# From IDA to Data Quality -Accessible Solutions to Promote Quality and Transparency

Carsten O. Schmidt, Elisa Kasbohm
Joany Marino Stephan Struckmann
March 2023

## IDA Workflow

https://www.semanticscholar.org/paper/A-Contemporary-Conceptual-Framework-for-Initial-Huebner/d6113f76f7c0b42df1a54e4917d5c9891f95d18e

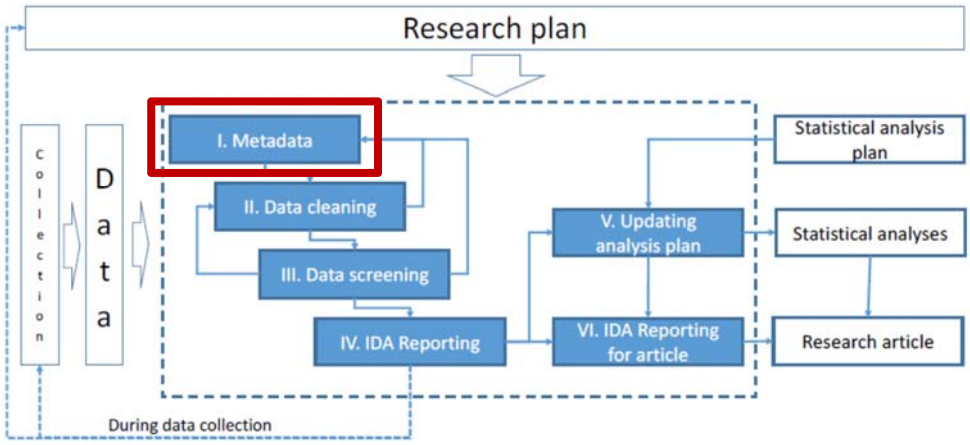



Figure 1: The main connections between the IDA steps and external components

## Data quality perspective on IDA

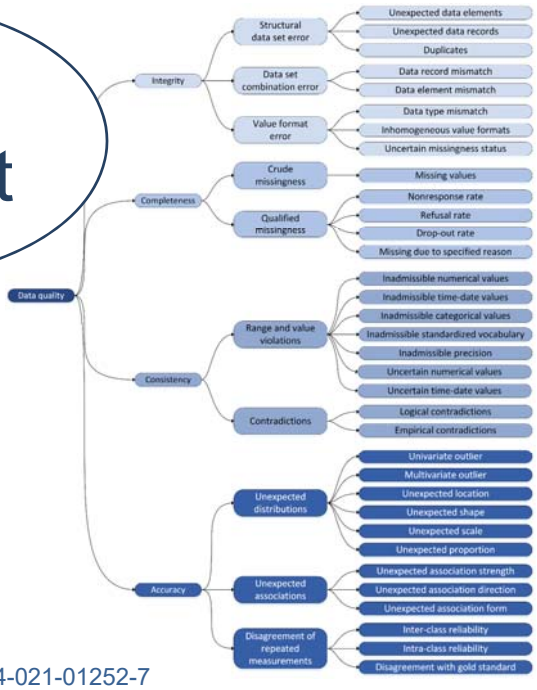
| | | |
|---|---|---|
| Intrinsic data quality | Inherent properties of data sets and data elements, e.g.: range violations, contradictions, formatting errors, missing values | ⟵ IDA |
| Extrinsic data quality | Findability<br>Accessibilty<br>Data safety<br>Use specific requirements | ⟵ IDA |

