How to manage big data (well)

A digitized process-chain

Inga Brentel
Heinrich Heine -
Universität Düsseldorf

CESSDA Training Day
27-28 November 2019, Cologne

cessda.eu @CESSDA_Data

cessda

How to manage big data (well) - a digitized process-chain

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

FORSCHUNGSVERBUND NRW
DIGITALE GESELLSCHAFT

Inga Brentel

This research was supported by the Digital Society research program funded by the Ministry of Culture and Science of the German State of North Rhine-Westphalia.

Ministerium für
Kultur und Wissenschaft
des Landes Nordrhein-Westfalen

Definition & Relevance

- Digitaliztion brings data tracks (of human behavior) = *datafication*
  - → ***Big-Data*** as unstructured, heterogeneous data bulk

- Exploitation and Management of data
  - → A huge chance for (social) science
  - → Nessecarity of high quality standards

Big-Data as a challenge & a chance

- *Exploitation of Big-Data* sources for CSS
  → setup of a new data infrastructure
- *Data management* for a sustainable access to research data of high quality
- *Technical interfaces* to combine process chains

- **Solution:** *digitized process chains for Big-Data exploitation and analysis*

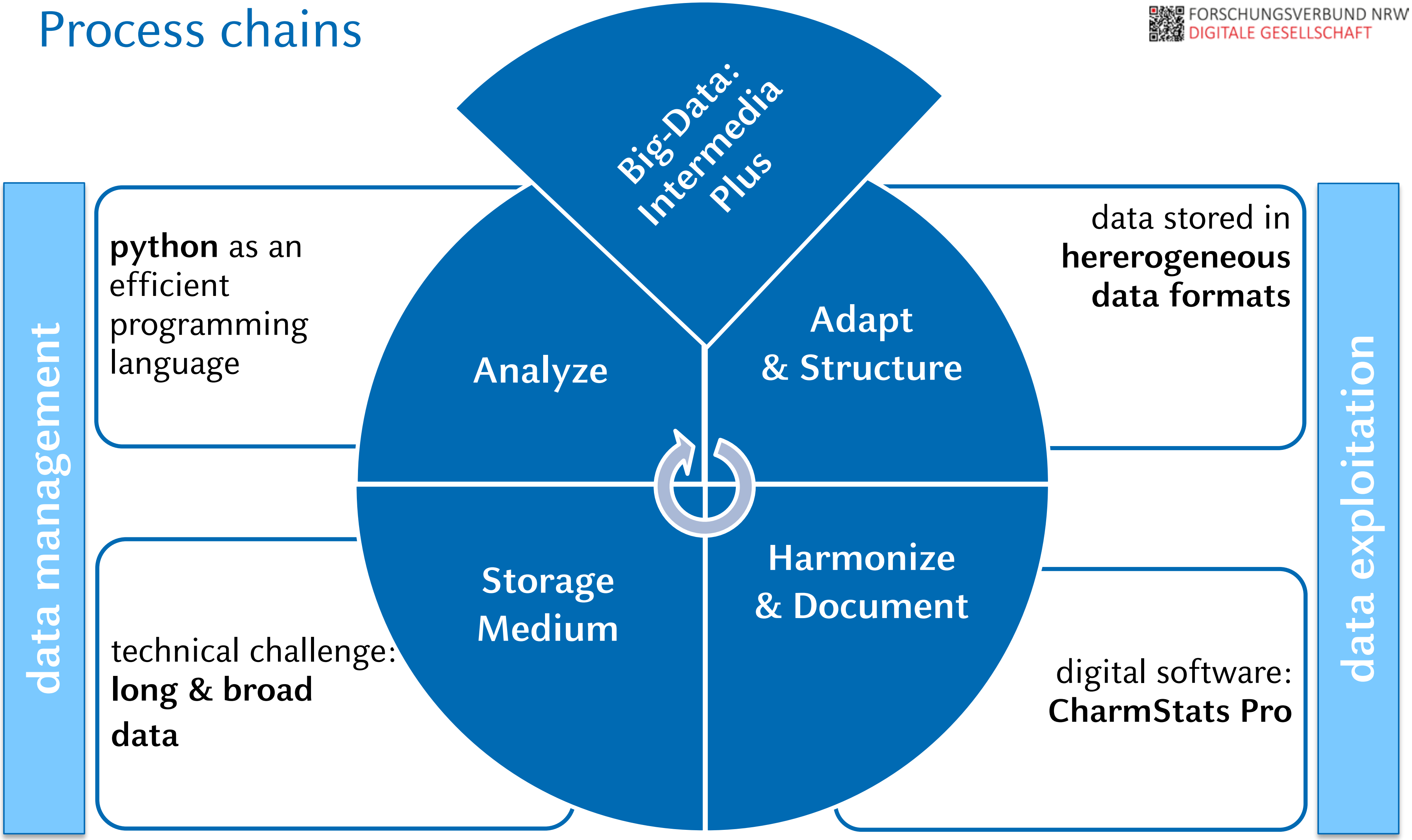
Case: MA IntermediaPlus data

Big-Data & Social Science

- Yearly gathered, cross sectional study for German media reach; conducted by commercial interests
- **Data typ:** survey & technical measures/tracking
- **Big-Data:** over 18.000 variables & 1,6 mio cases (for 2014 to 2016)
- **Content:**
  - Media reach for & media use of ~12.000 online offerings, 150 magazines, 100 newspapers, 100 radio channels, 10 tv-channels
  - Further info on respondents like socio demographics, daily routine, free-time activities, habits and household
- **Target:** *a structured, longitudinal dataset for academic research*

Solution Model I.

Process chains



data management

python as an efficient programming language

technical challenge: **long & broad data**

Big-Data: Intermedia Plus

Analyze

Adapt & Structure

Storage Medium

Harmonize & Document

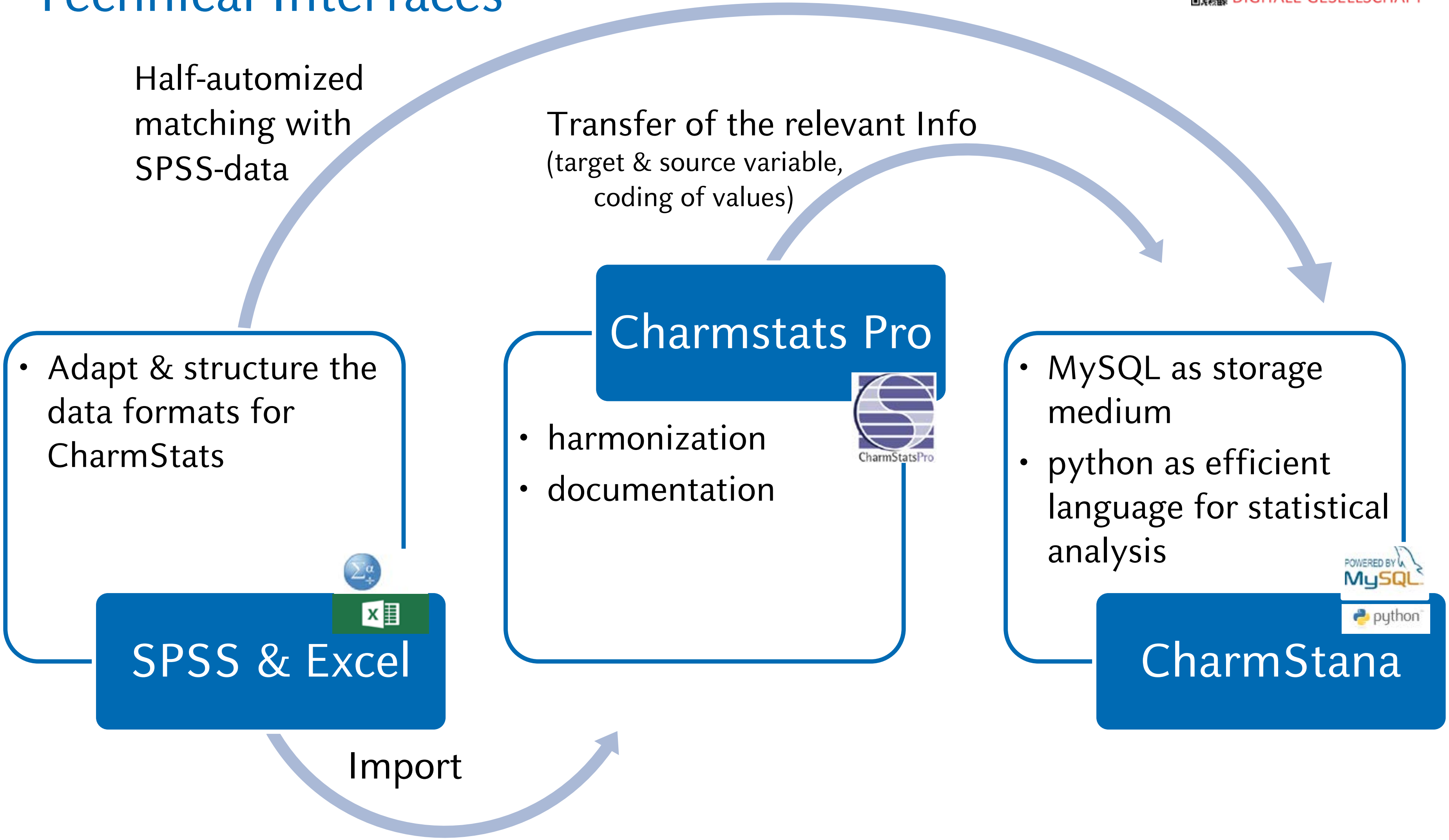
data stored in **hererogeneous data formats**

digital software: **CharmStats Pro**

data exploitation

Solution Model II.

Technical Interfaces



Half-automized matching with SPSS-data

Transfer of the relevant Info
(target & source variable,
coding of values)

- Adapt & structure the data formats for CharmStats

SPSS & Excel

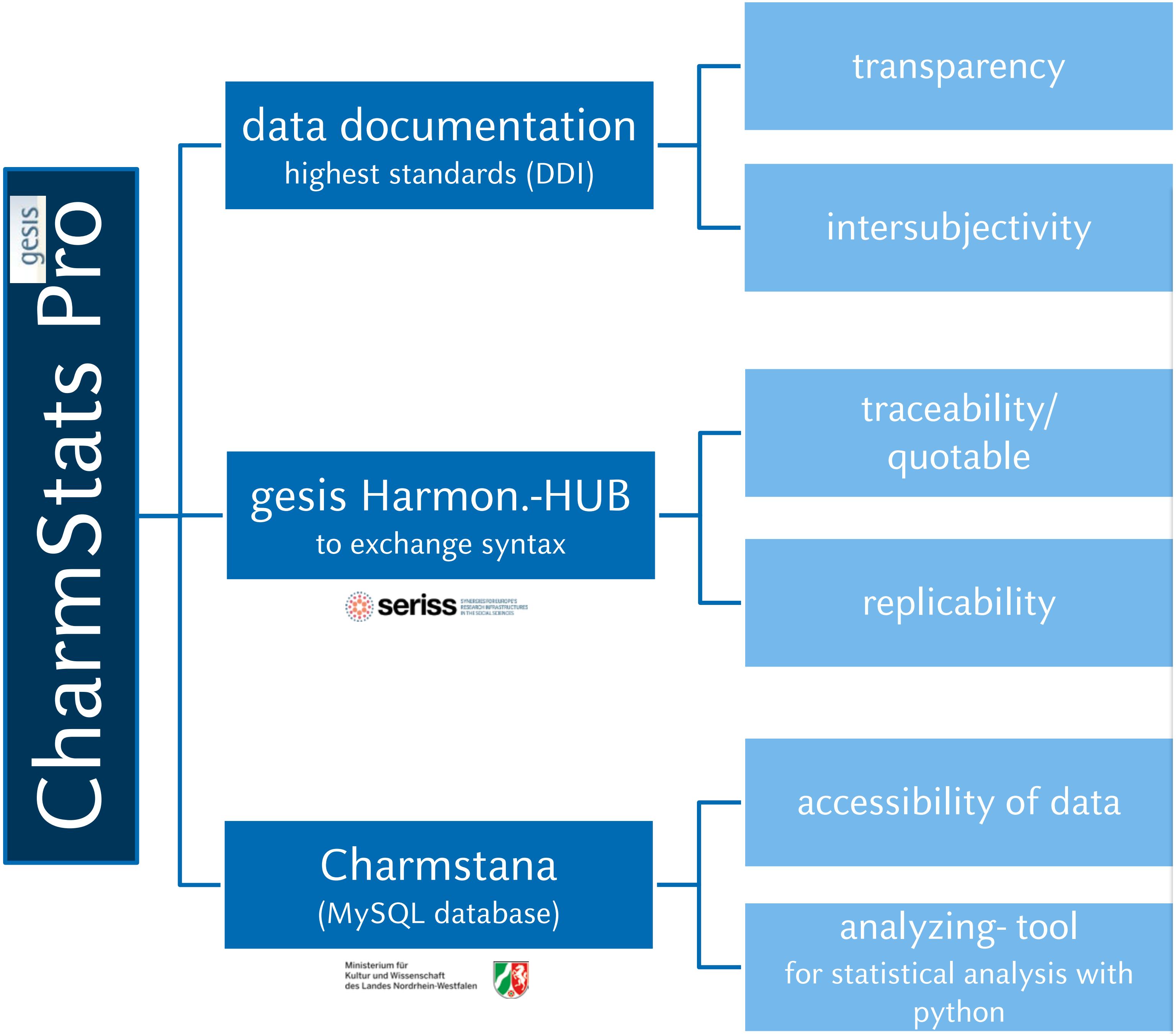
Charmstats Pro

- harmonization
- documentation

- MySQL as storage medium
- python as efficient language for statistical analysis

