outcomes in determining metadata and data model requirements majorly for complex comparative surveys by design. The recommendations focus the needs to facilitate metadata along with DDI 3 and the extension of contextual metadata.

As a consequence it is proposed to develop a DDI 3 compatible editor not yet present for the daily production needs of data providers. The modular editor allows the standardised, multilingual capture and extensive re-use of existing in-house metadata from study and question / variable level. Further functions are to support the management of new DDI facilities like persistent identifiers and versioning that provides long-term visibility of the data and its present documentation status.

Accompanied to the need of a DDI 3 compatible technical infrastructure it is recommended to coordinate data modelling for further complex data types with the development of a general data model for services in the cessda-ERIC.

The content of the report D8.1 is based on the work described at the DR 8.1.1 “Functional needs and specifications for metadata of cross-national surveys, time series and further complex data types”. This draft document is considered a specialised work paper in addition to the report D8.1 for use and potential extension in future development and implementation work in the cessda-ERIC.

Table of Contents

1 Introduction

The first core objective related to this report is to **determine the metadata and data model requirements** of the CESSDA RI in handling complex dataset types throughout the entire data life-cycle with special consideration to cross-national datasets and time series.

The second objective regards **the planning of the strategic developments required for metadata, data models and software upgrades** for data and metadata capture, management, processing, publishing and access within the CESSDA RI to support more complex dataset types. According to the division of labour between WP5 (data publishing) and WP 8 (data documentation) core needs are formulated on functions for data documentation in the next chapters.

To determine the particular requirements of metadata and data models in the cessda-ERIC **a conceptual analysis of the technical developments required to upgrade the present software** was conducted with strong focus on documentation needs for complex survey data.

Further references point to technical requirements and recommendations formulated at CESSDA-PPP work packages WP4 (Thesaurus), WP5 (Portal), WP9 (Harmonisation platform), WP11 (Grid), and WP12 (e.g. SSO, PID). The conducted analyses and specialised considerations at these work packages are highly related to the adoption and stepwise implementation of new DDI version 3.

Required tools development are to plan and realise as part of an integrated development programme to design and develop the cessda-ERIC technical infrastructure overall.

The major focus of the conceptual analysis in WP8 concerned the 'material' dedicated software has to deal with along the particular lifecycle phases (from project information up to the study ready made for public-reuse). Basic considerations on metadata needs from complex dataset types currently not or not to large extent present in the CESSDA RI portal were included in addition.

2 Core outcomes and recommendations

The major outcomes are presented and discussed initially under two central aspects as starting points: the lifecycle of studies and research data and requirements related to complex surveys data. At the end it this paper proposes to develop a modular, DDI 3 compatible, editor in line with the technical infrastructure recommendations as proposed at the WPL Coordination meeting in Essex, August 2009.

2.1 Lifecycle model and requirements for metadata and data models

With DDI version 3 the design of the documentation standards was consequently extended to enhance and support capturing management and publication of metadata along the whole lifecycle. It improves significantly the handling of complex dataset types in supporting:

- ‘Phase specific’ DDI instances the metadata production by specific user groups (data producer; data provider);
- At certain places of the (meta-)data production (research team and their local instances; integrating agency) and;
- Exchange between actors and places throughout the entire data life-cycle;

- Publication and public retrieval and re-use of data and metadata (data analysts).

To Identify key phases and outcomes in data work the draft "**DR 8.1.1 Functional needs and specifications for metadata of cross-national surveys, time series and further complex data types**" (present version 4) provides an Integrated Life cycle model with further information on (chapter 2)

- Main phase: Conception - Production - Repurposing and detailing each phase;
- Actors: Data producer - Data provider - Data analyst;
- Core objects (Project; Survey design; Study & dataset types), processes and events.

The model integrates different facets from life-cycle models presented during the last years at several conferences like IASSIST. As s 'blue-print' it supported specific work like on task 8.2 (DDI 3 evaluation; Preservation metadata) and data publishing issues in WP5 and WP12.

Recommendations:

- The provided model is recommended as starting point for specifications and detailing objects and process to expose particular metadata needs and data modelling in preparing proposals on software developments or upgrades to document complex survey data.
- The model is closely related to DDI 3 recommended as future mandatory technical documentation standard and the accompanying DDI 3 compliant technical developments.

