Title	**Functional and Technical Specifications of 3CDB (D9.2a)**
Work Package	WP9
Authors	Michael Bourmpos, Tolis Linardis (and others)
Date	05.08.2009
Dissemination Level	PU (Public)

Summary/abstract

Content:

Functional and technical specifications of a database for compiling and re-distributing concept, classification and conversion information (3CDB), including specifications for integrating such a database with QDB and concepts for linking the 3CDB with the general CESSDA RI infrastructure.

Contents

Figures

Tables

1 Introduction

This document aims to outline the functional and technical specifications of a Concepts, Classifications, and Conversions Database, which will be referred to in this document as CCCDB or 3CDB.

In chapter 1, the aim of CESSDA-PPP and especially of WP9 is laid out. The main definitions we come upon in this document are analysed and the theoretical background of a harmonization project is presented.

The next chapter talks about the main functional specifications of CCCDB. The contents and functionality of CCCDB are described and issues like maintenance, ownership, and security and user management are dealt with. A description of the basic procedures and requirements is also given.

Chapter 3 is about the technical specifications of CCCDB. The architecture of the database is presented. Finally the ways we can search and retrieve valid metadata are described. In the next chapter we can see how a use case harmonization project is being developed using a demo application named CHARMCATS. Finally, we present our conclusions in chapter 5.

We must mention that all material concerning the theoretical background of a harmonization project (chapter 1), the use case of the demo application (chapter 4) as well as the proposed database tables structure of a demo CCCDB (see the accompanying document "List of Tables, Charmcats, 2009, July") is product of work carried out by Markus Quandt, Martin Friedrichs and Alex Agache of GESIS.

1.1 CESSDA PPP and Work Package 9

The aim of the CESSDA-PPP is to plan the future development of the CESSDA RI and to focus on tackling and resolving a number of strategic, financial and legal issues in order to ensure that European social science and humanities (SSH) researchers have access to, and gain support for, the data resources they require to conduct research of the highest quality, irrespective of the location of either researcher or data within the European Research Area (ERA).

The project consists of several interlinked yet individually focused work packages, including work on developing the data portal to allow seamless access to data holdings across Europe, developing common authentication and access middleware tools, developing metadata standards, creating thesauri management tools, extending the coverage of the CESSDA RI, strengthening the CESSDA RI, investigating the potential of grid technologies, and improving data harmonisation tools.

The objective of work package 9 is the design and the requirements specification of two databases and related applications: the CCCDB and the Question Database (QDB). The main objective of the CCCDB is the implementation of conversion routines (see definitions below) in the most transparent, well-documented, and easy-to-use way. One, but not the only objective of the QDB is to support detecting information about the comparability of questions used in different surveys. Further goals could for example be to support the development of new questionnaires, or to help in data discovery based on question text search.

While the smallest structural element for the CCCDB is the Variable (of a usually rectangular social science data file), for the QDB it is the Question (of a measurement instrument such as a questionnaire). New Concepts, Classifications, Questions and Variables can be developed or entered when using these applications, but these structural research components can also be retrieved – independently from the study they belong to - from a central core system, such as the CESSDA Portal.

Both these databases are two new research infrastructures for CESSDA. These two new databases have a two-way role since they are “open” to the researchers outside

CESSDA to provide their own work and considerations. More concretely, CESSDA can provide to the research community additional comparative study documentation, to make completely comprehensible to the researchers how target variables are derived from source data, or how questions where translated, or how questions changed over the time and national axes. On the other hand the research community may provide to CESSDA new ‘projects’ containing the exact same type of information: such as new harmonization projects or even the development of new questionnaires. As a result, CESSDA provides an open platform through which the research community provides metadata on data harmonisation to the research community.

1.2 Definitions

HARMONIZATION WORK: The aim of the harmonization work is to produce comparable data. The CCCDB documents two harmonization procedures: the 'ex ante output harmonization' and the 'ex post harmonization'. Both these two procedures require harmonization routines to be implemented. The great difference between these two harmonization methods is that the achievement of comparability is guaranteed in the first case, since data are designed to be comparable ***within a study***, while in the second case the data comparability is not guaranteed but it is trying to be achieved ***within a research program,*** trying to combine data from different studies. In both cases, the conceptual background is ex ante agreed. In the first case, the conceptual background is agreed by the designers of the study while in the second case it is agreed by the secondary analysts carrying out the research program.

Ex ante output harmonization: comparable data by design

Lene Mejer (2003) refers:

«Ex ante output harmonization means to give a common internationally agreed definition for a variable and then leave to each single Member State to decide on its implementation. Each Member State decides what is the best national source for the variable (for example from already existing surveys and/or registers)».

According to the procedure described above, with 'ex ante output harmonization' the national members of an international study group agree in advance on the conceptual part of the measurement process, having in mind a broad international or a general standard of measurement. Then, every member measures the common concepts according to national measurements and converts the national measurements to the international or general standard, through conversion routines. The conversion routines, in fact, constitute part of the study documentation that is very helpful for the secondary analysts to understand how the internationally agreed variable has been derived from national variables. Nevertheless, conversion routines are not documented through DDI2 or even DDI3. That is why we need a new documentation model but also a new system, so as the conversion routines and all the relationships between appropriate study components they imply to be documented but also manipulated. So, in

this case “conversion routines” are essential for the documentation of a single, but internally comparative, study.

Ex post harmonization: potentially comparable data by ex post procedures

As already referred, ex post harmonization also uses conversion routines to implement the target variables within a research program by using different studies. The conceptualization of the procedures is agreed by the research group carrying out the research program and not by the study designers. So, in this case, “conversion routines” are essential for the documentation of a research program.

So, the goal of *harmonization* is to establish comparability across different data sources. Harmonization is also the name given to data-coding procedures which transform country-specific formats of *source variables* into a derived comparative variable (named *target variable*). Thus, the harmonization is here related to the broader concept of *equivalence* used in the literature which comprises all the conceptual and measurement implications for comparability. Accordingly, equivalence refers to the comparability of measures obtained in different population groups (universes). Nevertheless, the term of equivalence has different connotations in the social sciences. For example, Johnson (1998) counts in the survey related literature more than 50 different definitions of equivalence. But, there is consensus that conceptual equivalence is given by the identity of theoretical concepts across cultures and if this type of equivalence is not present, comparison is not possible at all (e.g., Vijver and Leung, 1997).

The harmonization work is based on a conceptual basis by applying definitions of existing CLASSIFICATION / SCALES/ INDEXES (C/S/I) to available data and usually results in one comparative indicator. Here the distinction between two types of variables is made: the harmonized variables named **1) TARGET VARIABLES (TV)** and **2) the SOURCE VARIABLES (SV)** used for constructing the TV. **The creation of a TV is called CONVERSION**. The harmonization work as a whole process and the documentation (publishing) of it in a system is called **HARMONIZATION PROJECT (HP).**

CONCEPTS: refer to broader theories which the C/S/I and CR aim to measure. So, to each concept a set of C/S/I and CR may correspond; different HPs could use the same concept.

CLASSIFICATION/SCALES/INDEXES: Comparative measurements of concepts that prescribe:

- How conceptual parts of one C/S/I should be measured.
- How generic source variables should be processed in order to produce the harmonized indicator.

UNIVERSES: Represents the target populations. Universes may refer to the universes of the C/S/I, the universe of a target variable or even the investigated under a study.

CONVERSION ROUTINES: The basic task of conversion routines consists in making one unique variable from a set of source variables. So, the conversion routine applies the generic prescriptions of C/S/I to existing specific source data.

HARMONIZATION PROJECT (HP): When a complete chain from concept to target variable will be documented in the system, this new published entity is called a harmonisation project.

1.3 Purpose of CCCDB

The **purpose of the CCCDB is to support the storage, access, distribution and contribution to the production of harmonized variables** developed in comparative (cross-cultural and longitudinal) social research.

The development of the new database and platform is based on the following goals:

1) To create a central database for harmonization routines (storage, distribution)

The system should enable:

- To store and publish descriptions of C/S/I;
- To store and publish harmonization (conversion) routines (CR);
- To connect harmonization routines to metadata on variables, questions and data files.

2) To supply and store documentation into the database (contribution)
- Modification of existing C/S/I's and/or related conversion routines
- Creation of new C/S/I's and/or conversion routines

3) To eventually assist in applying the harmonization routines to the data (data manipulation)

1.4 Three Layer Model of a HP

In an extensive workflow analysis of creating harmonized variables, three methodological layers or working steps were distinguished that can be labelled as the *Conceptual Step, Operational Step and Data Re-coding Step.* Before outlining these steps in an analytical way (1.4.2), a prototype workflow of harmonisation is presented in the following (1.4.1) as a basic enumeration of main conceptual and practical "tasks" to be fulfilled by the researcher in the harmonisation procedures.

1.4.1 Workflow

The following list portrays *which* tasks and decisions in a typical ex-post harmonisation project could be supported. Thus, this section does not aim yet to explain *how* the CESSDA platforms (3CDB or QDB) will support these procedures – for that, see the use case in section 4. However, some abstract terms have been used in this section that are also used to describe the attributes of the two planned databases (for example, the distinction between source and target variables; universe etc.)[1].

Scenario: A researcher wishes to analyze the relation between *education*, *class structure* and *material wealth.* A secondary comparative analysis of survey data is aimed with data for representative national samples including employed respondents across all European countries and USA, with latest available data to the year 2006.
Basic Tasks: • To choose between possible measurements for the concepts of interest; measurements: Classifications/Scales (C/S); • To locate data in the CESSDA archives corresponding to these measurements (find variables in data files/data repositories);

1 For examples on harmonisation materials used in basic workflows see D9.1

	• To transform the data and construct one variable with comparative values for the universe of interest = harmonisation (write conversion routines for re-codings in statistical programme, e.g, SPSS).
1	**Concept:** Definitions of *education, class structure, material wealth.*
2	**Universe:** All employed respondents within 25 European Countries and USA.
3	**Decide on C/S, applicable to the universe :** ISCED-97 (Education); Equivalized scales on household income (Material wealth) and Wright's classification/typology on class structure
4	**Store descriptive definitions of C/S** (lists with conceptual labels for classes and scales)
5	**Store the desired format of the Target Variable**
6	**Understand how the C/S could be derived:** make a list with required variables for every universe element (e.g., in case of ISCED-97 latest educational certificate obtained in every country).
7	**Understand the differences in meaning of these variables at the universe element level** (national and temporal differences in *Class Structure* and *Educational*, and *Tax* systems).
8	**Understand the alternative ways of measuring the required variables** (different measurement prescriptions within a country and across countries and time instances).
9	**Make a basic list with required transformations that can be applied to data** (e.g., coding rules and re-coding maps of national educational degrees into ISCED-97, transformations of household income using different weights)
10	**Data discovery process:** search variables connected to data sets and questions
11	**Organize search results:** basically by universe, data sets, and concepts (additional filters, e.g. for panel data)
12	**Inspect content of variables found:** access to question text, and questionnaire materials, variable sample frequencies, sample design;
13	**Compare the content of variables with the list of "ideal variables"** (identified at step 6-9). Variables are grouped according to the concept they suppose to measure
14	**Document deviations:** bias detected in the coding of variables, question texts, translations, sample characteristics
15	**Decide on final categories/values and format of the Target Variable**
16	**Prepare the variables for re-coding:** subset of variables rated as comparable and usable for re-codings
17	**Decide on final re-codings and write conversion syntax**
Additional tasks involved when writing the conversion routine:	
18	**Make a list with the variables in the data sets that identify/filter the samples of interest in the data** (e.g., variables and codes for countries and time periods)
19	**Make a list with variables required in the data management process** (e.g., personal identifier of respondents required when merging different data sets)
20	**Make a list with variables on sample weights necessary in the data analysis process**

21 Make a list with variables on sample weights necessary in the data analysis process

On the basis of this list, basic elements relevant for the harmonisation database listed below became evident.

A. Input elements for 3CDB

- Question and variables connected to concepts (For use cases on Questions, see also the MTG Report, May, 2009, pp.: 48- 49)
- Metadata from comparative studies by design (e.g., see Jensen , 2009: pp. 31-33)
- Identification of variables and questions measured ex-ante as part of harmonised measurement instruments within a study (e.g., item batteries/scales)
- Context information attached to variables (e.g., information on national educational systems)
- Contextual databases with indicators on aggregate levels (e.g., PPP rates)
- Documentation of the conceptual, methodological and data type bias involved in the recoding of each source code into target
- Methodological information on the measurement validity of specific source variables and questions (e.g., psychometric information from previous studies for items of a scale; results from cognitive interviews)

B. Tools for data re-coding

- Tools for comparing variables and questions (with full study information) within overview tables across universes data sets
- Tables with correspondence lists (pairs of values) of source variables into target variables
- Syntax generator for re-coding of source variables into target
- Technical documentation/guidelines for the application of conversion syntaxes to data

C. Documentation of Classification/Scales

- Lists with codes of classifications with descriptive labels
- Correspondence lists between different classifications codes or versions (e.g., ISCO68, ISCO88)

- Documentation of the theoretical basis of comparative classifications and scales
- Documentation of the coding procedures and variables required by the measurement (not connected to archived data)
- Documentation of previous applications in harmonisation

D. Controlled Vocabularies

- For types of similarity/equivalence of variables and questions
- For describing the conceptual and measurement elements of classifications and scales
- For concepts

E. Data citations

- Clear rules for data citations of conversion routines, versioning and authorship

1.4.2 The three working steps of harmonisation

The basic requirements outlined above centred on: (1) tasks of finding the data to be harmonised and (2) the documentation of the decisions a researcher took while designing/writing the conversion routine. Another basic requirement was to find a common framework for documenting different types of measurements (Classification, Scales and Indexes).