CharmStana

Import



gesis CharmStats Pro

data documentation
highest standards (DDI)

- transparency
- intersubjectivity

gesis Harmon.-HUB
to exchange syntax

seriss SYNERGIES FOR EUROPE'S RESEARCH INFRASTRUCTURES IN THE SOCIAL SCIENCES

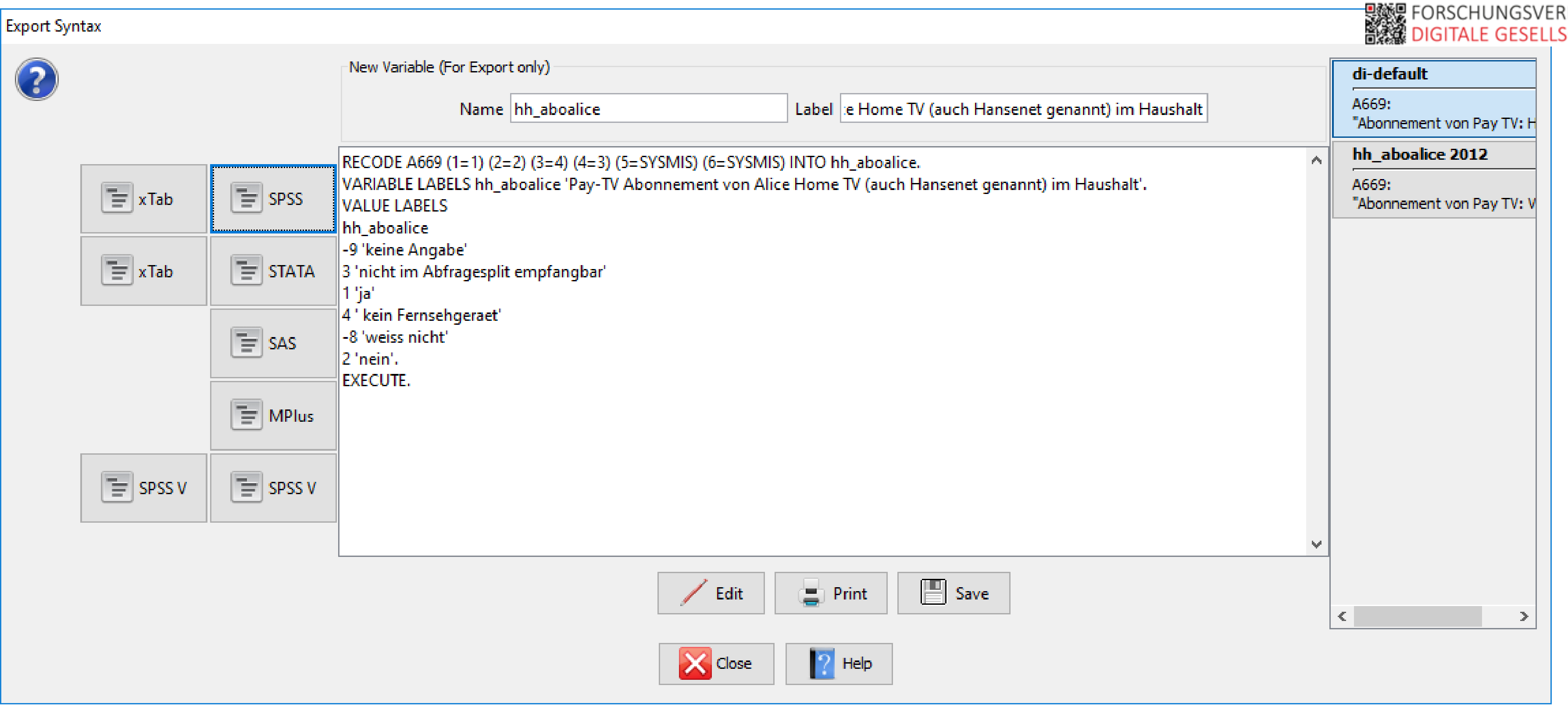
- traceability/ quotable
- replicability

Charmstana
(MySQL database)

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen

- accessibility of data
- analyzing- tool for statistical analysis with python

Interface I.: Data documentation



New Variable (For Export only)

Name: hh_aboalice

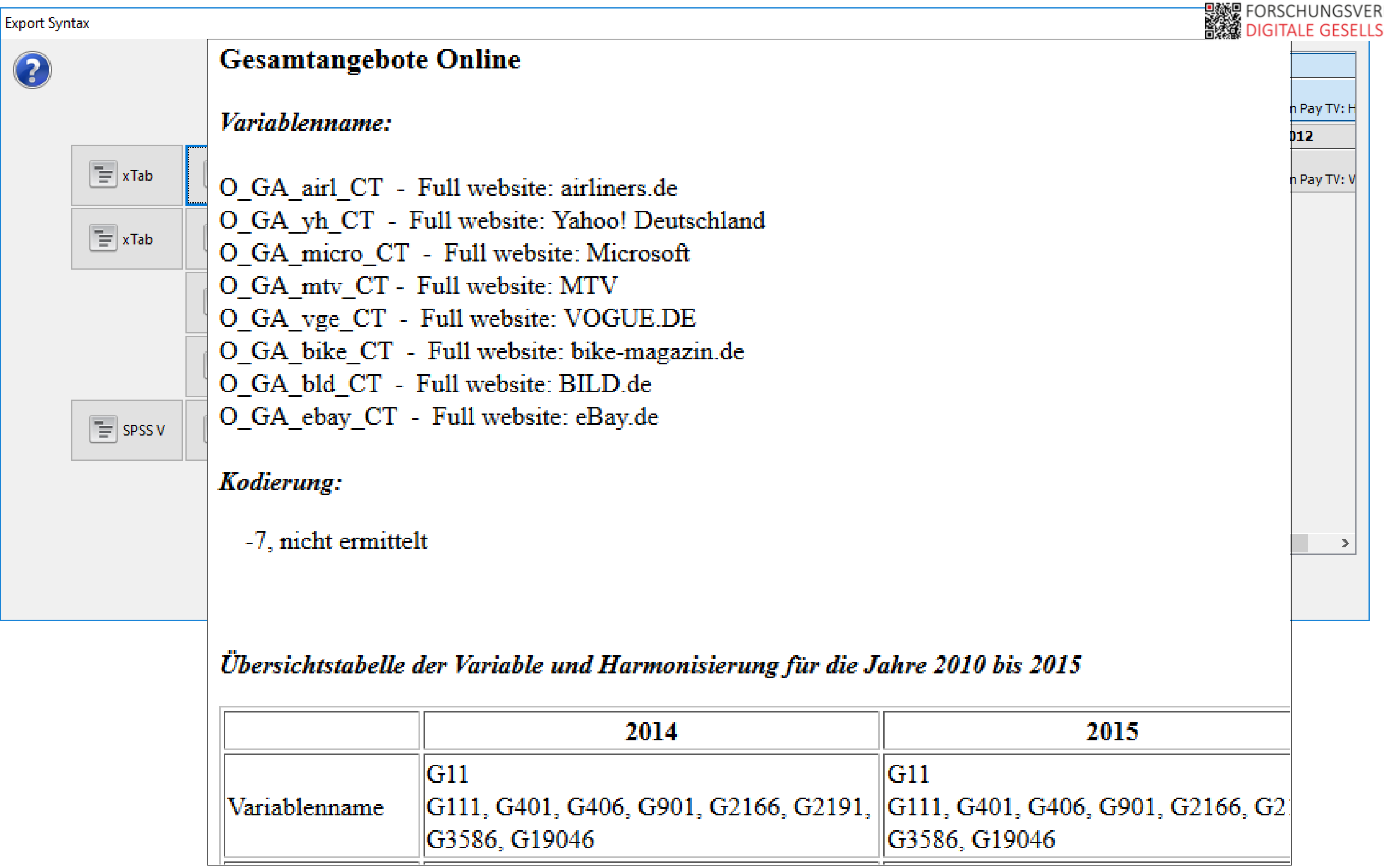
Label: e Home TV (auch Hansenet genannt) im Haushalt

```
RECODE A669 (1=1) (2=2) (3=4) (4=3) (5=SYSMIS) (6=SYSMIS) INTO hh_aboalice.
VARIABLE LABELS hh_aboalice 'Pay-TV Abonnement von Alice Home TV (auch Hansenet genannt) im Haushalt'.
VALUE LABELS
hh_aboalice
-9 'keine Angabe'
3 'nicht im Abfragesplit empfangbar'
1 'ja'
4 ' kein Fernsehgeraet'
-8 'weiss nicht'
2 'nein'.
EXECUTE.
```

di-default
A669:
"Abonnement von Pay TV: H

hh_aboalice 2012
A669:
"Abonnement von Pay TV: V

Interface I.: Data documentation



Gesamtangebote Online

Variablenname:

O_GA_airl_CT - Full website: airliners.de
O_GA_yh_CT - Full website: Yahoo! Deutschland
O_GA_micro_CT - Full website: Microsoft
O_GA_mtv_CT - Full website: MTV
O_GA_vge_CT - Full website: VOGUE.DE
O_GA_bike_CT - Full website: bike-magazin.de
O_GA_bld_CT - Full website: BILD.de
O_GA_ebay_CT - Full website: eBay.de

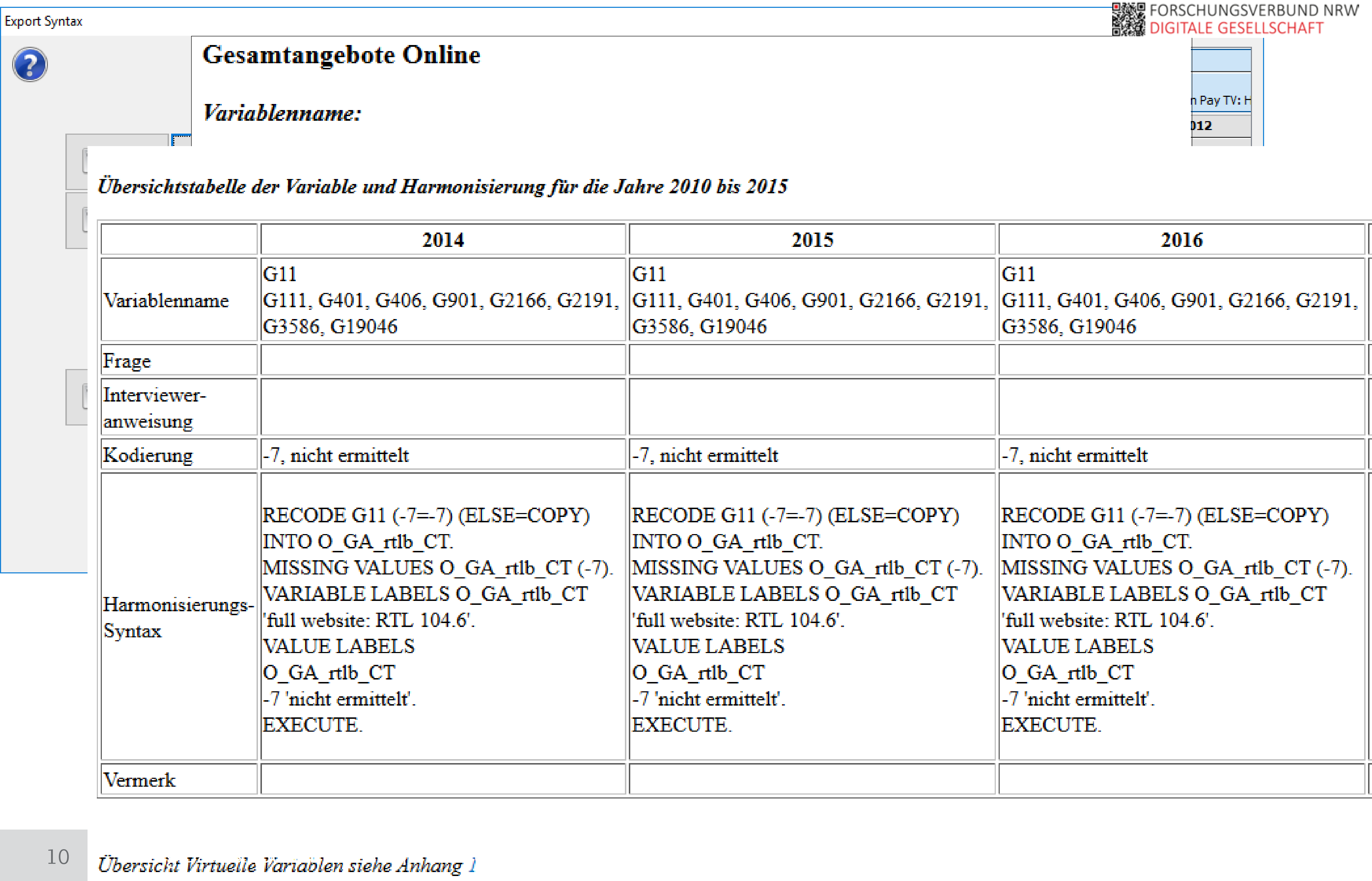
Kodierung:

-7, nicht ermittelt

Übersichtstabelle der Variable und Harmonisierung für die Jahre 2010 bis 2015

	2014	**2015**
Variablenname	G11 G111, G401, G406, G901, G2166, G2191, G3586, G19046	G11 G111, G401, G406, G901, G2166, G2 G3586, G19046

Interface I.: Data documentation



Gesamtangebote Online

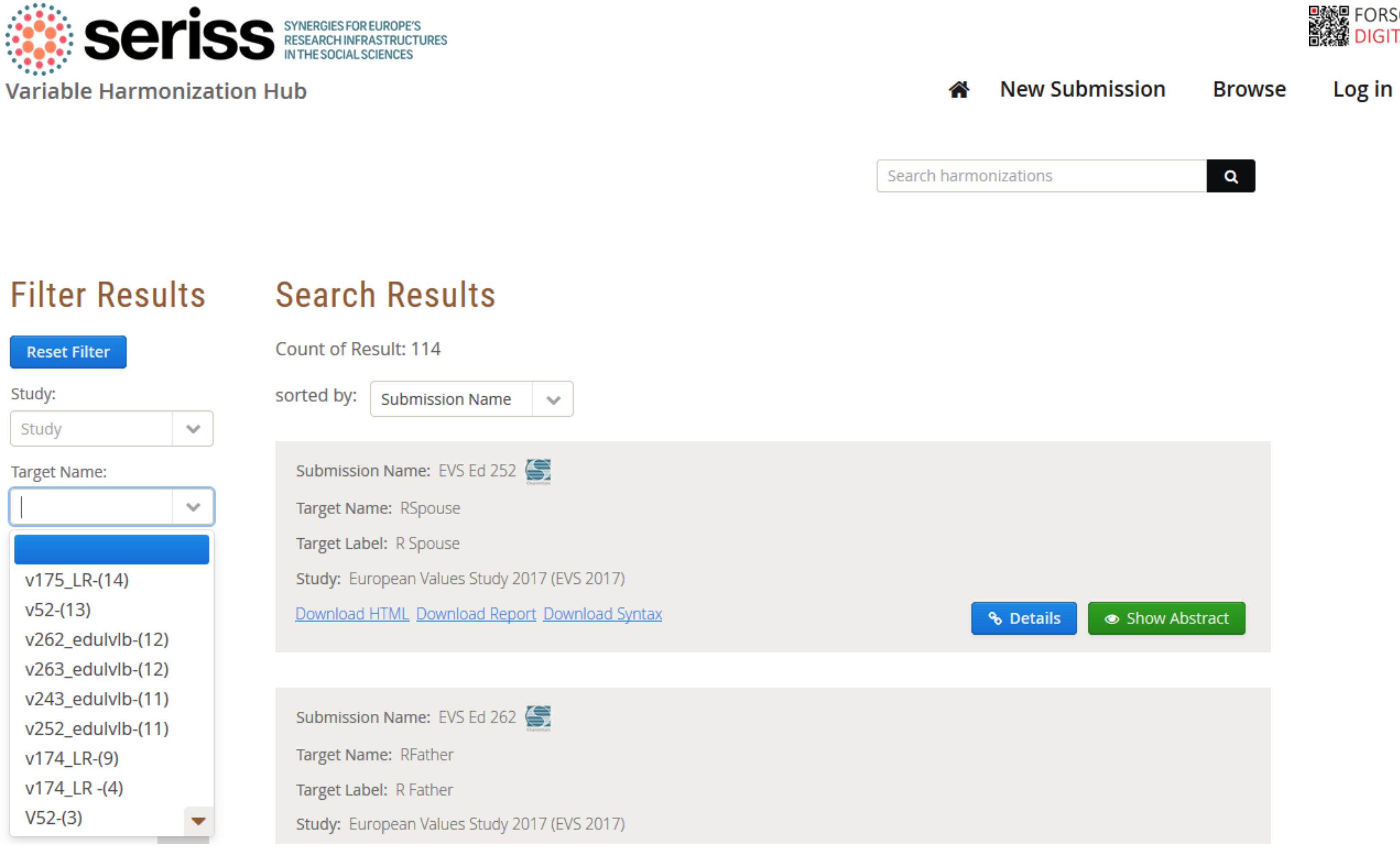
Variablenname:

Übersichtstabelle der Variable und Harmonisierung für die Jahre 2010 bis 2015

	2014	2015	2016
Variablenname	G11 G111, G401, G406, G901, G2166, G2191, G3586, G19046	G11 G111, G401, G406, G901, G2166, G2191, G3586, G19046	G11 G111, G401, G406, G901, G2166, G2191, G3586, G19046
Frage			
Interviewer-anweisung			
Kodierung	-7, nicht ermittelt	-7, nicht ermittelt	-7, nicht ermittelt
Harmonisierungs-Syntax	RECODE G11 (-7=-7) (ELSE=COPY) INTO O_GA_rtlb_CT. MISSING VALUES O_GA_rtlb_CT (-7). VARIABLE LABELS O_GA_rtlb_CT 'full website: RTL 104.6'. VALUE LABELS O_GA_rtlb_CT -7 'nicht ermittelt'. EXECUTE.	RECODE G11 (-7=-7) (ELSE=COPY) INTO O_GA_rtlb_CT. MISSING VALUES O_GA_rtlb_CT (-7). VARIABLE LABELS O_GA_rtlb_CT 'full website: RTL 104.6'. VALUE LABELS O_GA_rtlb_CT -7 'nicht ermittelt'. EXECUTE.	RECODE G11 (-7=-7) (ELSE=COPY) INTO O_GA_rtlb_CT. MISSING VALUES O_GA_rtlb_CT (-7). VARIABLE LABELS O_GA_rtlb_CT 'full website: RTL 104.6'. VALUE LABELS O_GA_rtlb_CT -7 'nicht ermittelt'. EXECUTE.
Vermerk			

Übersicht Virtuelle Variablen siehe Anhang 1

Interface II.: gesis Harmonization-HUB



Variable Harmonization Hub

seriss
SYNERGIES FOR EUROPE'S RESEARCH INFRASTRUCTURES IN THE SOCIAL SCIENCES
New Submission
Browse
Log in
Search harmonizations

Filter Results

Study: Study

Target Name:

- v175_LR-(14)
- v52-(13)
- v262_edulvlb-(12)
- v263_edulvlb-(12)
- v243_edulvlb-(11)
- v252_edulvlb-(11)
- v174_LR-(9)
- v174_LR -(4)
- V52-(3)

Search Results

Count of Result: 114

sorted by: Submission Name

Submission Name: EVS Ed 252

Target Name: RSpouse

Target Label: R Spouse

Study: European Values Study 2017 (EVS 2017)

Download HTML Download Report Download Syntax

Submission Name: EVS Ed 262

Target Name: RFather

Target Label: R Father

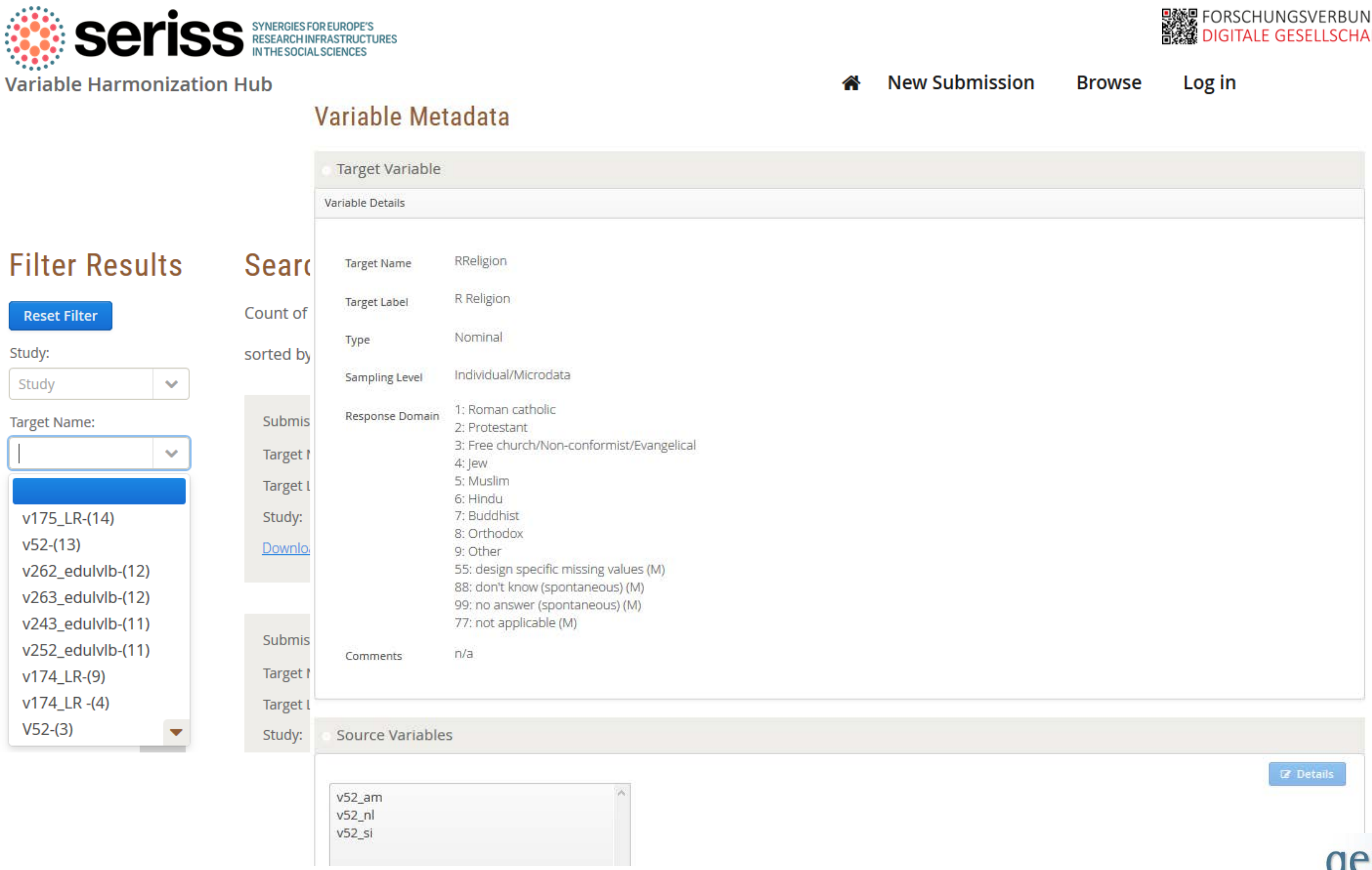
Study: European Values Study 2017 (EVS 2017)

FORSCHUNGSVERBUND NRW
DIGITALE GESELLSCHAFT

Interface II.: gesis Harmonization-HUB

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

FORSCHUNGSVERBUND NRW
DIGITALE GESELLSCHAFT



seriss SYNERGIES FOR EUROPE'S RESEARCH INFRASTRUCTURES IN THE SOCIAL SCIENCES

Variable Harmonization Hub

Variable Metadata

Target Variable

Variable Details

Target Name	RReligion
Target Label	R Religion
Type	Nominal
Sampling Level	Individual/Microdata
Response Domain	1: Roman catholic 2: Protestant 3: Free church/Non-conformist/Evangelical 4: Jew 5: Muslim 6: Hindu 7: Buddhist 8: Orthodox 9: Other 55: design specific missing values (M) 88: don't know (spontaneous) (M) 99: no answer (spontaneous) (M) 77: not applicable (M)
Comments	n/a

Source Variables

v52_am
v52_nl
v52_si

Filter Results

Study:

Study

Target Name:

v175_LR-(14)
v52-(13)
v262_edulvlb-(12)
v263_edulvlb-(12)
v243_edulvlb-(11)
v252_edulvlb-(11)
v174_LR-(9)
v174_LR -(4)
V52-(3)

Searc

Count of

sorted by

Submis

Target N

Target L

Study:

Downlo

Submis

Target N

Target L

Study:

gesis

Interface III.: Charmstana

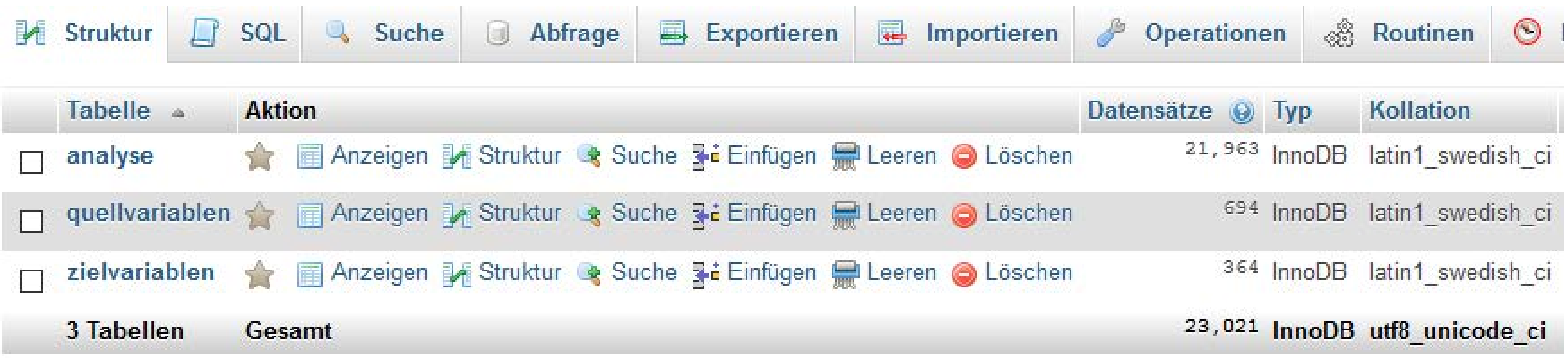
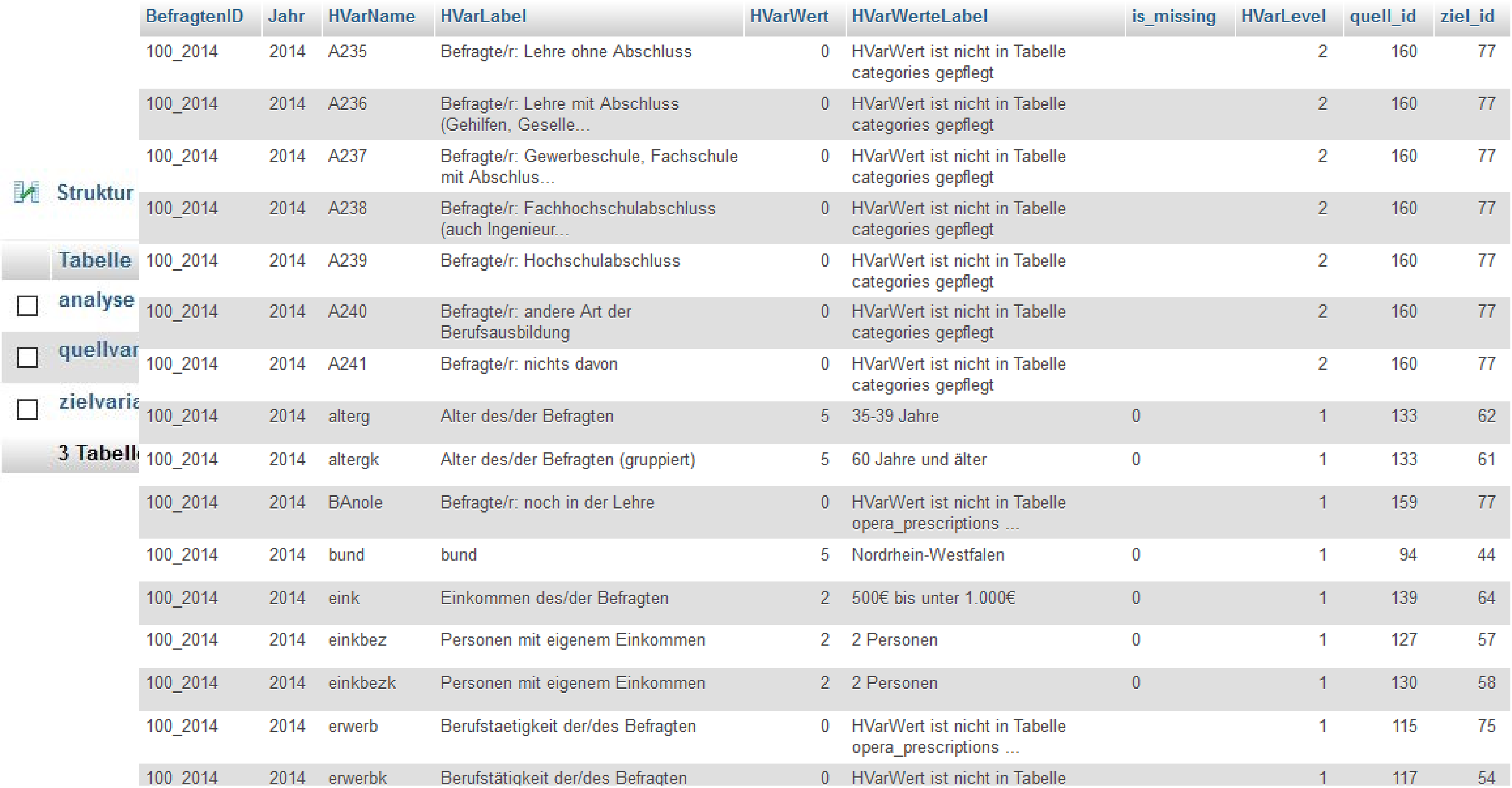


	Tabelle	Aktion	Datensätze	Typ	Kollation
□	analyse	☆ Anzeigen Struktur Suche Einfügen Leeren Löschen	21,963	InnoDB	latin1_swedish_ci
□	quellvariablen	☆ Anzeigen Struktur Suche Einfügen Leeren Löschen	694	InnoDB	latin1_swedish_ci
□	zielvariablen	☆ Anzeigen Struktur Suche Einfügen Leeren Löschen	364	InnoDB	latin1_swedish_ci
	3 Tabellen	Gesamt	23,021	InnoDB	utf8_unicode_ci

Entwickelt mit
Tanja Roeder (BA)

Interface III.: Charmstana



BefragtenID	Jahr	HVarName	HVarLabel	HVarWert	HVarWerteLabel	is_missing	HVarLevel	quell_id	ziel_id
100_2014	2014	A235	Befragte/r: Lehre ohne Abschluss	0	HVarWert ist nicht in Tabelle categories gepflegt		2	160	77
100_2014	2014	A236	Befragte/r: Lehre mit Abschluss (Gehilfen, Geselle...	0	HVarWert ist nicht in Tabelle categories gepflegt		2	160	77
100_2014	2014	A237	Befragte/r: Gewerbeschule, Fachschule mit Abschlus...	0	HVarWert ist nicht in Tabelle categories gepflegt		2	160	77
100_2014	2014	A238	Befragte/r: Fachhochschulabschluss (auch Ingenieur...	0	HVarWert ist nicht in Tabelle categories gepflegt		2	160	77
100_2014	2014	A239	Befragte/r: Hochschulabschluss	0	HVarWert ist nicht in Tabelle categories gepflegt		2	160	77
100_2014	2014	A240	Befragte/r: andere Art der Berufsausbildung	0	HVarWert ist nicht in Tabelle categories gepflegt		2	160	77
100_2014	2014	A241	Befragte/r: nichts davon	0	HVarWert ist nicht in Tabelle categories gepflegt		2	160	77
100_2014	2014	alterg	Alter des/der Befragten	5	35-39 Jahre	0	1	133	62
100_2014	2014	altergk	Alter des/der Befragten (gruppiert)	5	60 Jahre und älter	0	1	133	61
100_2014	2014	BAnole	Befragte/r: noch in der Lehre	0	HVarWert ist nicht in Tabelle opera_prescriptions ...		1	159	77
100_2014	2014	bund	bund	5	Nordrhein-Westfalen	0	1	94	44
100_2014	2014	eink	Einkommen des/der Befragten	2	500€ bis unter 1.000€	0	1	139	64
100_2014	2014	einkbez	Personen mit eigenem Einkommen	2	2 Personen	0	1	127	57
100_2014	2014	einkbezk	Personen mit eigenem Einkommen	2	2 Personen	0	1	130	58
100_2014	2014	erwerb	Berufstaetigkeit der/des Befragten	0	HVarWert ist nicht in Tabelle opera_prescriptions ...		1	115	75
100_2014	2014	erwerbk	Berufstätigkeit der/des Befragten	0	HVarWert ist nicht in Tabelle		1	117	54

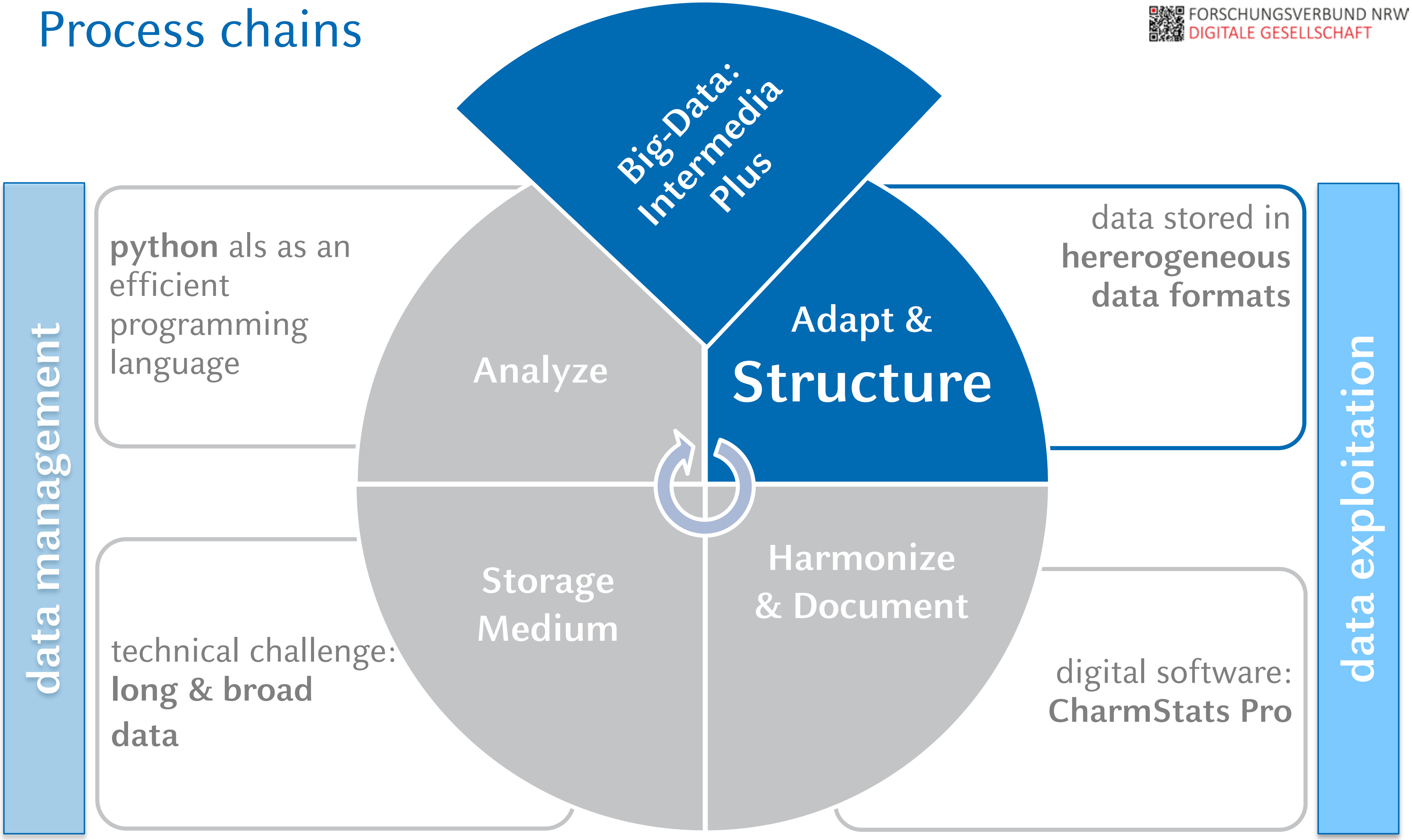
Entwickelt mit
Tanja Roeder (BA)

DEEP-DIVE

How to structure (Big-)Data?

Solution Model I.

Process chains



python als as an efficient programming language

data stored in **hererogeneous data formats**

technical challenge: **long & broad data**

digital software: **CharmStats Pro**

Media-Analysis Data: Intermedia Plus

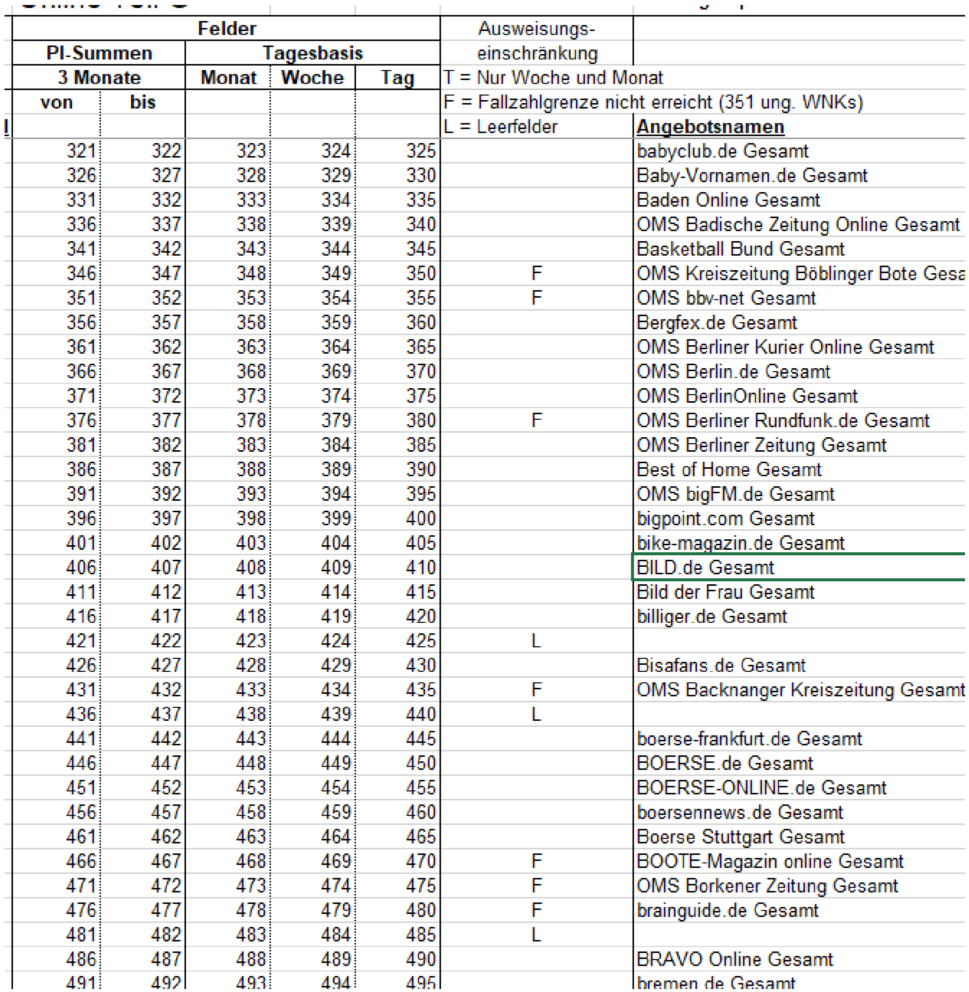
Variables for media use

- 100 radio broadcasts, usage per hour
- 150 magazins & 100 newspapers (incl. local newspapers)
- 10 TV-channels, usage per 30 minutes
- ~ 3000 Online sites (net coverage on a daily, weekly, monthly basis; cross coverage on a monthly and quarter basis)
  - ~ 733 overall online media
  - 335 single online sites
  - 52 homepages

→ **Goal:** harmonization of Intermedia Plus 2014 to 2017+

Step 1: Find a structure

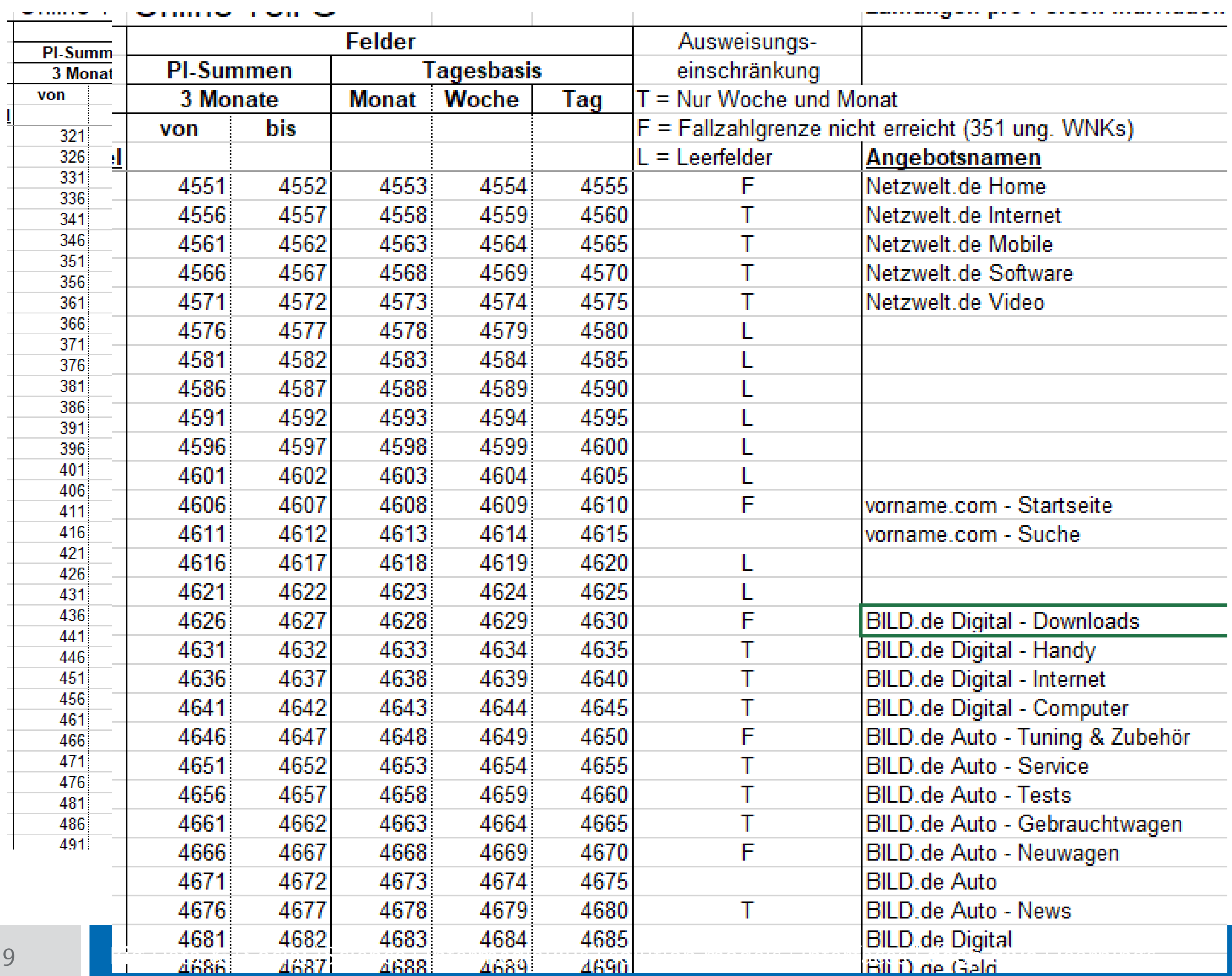
The case: MA Intermedia Plus, online-tranche



Felder					Ausweisungs-	
PI-Summen		Tagesbasis			einschränkung	
3 Monate		Monat	Woche	Tag	T = Nur Woche und Monat	
von	bis				F = Fallzahlgrenze nicht erreicht (351 ung. WNKs)	
					L = Leerfelder	**Angebotsnamen**
321	322	323	324	325		babyclub.de Gesamt
326	327	328	329	330		Baby-Vornamen.de Gesamt
331	332	333	334	335		Baden Online Gesamt
336	337	338	339	340		OMS Badische Zeitung Online Gesamt
341	342	343	344	345		Basketball Bund Gesamt
346	347	348	349	350	F	OMS Kreiszeitung Böblinger Bote Gesa
351	352	353	354	355	F	OMS bbv-net Gesamt
356	357	358	359	360		Bergfex.de Gesamt
361	362	363	364	365		OMS Berliner Kurier Online Gesamt
366	367	368	369	370		OMS Berlin.de Gesamt
371	372	373	374	375		OMS BerlinOnline Gesamt
376	377	378	379	380	F	OMS Berliner Rundfunk.de Gesamt
381	382	383	384	385		OMS Berliner Zeitung Gesamt
386	387	388	389	390		Best of Home Gesamt
391	392	393	394	395		OMS bigFM.de Gesamt
396	397	398	399	400		bigpoint.com Gesamt
401	402	403	404	405		bike-magazin.de Gesamt
406	407	408	409	410		BILD.de Gesamt
411	412	413	414	415		Bild der Frau Gesamt
416	417	418	419	420		billiger.de Gesamt
421	422	423	424	425	L	
426	427	428	429	430		Bisafans.de Gesamt
431	432	433	434	435	F	OMS Backnanger Kreiszeitung Gesamt
436	437	438	439	440	L	
441	442	443	444	445		boerse-frankfurt.de Gesamt
446	447	448	449	450		BOERSE.de Gesamt
451	452	453	454	455		BOERSE-ONLINE.de Gesamt
456	457	458	459	460		boersennews.de Gesamt
461	462	463	464	465		Boerse Stuttgart Gesamt
466	467	468	469	470	F	BOOTE-Magazin online Gesamt
471	472	473	474	475	F	OMS Borkener Zeitung Gesamt
476	477	478	479	480	F	brainguide.de Gesamt
481	482	483	484	485	L	
486	487	488	489	490		BRAVO Online Gesamt
491	492	493	494	495		bremen.de Gesamt

Step 1: Find a structure

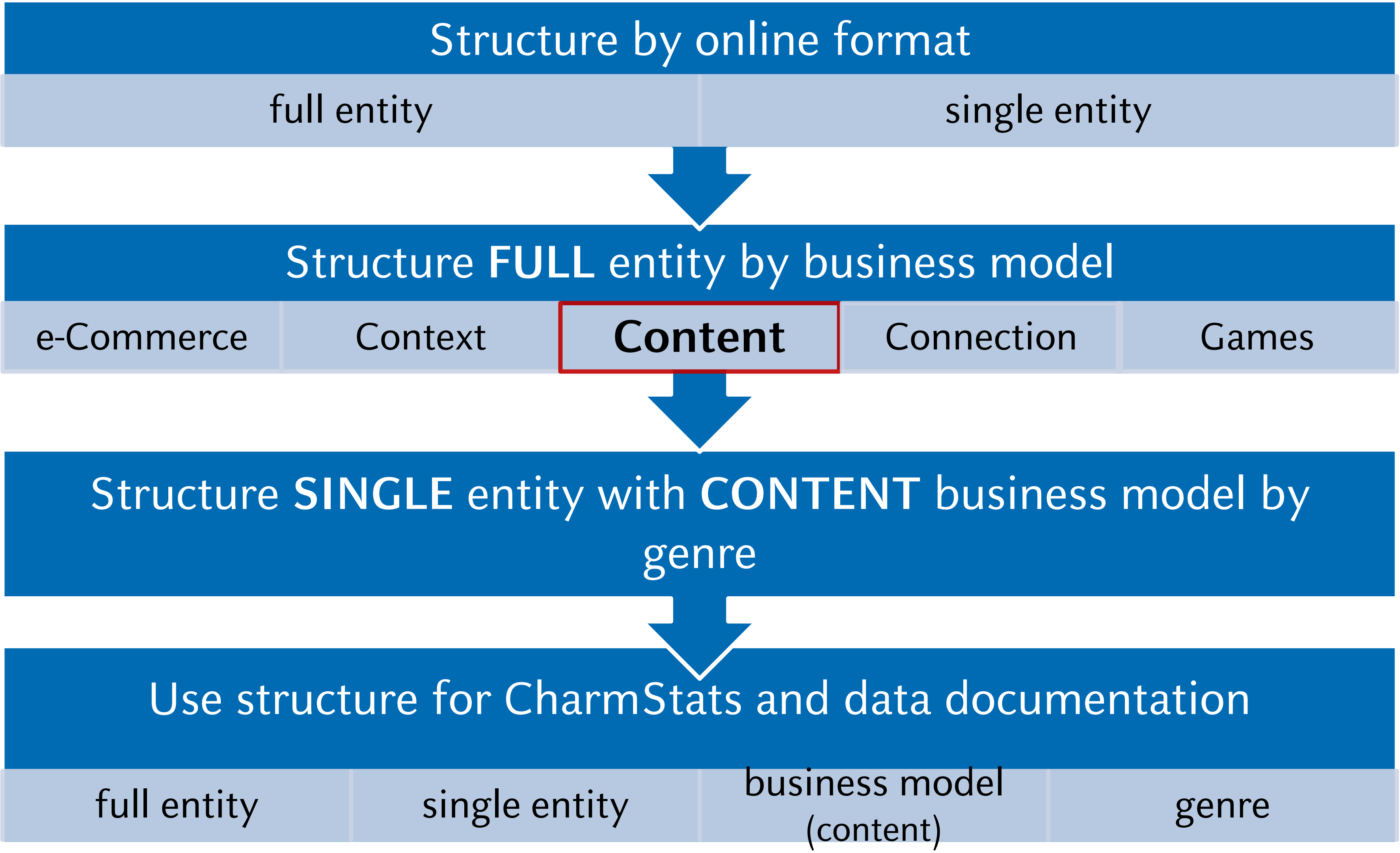
The case: MA Intermedia Plus, online-tranche



PI-Summ
3 Monat
von
321
326
331
336
341
346
351
356
361
366
371
376
381
386
391
396
401
406
411
416
421
426
431
436
441
446
451
456
461
466
471
476
481
486
491

Felder					Ausweisungs-einschränkung	
PI-Summen		Tagesbasis				
3 Monate		Monat	Woche	Tag	T = Nur Woche und Monat	
von	bis				F = Fallzahlgrenze nicht erreicht (351 ung. WNKs)	
					L = Leerfelder	**Angebotsnamen**
4551	4552	4553	4554	4555	F	Netzwelt.de Home
4556	4557	4558	4559	4560	T	Netzwelt.de Internet
4561	4562	4563	4564	4565	T	Netzwelt.de Mobile
4566	4567	4568	4569	4570	T	Netzwelt.de Software
4571	4572	4573	4574	4575	T	Netzwelt.de Video
4576	4577	4578	4579	4580	L	
4581	4582	4583	4584	4585	L	
4586	4587	4588	4589	4590	L	
4591	4592	4593	4594	4595	L	
4596	4597	4598	4599	4600	L	
4601	4602	4603	4604	4605	L	
4606	4607	4608	4609	4610	F	vorname.com - Startseite
4611	4612	4613	4614	4615		vorname.com - Suche
4616	4617	4618	4619	4620	L	
4621	4622	4623	4624	4625	L	
4626	4627	4628	4629	4630	F	BILD.de Digital - Downloads
4631	4632	4633	4634	4635	T	BILD.de Digital - Handy
4636	4637	4638	4639	4640	T	BILD.de Digital - Internet
4641	4642	4643	4644	4645	T	BILD.de Digital - Computer
4646	4647	4648	4649	4650	F	BILD.de Auto - Tuning & Zubehör
4651	4652	4653	4654	4655	T	BILD.de Auto - Service
4656	4657	4658	4659	4660	T	BILD.de Auto - Tests
4661	4662	4663	4664	4665	T	BILD.de Auto - Gebrauchtwagen
4666	4667	4668	4669	4670	F	BILD.de Auto - Neuwagen
4671	4672	4673	4674	4675		BILD.de Auto
4676	4677	4678	4679	4680	T	BILD.de Auto - News
4681	4682	4683	4684	4685		BILD.de Digital
4686	4687	4688	4689	4690		BILD.de Geld

Step 1: Find a structure



Structure by online format

full entity | single entity

Structure **FULL** entity by business model

e-Commerce | Context | **Content** | Connection | Games

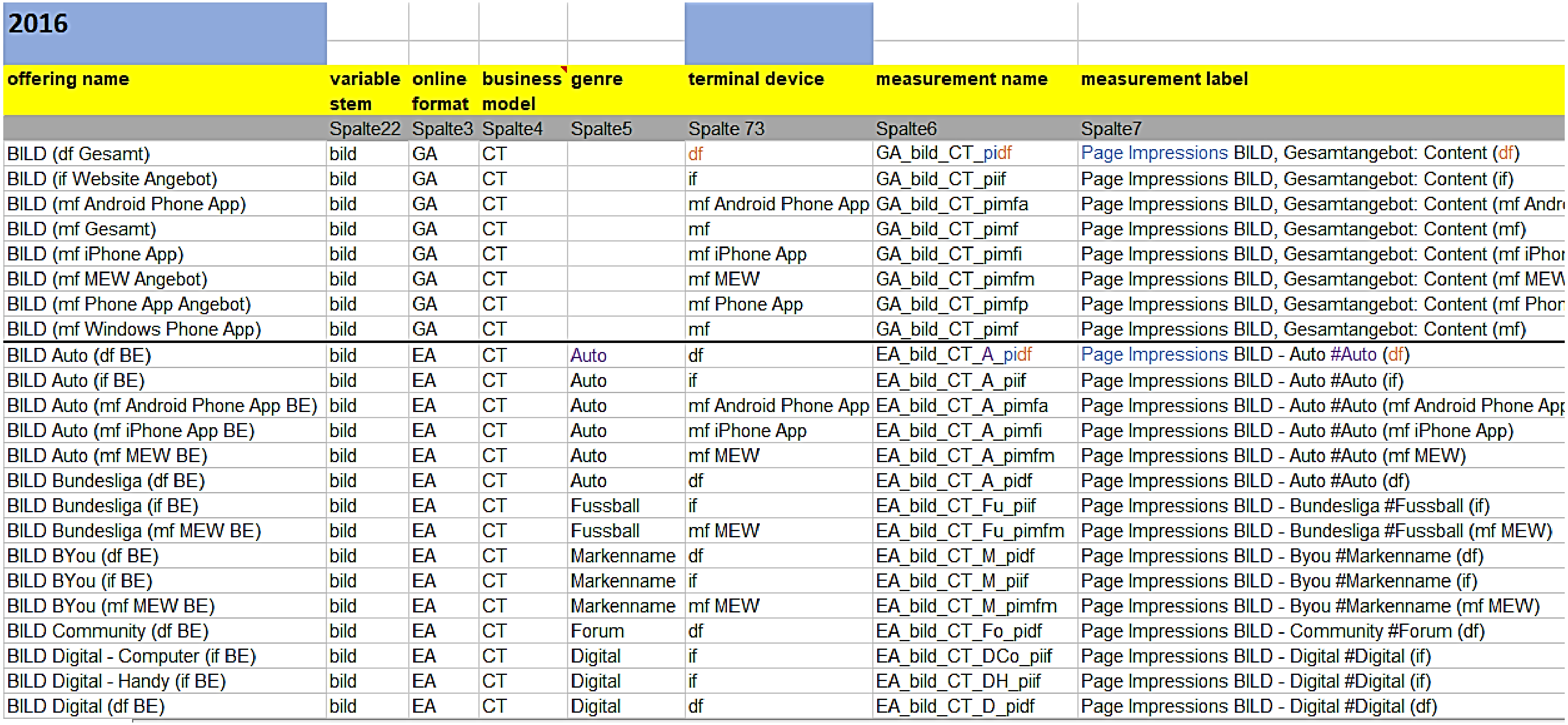
Structure **SINGLE** entity with **CONTENT** business model by genre

Use structure for CharmStats and data documentation

full entity | single entity | business model (content) | genre

Step 1: Find a structure

The case: MA Intermedia Plus, online-tranche



2016							
offering name	variable stem	online format	business model	genre	terminal device	measurement name	measurement label
	Spalte22	Spalte3	Spalte4	Spalte5	Spalte 73	Spalte6	Spalte7
BILD (df Gesamt)	bild	GA	CT		df	GA_bild_CT_pidf	Page Impressions BILD, Gesamtangebot: Content (df)
BILD (if Website Angebot)	bild	GA	CT		if	GA_bild_CT_piif	Page Impressions BILD, Gesamtangebot: Content (if)
BILD (mf Android Phone App)	bild	GA	CT		mf Android Phone App	GA_bild_CT_pimfa	Page Impressions BILD, Gesamtangebot: Content (mf Andr
BILD (mf Gesamt)	bild	GA	CT		mf	GA_bild_CT_pimf	Page Impressions BILD, Gesamtangebot: Content (mf)
BILD (mf iPhone App)	bild	GA	CT		mf iPhone App	GA_bild_CT_pimfi	Page Impressions BILD, Gesamtangebot: Content (mf iPhon
BILD (mf MEW Angebot)	bild	GA	CT		mf MEW	GA_bild_CT_pimfm	Page Impressions BILD, Gesamtangebot: Content (mf MEW
BILD (mf Phone App Angebot)	bild	GA	CT		mf Phone App	GA_bild_CT_pimfp	Page Impressions BILD, Gesamtangebot: Content (mf Phon
BILD (mf Windows Phone App)	bild	GA	CT		mf	GA_bild_CT_pimf	Page Impressions BILD, Gesamtangebot: Content (mf)
BILD Auto (df BE)	bild	EA	CT	Auto	df	EA_bild_CT_A_pidf	Page Impressions BILD - Auto #Auto (df)
BILD Auto (if BE)	bild	EA	CT	Auto	if	EA_bild_CT_A_piif	Page Impressions BILD - Auto #Auto (if)
BILD Auto (mf Android Phone App BE)	bild	EA	CT	Auto	mf Android Phone App	EA_bild_CT_A_pimfa	Page Impressions BILD - Auto #Auto (mf Android Phone App
BILD Auto (mf iPhone App BE)	bild	EA	CT	Auto	mf iPhone App	EA_bild_CT_A_pimfi	Page Impressions BILD - Auto #Auto (mf iPhone App)
BILD Auto (mf MEW BE)	bild	EA	CT	Auto	mf MEW	EA_bild_CT_A_pimfm	Page Impressions BILD - Auto #Auto (mf MEW)
BILD Bundesliga (df BE)	bild	EA	CT	Auto	df	EA_bild_CT_A_pidf	Page Impressions BILD - Auto #Auto (df)
BILD Bundesliga (if BE)	bild	EA	CT	Fussball	if	EA_bild_CT_Fu_piif	Page Impressions BILD - Bundesliga #Fussball (if)
BILD Bundesliga (mf MEW BE)	bild	EA	CT	Fussball	mf MEW	EA_bild_CT_Fu_pimfm	Page Impressions BILD - Bundesliga #Fussball (mf MEW)
BILD BYou (df BE)	bild	EA	CT	Markenname	df	EA_bild_CT_M_pidf	Page Impressions BILD - Byou #Markenname (df)
BILD BYou (if BE)	bild	EA	CT	Markenname	if	EA_bild_CT_M_piif	Page Impressions BILD - Byou #Markenname (if)
BILD BYou (mf MEW BE)	bild	EA	CT	Markenname	mf MEW	EA_bild_CT_M_pimfm	Page Impressions BILD - Byou #Markenname (mf MEW)
BILD Community (df BE)	bild	EA	CT	Forum	df	EA_bild_CT_Fo_pidf	Page Impressions BILD - Community #Forum (df)
BILD Digital - Computer (if BE)	bild	EA	CT	Digital	if	EA_bild_CT_DCo_piif	Page Impressions BILD - Digital #Digital (if)
BILD Digital - Handy (if BE)	bild	EA	CT	Digital	if	EA_bild_CT_DH_piif	Page Impressions BILD - Digital #Digital (if)
BILD Digital (df BE)	bild	EA	CT	Digital	df	EA_bild_CT_D_pidf	Page Impressions BILD - Digital #Digital (df)

What to think of ***before*** harmonizing (Big-)Data

Structure your (Big-) Data *smart*

- What is my research interest?
- What info in my data can I use as structure?
- What info do I need/want to add my data anyway?
- How should the final result look like?

→ CharmStats helps providing a formal structure

How to manage Big-Data (well)

- Data exploitation is relevant for social science
  - New data sources through digitalization
  - Need for a set up of digital-data infrastucture to support secondary data analysis and re-analysis

- Data management as key to Big-Data in social science, follwoing the quality standards of academia:
  - transparency,
  - intersubjectivity,
  - replicability, sustainability, etc.

→ **Solution:** *digitized process chains*

BACKUP-SLIDES

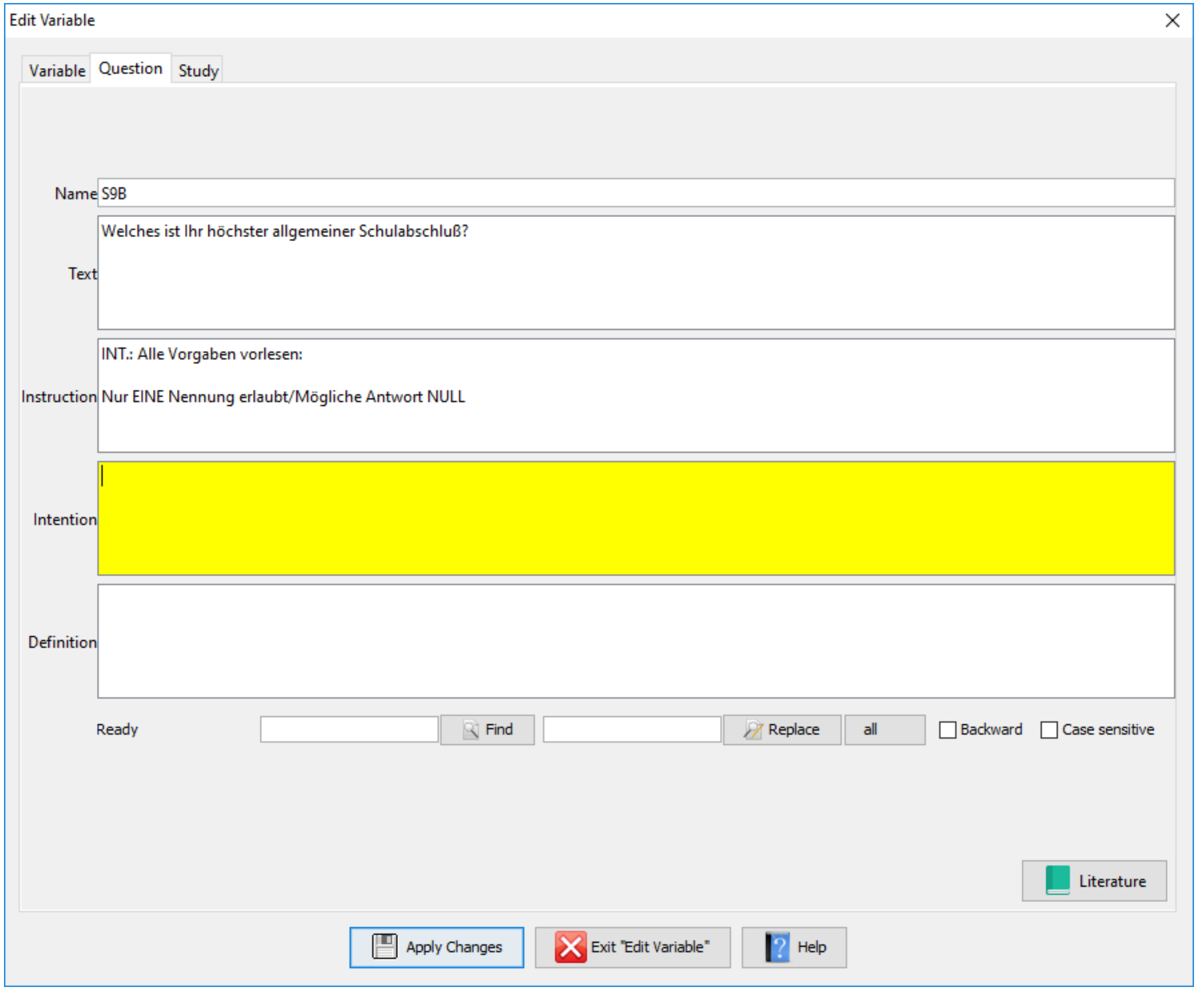
CharmStats

Step 2: Import to CharmStats

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

gesis

The case: add the information you want



Edit Variable

Variable | Question | Study

Name: S9B

Text: Welches ist Ihr höchster allgemeiner Schulabschluß?

Instruction: INT.: Alle Vorgaben vorlesen:

Nur EINE Nennung erlaubt/Mögliche Antwort NULL

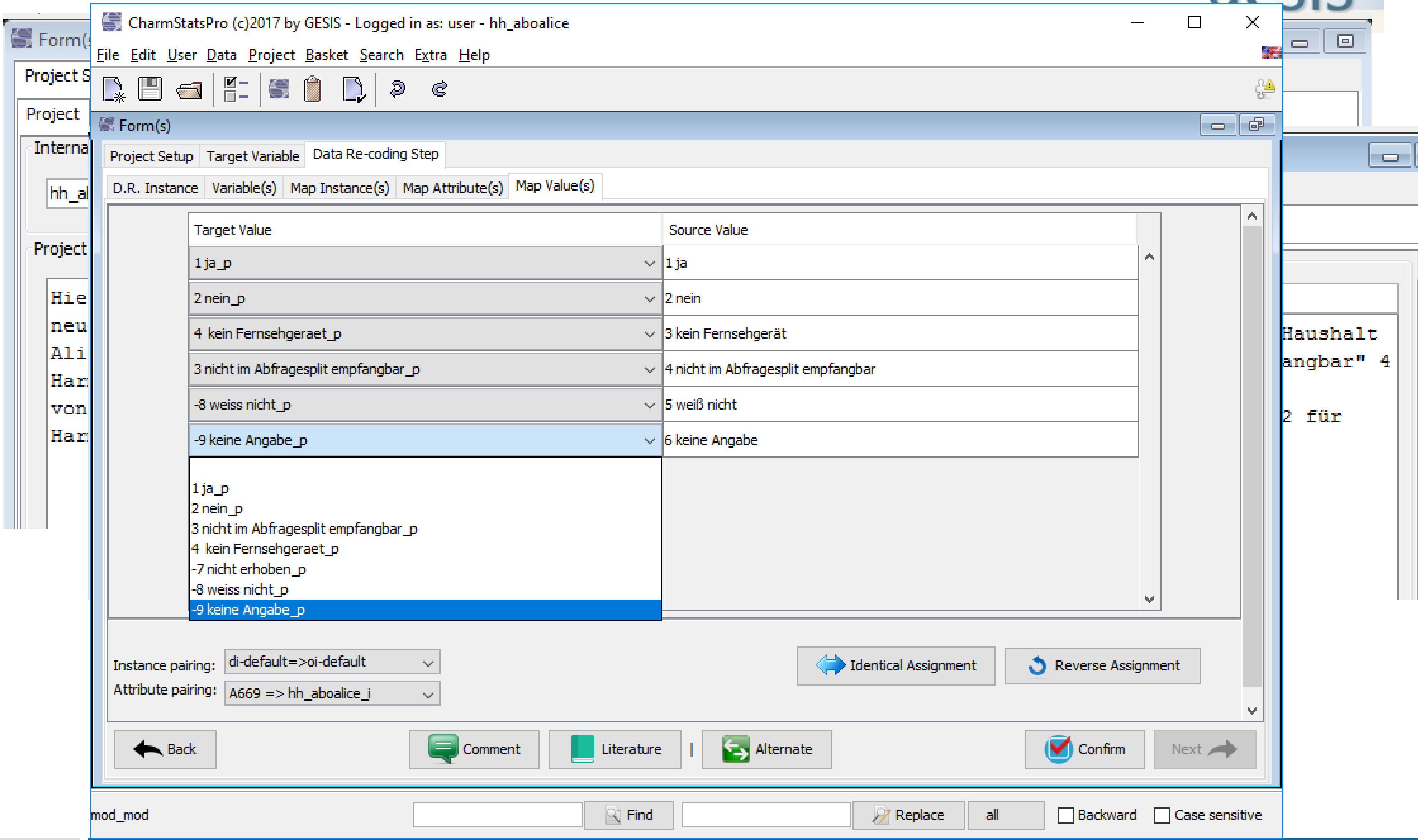
Intention:

Definition:

Ready

Step 3: data-processing with CharmStats Pro

Categorical & nominal Variables



Target Value	Source Value
1 ja_p	1 ja
2 nein_p	2 nein
4 kein Fernsehgeraet_p	3 kein Fernsehgerät
3 nicht im Abfragesplit empfangbar_p	4 nicht im Abfragesplit empfangbar
-8 weiss nicht_p	5 weiß nicht
-9 keine Angabe_p	6 keine Angabe

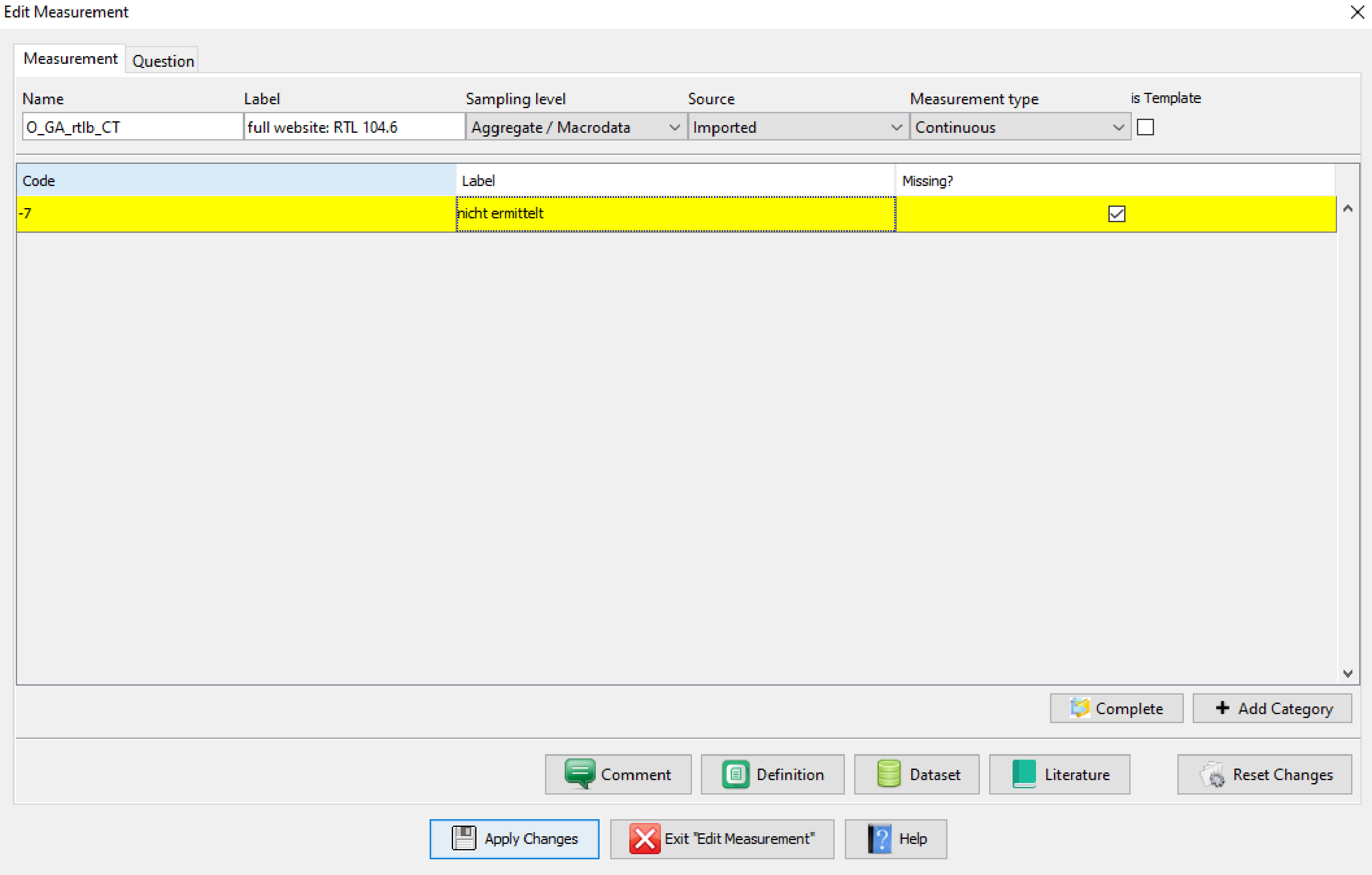
1 ja_p
2 nein_p
3 nicht im Abfragesplit empfangbar_p
4 kein Fernsehgeraet_p
-7 nicht erhoben_p
-8 weiss nicht_p
-9 keine Angabe_p

Instance pairing: di-default=>oi-default

Attribute pairing: A669 => hh_aboalice_i

Step 3: data-processing with CharmStats Pro

metric Variables



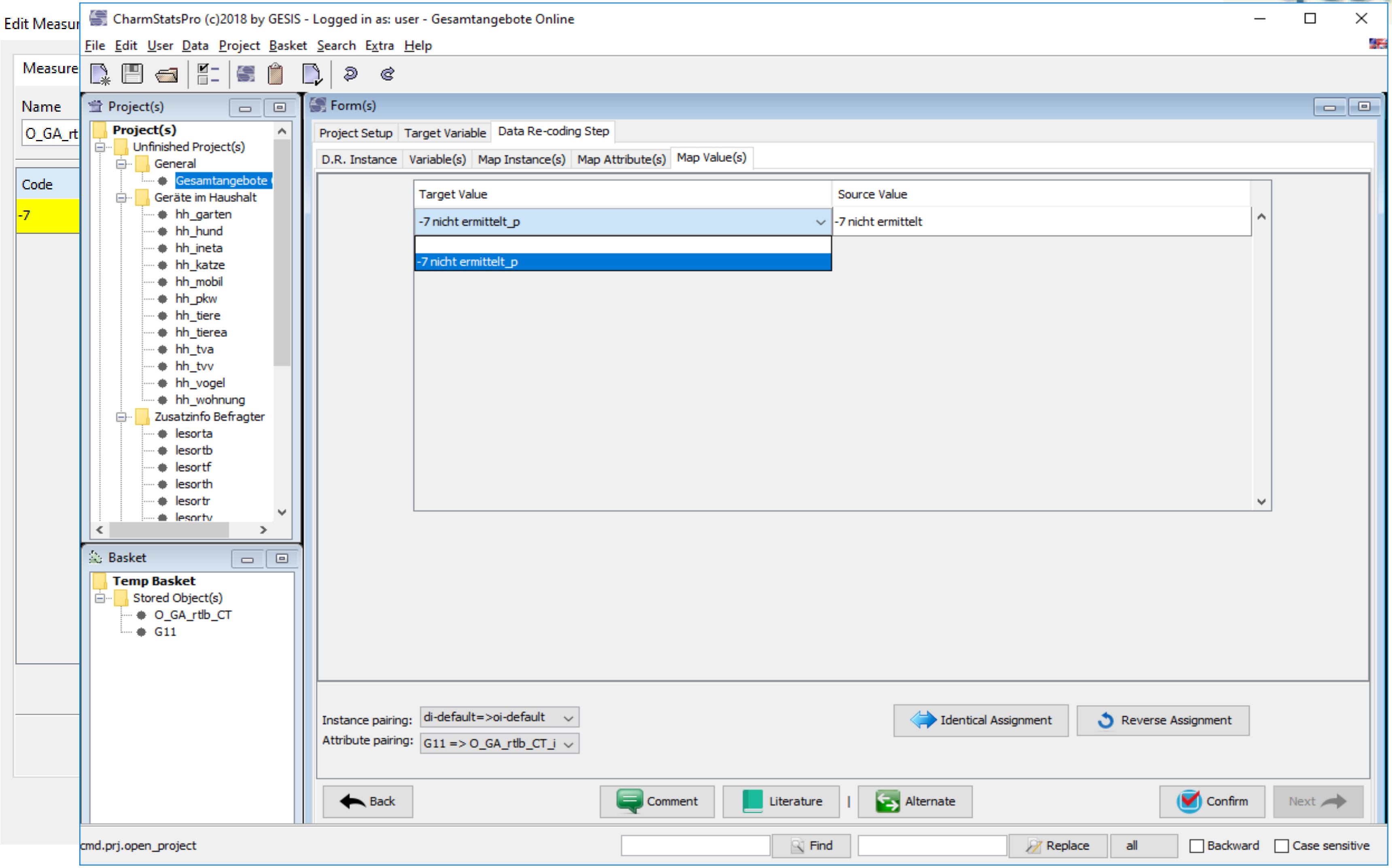
Edit Measurement

Name	Label	Sampling level	Source	Measurement type	is Template
O_GA_rtlb_CT	full website: RTL 104.6	Aggregate / Macrodata	Imported	Continuous	□

Code	Label	Missing?
-7	nicht ermittelt	☑

Step 3: data-processing with CharmStats Pro

metric Variables



CharmStatsPro (c)2018 by GESIS - Logged in as: user - Gesamtangebote Online

Project Setup | Target Variable | Data Re-coding Step

D.R. Instance | Variable(s) | Map Instance(s) | Map Attribute(s) | Map Value(s)

Target Value	Source Value
-7 nicht ermittelt_p	-7 nicht ermittelt

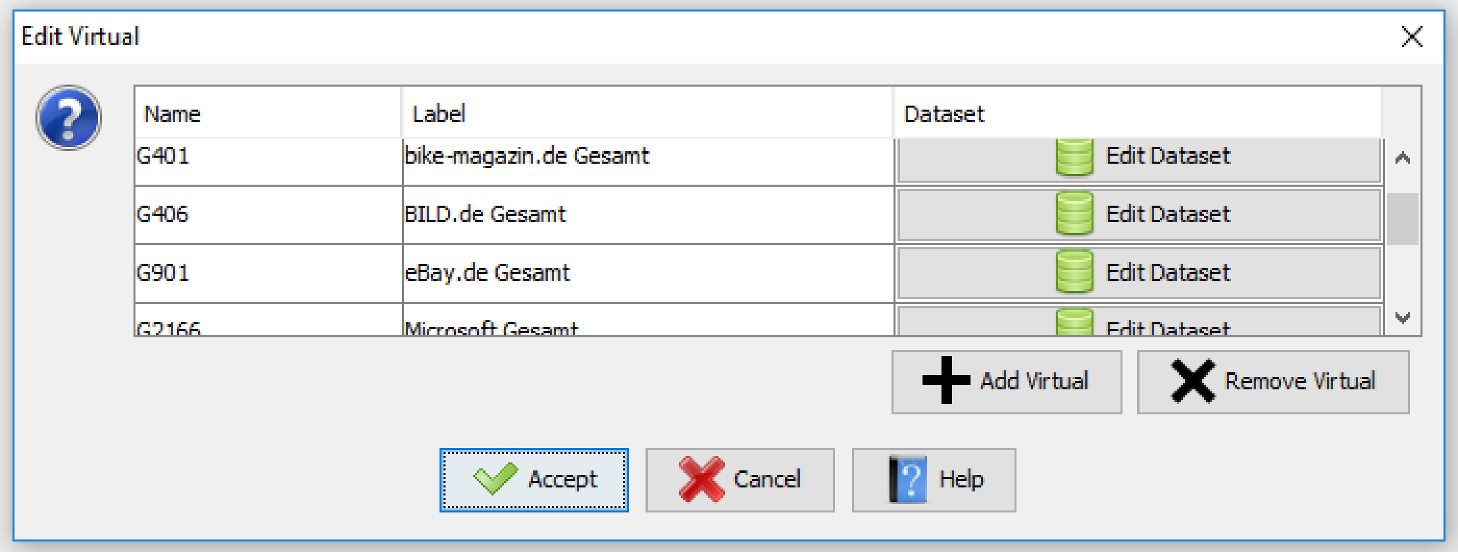
Instance pairing: di-default=>oi-default

Attribute pairing: G11 => O_GA_rtlb_CT_i

cmd.prj.open_project

Step 3: data-processing with CharmStats Pro

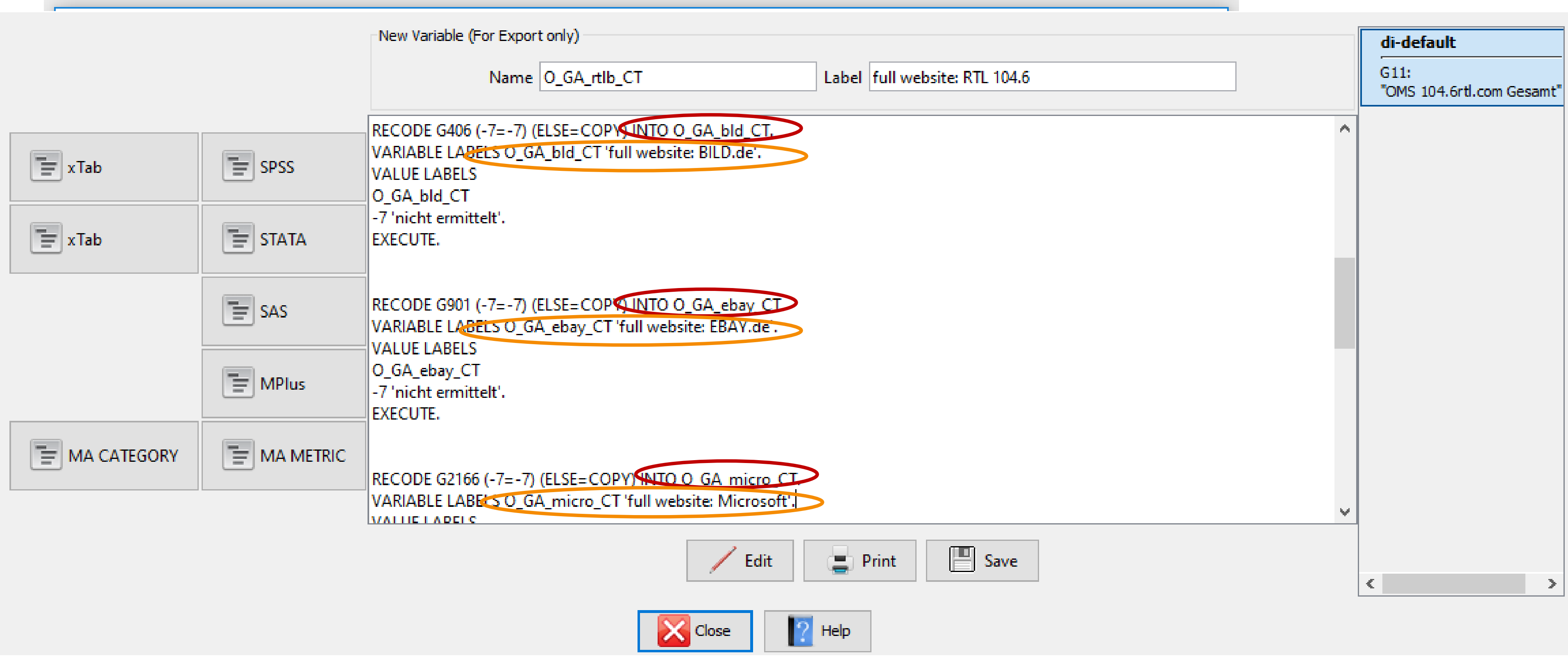
The case of virtual Variables (text based)



Name	Label	Dataset
G401	bike-magazin.de Gesamt	Edit Dataset
G406	BILD.de Gesamt	Edit Dataset
G901	eBay.de Gesamt	Edit Dataset
G2166	Microsoft Gesamt	Edit Dataset

Step 3: data-processing with CharmStats Pro

The case of virtual Variables (text based)



New Variable (For Export only)

Name O_GA_rtlb_CT

Label full website: RTL 104.6

```
RECODE G406 (-7=-7) (ELSE=COPY) INTO O_GA_bld_CT.
VARIABLE LABELS O_GA_bld_CT 'full website: BILD.de'.
VALUE LABELS
O_GA_bld_CT
-7 'nicht ermittelt'.
EXECUTE.


RECODE G901 (-7=-7) (ELSE=COPY) INTO O_GA_ebay_CT.
VARIABLE LABELS O_GA_ebay_CT 'full website: EBAY.de'.
VALUE LABELS
O_GA_ebay_CT
-7 'nicht ermittelt'.
EXECUTE.


RECODE G2166 (-7=-7) (ELSE=COPY) INTO O_GA_micro_CT.
VARIABLE LABELS O_GA_micro_CT 'full website: Microsoft'.
VALUE LABELS
```

di-default

G11:
"OMS 104.6rtl.com Gesamt"

Step 4: Produce output

Report as data documentation

- The **report** feature is used to output project content. Templates allow **user-defined** extraction of content, using an html formated document inter-spersed with keywords.

- Keywords and their interactions with the template are provided by plugins.

CharmStats Pro

Further Features

- CS Libary as an archive and exchange plattform
- Literature
- Work in Teams
  - In Mail
  - User-specified „tracking“ and reports
- Usage for experiments: Make Intensions, concepts and tretments visible

Synopsis

What is CharmStats for?

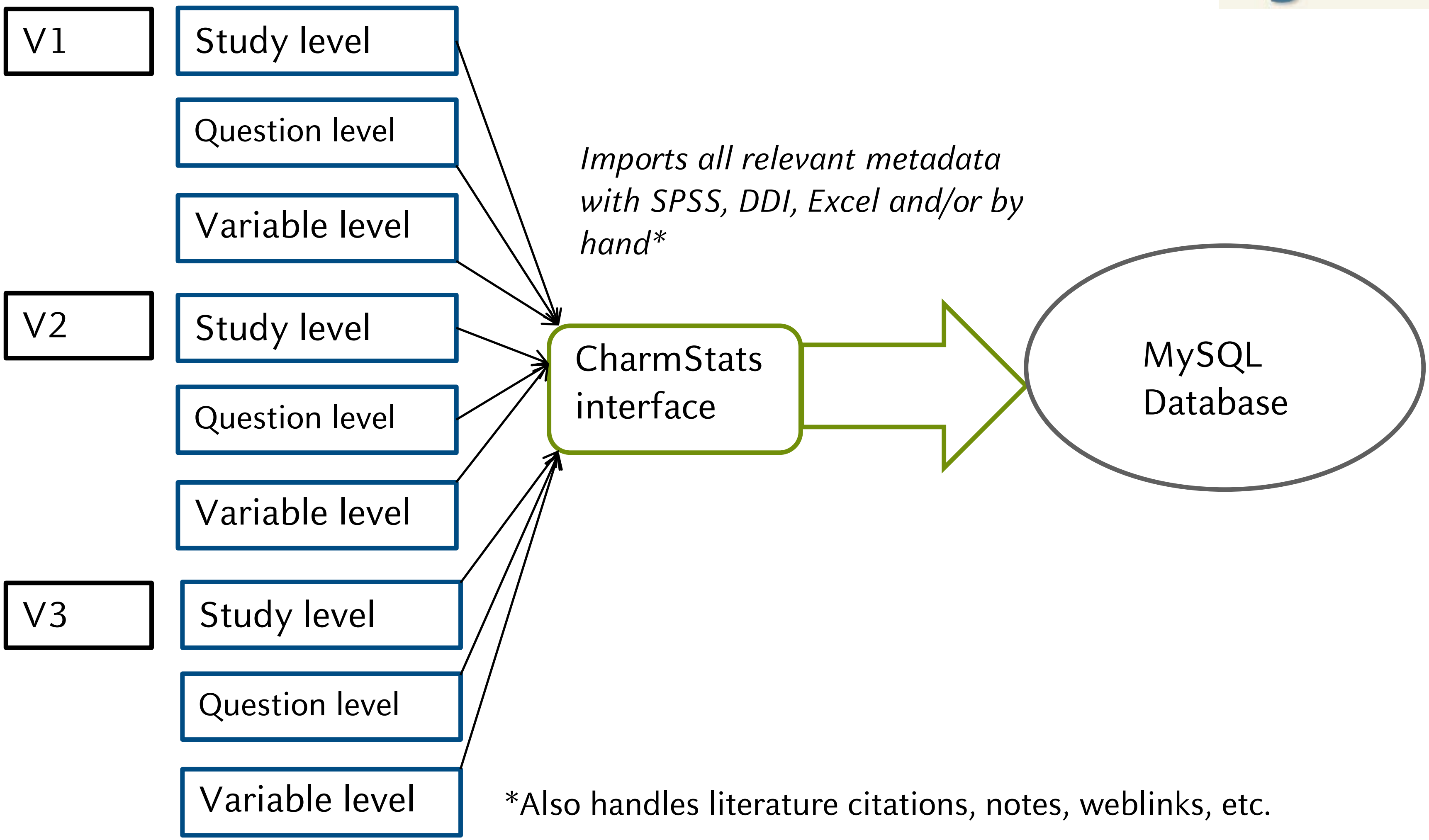
Situation in (social) science: **lack of tracability** through insufficient data-documentation, esp. in case of big data

Advantages of CharmStats:

1) Documentation of measurements *(shown today)*
2) Documentation of stimulus, treatments like pictures, questionnaire *(**not** shown today)*
3) Documentation of data processing and syntax *(shown today)*
   →esp. through CharmStats-Library *(launched for 2019)*
4) Output for different statistical programs *(shown today)*

The solution: CharmStats Pro

A digital solution



*Imports all relevant metadata with SPSS, DDI, Excel and/or by hand**

*Also handles literature citations, notes, weblinks, etc.

The Case: Media-Analysis Data

- yearly gathered cross-sectionally by ag.ma measuring media use for commercial purpose since 1954
- two main datasets: pressmedia (1954-2015) & radio (1977-2015)
  - Pressmedia: ~7.600 Variables, < 1,2 Mio cases
  - Radio: ~25.000 Variables, < 1,6 Mio cases
- One „new“ dataset: online
- One combined dataset: Intermedia (Plus) (1999-2014-2017)
  - since 2014: ~18.600 Variables (downsized), < 1,2 Mio cases
  - Intermedia Plus is a result of a joint venture of ag.ma, agof & AGF/GfK

Media-Analysis Data

Variables

- Socio-demographic variables
- variables regarding the daily routine (per hour and more detailed), e.g. sleeping, eating, driving to work, taking the bus, housekeeping, etc.
- free time activities, e.g. read books, do sports, go out, go to cinema, etc.
- habits of household and respondent, e.g. have a Laptop, a car, mobile, telephone, TV-Abo (e.g. Sky), a flat, pets, etc.
- media use, e.g. reading newspapers (SZ, FAZ, tz, ...), magazines (Automobil, AutoBild, Lisa, Bravo, ...), listening to radio per hour and more detailed (Antenne Bayern, Hessischer Rundfunk 1-3, WDR, 1 Live, KissFM)