Considerations towards further recommendations:

- Definition and use of preservation metadata requires the development of a congruent data model compliant with the OAIS reference model. Further details are provided with DR. 8.2.1 on Preservation Metadata and D8.2 (DDI evaluation and Preservation metadata).
- The congruence of the OAIS approach (SIP; AIP, DIP) with and relations to the DDI 3 model as backbone for structured metadata was discussed at IASSIST 2009.
- Overall data modelling to serve data documentation requirements are considered as part in developing an integrated data model for the whole CESSDA technical infrastructure.
- Core recommendations are formulated respectively in WP5 and WP9 concluding the Technical Specifications for a European Question Data Bank and recommendations for a web based Service oriented Architecture (SOA) as discussed in Essex, August 2009.

2.2 Metadata requirements and data modelling for complex comparative survey data

The second focus describes characteristics, structure and documentation standards and processes for comparative studies by design for cross-national studies and time series.

As major outcomes of the Conceptual Analysis of metadata requirements from complex survey data the draft report focuses on process analysis, process outcomes and particular use cases to describe complexities of different study types.

In summary the results in (chap 4-6) provide specifications for comparative surveys types on dimension space (use case ISSP), time (national trend series; no example yet included) and the combined space x time dimension (use case Eurobarometer trend). Particular metadata requirements concern the:

- Characteristics and documentation needs per dimension;
- Structural examples on relationship per instance;
- Substantial and technical documentation standards & documents used;
- Core processing events on data & metadata.

2.3 Metadata requirements and data modelling for further complex data types

Some basic metadata requirements are provided (chap 7) based on the criteria used for comparative surveys. Two examples were selected as initial use cases (British House hold panel; collection with mixed data types (cross-section data; three panel samples (with ten waves); four data files (Children, Adolescents, Adults 1991-1997).

The outcomes provide a rough overview on basic metadata needs for the BHPS and its related household and individual data. A broad analysis of the provided examples was not intended within the frame of this task. Further data types (like qualitative data, macro data) are beyond the scope of this work package.

WP10 provides recommendation on data from official statistics. Potential extensions of substantial metadata from micro and macro data are considered in section 2.5.

Recommendations related to outstanding questions:

- To specify needs in data documentation and publication for data types like Household panels requires additional investigation to determine particular requirements on data modelling and software upgrades.
- Further implications will be given considering extensions of data and metadata like to support interdisciplinary use of research results (compare e.g. INSPIRE - Infrastructure for Spatial Information in Europe)
- Respective specifications to support work with further complex data not currently present in the CESSDA RI portal are part of planning the future environment and services of the cessda-ERIC and a respective common data model.

2.4 Extensions of core metadata for complex survey data by design

Core metadata are presented as descriptive overview for comparative studies by design (chap. 8). The scope of core metadata and proposed extensions in data documentation regards the following levels considering the use of advanced DDI 3 features:

A.) Study level (beyond used metadata in present practice)

- Inclusion of project metadata
- Inclusion of citation rules along with the linking of data and literature
- Inclusion of PID and versioning (under work at WP5/WP12)
- Support for the bi-lingual study description (English / Country language) to foster data resource discovery in the cessda-ERIC (compare WP4 recommendation)
- Extended use of controlled vocabularies (compare WP4 recommendation)

B.) Question / variable level (beyond use in present practice)

- Particular requirements on PID and versioning
- Multi-lingual documentation of the question wording from country field questionnaires.

- Extended use of controlled vocabularies and ELSST based concepts (compare WP4 recommendation)

Further specified recommendations from other work packages are provided in chapter 3.

2.5 Contextual presentation of metadata and metadata for identification of comparative & comparable data

One of the major efforts in software upgrades is to allow documentation of the manifold relationships among single waves or space instances of particular study collections to support public retrieval, exploration and re-use of existing comparative data by design (chapter 9 on "Generalized options to present metadata from comparative surveys").

To support respective work in WP5 (Portal functionalities) a basic set of proposals on presentation of structural context aspects (study; questions / variables) was provided. Furthermore a set of metadata was accomplished to provide items useful to identify (in particular) comparable data.