Diverse uses, e.g. Wang et al. J Manag Inf Syst. 1996

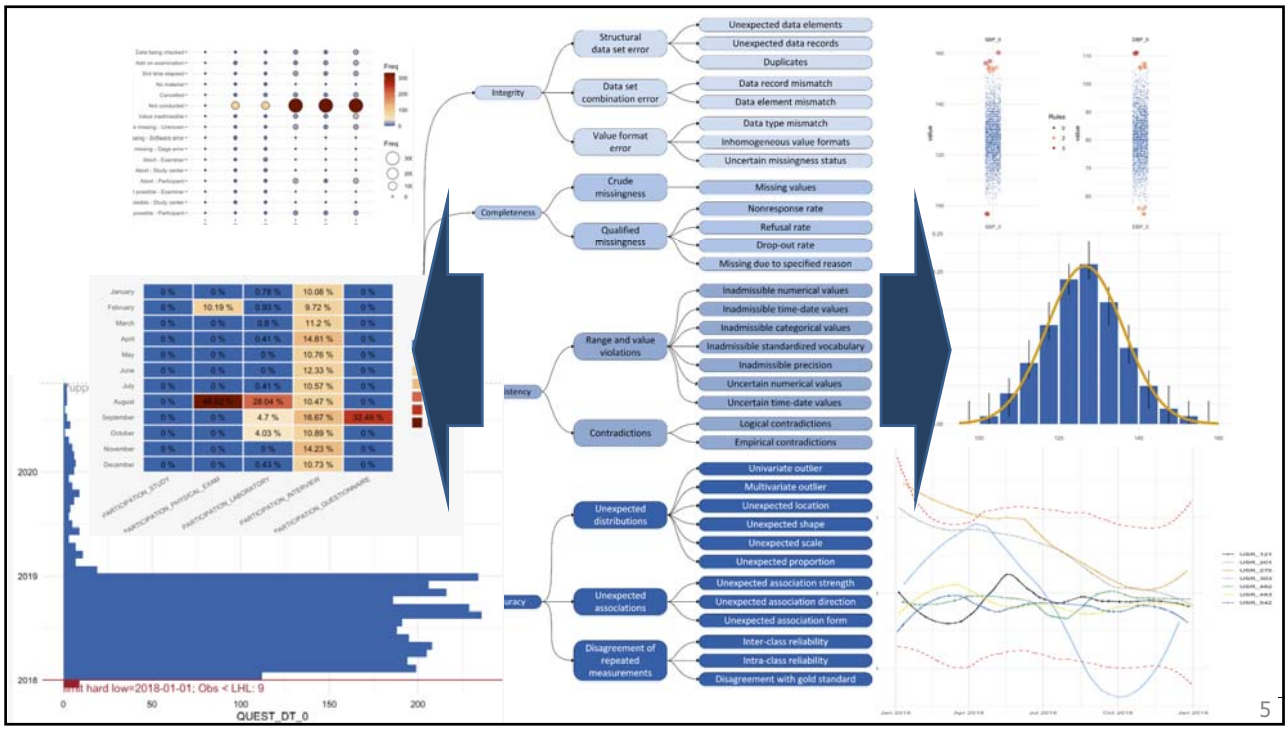


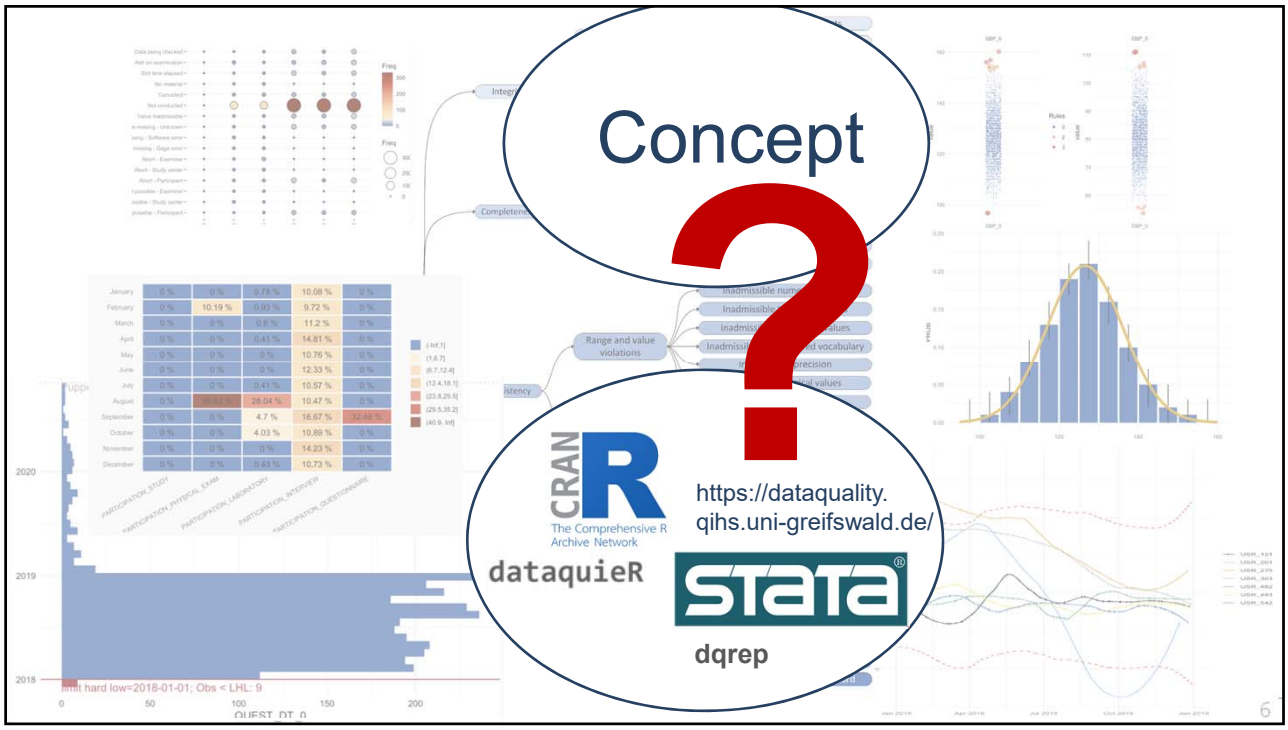


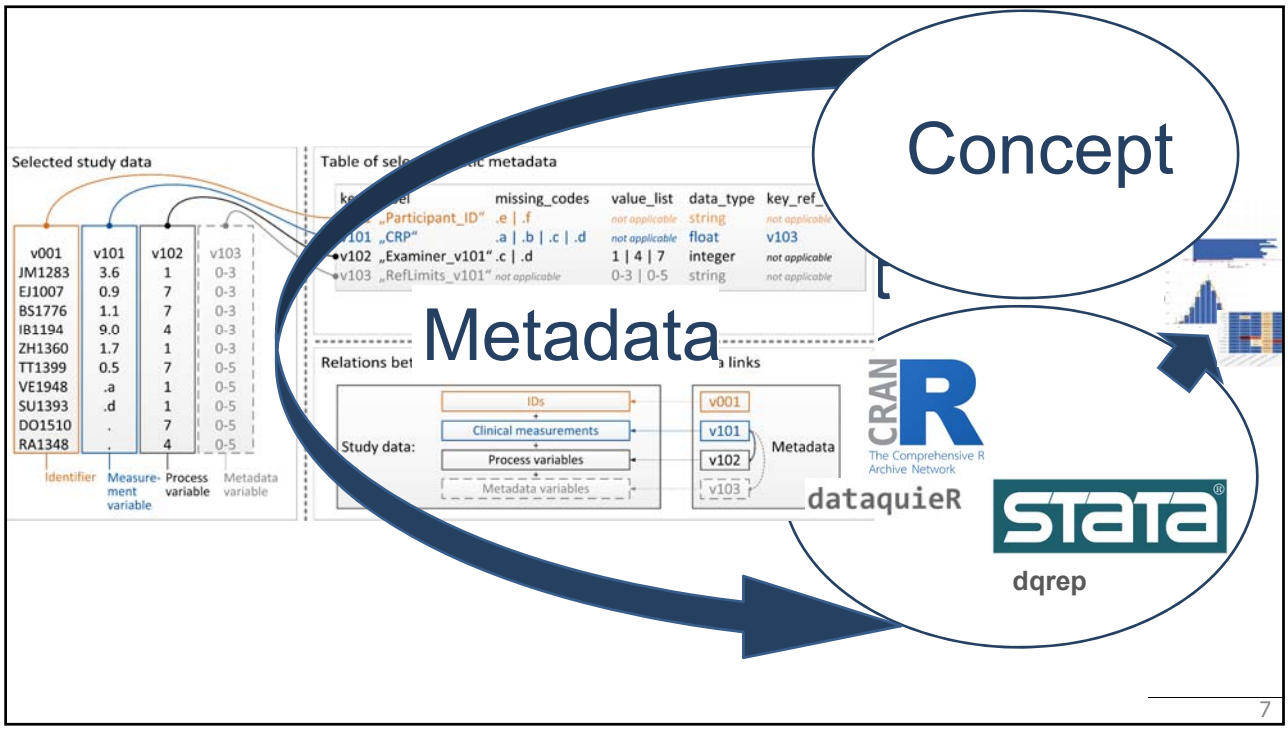
## DQ Concept



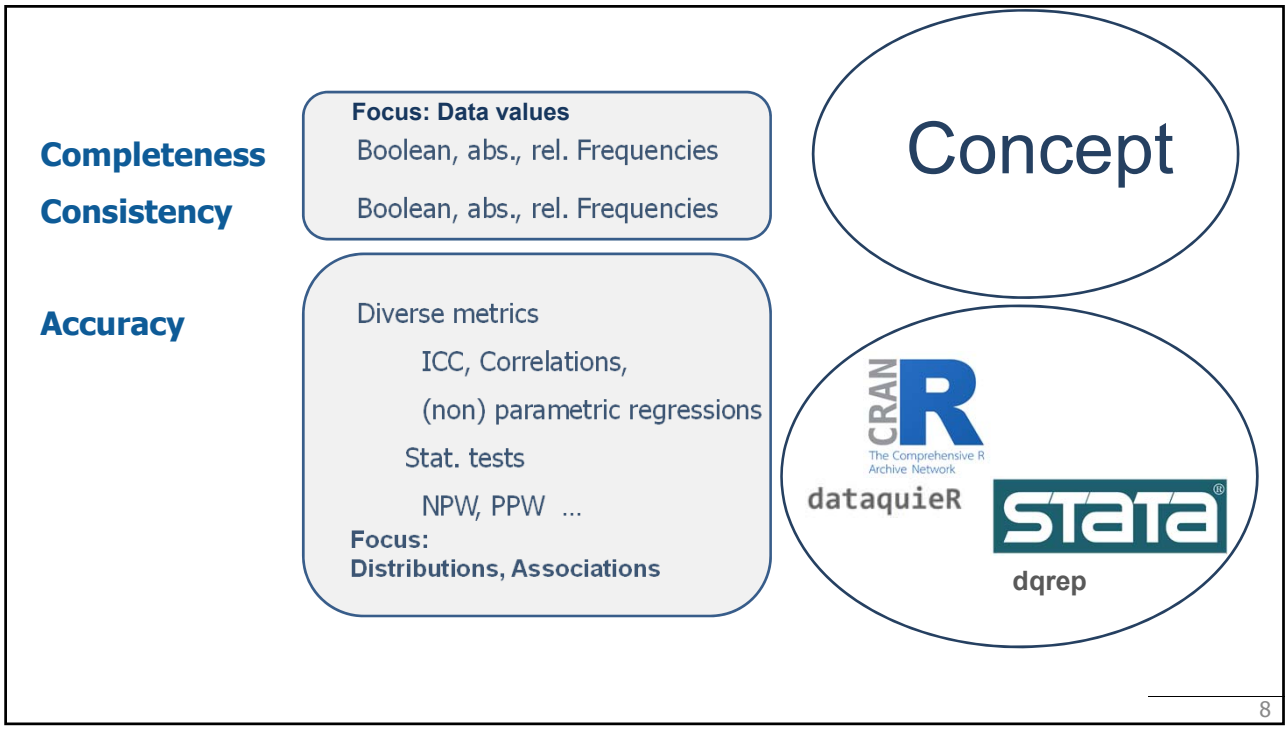

Schmidt et al. 2021 BMC
https://doi.org/10.1186/s12874-021-01252-7


Integrity
Completeness
Structural data set error
Data set combination error
Value format error
Crude missingness
Qualified missingness
Range and value violations
Contradictions
Unexpected distributions
Unexpected associations
Disagreement of repeated measurements
Unexpected data elements
Unexpected data records
Duplicates
Data record mismatch
Data element mismatch
Data type mismatch
Inhomogeneous value formats
Uncertain missingness status
Missing values
Nonresponse rate
Refusal rate
Drop-out rate
Missing due to specified reason
Inadmissible numerical values
Inadmissible time-date values
Inadmissible categorical values
Inadmissible standardized vocabulary
Inadmissible precision
Uncertain numerical values
Uncertain time-date values
Logical contradictions
Empirical contradictions
Univariate outlier
Multivariate outlier
Unexpected location
Unexpected shape
Unexpected scale
Unexpected proportion
Unexpected association strength
Unexpected association direction
Unexpected association form
Inter-class reliability
Intra-class reliability
Disagreement with gold standard
limit hard low=2018-01-01; Obs < LHL: 9
QUEST_DT_0
5


