

Curricular Reform Study in Thuringia (TH)

SUF Version 1.0.0

Data Manual [Supplement]:

Anonymisation

Tobias Koberg



SPONSORED BY THE



Federal Ministry
of Education
and Research

**Organizational Reform Study in Thuringia:
Anonymisation procedures and statistical
disclosure control**

Technical Report

Tobias Koberg

National Educational Panel Study
University of Bamberg

v 1.0 November 6, 2012

Preamble

This documentation gives an exhaustive explanation of all disclosure risk minimisation techniques applied before dissemination of the Organizational Reform Study in Thuringia. For a quick reference what is done to the datasets in detail and on which level you will find your desired information, please skip forward to appendix A, where all affected variables are listed.

Specifications

To ensure the best possible confidentiality protection of individuals and individual micro data, the National Educational Panel Study complies with strict international standards. Operationalise those, they have been abstracted to the following two criteria:

1. the disseminated data has been transferred to so called *de facto anonymous data*. Identifiable information is coarsened or cut off and kept securely to minimise the risk of statistical disclosure.
2. the use of data is strictly confidential and for statistical purposes only. The closed contract only grants access to members of the scientific community. This contract has a vast amount of legal stipulations, one of them being a large fine which applies for the realisation of re-identification on purpose. Therefore, the disseminated data is highly protected by law and allows a more flexible range of available data.

To pick up the latter, the NEPS has made a huge effort regarding legal regulations to offer as much analysis power of data as possible. This *paradigm of information esteem* reveals the fact that conducted measures of statistical disclosure control are few. Also, if there really was a need for modification, only non-perturbative methods were used.

Onion-shaped model

The NEPS grants the user three different modes of data access: (1) ***OnSite***, which stands for the opportunity to use the secured infrastructure made available at the NEPS in Bamberg, (2) ***RemoteNEPS***, which is a progressive remote access technology providing a virtual desktop, and finally (3) ***Download***, indicating the possibility to fetch data via a secure web portal.

These given access modes have been originated to allow anonymisation routines for a subtle differentiation of information. The three resulting levels of anonymisation define as follows:

- data provided ***OnSite*** is generally not further anonymised. However, even those data has been rendered *de facto anonymous*, for no disclosure risk to persist. All information contained remains completely sane. Although users have to deal with limited possibilities of data access (i.e. supervised import and export of their results), they are free to work with all data available at the NEPS in a secure environment.

- access via *RemoteNEPS* is considered equivalent to *OnSite*, hence most of the data stays complete.
- as *Download* is assumed to be the most hazardous access mode¹, some more anonymisation techniques are done to the dataset.

Obviously this approach results in three different versions of all involved datasets. To enable a consistent structure, these data files always contain the entire set of variables; it is their content which differs through the three levels.

As normally there is no need to resign aggregated variables in the higher levels (i.e. *OnSite* or *RemoteNEPS*), those are already defined as a surplus to the original variable in the *OnSite*-version. Stepping down to *RemoteNEPS* the content of related variables too sensitive for this level is overwritten with an exclusive missing code – an operation which we define as *purging*. Note that system missing values are not affected, allowing the user to differ between value existence and nonexistence. This still is a valuable additional information. Same applies to *Download*.

While there is no explicit documentation to this fact, it should remain clear that this procedure accumulates, i.e. purged content under *RemoteNEPS* is therefore neither included in *RemoteNEPS* nor in *Download*.

This *onion-shaped* model provides both ease of (1) use of different sensitivity models (e.g. preparing an analysis using the *Download* dataset and conducting it afterwards using the *OnSite*-data) and (2) documentation, for the subject of documentation is the most sensitive level (*OnSite*), with *RemoteNEPS* and *Download* levels being a subset of these data.

The fourth layer *master* depicted below contains every material which is needed during data processing by the NEPS, but is not meant for the scientific community to be usable.

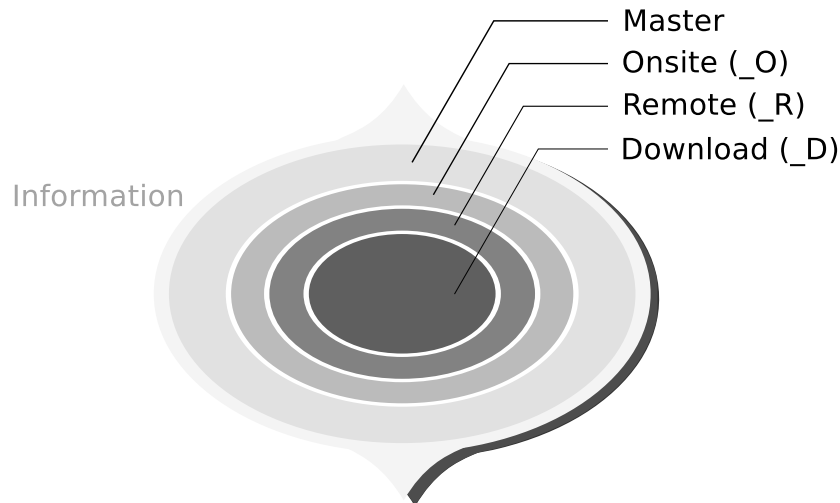


Figure 1: Onion-shaped model defining the different anonymisation levels

¹ 'hazardous' in terms of: the downloaded content is no longer under physical control of the NEPS

Technically, this model realizes in a single letter suffixed to dataset and variable names. All datasets available *OnSite* only are marked with an additional **_O**, those available via *RemoteNEPS* with **_R** and *Download* files with **_D**. The same procedure applies when it comes to variable differentiation. A variable which is only available *OnSite* has been suffixed with **_O**. In *RemoteNEPS*-access or *Download*, this variable is still present but purged. If there is an alternate version (mainly with coarsened content) for *RemoteNEPS* (suffix **_R**) or *Download* (suffix **_D**), those can be used. As said before, these are already integrated in the *OnSite* version.

Conducted measures

Keeping the usability and the paradigm of information esteem in mind, only very few alterations are actually done to the dataset. These modifications always account for the fact that information may never be lost completely, but aggregated into coarse categories or variables. Please note that all information is still available somewhere and that only *RemoteNEPS* and (mainly) the *Download* version are constraint in this matter. In fact, roughly 20 variables are modified in some way – which is about one percent of the whole dataset volume.

Please refer to appendix A for a complete overview of all variables which fell victim to anonymisation.

The following gives an explanatory overview of all measures conducted.

countries and languages All information corresponding to (international) localisation, nationality or languages is only available en full *OnSite* or via *RemoteNEPS*. Variables comprised in the *Download* Scientific Use File (SUF) are aggregated into german and non-german.

open ended strings All string variables containing actual text are purged in the *RemoteNEPS* version. The information remains accessible *OnSite*. However, all text entries have been reviewed by staff to ensure that absolutely no re-identificational material is included.

institutions For starting cohort 2 to 4, special focus of anonymisation has been directed to protection of institutional data, i.e. information about kindergarten and schools, but also educators and teachers. This includes the complete datafile *xInstitution*, but also basic structural details about the kindergarten group or school class. Furthermore, personal information about educators and teachers is treated more securely. You will find detailed information about these subjects from *RemoteNEPS* onwards.

regional Information Regional information is not available for NEPS data which has been surveyed in school context. This regards places of birth as well as work, school or residence. Only an indicator for west germany and east germany (including Berlin) is available. Please be aware that we still do offer macro indicators *OnSite* (see below).

number of employees Considering self-employed persons, information about the number of salaried employees has been censored to prevent effortless identification of

large entrepreneurs. Therefore, related variables are top-coded at 20 employees. Again, this information is still present via *RemoteNEPS* and *OnSite*.

macro indicators Additional information including structural topography and macro-economic measures has been made available only *OnSite*, also called *RegioInfas (infas geodaten)*. Please refer to the separate documentation describing those datasets for further information.

Topic	<i>OnSite</i>	<i>RemoteNEPS</i>	<i>Download</i>
International ¹	full data	full data	collapsed
String variables	anonymised	n/a	n/a
Age of teacher	month and year	years	years collapsed

¹ international geographical information (e.g., nation states, national languages)

² month of birth of educators/teachers and principals/headmasters is only available *OnSite*

³ national localisation is coarsened to west/east germany

Table 1: Availability of sensitive data

For enquiries or further information not covered in this document please feel free to contact userservice.neps@uni-bamberg.de.

A Anonymisation sheet

Instrument						
Name	Additional Study Thuringia					
Stage	5					
Study type	Main survey					
Study number	A70 / A71					
Dissemination	SUF					
Measures						
	data file	variable	label	On-site	RemoteNEPS	SUF-Download
Countries & Languages						
xParent		p18am	country of birth - mother			country aggregation
xParent		p18av	country of birth - father			country aggregation
xParent		p19m_g2	Mother: Nationality (response 1)			country aggregation
xParent		p19m_g3	Mother: Nationality (response 2)			country aggregation
xParent		p19v_g2	Father: Nationality (response 1)			country aggregation
xParent		p19v_g3	Father: Nationality (response 2)			country aggregation
xTarget		t100a	Country of birth - you			country aggregation
xTarget		t100b	Country of birth - your mother			country aggregation
xTarget		t100c	Country of birth - your father			country aggregation
xTarget		t102_g2	Nationality (response 1)			country aggregation
xTarget		t102_g3	Nationality (response 2)			country aggregation
xTarget		t102_g4	Nationality (response 3)			country aggregation
String variables (Note: all string variables have been approved for reidentificational material and anonymised where necessary)						
xCourse		e19a	assessment of changes - own statement		purged	
xCourse		e19b	assessment of changes - own statement II		purged	
xCourse		e19c	assessment of changes - own statement III		purged	
xTarget		t63b	Project paper - description: other		purged	
xTarget		t76ii	Sources of information choice of profession - other:		purged	
Other						
xCourse		e22a	date of birth - month		purged	
xCourse		e22b	date of birth - year			aggregation
xCourse		e23	start teaching			aggregation