- Browsing: Presentation of topical modules or trends from comparative study series in different structural contexts allowing for specific views on study and question/variable level;
- Search options to detect comparative and (potentially) comparable data in particular providing:
  - Basic metadata for universe (object), the concept (property) and the variable content (representation) (ISO/IEC 11179 compliance) and;
  - Extended metadata retrieval like for additional methodological aspects and options to select subsets of a study instance like for specific countries or fielding dates.

Apart from the place of presentation (on the web; at internal project tools), several options appear necessary to present a study collection under certain views on the manifold relationship. Based on the experience from data documentation projects carried out in cooperation with researcher groups it occurs sensible to link facilities to browse, search and present study and / or question - variable information to following basic contexts beyond further detailed metadata on deviations, filtering, routing, errata, embargoes etc. not explicitly presented in the following overview.

Structural context and relationships inform e.g. on:

- Cross-section study (part of a comparative study instance) and its dimensional level space, time, or time x space;
- Integration context of a comparative study instance comprising several cross-section studies (from study to question / variable level);
- Panel informs on waves context (from study to question / variable level);
- Original question texts presented for the basic language and the country specific language(s) including related show cards.

Content related context information is expected for e.g:

- Substantive topics organised in a hierarchy of groups / subgroups;
- Substantive context information (e.g. event history);
- Trends and content related grouping as well as structural context information on occurrence in countries and years.

From a methodological perspective metadata should inform e.g. on:

- Constructed variables and the documentation of the functional concepts;
- Harmonized variables and related conceptual guidelines and concepts;
- Scales should provide information on founding concepts and available test results.

2.6 Considerations to extend contextual metadata (micro and macro data)

The following considerations concern potential support for context information from multi-level-analysis approach. The Multi-level approach concerns systematic analysis of topics like social gravity, groups or school research for "individual data in context settings".

In short social science research endeavours earmark the growing demand to combine data from different levels on the micro-macro axis as shown by the three selective examples.

- **European Science Foundation – HumVIB - Cross-National and Multi-level Analysis of Human Values, Institutions and Behaviour**
  "The overarching objective of the HumVIB EUROCORES programme is the realization of the concept of Europe as a natural laboratory for the social sciences in which the diversity of institutions, practices, histories, and resources enables researchers to analyse how human values, attitudes and behaviour are affected by the characteristics of the multi-level systems or contexts in which they occur."
  Ref.: http://www.esf.org/index.php?id=3248
- **Example ESS – European Social Survey**
  To enrich the potentials in data analysis the project integrates substantial context information in addition to the fielded data:
    - First by event reporting hosted at an event database. http://www.europeansocialsurvey.org/index.php?option=com_eventlist&view=eventlist&Itemid=326
    - Secondly the development of attitudinal indicators is underway e.g. to link "life satisfaction" and "national economic indicator". http://www.europeansocialsurvey.org/index.php?option=com_content&view=article&id=83&Itemid=238
    - A third scope of context information regards the use of Socio-economic macro statistics ("The Macro Data Guide").

  Ref.: http://www.europeansocialsurvey.org/
- **Example Piredeu - Providing an Infrastructure for Research on Electoral Democracy in the European Union**
  The particular approach of this project is to combine different types of studies.
    - "… the pilot study will provide the basis for a fully-fledged study of the European Parliament elections of 2009, comprising a voters study, a candidate study, a media study, a manifestos study, and a contextual data study."

  Ref.: http://www.piredeu.eu/

Considerations to extend support for multi-level data

The report from WP10 T4, Draft v.1.1 provides some more insight to the European wide situation of 'Collections of data'. The report "review current CESSDA data collections in order to identify strengths, weakness, and areas of expertise across the various member organisations, and how they related to external data collections."

The following aspects are recommended for further conclusions to the strategic plan.

- The support of multi-levelled (quantitative and qualitative) data in the cessda-ERIC is considered in the strategic development plan;
- A dedicated workshop is initialised to explore substantive demands for specialised metadata and technical service needs to extent support for multi-level data in the European SSH research community;
- To initialise or support particular data service requirements and technical solutions it is proposed to make 'specialised data and metadata needs' a topic in strategic contacts, cooperation and meetings with social science projects and funding agencies.

3 Planning of the strategic developments required for metadata, data models and software upgrades

With respect to WPL coordination workshop in Essex August 2009 following general aspects are recommended cornerstones for further conclusions to the strategic plan and respective technical developments and implementations.