Concept
?
CRAN
R
The Comprehensive R Archive Network
dataquieR
https://dataquality.qihs.uni-greifswald.de/
STATA
dqrep
limit hard low=2018-01-01; Obs < LHL: 9
QUEST_DT_0
6


Concept
Metadata
Selected study data
Table of sele ic metadata
missing_codes value_list data_type key_ref_
„Participant_ID" .e | .f not applicable string not applicable
v101 „CRP" .a | .b | .c | .d not applicable float v103
v102 „Examiner_v101" .c | .d 1 | 4 | 7 integer not applicable
v103 „RefLimits_v101" not applicable 0-3 | 0-5 string not applicable
v001 v101 v102 v103
JM1283 3.6 1 0-3
EJ1007 0.9 7 0-3
BS1776 1.1 7 0-3
IB1194 9.0 4 0-3
ZH1360 1.7 1 0-3
TT1399 0.5 7 0-5
VE1948 .a 1 0-5
SU1393 .d 1 0-5
DO1510 . 7 0-5
RA1348 . 4 0-5
Identifier
Measure-ment variable
Process variable
Metadata variable
Relations bet a links
Study data:
IDs
Clinical measurements
Process variables
Metadata variables
v001
v101
v102
v103
Metadata
CRAN
R
The Comprehensive R Archive Network
dataquieR
STATA
dqrep




Completeness
Consistency
Accuracy
Focus: Data values
Boolean, abs., rel. Frequencies
Boolean, abs., rel. Frequencies
Diverse metrics
ICC, Correlations,
(non) parametric regressions
Stat. tests
NPW, PPW ...
Focus:
Distributions, Associations
Concept
CRAN
R
The Comprehensive R Archive Network
dataquieR
STATA
dqrep

# Metadata example

## Example item-level metadata file

| | VAR_NAMES | LABEL | DATA_TYPE | VALUE_LABELS | MISSING_LIST_TABLE | HARD_LIMITS | DISTRIBUTION | DECIMALS | TIME_VAR | STUDY_SEGMENT | PART_VAR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | v00000 | CENTER_0 | integer | 1 = Berlin \| 2 = Hamburg \| 3 = Leipzig \| 4 = Cologne \| 5 = Munich | | | | | | PART_STUDY | v10000 |
| 3 | v00001 | PSEUDO_ID | string | | | | | | | PART_STUDY | v10000 |
| 4 | v00002 | SEX_0 | integer | 0 = females \| 1 = males | | | | | | PART_STUDY | v10000 |
| 5 | v00003 | AGE_0 | integer | | | [18;Inf) | | | | PART_STUDY | v10000 |
| 6 | v00103 | AGE_GROUP_0 | string | | | | | | | PART_STUDY | v10000 |
| 7 | v01003 | AGE_1 | integer | | | [18;Inf) | | | | PART_STUDY | v10000 |
| 8 | v01002 | SEX_1 | integer | 0 = females \| 1 = males | | | | | | PART_STUDY | v10000 |
| 9 | v10000 | PART_STUDY | integer | 0 = no \| 1 = yes | | | | | | PART_STUDY | v10000 |
| 10 | v00004 | SBP_0 | float | | missing_table | [80;180] | normal | 0 | v00013 | PART_PHYS_EXAM | v20000 |
| 11 | v00005 | DBP_0 | float | | missing_table | [50;Inf) | normal | 0 | v00013 | PART_PHYS_EXAM | v20000 |
| 12 | v00006 | GLOBAL_HEALTH | float | | missing_table | [0;10] | uniform | 1 | | PART_PHYS_EXAM | v20000 |
| 13 | v00007 | ASTHMA_0 | integer | 0 = no \| 1 = yes | missing_table | [0;1] | | | | PART_PHYS_EXAM | v20000 |
| 14 | v00008 | VO2_CAPCAT_0 | string | A = excellent \| B = good \| | missing_table | | | | v00013 | PART_PHYS_EXAM | v20000 |
| 15 | v00009 | ARM_CIRC_0 | float | | missing_table | [0;Inf) | normal | 0 | v00013 | PART_PHYS_EXAM | v20000 |
| 16 | v00109 | ARM_CIRC_DISC | integer | 1 = (-Inf,20] \| 2 = (20,30] | missing_table | [1;3] | | | | PART_PHYS_EXAM | v20000 |
| 17 | v00010 | ARM_CUFF_0 | integer | 1 = (-Inf,20] \| 2 = (20,30] | missing_table | [1;3] | | | | PART_PHYS_EXAM | v20000 |
| 18 | v00011 | USR_VO2_0 | string | USR_101 \| USR_103 \| US | missing_table | | | | | PART_PHYS_EXAM | v20000 |
| 19 | v00012 | USR_BP_0 | string | USR_121 \| USR_123 \| US | missing_table | | | | | PART_PHYS_EXAM | v20000 |
| 20 | v00013 | EXAM_DT_0 | datetime | | | [2018-01-01 00:00:00 CET;) | | | | PART_PHYS_EXAM | v20000 |
| 21 | v20000 | PART_PHYS_EXA | integer | 0 = no \| 1 = yes | | | | | | PART_PHYS_EXAM | v20000 |





# Several levels of metadata possible

**Metadata - level**

Integrity

- Structural data set error → **Data frame level**, **Segment level**

Completeness

- Qualified missingness → **Item level**, **Missing table**

Consistency

- Range and value violations → **Item level**
- Contradictions → **Cross-item level**

Accuracy

- Unexpected distributions → **Item level**
- Outliers → **Item level**

## dataquieR: a quick intro

- Support for reading files
- New reporting engine

```
1  library(dataquieR)
2
3  dq_report <- dq_report2(
4    study_data = "study_data",
5    meta_data = "meta_data",
6    label_col = LABEL
7    )
8
9  dq_report
```

load package

run report command

call output




## dqrep

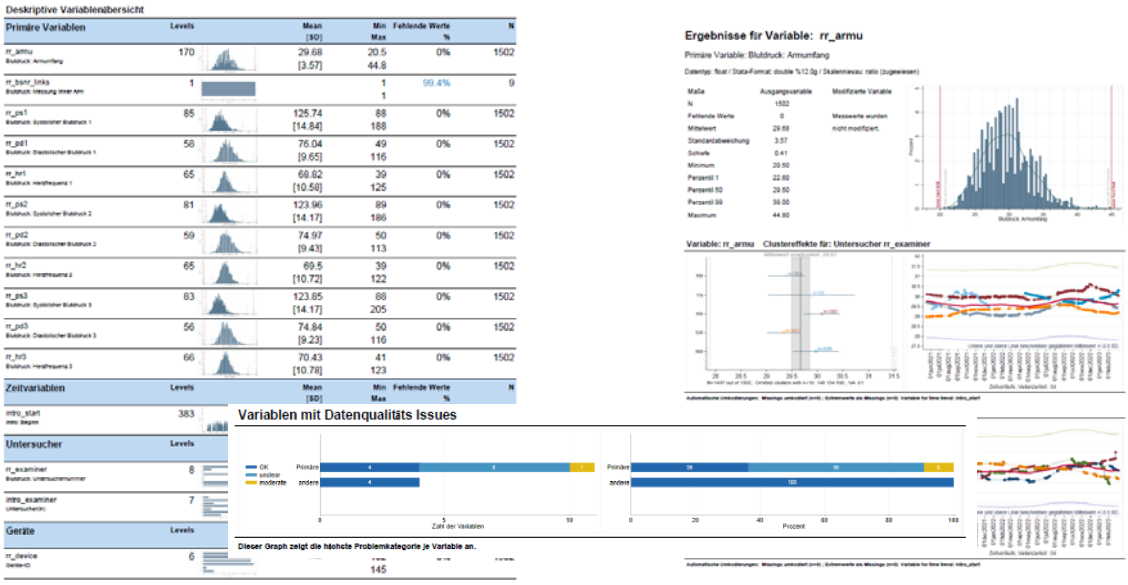
**Multiple Data Quality reports with metadata file**

```
dqrep, rd(Example5) metadatafile("SHIP_metadata.xlsx") ///
       reporttitle("SHIP-0 Data quality report") ///
       segmentselect(INTERVIEW LABORATORY SOMATOMETRY) segmentname(segments) benchmark(3)
```

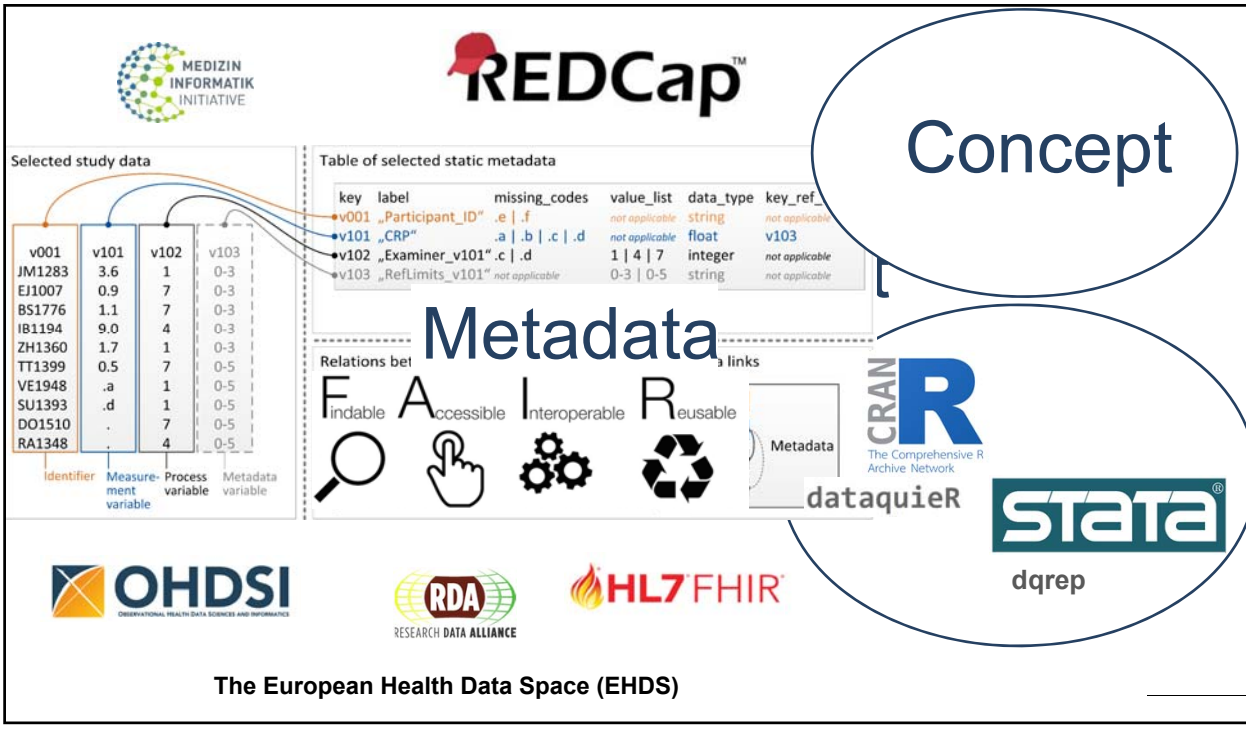
**..if no metadata available in a separate file….**

```
dqrep, rd(Example3) targetfiles("SHIP_study") ///
       itemmisslist(99900 99901 99902 99914) itemjumplist(99800 99801 99802) ///
       reportname("SHIP-Samplereport") reporttitle("SHIP-0 Data quality report") ///
       reportsubtitle("Report using anonymized SHIP-0 sample data") ///
       reportformat("docx") keyvars("sbp1 sbp2 dbp1 dbp2") ///
       minorvars(cholesterol stroke diab_known waist weight contraception) ///
       observervars(obs_bp)devicevars(dev_bp) controlvars(age sex) idvars(id) timevars("exdate") store
```

## dqrep






**The European Health Data Space (EHDS)**

## Outlook in STRATOS

Cooperation

- …in current DFG project, focus reporting (Willi)
- …with other Topic groups /Panels cooperate on papers to create use cases for DQ-reporting
- …on code development
- …paper on improved information work flow

STRATOS INITIATIVE





https://dataquality.qihs.uni-greifswald.de/

elisa.kasbohm@uni-greifswald.de

Thank you!

**Universitätsmedizin Greifswald . KöR**
Institut für Community Medicine, Abteilung SHIP-KEF

Prof. Dr. Carsten Oliver Schmidt
Walther-Rathenau-Str. 46
17475 Greifswald
Carsten.schmidt@uni-greifswald.de



euCanSHare
nfdi 4 health
Funded by: DFG Deutsche Forschungsgemeinschaft German Research Foundation