As the list of workflow showed, the main decisions a researcher takes are:

1. On a concept (even if vague formulated).

2. A classification (or scale) to be harmonised that is *assumed* to be **universal.**

3. When applying a classification to **different contexts**, there might be differences in the way the classification can be measured, operationalized (e.g, change of educational certificates over the time across countries).

4. When actually coding the values of, say, a classification, to a target variable with actual **survey data from different sources**, it is seldom the case that the desired information is to be found in exactly the same technical definition across sources; so this is another source of biases to be documented.

Considering these four points, a first conclusion of the analysis was that the full documentation of conversion routines spans over the three analytical steps mentioned at the beginning of this chapter: 1. Conceptual; 2. Operational and 3. Data Coding step.

The three steps are interlinked with a set of structural elements: Concepts (one per HP), dimensions (one or more per concept), indicators (one or more per dimension, in the operational step), and variables (one or more per dimension, in the data coding step). Some of these remain stable over all steps; others appear only in selected steps.

1. **Dimensions**: are used to understand, specify, and reduce the broad meaning of the single **concept.** Often, a concept comprises only a single dimension and the distinction of concept vs. dimension can be omitted. Dimensions are theoretical abstractions (they can never be actually measured). They have to have the same meaning or functions across all elements of the universe in depicting the concept of interest. The harmonization methods used in the subsequent steps should serve to guarantee that the identity of meaning is not lost in the process of measurement and coding. For example, in the ESeC Classification (Rose and Harrison 2007) two dimensions (among others) used in defining different forms of regulating employment situations are the distinction between *Monitoring problems* and *Human asset specifity* required in different job positions*;* it is assumed that all the skills and tasks labelled by these two dimensions have the same significance across all European country specific market situations.

2. **Indicators**: are measures of the dimensions that are required. Each dimension can be measured by one or multiple indicators. For every indicator the information given in the **operational step** defines, for example, how the indicator should be ideally expressed by the wording and categories of a question. Indicators may be specific for individual universe elements; e.g., in multi-national surveys using national languages for survey questions, country specific questionnaires trivially introduce specifity of the indicators per country. Beyond language differences, there may further be country specific wordings, or references to national institutions, etc.

3. **Variables**: variables are how indicators are represented in actual data sets, after data collection has been accomplished. In the **data coding step**, variables move into the place that indicators occupied in the operational step. Therefore, they carry on any possible specificity of indicators to the data coding step. Beyond that, they convey any

additional specificity that may have occurred in the data collection process for each element of the universe or data source.

The distinction between theoretical concepts, dimensions and their indicators, and between indicators and concrete variables has similar importance for different types of variables, either considering the so called "background", psychological or behavioural variables (gender, ethnicity or age could be examples of multidimensional/multi-indicator measures as well as political trust or cultural values).

The three layers or steps can be depicted like this (Figure 1, below):

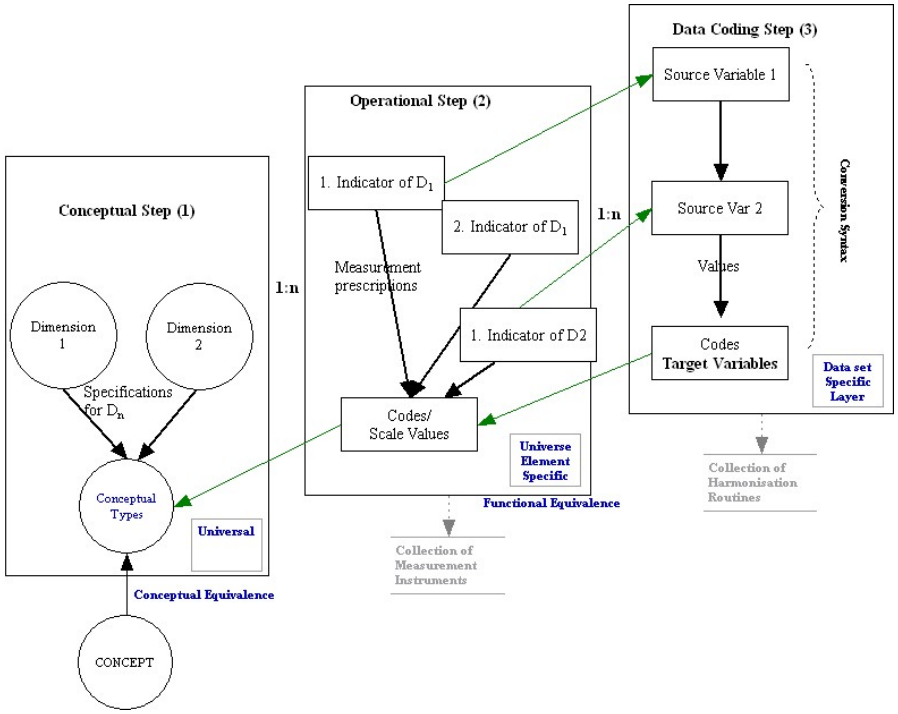

Figure 1: The three working steps or "layers" model of CHARMCATS

1) Conceptual Step: The first Layer incorporates the term "Conceptual" in its name, because it links a broader Concept and the theoretical basis involved in the construction of a harmonized variable. The concept and the dimension(s) cover the same meaning across all universe elements investigated. Consequently, this layer is "universal". The function of dimensions is to specify the broader concept in a particular way. Thus, the term "Conceptual Derivation Layer" denotes not that the concept itself is developed here in theoretical terms; rather the theoretical basis of the measurement, as condensed e.g. in a given classification, is stored. A list of "concept" terms and dimensions in form of a "controlled vocabulary", most likely ELSST or a subset

thereof, should be available in the application for a broad range of subjects and must be usable across different harmonization projects.

2) Operational Step: here, indicators of the dimensions at the level of universe elements are documented. Question Text, scaling of Items, re-coding tables used in classification (including coding instructions), etc. The combination of the properties of these indicators (values) results in the final operationalizations of the concept, without yet involving archived data. So, the materials entered on this layer supply the end users with 'ideal harmonized' definitions of the variables to be used by the conversion routine.

Seen across HPs, the products of this layer could be also understood as a collection of comparative measurement instruments.

3) Data Coding Step: here the data provided by CESSDA archives, through e.g. NESSTAR servers, are accessed and processed by the users, according to the operationalization procedure that was adopted on the previous layer. The source variables are explored and selected, and required re-codings on data are defined and documented.

The complete process implied in harmonization can be schematically represented by linking elements within and across Layers. Across Layers, all components have basically the same structure:

1. Broad content (Denominations of Dimensions, Indicators, Variables) and
2. Specification of their content (detailed definitions of dimensions, indicators, categories of Indicators and C/S/I; categories of source and target variables).

As the figure shows, these three components are structured sequentially, because the documentation of complete harmonization project typically starts with choosing a concept and ends with the construction of the harmonized variable. This does not mean that this order will be strictly imposed in the process of editing, because switching between "steps" might be necessary in the data discovery process.

To conclude, there are three types of published HPs, based on the completeness in the data model of CHARMCATS (sometimes, there might even be different user contribution to different steps of the same HP):

1. *Project of Harmonisation* This is the entirely complete status where all steps haven been worked out, which results in a harmonized variable connected with all elements of the three layers);
2. *Conceptual and operational Step*. These are projects completed for the measurement model, but not yet connected to archived data.
3. *Conceptual step only*. It may sometimes be worthwhile to just define the conceptual derivation of a C/S/I, with all its ideal codes. This may serve to just discussing the classification or scale or index on its own, or may be meant to be the starting point for later actual harmonisation work.

Further, it may happen that single components of a project such as sets or related questions or collections of variables are collected. By definition and construction of CHARMCATS, however, this cannot happen without having at least the conceptual step completed, too. If both the conceptual step and (some of) the variable selection in the coding step have been completed, CHARMCATS may be used to build a harmonisation rule in an explorative manner, driven by the availability of data. The ex post-derived harmonisation rule then takes the place of the operationalisation in the second step.

2 Functional specifications

2.1 Contents of CCCDB

CCCDB will actually be part of the wider CESSDA information infrastructure supporting the production and use of survey data. It will only contain all those required metadata needed for the creation of a Harmonization Project. Therefore, complete metadata records associated with the surveys are not stored within this database. However, CCCDB will be able to provide links to further sets of metadata and to data held in other CESSDA repositories as well.

We must point out that the CCCDB is proposed to begin as an empty database. However, as Harmonization Projects are created, the database will cumulatively fill with all the metadata used in the creation of HPs. These metadata will be stored in the projected database, within the tables belonging to the following groups of tables (see the List of tables in the *Database Model of Charmcats*):

- STUDY
- QUESTION
- VARIABLE
- CONCEPT
- UNIVERSE
- TEXT STORING (KEYWORDS)

Import of data might be required for certain tables of the proposed database. For example a predefined set of Concepts in terms of controlled vocabulary would be extremely helpful. Other metadata required include universes, concepts and keywords ingested by the ELSST thesaurus, questions coming from QDB and other metadata and data within the CESSDA Portal.

2.2 Functionality of CCCDB

CCCDB will have the following different areas of use:

- **To support users in the creation of new harmonization routines**: This could be perceived as the primary functionality of CCCDB. The users of the software application will be able to document and store within the database new harmonization projects and link them to existing concepts, variables, classifications, scales and actual surveys or datasets. Furthermore the users could actually apply these harmonization routines on desired datasets, as mentioned above.
- **To support ex-ante output and ex-post harmonization:** The documentation of ex-ante output harmonized data (comparable by design) and the production of ex-post harmonized data (which are potentially comparable, not comparable by design) is the primary goal of the CCCDB application. Once located, the metadata will be stored locally within CCCDB. The software application will then be employed to create, document and store within CCCDB the harmonization routines (including all the interconnected layers used in the process of building the HP).

- **To support data producers in their efforts to design new cross-national surveys**: With the use of CCCDB a data producer (an individual researcher as well as a regular data producing agency), will be able to search for C/S/I's that could be used to measure the concepts they are interested in. This could be done in order to locate well-proven and quality assured measurement instruments, or in order to eventually produce data that are comparable to already existing data.
- **To support users of data in their efforts to assess/understand a dataset:** When analyzing a dataset, a researcher will often need access to equivalent or harmonised measurements of a concept or variable. CCCDB could provide information on this harmonised variable and also help the user to apply the desired harmonization routine to the desired dataset.

2.3 Login, User Roles and Permissions

The first step a user of the CCCDB application will take is to login to the application. A typical authentication procedure will be carried out through a screen in the application.

There are two ways we could utilize the actual authentication procedure.

1) The first would be to **locally store all data** relevant to the user, including the authentication data, inside our database.
2) The second approach is to authenticate the user through a **central authentication system** responsible for authenticating all CESSDA users into all CESSDA applications and databases (the **single sign-on procedure as proposed by** WP5). Our database would still have to keep certain information on the users (while being synchronized to the central authentication system database) and, of course, hold all the roles granted to them and the permissions they have within CCCDB.

The potential users of CCCDB and its software application can be grouped in the following basic categories, identified by the roles they are granted:

1) **Contributing users**: These users are probably the most crucial for the operation, expansion and integrity of CCCDB. They will be able to search

and retrieve harmonization projects or other metadata required for the creation of a HP. They can also create and document a new HP and apply it to a desired dataset. Based on the above these users have extensive demands on metadata access and also permission to insert and modify data in the database. The contributing user role could be assigned to CESSDA affiliated and to external researchers (in the long term, the bulk of contributions is to be expected from/targeted to individual researchers outside the CESSDA network).

2) **Search users**: only have search and retrieve HP permissions. This role could also be assigned to CESSDA or outside researchers.
3) **Administrators:** This group of users will administer the database. They will be able to create, modify and delete users of CCCDB, as well as administer the complete database. They will be responsible for assigning user roles to each user. Based on the above, the administrator group should be comprised only by staff of CESSDA members.

For the group of contributing users and the managing and maintaining group of the DB (see further chapter 2.9 in this document) more detailed types and roles can be identified, but this would have no major technical implications. This was strongly suggested also by the T9.3 Report (2008, chapter 4), and we endorse it.