❖ Technical planning concerns the whole life-cycle for data and metadata capture, management, processing, publishing and access within the CESSDA RI. Web based.

❖ A constitutive part of a technical development plan in the cessda-ERIC concerns the step-wise implementation of technical standards in particular DDI 3.

❖ DDI 3 is recommended as mandatory target for the future technical documentation standard. This implies further recommendations:
- The development of singular tools and the distributed technical infrastructure must be DDI 3 compliant;
- It is proposed to follow the MT (Metadata Technologies) proposal for registry / repository based infrastructure (Technical Specifications for a European Question Data Bank) using a web-based Service Oriented Architecture;
- Development of a transition roadmap from DDI 2 to DDI 3;
- Close cooperation with the DDI TIC to get high-qualified support in related technical development and implementations issues within the cessda-ERIC;
- Support of DDI developments to support more complex dataset types.

❖ The OAIS (Open Archival Information System) reference model provides the base for professional data management, the related operational processes and respective guidelines. This implies further technical recommendations:
- Definition and use of preservation metadata which requires a more specific implementation strategy that aligns with the OAIS and towards enhanced 'harmonised' and 'interoperable' practises compared to present archival solutions.
- Agreed object model encompassing relevant studies and their content and a shared metadata model across the CESSDA distributed infrastructure.
- The standards PREMIS (Preservation Metadata Implementation Strategies) and METS (Metadata Encoding and Transmission Standard) are of particular technical importance.

Specified technical recommendation from further work packages (overview)

- WP4 on multi-lingual thesaurus and controlled vocabularies;

- WP5 on publishing & browse, search, access to data and metadata via the CESSDA portal;
- WP8 on DDI evaluation and preservation metadata;
- WP9 on Harmonisation platform and Question database;
- WP10 on End user license agreement and secure remote access systems;
- WP11 on the potential of grid technologies;
- WP5/WP12 on Persistent IDentifier system (PID) and versioning issues;
- Note: On complex technical data formats and new media.
- Technical data formats and media types with relevant data from the social science research (e.g. pictures, video, etc.) are not considered in this work package. However WP10 T3 and T 7 provide an extensive report on Data preservation in the Social Science and respective recommendations
  Ref.: WP10, T3 & T7, version 2.0, August 2009

The next chapter formulates first the functional core needs in metadata documentation and secondly propose to develop a related modular editor to capitalize the advanced DDI 3 features in the daily data documentation work.

3.1 Functional needs in metadata documentation

The following aspects describe in summary basic functionality required to document complex data from comparative research considering the future use of DDI 3.

1. Modular system for standardized documentation of metadata for:

- Project / Study / Dataset;
- Question / Variable documentation in a multi-lingual context;
- Provision of customized tools to support metadata documentation and exchange with research projects according to the lifecycle approach.

2. Reuse of (series specific):

- Metadata on study and question / variable level;
- Technical metadata standard (e.g. missing value definition, …);
- Substantial standards by global access to related repositories (ISCO, NUTS etc.)

3. Data manipulation functions to support basic data processing routines

4. Further functions are required to handle:

- Linkage of data & literature;
- Management of versioning and PIDs;
- Controlled vocabularies;
- Substantial context metadata.

3.2 Outline - Recommendation to develop a DDI 3 compatible Editor

To make use of the advanced DDI 3 features and to bridge the gap in missing a metadata production tool for daily data documentation work the development of a respective editor is recommended.

The first outline is provided for further practical specifications. In particular existing developments in this field and its potential for use, integration or extension are to consider in the detailed planning.

The proposal aims to develop and implement as soon as possible a fully functioning editor to work with DDI 3 in the daily production and for the dedicated purpose as mentioned below. The following modules appear considerable to meet these basic needs:

Pre-requisite
User tools and the corresponding middleware and the database layer must support the regular re-use of metadata in question / variable documentation (like question texts and concepts) for repeated surveys (Study collections), in processing trends as well as for compiling study descriptions.

1) Module Documentation on study level to capture and handle information on:

- Project and funding regulations
- Study concept and design
- Data set information
- Study citation
- Linking to documents closely to the study (questionnaire, reports etc.)
- Linking to substantial context information (event dB; Indicator systems)
- Interface to prepare in parallel the English version of the study description

Additionally the module should have a functional group to manage study collections and their single instances (Ingest control documentation, study no. and alike).

