The Case of CHARMCATS: Use of DDI3 for Publishing Harmonisation Routines

1st Annual European DDI Users Group Meeting:
DDI - The Basis of Managing the Data Life Cycle

IZA, Bonn, 04.12.2009

Martin Friedrichs (martin.friedrichs@gesis.org)
Markus Quandt (markus.quandt@gesis.org)
Alex Agache (alexandru.agache@gesis.org)

gesis
Leibniz Institute for the Social Sciences

Aims of this presentation

1. **What is:**
   Cessda HARMonisation platform of CATegories & Scales

2. **Conceptual issues:**
   Use of DDI3 elements for documenting Comparability - Harmonisation

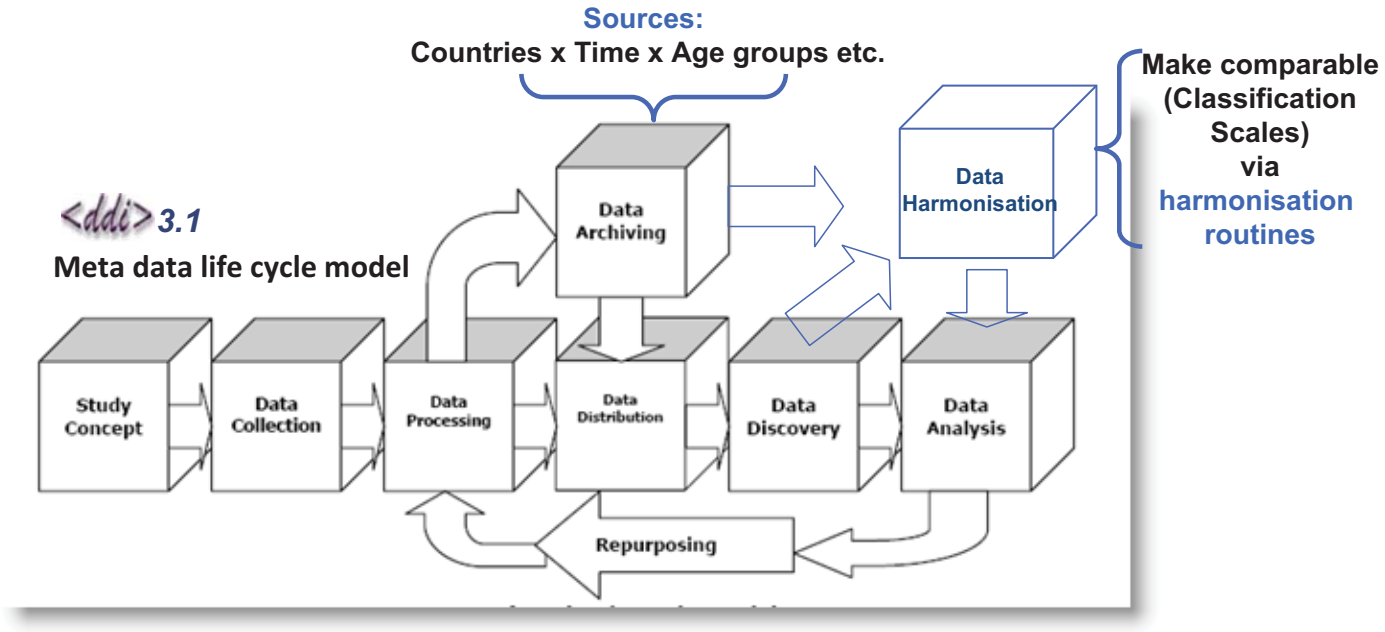
3. **Technical issues:**
   Managing source variables within the CHARMCATS - RDB

1. What is:
Cessda HARMonisation platform of CATegories & Scales

gesis
Leibniz Institute for the Social Sciences

CHARMCATS - Purpose

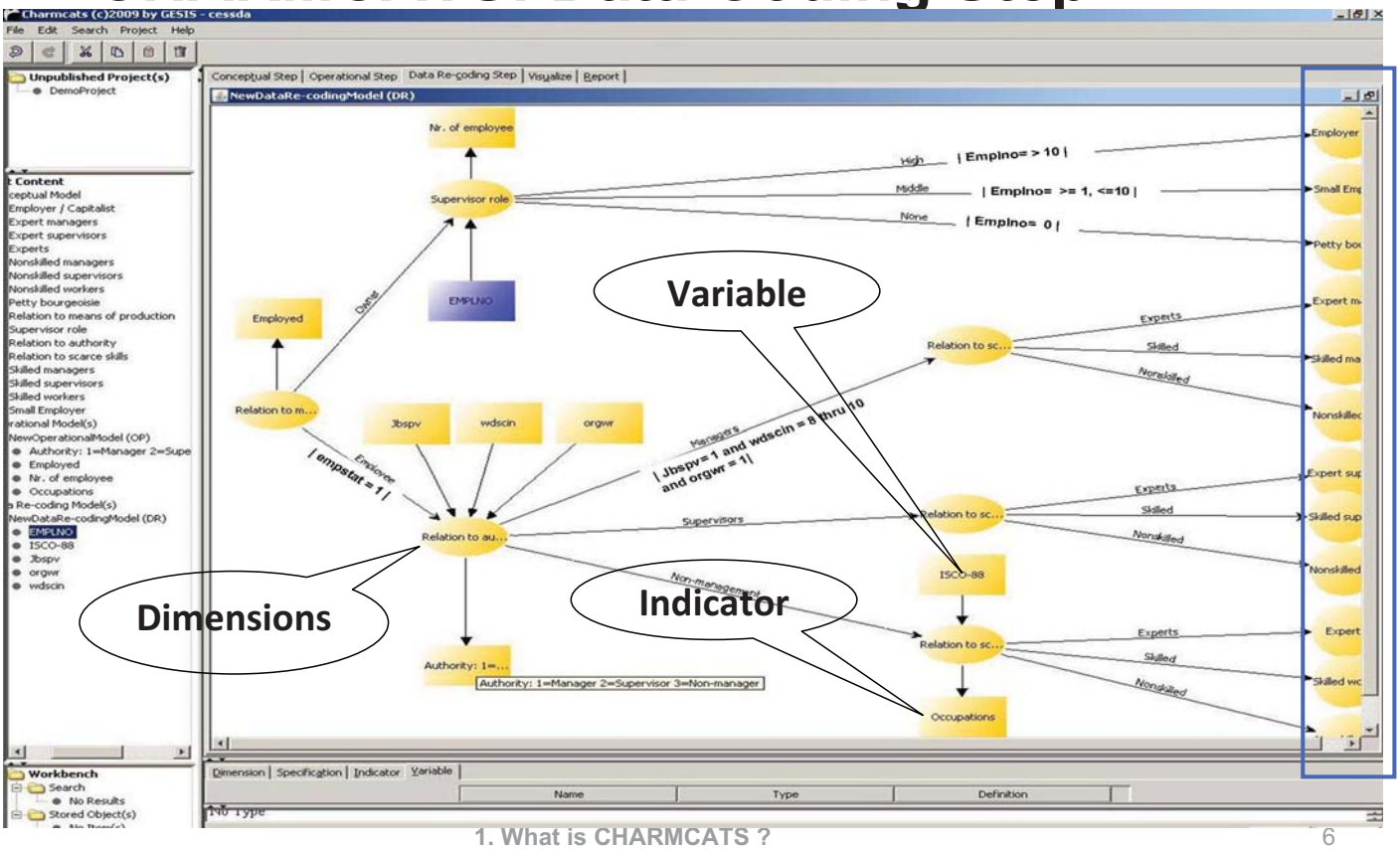
Document survey (ex-post) Data Harmonisation:
how data from different sources is made comparable (standard codes)



Harmonisation project: 3 interlinked working steps

1. **Conceptual Step**
   By assumption, a **classification** is applicable to all objects of a **universe** [e.g. Europe]:
   + **Concept**, **Dimensions**
2. **Operationalization Step**
   Bias at **operationalisations** to **elements** of **universe** [individual countries, e.g. France]:
   + **Indicators**, **Questions** (not connected to archived data)
3. **Data Coding Step**
   Bias **when applying operationalisation to actual data** (Compare variables to Indicators):
   + (source) **Variables (Questions)**, **Data files**
   -> Conversion Syntax
   -> Target / harmonized Variable

CHARMCATS: Data-Coding Step



Summary (1): CHARMCATS compliance with DDI3

Harmonisation Projects: **CHARMCATS data format**

CHARMCATS- metadata covered with DDI3:

- **Harmonised Variable: DDI3**
- **Simple re-coding commands: DDI3 (e.g., *generation instruction*)**
- **On source data side- compliance with DDI3 in CHARMCATS RDB model**

CHARMCATS- metadata 'covered but not covered' with DDI3:

- **Conceptual Schemes of source/target variables**
- **Variable (Item) Maps = Assessment of comparability via *comparative module***

gesis
Leibniz Institute for the Social Sciences

2. Conceptual Issues:

Use of DDI3 elements for documenting

Comparability - Harmonisation

Basic relationships between variables & concepts

Harmonized Variable:

< Universe > 'Europe'
- < Concept > 'Education achievment'
< Author > 'W. Müller'

Source Variable: 'EDLVADE'
< Universe > 'Germany, 2008'

< Concept > 'Highest Education in DE' < Concept > 'Socio-economic Status'

Use of 'Concepts' for comparison purposes

- **(Concept scheme) to (Var. Scheme) relationships:**
  - For **harmonized variables**:**1 to n** (as in DDI3)
  - For **source variables**: **m to n**

Study context (data collection) – where concept scheme reference made- cannot be imposed to harmonisation or analysis context (where same variable may indicate different concepts)

DDI3 Comparative Module: Overview

The Comparative Module:

- What can be mapped?:
  Concepts, **Variables**, Questions, Categories, **Codes**, Universes.

Correspondence: Difference

Commonality

CommonalityTypeCoded — **CV: Identical, Some, None**

CommonalityWeight

UserDefinedCorrespProperty

Pairwise comparison with DDI3:

- Among **source variables** (DDI3 User guide)
- **Source variables** with **target, harmonized variable**

(Sanda Ionescu, *Comparison Test Case,* 2009)

CHARMCATS: both solutions not recommended

gesis
Leibniz Institute for the Social Sciences

Use of *Comparative Module*

Employment of DDI3-based comparison within **CHARMCATS:**

- **In discovery phase:** for comparing (+ pre-selection) of **source variables**
- **What to compare ? :** workable only if pairwise comparison of **source variables** with **indicators** ("equivalent standards" in **CHARMCATS**) are employed

Summary (2)

- **DDI3 concept scheme**: currently limited applicability
- **Comparative module:** pairwise comparison of variables with an equivalent indicator (not harmonized variable in **CHARMCATS**)

gesis
Leibniz Institute for the Social Sciences

3. Technical Issues:
DDI3 and CHARMCATS - RDB

CHARMCATS DDI2 / DDI3 Functionalities

- Import DDI2
- Import DDI3
- Export DDI3 (limited to the variable level)
- Convert DDI2 to DDI3 (limited to HP)
- Map DDI3 to a RDB

Why a Relational DB?

- High efficient in case of large amount of data
- Supports a structured storage

Why not a XML-DB?

- Low efficient with large amount of data
- Only DDI Source is XML-based

Requirements on Mapping

Java objects and DDI have to be mapped into the relational database. Harmonisation-Objects (Variables etc) designed as close as possible to DDI3.

1. Size of DDI3:
   22 Schemas > 200 ComplexTypes
   > 50 SimpleTypes

2. A Table for each Type

VariableType:

```
- <xs:complexType name="VariableType">
  + <xs:annotation></xs:annotation>
  - <xs:complexContent>
    - <xs:extension base="r:VersionableType">
      - <xs:sequence>
          <xs:element ref="VariableName" minOccurs="0" maxOccurs="unbounded"/>
        - <xs:element ref="r:Label" minOccurs="0" maxOccurs="unbounded">
          + <xs:annotation></xs:annotation>
          </xs:element>
        - <xs:element ref="r:Description" minOccurs="0" maxOccurs="unbounded">
          + <xs:annotation></xs:annotation>
          </xs:element>
        - <xs:element ref="r:UniverseReference" minOccurs="0" maxOccurs="unbounded">
          + <xs:annotation></xs:annotation>
          </xs:element>
        - <xs:element ref="ConceptReference" minOccurs="0">
          + <xs:annotation></xs:annotation>
          </xs:element>
        - <xs:element ref="QuestionReference" minOccurs="0" maxOccurs="unbounded">
          + <xs:annotation></xs:annotation>
          </xs:element>
        - <xs:element ref="EmbargoReference" minOccurs="0">
          + <xs:annotation></xs:annotation>
          </xs:element>
        - <xs:element name="ResponseUnit" type="xs:string" minOccurs="0">
          + <xs:annotation></xs:annotation>
          </xs:element>
        - <xs:element ref="r:AnalysisUnit" minOccurs="0">
          + <xs:annotation></xs:annotation>
          </xs:element>
        - <xs:element ref="Representation" minOccurs="0">
          + <xs:annotation></xs:annotation>
          </xs:element>
        </xs:sequence>
      - <xs:attribute name="isTemporal" type="xs:boolean" default="false">
        + <xs:annotation></xs:annotation>
        </xs:attribute>
      - <xs:attribute name="isGeographic" type="xs:boolean" default="false">
        + <xs:annotation></xs:annotation>
        </xs:attribute>
      - <xs:attribute name="isWeight" type="xs:boolean" default="false">
        + <xs:annotation></xs:annotation>
        </xs:attribute>
      </xs:extension>
    </xs:complexContent>
  </xs:complexType>
```

VariableType: Ext. Base

<xs:extension base="r:VersionableType">

r:AbstractIdentifiableType
r:AbstractVersionableType
r:ActionCodeType
r:InternationalStringType
r:UserIDType
r:VersionableType

r:BaseDateType
r:BaseIDType
r:DDIURNType
r:NewVersionType

VariableType: Elements (1)

<xs:element ref="VariableName" minOccurs="0" maxOccurs="unbounded"/>
<xs:element ref="r:Label" minOccurs="0" maxOccurs="unbounded">
<xs:element ref="r:Description" minOccurs="0" maxOccurs="unbounded">
<xs:element ref="r:UniverseReference" minOccurs="0" maxOccurs="unbounded">
<xs:element ref="ConceptReference" minOccurs="0">
<xs:element ref="QuestionReference" minOccurs="0" maxOccurs="unbounded">
<xs:element ref="EmbargoReference" minOccurs="0">
<xs:element name="ResponseUnit" type="xs:string" minOccurs="0">
<xs:element ref="r:AnalysisUnit" minOccurs="0">
<xs:element ref="Representation" minOccurs="0">

VariableType: Elements (2)

ConcatenatedValueType
RepresentationType
r:CodeValueType
r:InternationalStringType
r:LabelType
r:NameType
r:ReferenceType
r:RepresentationType
r:StructuredStringType
r:URNType
r:IDType
r:VersionType

AdditivityMethodCodeType
AggregationMethodCodeType
r:BaseDateType
r:BaseIDType
r:CategoryRelationCodeType

gesis
Leibniz Institute for the Social Sciences

VariableType: Attributes

<xs:attribute name="isTemporal" type="xs:boolean" default="false">
<xs:attribute name="isGeographic" type="xs:boolean" default="false">
<xs:attribute name="isWeight" type="xs:boolean" default="false">

Types used by VariableType

ConcatenatedValueType
RepresentationType
r:AbstractIdentifiableType
r:AbstractVersionableType
r:CodeValueType
r:IDType
r:InternationalStringType
r:LabelType
r:NameType
r:ReferenceType
r:RepresentationType
r:StructuredStringType
r:URNType
r:UserIDType
r:VersionableType
r:VersionType

AdditivityCodeType
AggregationMethodCodeType
r:ActionCodeType
r:BaseDateType
r:BaseIDType
r:CategoryRelationCodeType
r:DDIURNType
r:NewVersionType

gesis
Leibniz Institute for the Social Sciences

```
<xs:extension base="r:VersionableType">

<xs:element ref="VariableName" minOccurs="0" maxOccurs="unbounded"/>
<xs:element ref="r:Label" minOccurs="0" maxOccurs="unbounded">
<xs:element ref="r:Description" minOccurs="0" maxOccurs="unbounded">
<xs:element ref="r:UniverseReference" minOccurs="0" maxOccurs="unbounded">
<xs:element ref="ConceptReference" minOccurs="0">
<xs:element ref="QuestionReference" minOccurs="0" maxOccurs="unbounded">
<xs:element ref="EmbargoReference" minOccurs="0">
<xs:element name="ResponseUnit" type="xs:string" minOccurs="0">
<xs:element ref="r:AnalysisUnit" minOccurs="0">
<xs:element ref="Representation" minOccurs="0">

<xs:attribute name="isTemporal" type="xs:boolean" default="false">
<xs:attribute name="isGeographic" type="xs:boolean" default="false">
<xs:attribute name="isWeight" type="xs:boolean" default="false">
```

Table „variables“ / MySQL

Field *	Type *	Null *	Key *	Default	Extra *
variable_id	int(11)	NO	PRI		auto_increment
versionable	int(11)	YES			
concept_reference	int(11)	YES			
embargo_reference	int(11)	YES			
response_unit	varchar(...)	YES			
analysis_unit	int(11)	YES			
representation	int(11)	YES			
is_temporal	tinyint(1)	YES		0	
is_geographic	tinyint(1)	YES		0	
is_weight	tinyint(1)	YES		0	

Handling „repeating groups“ (1)

"variables"	...
.	
.	
.	

"variable-name"	...
.	
.	
.	

"concept_scheme"	...
.	
.	
.	

"concept_scheme-name"	...
.	
.	
.	

"universe_scheme"	...
.	
.	
.	

"universe_scheme-name"	...
.	
.	
.	

"names"	...
.	
.	
.	

"concept"	...
.	
.	
.	

"concept-name"	...
.	
.	
.	

"universe"	...
.	
.	
.	

"universe-name"	...
.	
.	
.	

Handling „repeating groups“ (2)

bachemer_str	
number	...
40	...
42	...

liliencron_str	
number	...
6	...

access-phone		
table_id	table_entry_id	phone
1	40	1111
1	40	2222
1	42	3333
2	6	4444
2	6	5555

phone	
number	...
1111	...
2222	...
3333	...
4444	...
5555	...

tables		
table_id	name	...
1	bachemer_str	...
2	liliencron_str	...
3	phone	...

Handling „repeating groups“ (3)

reference_refs			
table_id	table_entry_id	element	reference_id
.	.	.	.
.	.	.	.
.	.	.	.

reference	
reference_id	...
.	.
.	.
.	.

```
<xs:element ref="r:UniverseReference" minOccurs="0" maxOccurs="unbounded">
<xs:element ref="QuestionReference" minOccurs="0" maxOccurs="unbounded">
```

variable_elements			
element_id	name	...	version
.	.	.	.
.	.	.	.
.	.	.	.

Tables needed for Mapping into RDB:

- „Type“-Tables
- „Elements“-Tables
- „Refs“-Tables

gesis
Leibniz Institute for the Social Sciences

Functions of „Refs“-Tables:

- They are not only used as a means to an end in the process of mapping DDI.
- They are also used prevalent for storing the structural information of a HP.
- The same „Refs“-Table may store DDI-based entries and CharmCats-based ones. (e.g. assign an object like a variable to an indicator)

Conclusions & Outlook

- DDI3 supports Harmonized Variables and simple re-coding commands
- The way DDI3 handles Comparison and Concepts is not as supportive as it could be
- CharmCats-RDB is designed as DDI3-compliant for a subset of the full standard

- Only selected parts of a HP will be exported in DDI3

gesis
Leibniz Institute for the Social Sciences

Use of *Comparative Module*

Employment of DDI3-based comparison within **CHARMCATS:**

- **In discovery phase:** for comparing (+ pre-selection) of **source variables**

- **What to compare ? :** workable only if pairwise comparison of **source variables** with **indicators** ("equivalent standards" in **CHARMCATS**) are employed

- **Types of comparability:** DDI3 commonality type '**similar**'- is more exhaustive within **CHARMCATS**

gesis
Leibniz Institute for the Social Sciences

Pairwise comparisons: indicators- variables

	Source Indicator: Education Qualification			Target Indicator: ISCED-97 Classification	Source Variables	Target Variable: Harmonized Education, ISCED-97
Universe	Code	ame of the education	English Labels	Categories	EDLVAFR Dataset: ESS3.sav Variable Values	
France 2006	1.00	Enseignement préélém	Pre-school education	0	01 Sans diplôme	0
	2.00	Enseignement primaire	Primary education	1	04 Certificat d'études primaires	1
	3.00	Enseignement du prem	Secondary education 1st cycle	2A	07 Brevet élémentaire, brevet d'étude d	2A
	4.00	Enseignement dans le	Vocational training for young people without qualification secc	3C	05 CAP, examen de fin d'apprentissage	3A; 3C
	5.00	Enseignement de secc	Secondary education 2nd cycle, vocational training under sch	3C	05 CAP, examen de fin d'apprentissage	3C
	6.00	Enseignement de secc	Secondary education 2nd cycle, vocational training under sch	3C	05 CAP, examen de fin d'apprentissage	3C
	7.00	Enseignement de secc	Secondary education 2nd cycle, vocational training, programs	3C	05 CAP, examen de fin d'apprentissage	3C
	8.00	Enseignement de secc	Secondary education 2nd cycle, vocational training, programs	3C	06 BEP, BP, BEA, BEC, BEI, BES	3A
	9.00	Enseignement de secc	Secondary education 2nd cycle, vocational training, second le	3C	05 CAP, examen de fin d'apprentissage	3A; 3C
	10.00	Enseignement de secc	Secondary education 2nd cycle, vocational training, second le	3C	05 CAP, examen de fin d'apprentissage	3A; 3C
	11.00	Enseignement des écc	Schools of health and social specific schools	3C	06 BEP, BP, BEA, BEC, BEI, BES	3A
	12.00	Enseignement de secc	Secondary education 2nd cycle, vocational training, second le	3C	06 BEP, BP, BEA, BEC, BEI, BES	3A
	13.00	Enseignement de secc	Secondary education 2nd cycle, vocational training, second le	3B	09 Brevet de technicien, baccalauréat	3A
	14.00	Enseignement de secc	Secondary education 2nd cycle, vocational training, second le	3B	09 Brevet de technicien, baccalauréat	3A
	15.00	Enseignement de secc	Secondary education 2nd cycle, general	3A	08 Baccalauréat général, brevet supér	3A
	16.00	Enseignement de secc	Secondary education 2nd cycle, technology	3A	08 Brevet de technicien, baccalauréat	3A
	17.00	Enseignement des écc	Schools of health and social specific schools	4C	09 Brevet de technicien, baccalauréat	3A
	18.00	Enseignement pré-univ	Pre-university education	4A	09 Brevet de technicien, baccalauréat	3A
	19.00	Enseignement dans le	Vocational training for young people without qualification level	5B	10 Diplôme universitaire du premier cyc	5A
	20.00	Enseignement en instit	Specific vocational training university	5B	10 Diplôme universitaire du premier cyc	5A
	21.00	Enseignement d'écoles	Courses in specialized higher schools short teaching, leading	5B	10 Diplôme universitaire du premier cyc	5A
	22.00	Enseignement des cla	Courses in the classes of the sections of high-level technicier	5B	10 Diplôme universitaire du premier cyc	5A
	23.00	Enseignement des cla	Courses in the classes of the sections of high-level technicier	5B	10 Diplôme universitaire du premier cyc	5A
	24.00	Enseignement des cla	Courses in the preparatory classes at "grandes écoles" speci	5A	10 Diplôme universitaire du premier cyc	5A
	25.00	Enseignement de prem	University education, 1st cycle	5A	10 Diplôme universitaire du premier cyc	5A
	26.00	Enseignement de deux	University education, 2nd cycle, 1st year	5A	11 Dipl. univ. du deuxième cycle, CAPE	5A
	27.00	Enseignement des écc	Higher engineering school	5A	11 Dipl. univ. du deuxième cycle, CAPE	5A
	28.00	Enseignement des écc	Higher business school	5A	11 Dipl. univ. du deuxième cycle, CAPE	5A
	29.00	Enseignement de deux	University education, 2nd cycle, 2nd year	5A	11 Dipl. univ. du deuxième cycle, CAPE	5A
	30.00	Diverses formations: ar	Various training: architect, veterinary surgeon, art, etc. Speci	5A	11 Dipl. univ. du deuxième cycle, CAPE	5A
	31.00	Enseignement en instit	Teaching in university institute of training of Masters universit	5A	11 Dipl. univ. du deuxième cycle, CAPE	5A
	32.00	Enseignement de trois	University education, 3rd cycle	5A	12 Diplôme universitaire du troisième cy	6
	33.00	Enseignement dans le	Teaching in universities with a pharmacy speciality	5A	11 Dipl. univ. du deuxième cycle, CAPE	5A
	34.00	Enseignement dans le	Teaching in universities with medicine and odonthology speci	5A	11 Dipl. univ. du deuxième cycle, CAPE	5A
	35.00	Enseignement de spéc	Teaching of health specialization	5A	11 Dipl. univ. du deuxième cycle, CAPE	5A
	36.00	Enseignement de trois	University education, 3rd cycle, doctorate	6	12 Diplôme universitaire du troisième cy	6
	37.00	Enseignement de trois	University education, 3rd cycle, 1st year	6	12 Diplôme universitaire du troisième cy	6

Example ISCED-97 - Mapping

gesis
Leibniz Institute for the Social Sciences

Paiwise comparisons: indicators- variables

Country specific variables	Country specific values	Country specific labels	ISCED-97: ESS coding* (variable: eduvl***)	ISCED-97: Final Wp9.2 Coding**	ISCED-97 First Coding	Comments (VM)	Comments (AA)	Comments (AK)
France (edlvafr)	1	Sans diplôme	0	0	0	o		
	2	Non diplômés jusqu'à la fin 3ème, 2de, 1ère filière générale	2	2	1	ISCED0 or ISCED1 ?; DataDoc (ESS3): "2) lower secondary or second stage of basic".		*categ 2. i would classify this isced2 as proposed by ESS since 2de 1ère belong to lower secondary. **The best would be 1;2 depending on the age: the older droped out before the end of lower secondary, the younger after.** It is possible to enter upper secondary without having BEPC or brevet des collèges since this diploma is not a prerequisite for access to upper secondary.
	3	Non diplômés du CAP BEP filière professionnelle	1	2	1	1, 2A, DataDoc (ESS3): 1; Weil 7 schon 2A ist, denke ich, hat 2 nur eine 1 verdient		*categ3 same remark. They can having nothing. the most important is the level of education here. In term of duration of study they have the same level as BEPC. The level is 2, but not 2A (this track lead to 3C). Contrat de qualification is something different, it is possible that they are classify in this category, but most people prepared CAP, even in that frame.
	4	Certificat d'études primaires	1	1	1	1	no certificate of CAP/BEP; does this automatticaly mean they have only BEPC?	
	5	CAP, examen de fin d'apprentissage artisanal	2	3C	3C	3C		*categ 5 depending on the age, they can be classified 2 or 3C. I would classify 3c.
	6	BEP, BP, BEA, BEC, BEI, BES	2	3A	3C	3C	5/11/2008: No clear distinction betwen subcategories was possible.	*categ 6 BEP is 3C like CAP, but BEA, BEC, BEI, BT have been tranformed in baccalauréat technologique. They concern only the older generation. Some survived recently. I would classify them 3A.
	7	Brevet élémentaire, brevet d'étude du premier cycle, brevet des colleges	3	2A	2A	2A		
	8	Baccalauréat général, brevet supérieur	4	3A	3A	3A		
	9	Brevet de technicien, baccalauréat de technicien,...	4	3A	3A	3A		
	10	Diplôme universitaire du premier cycle (DEUG, nouvelle licence)	5	5A	5A	5A		
	11	Dipl. univ. du deuxième cycle, CAPES, Diplôme des grandes écoles	5	5A	5A	5A		
	12	Diplôme universitaire du troisième cycle (DEA, DESS), Agrégation, Doctorat	6	6	6	6	no distinction between 5A; 5B; and 6- ISCED levels is possibile	

Example ISCED-97 - Mapping

Additional Commonalitytypes -“similarity” (Possible Examples):

Criteria: Content covered by variables’ codes

- Type 1: Part of desired codes (indicators) are missing in source variable -> an additional variable from the same data set is required for successful recoding

Criteria: Re-coding procedures

- Type 2: Collapsing of categories;
- Type 3: Inversing values
  (…)
- **Types of comparability:** DDI3 commonality type ‘**similar**’- is more exhaustive within **CHARMCATS**