2.4 Search and Retrieve Metadata

After logging in to the application, the user will have the option to search and retrieve different kinds of metadata. These different types of retrieved metadata and their locations (sources) over which the search will be performed will be discussed in the following two sections.

2.4.1 Types of retrieved metadata

The two main types of metadata that a researcher will be searching for are:

a) Study components that could be used in a Harmonization Project e.g. C/S/I's, Studies, Questions, Universes etc (see section 1.4.1). For these

search criteria, free text as well as keyword supported (if the sources are indexed with such) search could be used.

b) Harmonization Project components that can be found within already created HPs e.g. Concepts (Conceptual Layer), Dimensions (Conceptual Layer), Indicators (Operational Layer), Variables (Data Coding Layer) etc.

2.4.2 Sources

The first thing we must define when we search for relevant metadata in order to construct a harmonization project is the location of our search. Within CESSDA the only currently available resource is the **CESSDA Portal** where studies are documented using the DDI2 prototype. So study documentation metadata can be retrieved either by the CESSDA Portal or (in case there is no documentation of the study the user wants to deal with) directly by a user of the system. The user has to document the study s/he wants to deal with in DDI2/3 xml format since CHARMCATS is DDI3 compliant. Then CHARMCATS has to import this documentation.

Besides the CESSDA portal, several **ongoing projects** will lead in the near future in the development of **other resources** that could also be of use. One of them is the Question Database. **QDB** could be conceived as a large bank of all available questions and questionnaires currently residing in the different databases of the archives. A search on existing questions and questionnaires would be of great benefit when building a harmonization project.

From our point of view, we should also be prepared to be able to retrieve data from any other types of database (repositories of source metadata), regardless of its structure. It is possible to deal separately (communication issue, types of data acquired issues) with each and every one of these databases from within the search application. However, a better approach would be to set some standards regarding **heterogeneous data sources** in general. Thus, we should set some communication standards (web services or direct data access) and standards on types of data exported from these data sources in order for the final search application to be able to use them.

As mentioned before, a well known thesaurus that could serve the CCCDB application as a metadata source is the **ELSST thesaurus**. The ideal route of access to that would be via web services, and CCCDB should also provide a complementary web service to send a search request and receive the returned metadata.

In all the above sources of metadata, we are searching for study components to help us build a HP. The basic source for searching and retrieving Harmonization Project components is the **CCCDB** itself. CCCDB is responsible for keeping all parts of an HP documentation. Therefore searching, retrieving and reusing these documentation components (e.g. Conceptual – Operational – Data Coding Layers, Dimensions, Indicators etc.) is straightforward and easy to achieve.

2.5 Insert, Update, Delete HP components

During the process of building a Harmonisation Project, the user will create systematically the three layers described in chapter 1.4. Starting with the Conceptual step, moving on to the Operational step and finally building the Data coding step, the researcher will effectively have documented a complete harmonization project.

While building these three layers, the user will be assisted by the search utility of the CCCDB application to discover and use various study components (e.g. universes, concepts, questions etc). Furthermore, the user will be able to reuse harmonization project components, created in previous published HPs (e.g. Dimensions, Indicators, complete HP layers, etc). However, there will be many situations, when the user will have to create a harmonization project layer or component by himself. For these cases, the CCCDB application should assist him to insert, update and delete HP components.

For instance, let us assume that a user decides to build a complete conceptual layer. After discovering several study and HP components that help him to create a part of the layer she/he comes to the point where a new HP component (e.g. an Indicator, which is an ideal question text) has to be created. The CCCDB application should allow the user to create, document and save the new HP component (Indicator) to the database. Moreover, the application should allow the creator – the owner of this HP

component -, to edit or delete this new HP component. However, the full HP must be frozen and locked against any further edits once that it has been published, because it is intended to be citable in e.g. published papers, or in other HPs, and therefore must remain stable after its first publication.

Finally, the application should allow the user to insert into the database the complete layer. If the layer is not finalized and not used in any other HP then the user will be allowed to update and delete it.

2.6 Build the harmonization routine

Using the above procedures, the end user of the CCCDB application is lead to the creation of all three layers that will help him document his harmonization project. After completing the conceptual, operational and data coding layer, the researcher comes to the point of building the actual harmonization routine.

To actually develop a harmonisation routine that is tailored to the specific data sets relevant to her/his harmonisation project, the user must search and discover the suitable source data sets. The CCCDB application should provide a search environment to assist the user in this quest. The locations of this search currently would be the NESSTAR Servers accessed through the CESSDA portal. However, the application should be easily modified to include more data sources of different forms (non-NESSTAR server local archives or a central data bank, e.g. a QDB put on top of the current CESSDA infrastructure).

After selecting the appropriate variables from the data sets found above, the user has to write the harmonization routine. A first approach is to write the routine using a language introduced by 3CDB. This approach has the advantage of being independent of any statistical package (except data input at the beginning of the process) as well as that a user may finish the whole HP being on line without having to change between different environments. On the other hand this requires a lot of effort since a new language has to be implemented and data have to be manipulated through commands. Additionally, the routine can be applied to data only through 3CDB environment.

A second approach is that CCCDB should allow the insertion in the database of routines written in different statistical packages such as SPSS, SAS, STATISTICA, NSDStat (statistical packages with enough overall popularity and good availability within CESSDA). Therefore a conversion routine may have n syntaxes. This approach is easily implemented if the user has the data locally, but is very hard to be implemented when the harmonization procedure takes place “on line” through CHARMCATS environment (see 2.7 Apply the routine).

Ideally, the complete procedure could be performed on line since metadata can easily be captured and checked this way, avoiding errors or modifications. The application itself can also transform the data-coding layer to the desired language or syntax giving a first version of the routine. Then the user can use an editor provided by the CCCDB application in order to to make minor changes in the harmonization routine, or even to create a harmonization routine from scratch, using 3CDB language or any other statistical package syntax.

2.7 Apply the Harmonization routine

Before applying the harmonization routine the user has to ask for the data. 3CDB, should help the user to send a coordinated request to all archives the data of interest belong to, explaining the reasons why the HP takes place. For example, a web form might be included, through which the user could submit the request for downloading the dataset and then a coordinated response comes from the relevant data archives. The coordinated response could be an email to the user creating the HP, simply stating if permission is granted or not, defining the way of data access.

Applying the harmonization routine implies that there should be an appropriate mechanism in order the language or syntax to run. There can be three different approaches on how a harmonization routine could be actually applied to data, based on the way of data availability, the degree of complexity we desire to include in the CCCDB application and the modifications required on the side of the CESSDA Portal. The data may be available in three ways: a) available for download, b) available

only on line through the 3CDB, and c) indirect data availability by applying the harmonization routine by another person or agency or automated procedure. More analytically:

a) If the user has the source data locally, after **downloading,** then s/he has to apply the routine locally, using a statistical package like SPSS, STATISTICA,etc.. At least in a world without a close technical integration of the CESSDA institutes, this approach may pose problems related the process of downloading the data sets. Matters of ownership and permissions for downloading lie within each local archive.

b) The second approach (the most difficult to be implemented but also the most practical) would be to **have the CCCDB perform the routine** on the data set(s). As a prerequisite, we must assume that the CCCDB will be granted access to all local data archives. The end user selects the desired source data set and CCCDB takes care of the rest. 3CDB takes the data set (or the required subset thereof) into its system, performs the routine and returns only the outcome of the routine, the Target Variable. It is clear that the end user has in this way no direct access to the source data set. Yet, there are other matters that have to be addressed if this approach is adopted. The CCCDB must either be able to run the routine using a new established syntax language created ad hoc, or it must be able to communicate at least with one, if not with every, statistical package (see above on the generation of programme specific syntaxes). The statistical package(s) would then be called to perform the actual data manipulation. The manipulation of data of different statistical packages requires a lot of effort. Nevertheless, Nesstar Server is currently doing that by applying “on line” several statistical procedures (compute, regression, correlation, etc) to different statistical formats. Possible grant to this technology would be desirable.

c) The third approach is to **use the CESSDA Portal as coordinator for the harmonization process**. This approach implies indirect data availability of the end user, since an agency, a person or an automated procedure (found at the top of all Nesstar Servers, that is CESSDA Portal) is that one that applies the harmonization routine to data found in Nesstar servers and returns the TV The CCCDB application uploads the routine to CESSDA Portal Then the CESSDA Portal will be responsible for executing the routine, returning the target variable to 3CDB.

2.8 Store the TV within CCCDB: Link to the source data

After the harmonization routine has been run on the source data set, the user receives the target variable data set. One option for that is that the target variable data set is stored within the CCCDB, having a link to the source data set attached to it. This way the researcher has all the information needed to effectively use the harmonization process results.

Besides having the target variable data being stored within CCCDB, the user should be also able to download them locally to his computer. This way he/she has the opportunity to further analyze these data using the preferred statistical package.

2.9 Maintenance, Ownership and User Management

One of the problems that must be addressed in the creation of applications that serve an entire community is the handling of maintenance and ownership issues around the shared content. CESSDA will be responsible to house, maintain and operate the database.

There are several fundamental requirements that have to be met:

- The contents must be maintained, including corrections, changes.
- The ownership of each metadata resource must be clear, so that questions around its use and provenance can be answered.
- Attribution may be required in cases where intellectual metadata resources are reused.
- New content has to be approved in some way before insertion in the database. This problem is solved by simply granting permissions of insertion to the database to certain people. This group of ‘certified’ researchers will be granted the contributing user role. For details, see the report of T9.3 (Krejci, Orten, Quandt, 2008).

3 Technical specifications

3.1 CCCDB architecture

The following figure depicts an overview of the actual proposed architecture of CCCDB.

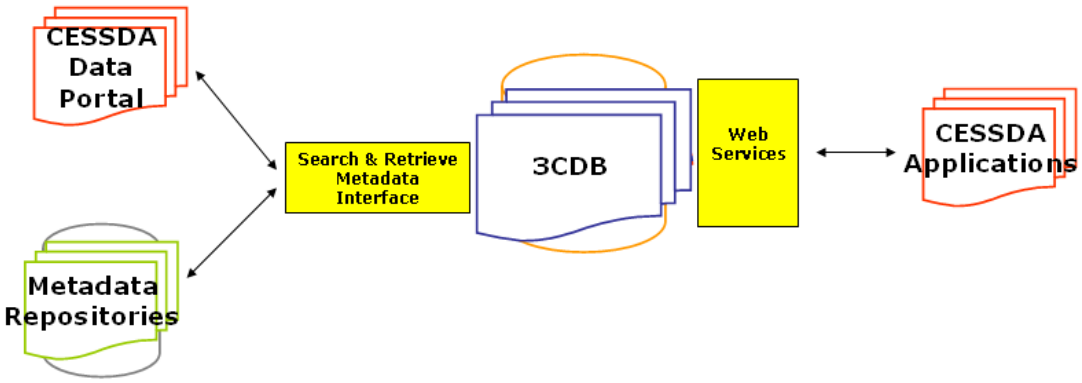

Figure 2: CCCDB architecture

As mentioned in the previous chapter the **contents** of the database can be summarised as:

- Metadata study components retrieved from other metadata repositories when a search is performed.
- Harmonization project components which have been produced by users of the application when building a harmonization project and a harmonization routine.
- Links to actual data files or other not locally required metadata.
- Target variable data files.

In terms of **interfaces,** we can distinguish between the following:

- Local interfaces that support the creation of local objects. These interfaces will help the user create the three layers of a harmonization project, as well as to create certain HP components e.g. Classifications, Indicators etc. and store them within CCCDB.
- Local search interfaces for data stored within CCCDB. This interface will perform the search within CCCDB. Simple SQL queries, varying only depending

on the choice of the database used, would be sufficient for this interface to produce valid results.

- A proxy of other repository interfaces (maybe a registry interface as proposed in the Tender Report for QDB, 2009), so that the search and retrieve valid metadata procedure is completely transparent to the end user. A web services approach is proposed for this interface. It should be able to initiate communication with the metadata repositories web service, regardless if this would be a central registry (as proposed by the MTG), the CESSDA Portal, local NESSTAR servers, or even heterogeneous data archives with known web services enabled. Furthermore, it should be flexible enough to be able to introduce in the search new sources of metadata when they become available in case these repositories provide the necessary corresponding web service.
- Admin and security interfaces. These interfaces should handle the sign in requests of CESSDA users, as well as their requests for metadata components and data sets in respect with the roles they have been granted and therefore with their permissions on the available information. They should be able to communicate with a central authentication server and coordinate the user's authentication if the single sign-on on all CESSDA systems approach is adopted.
- Web services interfaces so that CCCDB can communicate and deliver metadata to other applications or databases. Finally, the CCCDB application should be able to feed information to all other CESSDA applications and databases, when requested. This could be utilized by a web service that receives the requests for information, checks the permissions on the requested data and provide a suitable response to the requesting service.

Proposed software and technologies:

In the Tender Report of WP9 on QDB (Gregory et al., 2009, pp.:70-80) a list of software tools and technologies that could be used in the implementation phase were provided. These products could be also considered for the realization of 3CDB. Because the implementation phase of 3CDB is targeted to begin for about two years from now, and in this time further developments of these technologies can be expected, an ex-

haustive list with software and recommendations is not provided here. Three additional potential software resources that would be relevant for 3CDB per current knowledge can however be named: JBossAS5, JBoss SSO und JGraph (see short description in the Appendix 6.1.). The JGraph API allows the visualizing and interacting of graphs and it was used in the development of the prototype application (CHARMCATS). A more detailed description of JGraph will be included in the technical documentation of the prototype.

3.2 Search and Retrieve Data within an HP application

We can see core system architecture as proposed by the MTG (Gregory et al., 2009), in the following schema:

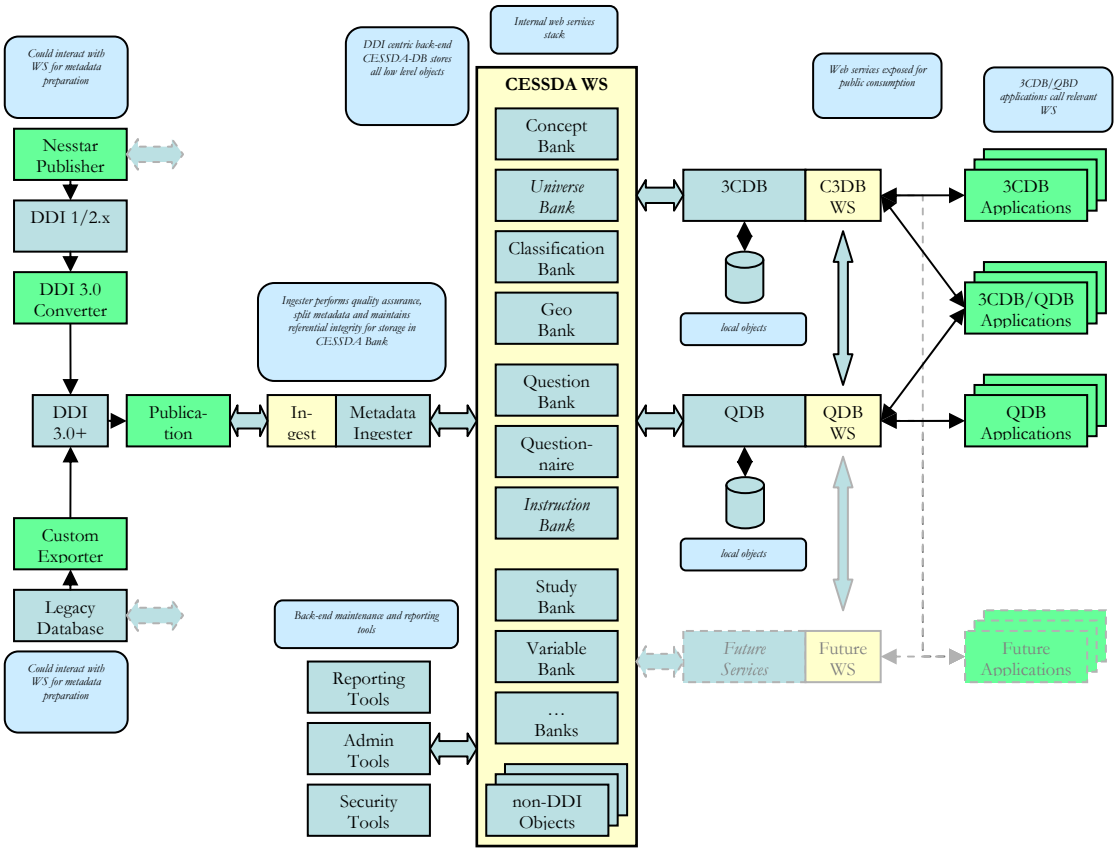

Figure 3: System architecture as proposed by MTG (May, 2009)

Each local repository (Nesstar or other) has to create valid DDI3 components and register them using the Registration Service.

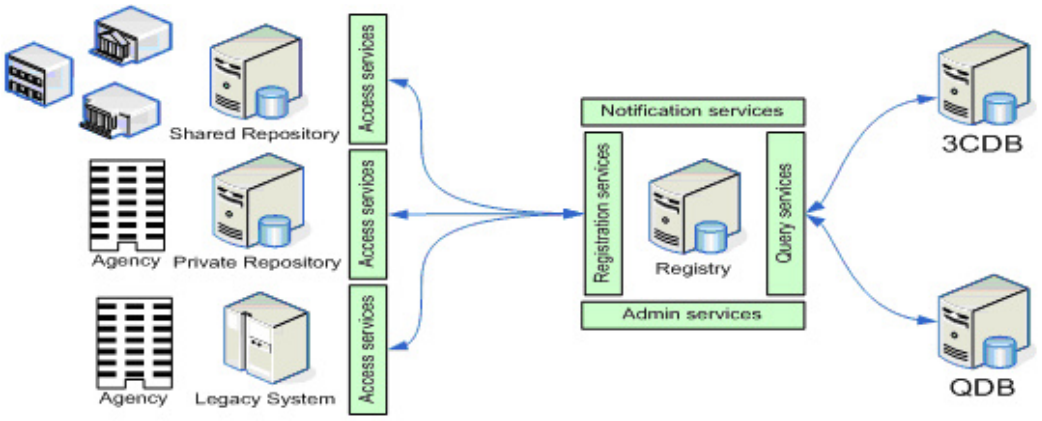

Figure 4: Registration Services 3CDB and QDB

In order for the registry to supply information, the resources must be registered. Applications can then query the registry to find the resource required, and to discover which resources reference other resources.

In order to support applications that require access to components, a set of **core services** is envisaged. These services will perform the following functions:

- Generate queries for the Registry
- Generate queries for queryable sources
- Retrieve components
- Resolve referenced components if required (e.g. obtain referenced components)
    - Embedded in the same source/repository
    - Referenced as external
    - Not referenced as external - this may involve further registry queries to locate the component
- Validate the components returned
- Resolve any references in these returned components (this can be iterative)
- Transformations between versions and standards

Using the above core services metadata could be retrieved for use when building a Harmonization Project.

However, if the MTG architecture is not adopted then another approach must be followed in order to search and retrieve valid metadata. A **web service on the existing CESSDA Portal** could be the solution for feeding metadata to all applications in need, as is the CCCDB application for building Harmonization Projects.

Another approach would be the **direct access to NESSTAR servers** within the CESSDA RI, but then several problems of ownership and security might arise. Therefore, this approach is not recommended.

Whatever the decision for the final architecture of the CESSDA RI, the requirement for searching and retrieving metadata can be fulfilled. In terms of the actual CCCDB application, it only has to talk to the responsible web service, depending on the adopted architecture (core registry services or CESSDA Portal web services).

The search engine tool that will help the user perform his searches should support a mixed search using controlled vocabulary keywords and free text keywords. If the user enters controlled vocabulary term used in ELSST then a multilingual search is feasible. If a free text keyword is used, the application searches as free text in every metadata source available. Keywords may used for study level, question level, variable level etc.

For those sources that provide metadata in the form of DDI2/3 xml files a transformation tool is required. This application – tool will parse the xml tags, retrieve the actual metadata and store them in the correct relevant tables in our database.

3.2.1 DDI2/3 elements to be used

DDI2 is the present format of most studies within CESSDA member archives. However, the recent release of version 3.0 has broadened the abilities of DDI to document the entire survey life cycle and maximize metadata reusability (DDI Alliance, 2008). Most modules of DDI3 will be required in CCCDB, as summarised below:

1. Basic packing modules:
   - Instance
   - Group
   - Study Unit

2. Scheme-based modules:
   - Data Collection
   - Conceptual Components
   - Logical Product
   - Physical Data Product
   - Archive
3. Non-scheme based modules:
   - Physical Instance
   - Comparative
   - DDI Profile
4. Sub- modules:
   - Dataset

5. Shared content:
   - Reusable
   - Dcelements

The corresponding xml schemes for the above modules can be found within the DDI3 documentation (other modules that are standard were not listed here). The prototype database of 3CDB was designed to be compliant with DDI and therefore the metadata on variables and questions should be provided in DDI3 format. Table 2 in Appendix 6.2 lists all recommended DDI3 elements that should be provided by the CESSDA repositories.

For data sources that provide valid DDI2 documentations, as are the NESSTAR servers, we must point the main section and tags that must be searched and retrieved from the xml files. We encounter five sections and the corresponding tags, as follows (example from Martinez, 2008):

1) Document description

- <titl> - Document title
- <subtitl> - Document subtitle
- <biblcit> - Bibliographic citation

and more …

2) Study description

<subject>
<keyword source="archive**">**Common Market**</keyword>**
<keyword source="archive**">**European Community**</keyword>**
<keyword source="archive**">**Europe**</keyword>**

…
<topcClas vocab="ICPSR Subject classifications" Source="archive">
3. Attitudes Toward Regional Integration
</topcClas>
</subject>

<abstract> EURO-BAROMETER 10 WAS CONDUCTED BY JACQUES-RENE RABIER, SPECIAL ADVISER TO THE COMMISSION OF THE EUROPEAN COMMUNITIES, AND BY RONALD INGLEHART OF THE …
</abstract>

<sumDscr>
<collDate date="1978-10" event="start">October 1978</collDate>
<collDate date="1978-11" event="end">November 1978</collDate>
<nation abbr="FRA">France</nation>
<nation abbr="BEL">Belgium</nation>
…
<geogCover>nine countries forming the European Community in 1978: France,
Germany, Great Britain, Italy, the Netherlands, Belgium, Denmark...
</geogCover>
<geogUnit>country</geogUnit>
<anlyUnit>individuals</anlyUnit>
<universe clusion="I" level="study">the population, aged fifteen years or older, of
nine nations members of the European Community: France, Germany..
</universe>
<dataKind>survey data</dataKind>
</sumDscr>

and more …

3) Data files description

- <filename> - Data file name
- <dimnsns> - Dimensions
- <software> - Software used

and more …

4) Variable description

- <var> - Variable
- <catgry> - Category

- <labl> - Label
- <qstn> - Question
- <valrng> - Value range

and more …

5) Other study related material

- <relStdy> - Related study
- <relPubl> - Related publication

and more …

The complete list of tags for DDI2 can be found at the DDI Alliance website.

3.3 Compliance with CESSDA RI

A minimum standard should be set for the data sources that CCCDB could interact with. DDI3 compliance might be a lot to ask for, keeping in mind that almost all of the published archived studies have been documented with DDI2. Furthermore, the CESSDA Portal and all the NESSTAR servers are only DDI2 compliant for the time being. Thus, DDI2 compliance seems to be a reasonable minimum standard for these data sources in the short time perspective. At least the DDI2 elements corresponding to the list of DDI3 elements provided in Table 2 (Appendix 6.2) should be made available.

In addition, we must ensure that all interacting data sources comply with the technical requirements of the 3CDB application that supports the creation of a Harmonization Project. Such technical needs are:

- **Communication.** Typical network communication protocols as TCP/IP are proposed for the interaction of all CESSDA databases and applications. The web services approach is the one recommended.
- **Metadata access.** Permission to access the necessary metadata that reside on local archives should be granted to CCCDB.

- **Source Data set access or request for manipulation.** The first approach is to directly access the source data sets residing in local archives, so the relevant permissions should be granted for CCCDB. A second approach is manipulation of these data sets by either the local archives themselves, or by a third trusted party e.g. the CESSDA Portal. In this case communication and upload of the manipulation code (harmonization routine) is required.
- **Verification and authenticity of data.** Both metadata and data should be verified for their completeness and authenticity.
- **Persistence.** Both data and metadata should be accompanied by a unique persistent identifier. A standardized use of PIDs would be the solution. The use of URN is supported by DDI3 and seems like a good candidate for PID. However the final decision on these matters will be a result of collaboration with other WPs as well as with other Institutes. The final decision on the adoption of PIDs should be taken into account so that the CESSDA Portal and the rest data sources interacting with CCCDB comply with it.
- **Versioning.** As with any documented study, versioning of harmonization projects should be supported by both the database and the CCCDB application. The degree of change of an HP that would justify a new version of the HP should be included in a general approach on versioning rules that should be established by CESSDA.
- **User authentication and security.** The approach of a single sign-on for all CESSDA applications and databases if adopted and implemented should be included as a compliance requirement for CCCDB.

All the above matters should be addressed and answered by CESSDA in a uniform way, so that each database and every application could act accordingly.

4 Use case and screenshots of a demo application (CHARMCATS)

In the present chapter following use case scenario is presented: the progress of a harmonization project on Wright's *typology of social structure* (e.g., Wright, 2005; Wright and Cho, 1992) on the basis of ESS round 3 and ISSP 2005 data. This is a

typical example for a multidimensional classification (see 4.1 below). The described workflow within CHARMCATS could be also applied for other classifications as well (i.e., the different variants of EGP, ESeC, ISCED-97) or for Scales (the case of scales will be not discussed here). The conversion syntaxes referred here were adapted after Leiulfsrud et al. (2001). The presented workflow (4.2 below) shows only the main functionalities and features of the interface and points to the main elements of CHARMCATS that are under development.

4.1 Wright's class structure typology

The "typology of class structure" distinguishes between 12 Classes displayed in Table 1 below. Three main dimensions can describe this classification: the *property* dimension, *the authority* dimension, and the *expertise* dimension. The property dimension is further differentiated into three types of owner based on the dimension of being *employer/property of labor*. These dimensions were heuristically combined to form a "basic class structure typology" that is divided into two parts, a three celled typology for "owners", and nine- celled typology for "employees".

Table 1: Wright's class structure typology

Classes	
1.	Employers/capitalist: Self-employed, with authority and expertise
2.	Small Employers: Self-employed, with lesser degree of labour property than employers
3.	Petty bourgeoisie: Self- employed, without being properties of labour of other employees
4.	Expert managers: Employed workers, with high authority, and expertise level
5.	Skilled managers: Employed workers, with high authority and skilled level
6.	Non-skilled managers: Employed workers, with decision power but with low skills
7.	Expert supervisors: Employed workers with decision power, supervising tasks and experts
8.	Skilled supervisors: Employed workers with decision power, supervising tasks and skilled
9.	Non-Skilled supervisors: Employed workers with decision power, supervising tasks and non- skilled
10.	Experts: Employed workers without authority, experts
11.	Skilled workers: Employed workers without authority, skilled
12.	Non-skilled workers: Employed workers without authority, non-skilled

4.2 CHARMCATS interface: The use of decision graphs

The main feature of the interface is the use layer specific decision diagrams. Following assumptions were made in designing the diagram: In case of multidimensional classifications, the process of reduction and subtraction of different dimensions into classes could be represented by diagram that was developed in analogy with decision trees (not to be confound with decision tree/graphs analysis employed as a predictive technique, as presented e.g. in Tan, 2006). This could clearly enlighten the derivation

of the final types (first conceptually and then operational) that are measured by the harmonized variable and support the recoding process because: (1) one hast not to order the source variables in a multidimensional table which is unpractical when more than 2 dimensions are involved; and (2) has a clearer structure of the order of dimensions to be considered for classifying indicators or variables.

As shown in Chapter 1.4 the basic components of multidimensional typologies to be represented within the graph are:

1. *Dimensions*: in case of typologies this are abstract definitions which are usually very close to the operational definition given by the indicator; they group the properties of an indicator or variable.

1.1. Dimensions may be devised according to theoretical considerations or decision taken after inspecting empirical data in classes of their own.

2. *Classes and subclasses*: the final typology is represented by *definitions* for 12 types/classes (e.g., Table 1).

Practical use in harmonization: the graph was thought as an aid (simple to understand and interpret) for the final grouping variables into the harmonized classification of interest. Basic characteristics of this graph are revealed in the description of the workflow below.

4.3 Working steps in Charmcats

- **Create a project**

First, the user has to login to the demo (desktop) application. Next, she/he has the possibility to browse the system and load different projects or components (e.g., variables, questions); in the following, the simplest case is sketched where a harmonisation project is produced from scratch.

- **Conceptual Step**

After assigning a working title for the project, the concept (social structure) to be measured is defined (or selected from a controlled vocabulary) for a universe (24 European Countries and USA for the year 2006)

As figure 5 (on p. 36) shows, the dimensions mentioned above are created by drawing ellipse nodes within the graph area and circle shaped nodes represent the

classes. These are conceptual definitions with no operational instructions attached that the users structures here. Figure 6 shows the “full” image (created by the application as a jpg export file) of a conceptual graph for classifications. The nodes are connected through edges that are depicted as paths; as the graph shows every classification has a root node that will end following one branch in a leaf node (nodes without edges) that depict the classes/subclasses. Few restrictions are only imposed here, like: the nodes cannot have income directed paths from nodes that they connect or subsequent nodes, edges can be drawn only if they connect nodes or leaf nodes; content and position of nodes within the graph is arbitrary and flexible (e.g., before publishing, adjacent nodes and edges can be added, nodes deleted, labels can be changed).

Turning back to figure 5 it can be seen that nodes created in the graph are displayed in the navigator bar on the left side of the screen. The lower left side of the navigator bar structures the project elements into a so called “workbench” or “basket” folder, where the user stores components from other published projects, data sources and QDB. In the lower part of the screen, below the graph, input dialogues were placed for every created node/dimension. Below this table with input dialogues, the full content of single project components is shown in a separate window (because the connection between GUI and database is not completely implemented this window is empty in figure 5).

Besides using the graph for storing classes, a table editor can be used, where lists with codes, classes and definitions are created, copied or imported (this may be especially useful in case the classification is one-dimensional, or the lists of classes is high enough that the use of the graph alone becomes unpractical).

- **Operational Step**

The next step is then to apply this structure for a set of universe elements (in our case countries and years) at the operational level (figure 7): a set (in case the same operational definitions applies for all) or single countries for which we want to define indicators and questions may be selected. The graph structure created at the conceptual step is copied here: the structure of dimensions and classes are inherited. To each node, country specific indicators may be attached; they are depicted as square nodes.

The question at this step is, what indicators are to be used to measure each of these dimensions? To measure the authority dimension, there are many possibilities (see also Wright, 1997, pp.80-90): formal positions within the authority hierarchies as indicated by organisational diagrams; the nature of the decisions the individuals can make at the workplace; different kinds of power the individual has over subordinates, etc. For example, in USA following question may be used (Wright & Cho, 1992): "Which of the following best describes the position which you hold within the business or organization in which you work? Would it be a…1. managerial position, 2. a supervisory position, or a 3. non management position?". For Germany, an additional answer category may be asked that measures also the country specific forms of management positions for "officials" ("Beamten"). Another possibility is that two different sets of standardized indicators may be used across countries (see the so called simplified and "full version of operationalizing the Wright Classification, presented in Leiulfsrud et al., 2001). Besides deciding on indicators, the next question is on how should these indicators be combined to generate operational categories of a classification? The answer for this would be described for each class by its branch.

To summarize, different indicators with specific answer categories across the subset of countries chosen from the CS may be here represented. The Indicators/Questions and their values assigned across different countries to the same inherited conceptual node/edge are functional equivalent.

The edges will depict here a subset of the indicators values attached to indicators nodes (incl. values transformations) and with the conceptual definitions (dimension) at "hand"; that means that the edges will inherit here the definitions created at the conceptual label. For defining the edges the user should specify:

1. Formula: Mathematical/statistical functions or simple recoding commands for transforming values).

1.1. Other Indicators (defined within layer), constant values or weights used in the formula;

2. Transformations of indicator values using the formula- this information should be attached to the edges as a structural element in the graph.

The specific values and value transformation for each node may be edited by a special editor accessible through the graph or through the specific input dialogues of

indicators. In the prototype version of the application these operational definitions will be shown on the edges in brackets, next to the conceptual labels. Value transformations will be created by a free text editor, but should be supported in the future by a formula editor. In the simplest case, values of indicator (or variables in the next step) are assigned to the edge without additional transformations, as shown in figure 7.

As the general first feature of the graph across all layers, within a branch, all paths from the root node to the leaf node proceed by way of conjunctions (AND). That means that, for example the Class of Employer could be now defined in a more formal expression as: ***if*** *Indicator- Employed (1)* ***and*** *Nr. of employees (1< >=9)* ***then*** *Classification* ***(Employer)***. The latter is only an example, because a variety of complex arithmetical operations and transformation of indicators values could be imagined at each node/edge. A general second main feature of the graph is that from the root node to the leaf node, the classification may operate by way of subtraction. This two main feature are taken now into account in the developing of Syntax to describe the coding algorithm of classifications. The assumption is that this syntax, "neutral" to any statistical program, could be used as a template for data re-codings in the next step.

- **Data Re-coding Step**

Harmonisation of variables can now be performed for all the universe elements or only a subset of it for which the operational step was created.

Creating, or importing from other projects the Classification with operational definitions plays now a crucial supporting role when creating harmonized variables:

1. Having indicators and coding procedures defined, the user can operate through the graph interface search queries across data sources; variables and questions may be searched using the information stored at each indicator node and filtered for each universe element;
2. A set of competitive archived variables found in the CESSDA portal could now be compared to the desired reference indicators. Comparison of variables will be realized within a special windows where the reference indicators are displayed and different can be "dragged" in for pairwise comparisons and detailed inspection.

Assignment of source variables to dimensions and indicators is required in order create the values for every edge. For example, for the indicator labelled *Authority* one Question was attached on the OS with three possible answer: *Managers, Supervisor, and non-managers*; with the available data from ESS Round 3 values of three variables are re-coded to measure these categories assigned to the edges defined in the OS: **Jbspv** *Responsible for supervising other employees;* **Orgwrk**, *To what extent the respondent organize his/her own work* and **Wrkdscin** *Allowed to influence decisions about work direction*"; Whereas with data from ISSP only ISCO-88 codes can be assigned to this Indicator. Figure 8 shows that variables are depicted as squared nodes with paths directed to dimensions. The representation of indicators defined at the OS should be allowed to be represented in the graph also at this working step; because "indicators" are not attached to data, and should not be confused with source variables, the shape of nodes for indicators should be different and it would be useful if a function will be implemented that will show/hide the indicator nodes upon request. Edges contain the block of conversion routine for the subset of source variable values/transformations they represent. In a separate window, a syntax editor can be used, and only the lines with re-coding of variables used at a specific node/edge could be highlighted. This would allow a structured way of writing and documenting the conversion syntax, with interactive access to the metadata of variables/questions. However, as mentioned above, this part of the application is now still under development.

It is highly probable that at this step, data will not be available for all nodes/edges copied from the OS. If edges and nodes will remain without "assignments" of variable values, the probably most frequent consequence will be that classes of the target variable will be collapsed and the classification will be available only in a reduced/simplified form. Using the graph, this "partial" measurement of the classification defined at the conceptual or operational step in the harmonized variable becomes evident, and is not only grounded on intuitive/ad hoc decisions.

After the final re-coding commands have been written, the project may be published with a complete documentation starting from concept, country specific measurement procedure and documentation of restrictions imposed by available data.

At every step, specific forms of conceptual, measurement and source data types of equivalence and bias are documented that were not mentioned here in detail.

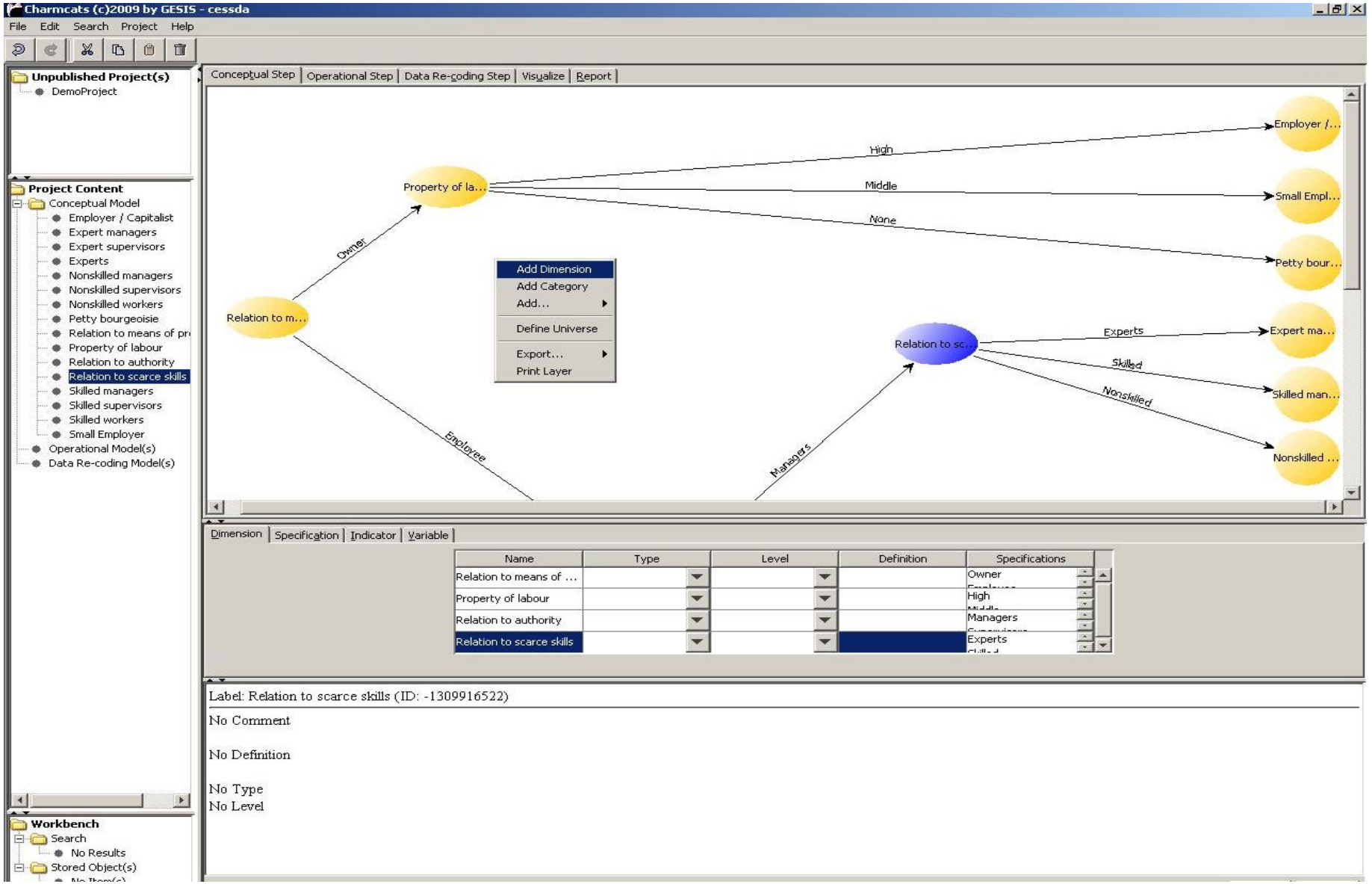

Figure 5: Screenshot of the CHARMCATS application, Version 0.4: Conceptual Step

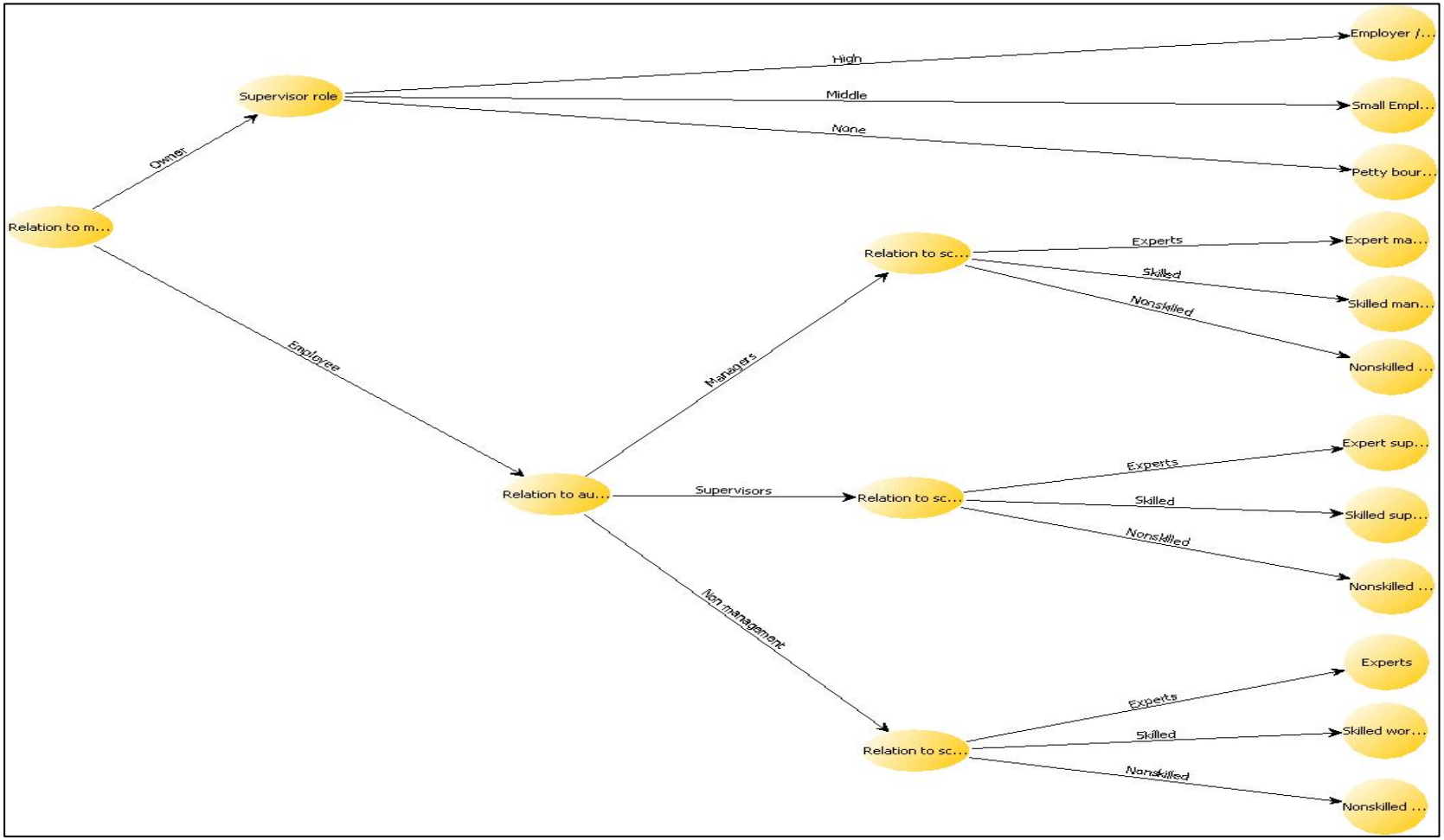

Figure 6: Jpg export image of the conceptual diagram in CHARMCATS

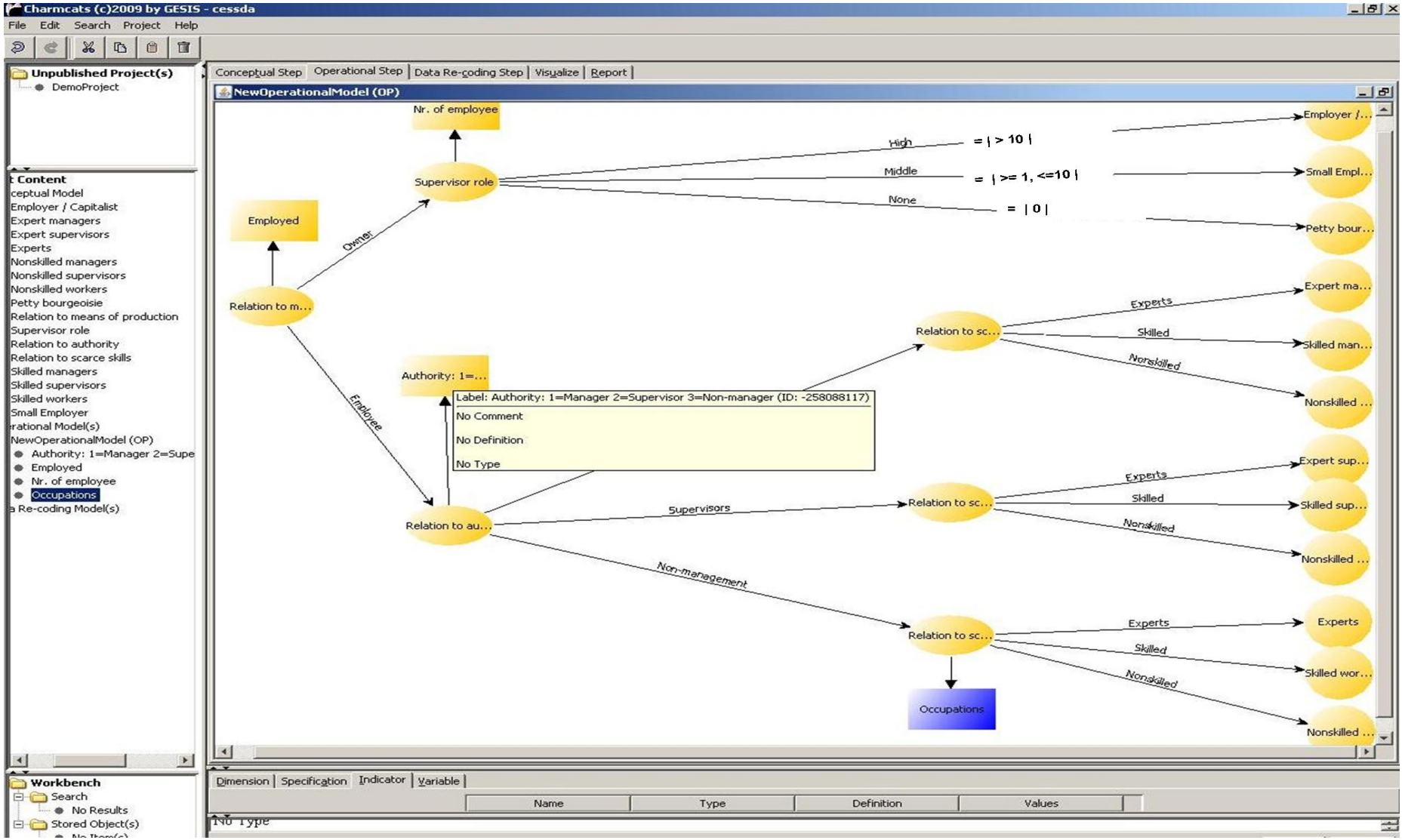

Figure 7: Screenshot of the CHARMCATS application, Version 0.4: Operational Step

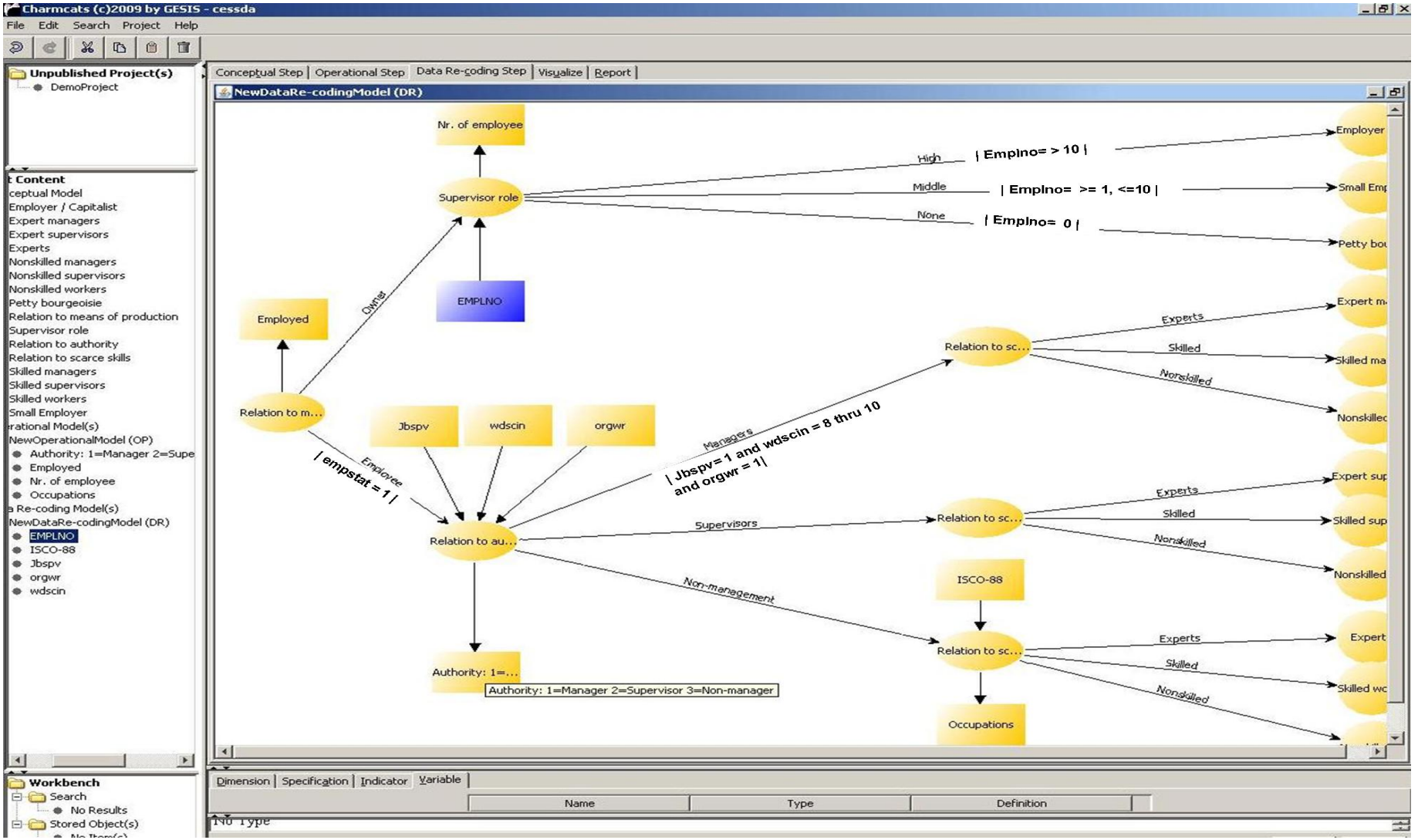

Figure 8: Screenshot of the CHARMCATS application, Version 0.4: Data Re-coding Step

5 Conclusions

As part of the general work being done within WP9, this document has aimed to describe the basic functional and technical specifications of the proposed Constructs, Classifications and Conversions Database and the corresponding software application. Our analysis suggests that the three-layer or working step model of a harmonization project that has been presented is helpful to the researcher in the process of building and documenting a harmonization routine.

CCCDB holds these harmonization routines as well as the target variable data and the data required for the graphical representation of the three layers created in the process of building the actual routine. This would help researchers to better understand these routines and maybe reuse them by transforming another source variable into a target variable. Thus, it is evident that reusability of this routine is supported by the documentation of the complete harmonization project.

Some basic decisions should be taken in collaboration with the other PPP workpackages, as well as with the manufacturers of NESSTAR (or any other data repository used in CESSDA). It seems that data dissemination systems that incorporate routines for manipulating data sets and provide a language for building these routines would be a helpful tool for CCCDB to use when creating the actual harmonization routine. Moreover, it would provide the ability to perform the harmonization routine regardless of the format the source data set is in.

The general CCCDB architecture could easily stand within the proposal of Gregory et al., 2009 for the general CESSDA architecture. Even if this architecture is not adopted, CCCDB could also stand on its own, since the proposed CCCDB architecture is unaffected from the overall architecture. It would however be necessary to have a decision on the infrastructure architecture before starting to implement CCCDB, since there must be made certain changes to the database and to the web services. This flexible, service-based architecture also makes CCCDB available to other CESSDA projects and software applications.

The benefits of CCCDB would be great for CESSDA's user community, since the researchers will have in their disposal an assisting tool to transform and interpret already published studies as well as to easily build new cross-national studies.

Finally, we must point out that even though the current purpose of CCCDB is to provide a context for building harmonization projects, the architecture, software tools and the specifications proposed make it possible for further expansion and use of the database. More functionality can therefore be added in the future without having to change anything significant in the technical, functional and security requirements described in this document.

6 Appendix

6.1 List of software products

JBossAS5

http://www.jboss.org/jbossas/

This is the final release of the JBoss 5.0 series for the Java EE5 codebase that fully complies with the Java EE5 conformance testing certification requirements. JBossAS 5 provides a healthy foundation and the most advanced and fully extensible, cross component model, aspect integration, server runtime environment. For information on the APIs that make up Java EE5, see Java EE APIs . JBossAS 5 is the next generation of the JBoss Application Server build on top of the new JBoss Microcontainer. The JBoss Microcontainer is a lightweight container for managing POJOs, their deployment, configuration and lifecycle. It is a standalone project that replaces the famous JBoss JMX Microkernel of the 3.x and 4.x JBoss series. The Microcontainer integrates nicely with the JBoss framework for Aspect Oriented Programming, JBoss AOP. Support for JMX in JBoss 5 remains strong. Further, it lays the groundwork for JavaEE 6 profiles oriented configurations and JBoss AS embedded that will allow for fine grained selection of services for both unit testing and embedded scenarios.

JBossAS 5 is designed around the advanced concept of a Virtual Deployment Framework (VDF) that takes the aspect oriented design of many of the earlier JBoss containers and applies it to the deployment layer. Aspectized Deployers operate in a chain over a Virtual File System (VFS), analyze deployments and produce metadata to be used by the JBoss Microcontainer, which in turn instantiates and wires together the various pieces of a deployment, controlling their lifecycle and dependencies. The VDF allows for both customization of existing component modules including JavaEE and JBoss Microcontainer, as well as introduction of other models such as OSGi and Spring.

Many key features of JBoss 5 are provided by integrating other standalone JBoss projects:

JBoss Microcontainer is the next generation POJO based kernel that is used as the core of the server. It supports an extensible deployment model and advanced dependency relationships.

The definition of the non-kernel deployers and deployment is now defined a Profile obtained from the ProfileService. The ProfileService also provides the Management-View for ManagedDeployments/ManagedObjects used by the OpenConsole admin tool.

JBoss EJB3 included with JBoss 5 provides the implementation of the latest revision of the Enterprise Java Beans (EJB) specification. EJB 3.0 is a deep overhaul and simplification of the EJB specification. EJB 3.0's goals are to simplify development, facilitate a test driven approach, and focus more on writing plain old java objects (POJOs) rather than coding against complex EJB APIs.

JBoss Messaging is a high performance JMS provider in the JBoss Enterprise Middleware Stack (JEMS), included with JBoss 5 as the default messaging provider. It is also the backbone of the JBoss ESB infrastructure. JBoss Messaging is a complete rewrite of JBossMQ, which is the default JMS provider for the JBoss AS 4.x series.

JBossCache that comes in two versions. A traditional tree-structured node-based cache and a PojoCache, an in-memory, transactional, and replicated cache system that allows users to operate on simple POJOs transparently without active user management of either replication or persistency aspects.

JBossWS is the web services stack for JBoss 5 providing Java EE compatible web services, JAX-WS-2.0.

JBoss Transactions is the default transaction manager for JBoss 5. JBoss Transactions is founded on industry proven technology and 18 year history as a leader in distributed transactions, and is one of the most interoperable implementations available.

JBoss Web is the Web container in JBoss 5, an implementation based on Apache Tomcat that includes the Apache Portable Runtime (APR) and Tomcat native technologies to achieve scalability and performance characteristics that match and exceed the Apache Http server.

JBoss Security has been updated to support pluggable authorization models including SAML, XACML and federation.

JBoss SSO

http://www.jboss.org/jbosssso/

The JBoss SSO Framework is a collection of components that software developers can easily integrate within their existing web applications to create a federation of

trusted web sites. The framework has support for important SSO standards such as SAML. The system consists of the following components:

- **Federation Server -** A Federation Server is used for securely propagating the **Federation Token** across web applications located in different security domains
- **Token Marshalling Framework -** This is a flexible/pluggable Java API to marshal/unmarshal a **Federation Token**. The system ships with a **SAML-compliant** Marshaller
- **Identity Connector Framework -** This is a flexible/pluggable Java API to connect to central identity stores. The system ships with a Provider to connect to **LDAP based Identity Stores**

JGRAPH

JGraph is a graph visualization library written in Java. Its use requires an installed Java version 1.4 or later. JGraph is based on the Swing MVC (Model/View/Controller) pattern and designed to be fully compatible with Swing. It is an Open Source Software using the LPGL.

The JGraph API allows the visualizing and interacting of graphs, graphs, as in the mathematical graph theory as structures who model pairwise relations between objects.

As such it is used in the prototype application for visualizing and editing the main structural elements in the three working step model of CHARMCATS: Dimensions, Indicators and Variables.

6.2 Minimal DDI3 requirements for 3CDB and QDB input data

Relevant chapters in the D9.2 Report on required input metadata:

- 1.4: Workflow and the working steps/layers of CHARMCATS
- 2.1: Contents of 3CDB
- 2.2: Functionality of 3CDB
- 2.4 Search and retrieve metadata; 2.4.1 Types of retrieved metadata and 2.4.2 Sources (pp.: 16-17).
- 3.2.1: DDI3/2 elements to be used

Throughout this report, following questions were addressed regarding minimal recommended DDI3 elements to be included in 3CDB and QDB:

1) What minimal metadata must be provided to 3CDB and QDB?
This will be mainly be answered here from the perspective of 3CDB functionalities/ creation of harmonised data. Therefore, first an overview of the required input metadata is given below.

2) Where should these metadata be located?
Before going into detailed description of required metadata and format it is important to clarify for whom these requirements are made. This is connected to the question: where will the data be located.

3) Recommended DDI3 elements?
Next, a description of all the recommended DDI3 elements is listed in Table 2.

6.2.1 Overview of input metadata required by 3CDB

Researchers using QDB and 3CDB will be primary interested in the availability of **specific** survey questions (or **batteries** of questions) and variables for a **defined concept and universe.** That means, that the **basic level** of metadata both databases need will be the question and variables rather than the whole set of metadata attached to each survey/study the question/variable originates (it will be rarely the case the researchers are interested in the whole study/survey). As it was shown metadata of a survey- including methodological information (e.g., information on data collection and sampling procedures) will be required in the stage of harmonizing and matching equivalent variables/questions, but primarily in the stage of **exploring** available data this is not the case. Thus, survey metadata should be always connected on the level of variables/questions when provided to QDB and 3CDB.

What input data is required in 3CDB and for what application? The data required for input nodes of 3CDB that are primarily created outside 3CDB will be the variables (connected to questions).
The minimum source data required for the input nodes of variables within Charmcats are outlined in figure 9, below. The arrows between the data boxes indicate that variables are connected to questions and to descriptions of the survey/datafiles they originate. All metadata should be provided in original language and with English translations. The elements enumerated in the primary input data boxes may be seen as the

minimum “must” required by both databases, regardless of the existing metadata format.

ELLST could be used for indexing variables and questions and will be necessary integrated into 3CDB and QDB; another question will be how to integrate it into the CESSDA portal for **search** purposes, and as a thesaurus resource for **concepts, keywords, and subjects** (topics).

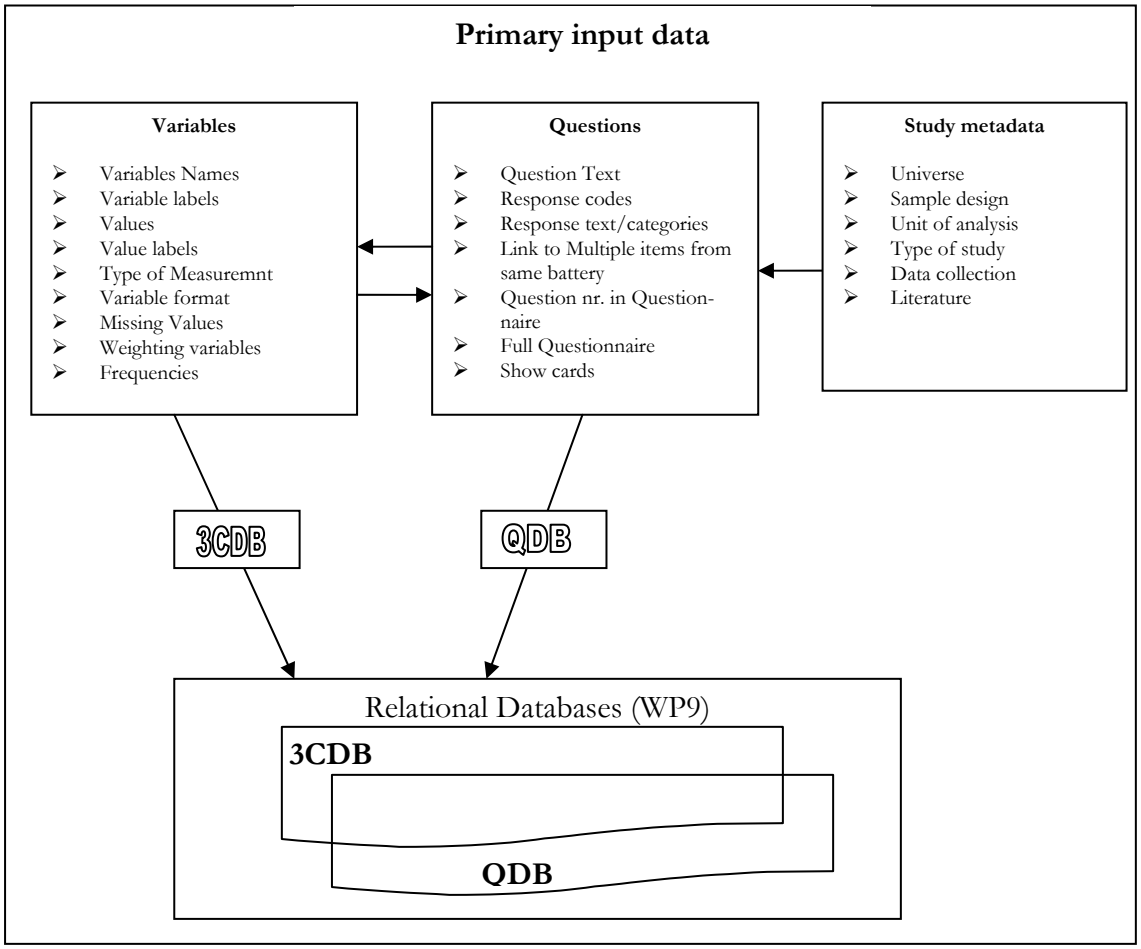

Figure 9: Primary input data flow into the relational databases of 3CDB and QDB

6.2.2 Location of input metadata

1) Ideally, the source metadata on questions will be provided and organized in the relational DB of QDB. Hypothetically the data on variables and questions will be provided to 3CDB only from QDB, only by CESSDA SOA repositories, and by repositories located outside of CESSDA or by a combination of all (QDB and repositories)
2) Second, it is envisioned that 3CDB and QDB will make possible to store metadata on questions and variables of studies not archived within CESSDA via a basic digital library storage system where users can upload/edit their own (meta-) data.
3) Third, in 3CDB Indicators and Question documentation will be provided that is not connected to an archived survey study but that could be (re)used into a survey making the data flow from 3CDB into QDB and into repositories imaginable.

But, the bulk of necessary source data for QDB and 3CDB will be primarily located in the CESSDA repositories embedded in the CESSDA portal (regardless of the adopted architecture). The focus relies on metadata without methodological information on comparability attached (since this will be realized within 3CDB/QDB).

To sum up, for a first working version of both databases, it must be assumed that:

- QDB pulls all available questions records from CESSDA repositories – organize them for purposes of comparability within a relational database (see conceptual DB in Gregory et al, 2009, p. 32).
- 3CDB pulls questions- variables metadata records from QDB and CESSDA repositories; performs harmonisation and comparability information that may supplement QDB records; equivalent questions are created within 3CDB that may supplement in the long term the QDB.

6.2.3 Recommended DDI3 elements

The prototype database of 3CDB was designed to be compliant with DDI and therefore the first requirement is that the metadata on variables and questions should be provided in **DDI3 format**. The list in **Table 2** provides a set of minimal elements. Since at the current stage DDI3 is not being implemented at larger scale within CESSDA, a second requirement would be to provide at least the corresponding **DDI2 elements** (matching of the DDI3 elements presented in Table 2 and corresponding DDI2 elements are presented in DDI3, 2008 (pp. 86-162) and are not reproduced here). However migration from DDI2 to DDI3 standard should be kept in mind for the long term strategy (see also Alvheim, August 2009; recommended also by the QDB evaluation of tendered report).

Recommended DDI3 elements (additional notes on Table 2):

- **Multilingual versions of elements and English translations**: original version plus English translations
- **For all variables- coding schemes and referenced category and value schemes** (use case: variable codes constructed from several questions/response domains).
- **Coding schemes and metadata on re-codings:** in case a harmonized variable is part of the original source datafile (e.g., created by the archive/study team), documentation of source variables and re-codings (documented with *generationinstruction* in DDI3) together with a *Text description* should be provided.
- **Variables as part of item/question batteries:** it is recommended to have questions and variables groups (as in DDI3) to tap items batteries.

Table 2: Required DDI3 elements

***MAPPING DDI3 IDS - CHARMCATS IDS (NA)**: reusable.xsd (r)*	
IDENTIFIABLE_IDS	*DDI3: r:IdentifiableType*
VERSIONABLE_IDS	*DDI 3: r:VersionableType*
MAINTAINABLE_IDS	*DDI 3: r:MaintainableType*
DDI 3 COMPLIANCE: reusable.xsd (r), datacollection.xsd (d), logicalproduct.xsd (l)	
INTERNATIONAL_STRINGS	*DDI3: r:InternationalStringType*
LABELS	*DDI3: r:LabelType*
TYPED_STRINGS	*DDI3: r:TypedStringType*
STRUCTURED_STRINGS	*DDI3: r:StructuredStringType*
IDENTIFIED_STRUCTURED_STRINGS	*DDI3: r:IdentifiedStructuredStringType*
DYNAMIC_TEXTS	*DDI3: d:DynamicTextType*
TEXTS	*DDI3: d:TextType*
LITERAL_TEXTS	*DDI3: d:LiteralTextType*
CONDITIONAL_TEXTS	*DDI3: d:ConditionalTextType*
CODES	*DDI3: r:CodeType*
STRUCTURED_MIXED_RESPONSE_DOMAINS	*DDI3: d:StructuredMixedResponseDomainType*
OTHER_MATERIALS	*DDI3: r:OtherMaterialType*
RELATIONSHIPS	*DDI3: r:RelationshipType*
DATES	*DDI3: r:DateType*
BASE_DATES	*DDI3: r:BaseDateType*
HISTORICAL_DATES	*DDI3: r:HistoricalDateType*
INTERVIEWER_INSTRUCTION_SCHEMES	*DDI3: d:InterviewerInstructionSchemeType*
INSTRUCTIONS	*DDI3: d:InstructionType*
CONTROL_CONSTRUCT_SCHEMES	*DDI3: d:ControlConstructSchemeType*
CONTROL_CONSTRUCTS	*DDI3: d:ControlConstructType*
EXT_INTERVIEWER_INSTRUCT_REFS	*DDI3: d:ExternalInterviewerInstructionReferenceType*
INTERVIEWER_INSTRUCT_REFS	*DDI3: d:InterviewerInstructionReferenceType*
LOOPS	*DDI3: d:LoopType*
IF_THEN_ELSES	*DDI3: d:IfThenElseType*
REPEAT_UNTILS	*DDI3: d:RepeatUntilType*
REPEAT_WHILES	*DDI3: d:RepeatWhileType*
SEQUENCES	*DDI3: d:SequenceType*
COMPUTATION_ITEMS	*DDI3: d:ComputationItemType*
STATEMENT_ITEMS	*DDI3: d:StatementItemType*
CREATORS	*DDI3: r:CreatorType*

CONTRIBUTORS		*DDI3: r:ContributorType*
CODINGS		*DDI3: d:CodingType*
GENERAL_INSTRUCTIONS		*DDI3: d:GeneralInstructionType*
GENERATION_INSTRUCTIONS		*DDI3: d:GenerationInstructionType*
COMMANDS		*DDI3: r:CommandType*
STRUCTURED_COMMANDS		*DDI3: r:StructuredCommandType*
COMMAND_FILES		*DDI3: r:CommandFileType*
CODE_VALUES		*DDI3: r:CodeValueType*
ACTION_CODES		*DDI3: r:ActionCodeType*
ADDITIVITY_CODES		*DDI3: l:AdditivityCodeType*
AGGREGATION_METHOD_CODES		*DDI3: l:AggregationMethodCodeType*
CATEGORY_RELATION_CODES		*DDI3: r:CategoryRelationCodeType*
CONCATENATED_VALUES		*DDI3: l:ConcatenatedValueType*
EXCLUDES		*DDI3: r:ExcludeType*
IDS		*DDI3: r:IDType*
URNS		*DDI3: r:URNType*
VERSIONS		*DDI3: r:VersionType*
UNIVERSE: conceptualcomponent.xsd (c)		
UNIVERSES		*DDI3: c:UniverseType*
UNI_SCHEMES		*DDI3: c:UniverseSchemeType*
CONCEPT : conceptualcomponent.xsd (c)		
CONCEPTS		*DDI3: c:ConceptType*
CONCEPT_SCHEMES		*DDI3: c:ConceptSchemeType*
CONCEPT_GROUPS		*DDI3: c:ConceptGroupType*
CITATIONS	*440*	*DDI3: r:CitationType*
REFERENCE: reusable.xsd (r)		
VARIABLE: logicalproduct.xsd (r)		
VARIABLES		*DDI 3: l:VariableType*
VARIABLE_SCHEMES		*DDI 3: l:VariableSchemeType*
VARIABLE_GROUPS		*DDI3: l:VariableGroupType*
QUESTION : datacollection.xsd (d)		
QUESTION_ITEMS		*DDI3: d:QuestionItemType*
MULTIPLE_QUESTION_ITEMS		*DDI3: d:MultipleQuestionType*
QUESTION_SCHEMES		*DDI 3: d:QuestionSchemeType*
QUESTION_GROUPS		*DDI3: d:QuestionGroupType*
SPECIFIC_SEQUENCES		*DDI3: d:SpecificSequenceType*

QUESTION_SEQUENCE_TYPES	*DDI3: d:QuestionSequenceType*

Suggested DDI3 elements:

Grouping
Group allows you to define which parts of the major components are shared, where overrides take place, and how to relate or link data in one study to data in a subsequent survey.

In DDI3 the schema group.xsd was developed for capturing the life cycle of data across similar studies. Basically, there are two forms of grouping in DDI3: a) by design and b) ad hoc.

a) **by design**: repetition of series of studies; second and subsequent studies inherit features of the first study (base structure of concepts, question, variable)for the purpose of comparability. Inheritance is realized from a item by item comparison structure. For following types of studies grouping by design should be used:

- Cross-sectional
- Repeated survey
- Panel data

Grouping by design may be used also in case of household surveys.

b) **Ad hoc groups:** are realized without the use of inheritance, comparability must be described explicitly using the schema Comparative. This could be used in post-hoc harmonisation. Commonalities and differences are described by pair-wise comparisons of study units.

Components of the comparison module

COMPARISON: *comparative.xsd (c)*

COMPARISONS	*DDI3: c:ComparisonType*
GENERIC_MAPS	*DDI3: c:GenericMapType*
CORRESPONDENCES	*DDI3: c:CorrespondenceType*
USER_DEFINED_CORR_PROPERTYS	*DDI3: c:UserDefinedCorrespondencePropertyType*
ITEM_MAPS	*DDI3: c:ItemMapType*
CODE_MAPS	*DDI3: c:CodeMapType*

7 References

Alvheim, A. (2009, August). *D5.3. A CESSDA Common Data portal.* Accessible via the CESSDA PPP intranet: *http://www.cessda.org/ppp/wp05/WP5_Common_Data_Portal_v.2.0.pdf*

CESSDA (2008-2009). *CESSDA PPP - Preparatory Phase Project for a Major Upgrade of the Council of European Social Science Data Archives (CESSDA) Research Infrastructure*. Online resource, last accessed 2009-05-18: http://www.cessda.org/project/

CESSDA PPP - Work Package 9 (2008). *Building an Infrastructure for Content Harmonisation and Conversion*. Online resource, last accessed 2009-05-18: http://www.cessda.org/project/doc/wp09_descr2.pdf

DDI Alliance. (2008). *Data Documentation Initiative (DDI), Technical specifications. Part 1: Overview. Version 3.0.* Retrieved from www.ddialliance.org, on 2009-08-05.

Friedrichs, M. (2009, August). *CHARMCATS: List of Tables*. Available at: http://www.cessda.org/ppp/wp09/WP9_Database_Models_Aug09.pdf

Gregory, A., Heus, P.; Nelson, C., and Ryssevik, J. (2009): *Technical Specifications for a European Question Data Bank. Tender Report to the CESSDA-PPP*, available at: http://www.cessda.org/project/doc/CESSDA_PPP_QDB_May09.pdf

Jensen, U. (2009). DR 8.1.1 *Functional needs and specifications for metadata of cross-national surveys, time series and further complex data types*. Draft 2. Available at: http://www.cessda.org/ppp/wp08/DR.8.1.1_v2.0.pdf

Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In J. Harkness (Ed.), *ZUMA-Nachrichten Spezial: Cross-cultural Survey Equivalence* (Vol. 3, pp. 1-40). Mannheim: ZUMA.

Krejci, J.; Orten, H., & Quandt, M. (2008): Strategy for collecting conversion keys for the infrastructure for data harmonisation, http://www.cessda.org/ppp/wp09/wp09_T93report.pdf

Leiulfsrud, H., Bison, I., & Jensberg, H. (2005). Social class in Europe. European Social Survey 2002/3. Trondheim: NTNU Social Research Ltd., Retrieved from http://ess.nsd.uib.no/files/2003/ESS1SocialClassReport.pdf , on 2009-08-05.

Mejer, L. (2003). Harmonisation of socio-economic variables in EU Statistics. In J. Hoffmeyer-Zlotnik & C. Wolf (Eds.): *Advances in cross-national comparison. A European working book for demographic and socio-economic variables* (pp. 67-85). New York: Kluwer Academics/ Plenum Publishers.

Martinez, L. (2008). *The Data Documentation Initiative (DDI) and Institutional Repositories.* Retrieved from http://www.disc-uk.org/docs/DDI_and_IRs.pdf, on 2009-06-01.

Rose, D., & Harrison, E. (2007). The European socio-economic classification: A new social class schema for comparative European research. *European Societies*, 9(3), 459-490.

Tan, P., Steinbach, M. & Kumar, V. (2006). *Introduction to data mining.* Boston, Mass. [u.a.]: Pearson/Addison-Wesley.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research* Thousand Oaks: Sage.

Wright, E. O. (1997). *Class counts. Comparative studies in class analysis*. Cambridge: Cambridge University Press.

Wright, E. O. (2005). *Approaches to class analysis*. Cambridge: Cambridge University Press.

Wright, E. O., & Cho, D. (1992). The relative permeability of class boundaries to cross-class friendships: A comparative study of the United States, Canada, Sweden and Norway. *American Sociological Review, 57*(1), 85-102.