2) Functional modules of a Question / Variable editor

2.1 Module to capture of standard questions and definition of the variable structure
It should include all question elements (from reference questionnaire) and support automated (parallel) creation of data set structure based on available question typology.

a. It would be desirable to develop an interface to import question from Questionnaire developments tool
b. To import and export standards questions to CAI systems

2.2 Module to capture country specific questions (field questionnaire)
This module should provide stand question text allowing to add the questions used in field work in the respectively used country languages

2.3 Module to compare fielded metadata in relation to the metadata by design and documentation of the results

3) Extension: Data Harmonisation of singular variables or variable group
Respective specifications can be derived from 3CDB developments for integration in the standard procedures of internal work flows (see WP9 internal website).

4) Required functions to extend use of substantial and structural metadata

- Assign concepts to question / variables and support DDI 3 Grouping options;
- Assign metadata to scales of particular scientific interests;
- Assign standardised metadata to specific variable types (like Trends and it's linkage);
- Allow access and internal re-use of to central repositories with internal standard classifications.

4 References

4.1 Sources documents from the MetaDater Project

Work Package 3 – Modelling the Database

- Supplement 1: Data model for handling the repeated cross-national study
  see http://www.metadater.org/Datamodels/RCNS_datamodel_06-07-2005.zip
- Supplement 2: "A typology of questions and the related data model"
  see http://www.metadater.org/Datamodels/QVTypes_datamodel_07-07-2005.zip
- D3.2 - The Conceptual Metadata Model. Entity Relationship Schema Report see http://www.metadater.org/Datamodels/MetaDater_Conceptual-Metadata-Model_D32-edition2_public_v1.zip
- D3.3 – final version: The Relational Data Model Report (project internal document)

Work Package 4 – System Architecture and specification

- D4.3 – 1st edition: Functional Specification Report (project internal document)
  Extensive description and use cases on core processes for the MetaDater like:
  - Process 3 Data and documentation control
  - Process 4 Study extension, harmonization and variable standardization
  - Process 5 Indexing and classification
  - Process 6 Data and metadata publication and dissemination

4.2 Resources on Life-cycle models for Metadata, Data and Study

MetaNet: A network of excellence for harmonising and synthesising the development of statistical metadata:

- Work Package 1: Methodology and Tools
  Deliverable D4 "Overview of technical aids to Metadata representation"
  Chapter 2.2 "The Metadata Life Cycle"
  http://www.epros.ed.ac.uk/metanet/deliverables/D4/IST_1999_29093_D4.pdf
- Work Package 2: Harmonisation of metadata - structure and definitions
  Deliverable D5 "The Concept of Statistical Metadata"
  http://www.epros.ed.ac.uk/metanet/deliverables/D5/IST-1999-29093_D5.doc

IASSIST 2009 - Thu 28 May - Session D1
"Tag - You're it! DDI Applications and Experiences " - Presentation
"Managing the Metadata Life Cycle: The Future of DDI At GESIS and ICPSR" slide 14
Chuck Humphrey Research Life-cycle 2006
e-Science and the Life Cycle Model of Research retrieved from
datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc
IASSIST Edinburgh May 2005 - Wong / Humphrey 2005
Fitting a survey life cycle in the DDI
DDI Alliance "DDI Version 3 Conceptual Model" (draft 2004)
http ://www.icpsr.umich.edu/DDI/committee-info/Concept-Model-WD.pdf
IASSIST Communiqué - Comprehensive overview on further models
Conceptualizing the Digital Life Cycle 2006

4.3 Further Sources

ISSP use case material sources

- ISSP modules in general
- ISSP Module Role of Government I-IV and the Cumulation

Technical Specifications for a European Question Data Bank. Final Version May 2009
Arofan Gregory / Pascal Heus / Chris Nelson (Metadata Technology) and Jostein Ryssevik (Ideas2evidence)
http://www.cessda.org/project/doc/CESSDA_PPP_QDB_May09.pdf
Reference Model for an Open Archival System (OAIS)
Consultative Committee for Space Data Systems - CCSDS 650.0-P-1.1 (May 2009)
http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf